Appendix: OCR'ing Early Modern Texts

TABLE OF CONTENTS*

Commitment Letters

Contracts

Subaward Contracts

Curriculum Vitae

Additional Information

* In Adobe Acrobat, open Bookmarks for a clickable Table of Contents. The Bookmark icon is to the left—🔖

EASTERN CONNECTICUT STATE UNIVERSITY
83 WINDHAM STREET • WILLIMANTIC, CONNECTICUT 06226 • 860-465-5000

May 30, 2012

To the Mellon Foundation:

As Secretary of the Defoe Society, I proposed to the board at our recent meeting in San Antonio in March that the Society work to organize and facilitate OCR correction for works by Defoe in 18thConnect's TypeWright, with an eye towards eventually preparing a digital edition of Defoe's works. The board was quite interested in the prospect, and appointed a committee to explore how best to proceed.

At this point, we're treating the correction of OCR and the preparation of a digital edition as two distinct but, we hope, related steps. We see the improvement of searchable text in ECCO and at 18thConnect as a good in itself, since it benefits all scholars and students of Defoe's works, and the Society would like to help make that happen. Though we haven't yet decided exactly how we want to organize the work, I strongly suspect that 18thConnect offers all the infrastructure we need. There are currently 484 texts attributed to Defoe available for correction in TypeWright, so this will be a pretty large endeavor, likely involving many different hands before it's done. TypeWright's group capabilities will, I think, be perfectly suited for discussion among groups of editors, and the fact that the platform allows one to see at a glance how much of a text has been edited should also prove incredibly valuable.

The further possibility of a digital edition is a very intriguing one to many of us on the board of the Defoe Society. One of our members is the managing editor for the Stoke Newington edition of Defoe's works for AMS Press, and she (rightly, I believe) encourages us to keep our textual editing standards as high for a digital edition as they would be for that print edition. I believe that, in a great many cases, we could treat page images from ECCO as our (virtual) copy text and, using TypeWright, arrive at good base texts for a TEI-compliant edition. For a full-dress digital edition, however, we would obviously need to collate multiple instances of a work, and so I am very intrigued by the Cobré software developed at Texas A&M, and by the possibility of its functions being adapted for work in 18thConnect and TypeWright.

I expect that the preparation of a digital edition of Defoe's works today is going to involve working with both physical books and digital surrogates—and even, perhaps, making new digital surrogates of physical copies that haven't yet been digitized, so as to facilitate comparison and collation. I definitely imagine a place for a tool like Juxta in our work, for collating and editing texts once they've been transcribed. Cobré seems to offer a different and highly valuable benefit, however, as it could assist in the visual comparison of page images even before the texts have been transcribed or OCR'ed and corrected. It seems like a very promising tool for the kind of bibliographical and textual work that a digital edition will require.

I don’t think the Defoe Society would be contemplating the creation of a digital edition of Defoe’s works if it weren’t for the tools 18thConnect provides. The task simply of gathering the raw materials for doing so would have seemed so unimaginably large that I don’t think the idea would even have come up. Even if it had, the prospect of building the digital infrastructure needed to facilitate the work of a large and distributed editorial team would have been very daunting, to say the least. By gathering such a sizable body of material in one place through its agreements with Gale-Cengage, 18thConnect makes this project seem much more manageable. Even if we find that we need to create further digital surrogates, TypeWright provides a critical mass of material to start with.

I must stress that our thinking about this project is still at an early stage, and so I don’t know for certain whether a full edition of Defoe’s works will come to fruition. Even if it doesn’t, however, the correction of the OCR for even a meaningful fraction of Defoe’s works through TypeWright would be a boon, and that, I believe, is eminently achievable. When I described to the board how TypeWright worked and what it would yield—clean, searchable text of Defoe’s works, created through a simple, intuitive interface—the consensus was immediate: why *wouldn’t* we want to get involved in fixing the OCR when somebody has made it so easy to do? What 18thConnect has already developed is impressive. What it could offer with the incorporation of Cobré-like visual comparison tools is still more exciting.

Yours sincerely,

Benjamin F Pauley

Benjamin F. Pauley
Associate Professor, English
Eastern Connecticut State University
Secretary, The Defoe Society

COLLEGE OF SCIENCE
Academy for Advanced Telecommunications and Learning Technologies

TEXAS A&M UNIVERSITY

The ATTLT or "Academy," Directed by Guy Almes, hosts the Brazos High Performance Computing Cluster at Texas A&M.

June 1, 2012

Dr. Laura Mandell
Initiative for Digital Humanities, Media, and Culture
Texas A&M University

Dear Dr. Mandell:

On behalf of the Academy for Advanced Telecommunications and Learning Technologies, I am pleased to write in support of your proposal "OCR'ing Early Modern Texts" to the Mellon Foundation.

The Academy provides a variety of cyberinfrastructure in support of research at Texas A&M University. Much of it centers on advanced computing and data management, with particular emphasis on the efficient processing of data-intensive computations, such as the massive set of OCR computations contemplated in your proposal. With respect to your project, we anticipate supporting you in several specific ways:

- We will support more than a dozen "virtual machines" with specific roles in your project. These VMs will be tailored to the needs of your research, yet enjoy access to the large parallel file systems of the Brazos cluster. For one of these VMs, project colleagues at Performant Software will have sudo access, supported by our technical staff.
- We will provide access to large computing and data storage via our Brazos cluster, currently with an aggregate of 2,600 computing "cores" and more than 200 TBytes of storage. This system was designed, primarily, to support "high throughput" computing, in which researchers run hundreds, or many thousands, of independent computing jobs, often differing primarily in just their specific data inputs. Our work with A&M's high-energy physics group has made us quite experienced in supporting this style of computation and, to a remarkable extent, this same style of large numbers of independent jobs running against large numbers of data sets (particle physics data for them and document images for you) applies.
- We will support your group in the careful storage and management of multiple terabytes of data, both the original image data and the processed data.
- Based on years of experience in computer systems design and operations, I will consult with your group on the most effective technical approaches to emergent problems.

We estimate that this will take some of my time plus, specifically, up to 20-25% of the time of Mr. Trey Dockendorf, one of our staff highly skilled in Linux system administration and in the needs of humanities researchers. We look forward to working with you on this exciting research.

Sincerely,

Guy T. Almes
Director

JISC Historic Books Advisory Board

JISC Historic Books is managed by JISC Collections in partnership with one of the JISC datacentres called Mimas (www.mimas.ac.uk).

The JISC Historic Books Advisory Board, made up of expert academics and library representatives from institutions from including:

- Joanna Ball, Research Liaison Manager, University of Sussex
- Simon Bell, Head of Partnerships and Licensing, The British Library
- Giles Bergel, JPR Lyell Research Fellow in the History of the Book, Merton College, University of Oxford
- Laurel Brake, Professor Emerita of Literature and Print Culture, Birkbeck, University of London
- Justin Champion, Professor of the History of Early Modern Ideas Royal Holloway, University of London
- Godfried Croenen, Reader in French Historical Studies, University of Liverpool
- Adrian Edwards, Lead Curator, Printed Historical Sources, The British Library
- Jess Edwards, Head of the Department of English, Manchester Metropolitan University
- Gabriel Egan, Reader in Shakespeare Studies, Loughborough University (Chair)
- Simon Elliot, Chair in the History of the Book, Deputy Director, Centre for Manuscript and Print Studies, School of Advanced Studies, University of London
- Lorraine Estelle, CEO, JISC Collections
- Scott Gibbens, Service Representative, JISC Collections
- Jonathan Gibson, Academic co-ordinator, HEA English Subject Centre Royal Holloway, University of London
- Stephen Gregg, Senior Lecturer in English, University of Bath
- Jerome de Groot, Director of Research Training in the Arts, University of Manchester
- Catherine Grout Programme Director, e- Content, JISC
- Tracey Hill, Head of Department of English & Cultural Studies, Bath Spa University
- Diarmund Kennedy, Subject Librarian, Queen's University Belfast
- Vic Lyte, Senior Manager, Mimas
- Elizabeth McHugh, Electronic Resources Manager, University of the Highlands and Islands
- Caren Milloy, Head of Projects, JISC Collections
- Chris Mounsey, Lecturer in English, University of Winchester
- Andrew Murphey, Head of the School of English, University of St Andrews
- Beth Palmer, Lecturer in English Literature, University of Surrey
- Michael Poppham, Head of Digital Initiatives, Bodleian Digital Library Systems & Services, Bodleian Libraries, University of Oxford
- Paul Rayson, Director of UCREL and Lecturer in Computer Science School of Computing and Communications, Lancaster University
- Elizabeth Scott-Baumann, Lecturer in Early Modern English Literature, Wadham College, University of Oxford
- Matthew Steggle, Reader in English, Sheffield Hallam University
- Adrian Streete, Senior Lecturer, School of English, Queen's University Belfast
- Ceri Sullivan, Professor, School of English, Bangor University
- Erica Swain, Subject Librarian – English, Communication & Philosophy / Religious Studies & Theology, Cardiff University
- Samantha Tillet, Project and Service Manager, The British Library

- Mark Townsey, Coordinator of Postgraduate Research, Lecturer in Modern British History, University of Liverpool

The advisory board is in charge of the platform and its future development – the community that uses it, has control. This is reflected in the board's terms of reference which are available at: http://www.jiscecollections.ac.uk/advisory-board/jhbadvisoryboard/

Improving OCR through crowdsourcing

JISC Collections has been following your work and the earlier Mellon Foundation funded project with great interest. In 2011, you kindly invited us to attend your OCR Summit. The chair of the JISC Historic Books Advisory Board, Gabriel Egan, and my colleague Scott Gibbens attended and returned full of enthusiasm. I had asked them to ascertain whether there was scope for collaboration with your project and to assess the potential value to UK academia.

In January 2012, following conversations with you, Gabriel Egan, Scott Gibbens and I presented a proposal for collaboration with your project to the JISC Historic Books advisory board. The advisory board members were fully supportive of the idea, highlighting how valuable the work would be for their research and teaching and requested that JISC Collections work to take the proposal forward.

The Proposal

JISC Collections would like to work with you and the relevant publishers to:

1. Ingest the OCR you create as part of your project into JISC Historic Books. The OCR would be for EEBO images and ECCO images.

2. Install TypeWright on JISC Historic Books to enable the crowdsourcing of corrections to the OCR by the UK academic community.

3. Share the corrections made to the OCR with you and your platforms (18th Century Connect) and with the publisher's platforms (ProQuest and Cengage).

4. Work with you to develop the technical infrastructure to support a centralised system that will

- ➢ support the import of corrections from multiple platforms – it is likely that each platform will export corrections in TEI-A format from the backend MySQL database to this central location
- ➢ track submissions and versions – it is a requirement that the centralised system can cope with and distinguish between the same corrections occurring simultaneously, account for the potentially different versions of OCR being used on platforms and logs everything clearly
- ➢ enable an international editorial board to easily interact with the corrections and validate them
- ➢ export the validated corrections to each platform for ingest into each platform

5. Work with you to develop a methodology for ingest of corrections back into the platforms – this will need to support different backend systems and ensure that corrections overwrite with some sort of identifier being maintained.

6. Explore the development of a widget that would allow students in particular to correct obvious and easy mistakes in the OCR that integrates with the TypeWright software or to undertake specific exercises led by their teachers that help increase their knowledge of early books and the pitfalls of OCR

7. Explore the possibility of some reward and recognition system to encourage students and academics

8. Provide our advice and guidance as digital content licensing experts and negotiators

9. Offer the expertise of the JISC Historic Books advisory board members who wish to be involved in all aspect of the project, technical and editorial

It should be noted, that agreement from the publishers will be required to enable JISC Collections to proceed with this collaboration. Initial meeting held with the publishers indicate their willingness to participate in this initiative.

In addition, there is a requirement that the IPR of the OCR, the corrections and the final versions are clearly articulated in agreements and understood by end users of the TypeWright tool.

JISC Collections is extremely keen to work with you on this initiative. There is no doubt, that working collaboratively would be of great benefit to the scholarly community. Crowdsourcing corrections across the US and the UK would be an amazing initiative and the enthusiasm of academics, as evidence by the JISC Historic Books advisory board, would help ensure it is a success.

I hope that your proposal is successful and that we may work together to take this forward.

Kind Regards

Caren Milloy
Head of Projects
JISC Collections

M Library

UNIVERSITY OF MICHIGAN LIBRARY

Harlan Hatcher Graduate Library
913 South University
Ann Arbor, Michigan 48109-1190

February 13, 2012

Donald J. Waters, Program Officer
Helen Cullyer, Associate Program Officer
Mellon Foundation
The Andrew W. Mellon Foundation
140 East 62nd Street
New York, NY 10065

To Whom it May Concern:

On behalf of the Text Creation Partnership (TCP), a non-profit scholarly project based at the University of Michigan and University of Oxford libraries, I am pleased to express support for the proposal before the Mellon Foundation to develop optical character recognition (OCR) software that is capable of generating accurate electronic text from digital images of early modern books.

Since 1999, the TCP has produced more than 40,000 manually keyed, SGML-encoded electronic texts based on the titles in the Early English Books Online (EEBO) database. The TCP will gladly share these text files with the Initiative for Digital Humanities, Media, and Culture (IDHMC) at Texas A&M University that they may use this text for the purposes of creating "ground truth" in order to create and successfully train early modern OCR engines. We have already shared a similar set of documents that we keyed from the Eighteenth Century Collections Online (ECCO) collection, since we were given permission by Gale-Cengage Learning to release those files to the public. We will distribute our keyed texts in their original SGML format: it is the responsibility of Texas A&M's IDHMC to convert them to a format that is usable for their purposes.

The TCP believes deeply in the value of generating a comprehensive, accurate electronic corpus of early modern texts. However, until better OCR tools exist, manual keying is the only way to carry out this work. We therefore wholeheartedly support this proposal, whose results have the potential to change the landscape of the TCP's work. We would be glad to assist with testing the proposed OCR engines, or with proofing and reviewing the texts they produce. Further, we would be interested in using such OCR engines, and in incorporating accurate early modern texts produced by them into the EEBO-TCP corpus. We are glad for the opportunity to support this important work.

Please feel free to contact me with any questions or for more information about the TCP's interest and involvement in this project.

Yours sincerely,

Rebecca Welzenbach
Project Outreach Librarian
Text Creation Partnership

(734) 615-0038
rwelzenb@umich.edu

Subject: REKn, ARC, INKE, & Textual Studies
Date: Wednesday, May 23, 2012 3:39:57 PM CT
From: Richard Cunningham
To: Mandell@tamu.edu
CC: Ray Siemens, martinmueller@northwestern.edu

Hi Laura,

I'm wearing my INKE Textual Studies leader hat to write you in your capacity as Director of the Advanced Research Consortium. I know you're going to be in Victoria for the DHSI—it'll be great to see you again. But I want to move along my own planning document for the next year of the INKE TS team, and I would like to suggest we think about creative and productive ways to incorporate existing REKn resources into the TS year 4 plan. Once the TS plan is approved, I will fund an English MA student at my university, and if there is some way we could put her to work such that she adds value to REKn and thereby ARC while simultaneously contributing to my own work as it relates to the scholarly edition and the social edition, I think that would be great. That was kind of an arduous sentence. Let me try again: I'll have a GRA working for me; how can we plan for her to make use of REKn resources in such a way that helps you?

That is, I am not writing to ask for money, or for anything more than access to texts. I hope that we can use my GRA funding this year to create a productive link between INKE and ARC's recently "acquired" knowledge-base, REKn.

I look forward to hearing back, and later to seeing you in Victoria.

Cheers,
Richard

Subject: RE: Invitation
Date: Monday, March 19, 2012 7:04:35 PM CT
From: Rob Hume
To: 'Mandell, Laura'

Dear Laura (if I may),

Thanks for this. I am definitely interested in this project. Improving the texts in ECCO (and especially in EEBO) is something I am very anxious to see happen. Also to see corrections made to ESTC (with the hope that EEBO and ECCO can be improved as we chip the disasters out of ESTC). I am working right now on play publication 1660-1800 and suffering badly from uncertainty as to what ESTC lists as (say) five different editions are really five, or perhaps four, or three, or two, or just one. I could certainly spend three days in College Station on the terms you describe. Spring 2013 would be a good time for me: I haven't taught in the spring in 30 years, and my schedule is fairly flexible. I think overlapping with James Raven would be a very good idea. We don't "work together," but we do similar kinds of work (a couple of my current thesis students are using his scholarship extensively) and I suspect that we would have productive conversations. Best wishes, Rob Hume

Robert D. Hume
Evan Pugh Professor of English Literature
33 Burrowes Building
Penn State University
University Park, Pa 16802
USA
www.personal.psu.edu/hb1 [h-b-arabic one]

-----Original Message-----
From: Mandell, Laura [mailto:mandell@tamu.edu]
Sent: Monday, March 19, 2012 5:14 PM
To: Rob-Hume@psu.edu
Subject: Invitation

Dear Robert Hume:

I'm director of 18thConnect.org - I met you at last year's ASECS while we were both speaking with Gale Cengage Learning. 18thConnect is embarking on a big OCR project to improve the texts in both amount and readability of ECCO and EEBO. I'm writing to ask whether you would be willing and able to participate in a Mellon-funded project during Spring 2013? The project is called, "OCR'ing Early Modern Books," and we are trying to adapt a tool called "Cobre" so that it can be used to crowd-source corrections to OCR. The "crowd" we imagine are experts in early modern and eighteenth-century culture, like yourself, and we will be asking them to disambiguate entries for books in the ESTC Catalogue. This tool allows for looking at and correlating multiple editions of the "same" text. Obviously, what has been called "the same book" during the early modern period is quite problematic, and we would like people to work comparing not just title pages but entire book contents. While the tool has already been developed to use in one project, the Primeros Libros project (http://www.primeroslibros.org/), we will be adapting it for working with EEBO and ECCO texts as well as newer page-image scans made by google books and residing in the HathiTrust. We need expert advice, like yours, about how to design the tool to capture the most bibliographic information and to make it enjoyable for book-history and early-modern-culture experts to use. We would invite you to come for 3 days to College Station, TX, to be trained in using the tool and to give us your feedback. I include a picture of it below. We can offer you an honorarium of $1,000, if that is enough, plus of course pay for all travel and subsistence expenses.

I have invited James Raven as well who has said that he can come; ideally we will be able to coordinate your visits, as I think you said that you work with him?

Best, Laura
--
Laura Mandell
Professor of English
Director, Initiative for Digital Humanities, Media, and Culture Texas A&M University
237 Blocker, MS 4227
College Station, TX 77843-4227
(979) 845-8345
mandell@tamu.edu

Subject: FW: query
Date: Monday, March 19, 2012 5:24:56 PM CT
From: Todd Samuelson
To: Mandell, Laura

Laura,

Here is the response I received from James. As you can see from my initial letter, I did not get too far into the logistics of his visit, but the good news is that he seems interested in and open to coming. I think that it's safe to include him in the grant proposal. I am happy to write back to him with some of the specific details about a visit, but I wanted to check with you to see what would be best. His interest seems to be primarily in the technical problem of recognizing particular types, but I suppose that if we train the engines on specific typefaces, one outgrowth of the project may help with that issue. Also, how would you feel about potentially trying to book him for a lecture to be held in Cushing Library (or another venue)?

Thanks,
Todd.

-----Original Message-----
From: James Mosley [mailto:james.mosley@nonpareil.demon.co.uk]
Sent: Friday, March 16, 2012 8:44 AM
To: Todd Samuelson
Subject: Re: query

Dear Todd,

It's good to hear from you. I enjoyed last year's teaching, which I felt went well. For a number of reasons, including such simple ones as a lack of enthusiasm for too many long-haul flights, I am not doing a follow-up class at Charlottesville, at least for the summer of this year.

But the project you outline has some intriguing aspects. For some time I have wondered if current imaging technology would eventually let us recognise specific types. At all events I should be glad to discuss the possibilities of the project that you mention.

With all good wishes,

James

In message
<9C3A9B2D7DC5DA4EB4941FB703890DC81B7ECE9C@EXCH-MAILBOX3.library.tamu.edu>
, Todd Samuelson <toddsamuelson@library.tamu.edu> writes

> Dear James,
>
> I hope that you have been well since last year?s Rare Book School course. It was a great pleasure to meet you then. I very much enjoyed the class, and have continued to feel enriched from the opportunity to

learn from your expertise.

I?m writing because I?ve become involved in a book history project which I hoped to discuss with you. Texas A&M University has a Digital Humanities center whose director, Laura Mandell, has been putting together a large collaborative project drawing upon handpress-period typography. Essentially, she is attempting to use various technical processes to improve optical character recognition software to read 16th-18 th century books at a high level of accuracy. She has partnered with representatives from companies such as Google and Proquest, academic institutions like King?s College London and the University of Chicago, and libraries such as the National Library of the Netherlands. Ultimately, she has produced a proposal for a $600,000 Mellon Grant to examine the problem and produce technical solutions. Mellon has responded enthusiastically with a request for revisions, and will respond early this summer with their decision regarding funding.

The reason that I?m writing to you is that Laura is interested in inviting you to visit Texas A&M for three or four days, potentially in September or later this fall, to discuss some of the historical typographical issues with the team. Your visit might include, if you found the possibility agreeable, an event in which you could give a talk about some issue of typography ? but it could also involve simply the opportunity to consult with Dr. Mandell and other faculty members about some of the issues related to the project.
Before discussing issues of logistics such as travel expenses and honorarium, I wanted to check with you to see if this visit might fit into your plans for the Fall, and whether you would consider the invitation. We would be very excited to have you on campus; I understand that you gave a lecture at the Harry Ransom Center in 1977, but have you been to Texas since?

Thank you for considering the possibility; I look forward to hearing from you once again,

Best,

Todd.

Todd Samuelson, Ph.D., C.A.

Curator of Rare Books and Manuscripts

Director, Book History Workshop

> Cushing Memorial Library & Archives
>
> Texas A&M University
>
> 5000 TAMU
>
> College Station, Texas 77843-5000
>
> Tel. 979.845.1951
>
> Fax. 979.845.1441

--

Subject: RE: Invitation
Date: Tuesday, March 13, 2012 5:39:52 PM CT
From: Raven, James
To: Mandell, Laura

Dear Laura
Thank you for this - forgive my brief reply as I am just off to bed to prepare for a flight to Los Angeles/Santa Barbara tomorrow, but yes, this sounds fascinating and I would certainly be happy in principle to help you over this on the terms you suggest - it depends a little on the dayes you have in mind but I should be able to work the diary. I am very interested in this type of project
All good wishes
James Raven

From
Professor James Raven MA (Cantab) MA (Oxon) PhD LittD (Cantab) FSA FRHistS

From: Mandell, Laura [mandell@tamu.edu]
Sent: 13 March 2012 21:48
To: Raven, James
Subject: Invitation

Dear James Raven:

I'm writing to ask whether you would be willing and able to participate in a Mellon-funded project during Spring 2013? The project is called, "OCR'ing Early Modern Books," and we are trying to adapt a tool called "Cobre" so that it can be used to crowd-source corrections to OCR. The "crowd" we imagine are experts in early modern and eighteenth-century culture, like yourself, and we will be asking them to disambiguate entries for books in the ESTC Catalogue. This tool allows for looking at and correlating multiple editions of the "same" text. Obviously, what has been called "the same book" during the early modern period is quite problematic, and we would like people to work comparing not just title pages but entire book contents. While the tool has already been developed to use in one project, the Primeros Libros project (), we will be adapting it for working with EEBO and ECCO texts as well as newer page-image scans made by google books and residing in the HathiTrust. We need expert advice, like yours, about how to design the tool to capture the most bibliographic information and to make it enjoyable for book-history and early-modern-culture experts to use. We would invite you to come for 3 days to College Station, TX, to be trained in using the tool and to give us your feedback. I include a picture of it below. We can offer you an honorarium of $1,000, if that is enough, plus of course pay for all travel and subsistence expenses.

Thank you so much for considering this request.
Sincerely,
Laura Mandell

--
Laura Mandell
Professor of English
Director, Initiative for Digital Humanities, Media, and Culture
Texas A&M University
237 Blocker, MS 4227
College Station, TX 77843-4227[cid:4F469635-A9B1-4A16-86F5-F130AE36CF51]
(979) 845-8345
mandell@tamu.edu

Laura Mandell <laura.mandell@gmail.com>

OCR project

1 message

Ted Underwood <ted.underwood.3@gmail.com> Tue, Jun 5, 2012 at 7:17 AM
To: Laura Mandell <laura.mandell@gmail.com>

Dear Laura,

I was very glad to hear your proposal for improving OCR on eighteenth-century text, and am writing now to express my willingness to help with the project.

My contribution will begin after the first OCR pass, when we have an initial version of the text, and need to identify options for further improvement. We will probably want to consider both conservative and proactive kinds of improvement. Conservatively, we can identify specific corrupt passages for rescanning or crowdsourced correction. More proactively, we can begin a versioning process that automatically corrects the whole collection for immediate research purposes (without ever displacing the underlying, conservatively corrected text).

Both of these approaches will require lexical information and scripts that are specific to the eighteenth century. (For instance, the frequency of apostrophe-d is much higher in eighteenth-century texts than elsewhere, because of syncopated forms like "beg'd." When we're assessing OCR quality, we will need to realize that this character combination does not necessarily indicate an OCR error.)

By next summer (2013), I will be able to contribute a fully-developed workflow that can a) conservatively, assess the quality of eighteenth-century OCR to identify passages for rescanning or crowdsourcing and b) more proactively, generate a corrected version of the text for immediate research purposes.

Mike Black and I are currently developing these workflows to correct existing, public-domain 18c OCR. We have already developed Python scripts that assess OCR quality, with special attention to the peculiarities of eighteenth-century English. We are now developing a more ambitious set of Python and Java modules that

a) Identify period spellings and proper nouns peculiar to the eighteenth century (tokens that should not mistakenly be flagged as errors),
b) Generate a full ruleset for correction of common, unambiguous OCR errors ("expreflion" => "expression," "corre6ed" => "corrected").
c) Use local context to assess possible errors that are too ambiguous to be handled by correction rules. For instance "flips" is a real word. But in 18c OCR it is very often an error for "ships." The only way to correct errors like this is to examine context. If "flips" is preceded by "the" and followed by "sailed" or "failed," then we're probably looking at "The ships sailed." Last summer, I developed a Python script that does this contextual correction using 2gram frequencies; as we generate a more extensive library of 18c ground truth this summer, we will refine its performance.

I have found that this workflow can generally produce a 10% improvement on existing 18c OCR, taking it from 80% accuracy to 90% accuracy. When you have developed an OCR process that is more suited to eighteenth-century pages and fonts, we will have much better underlying text (at, say, 85-90% accuracy); with that kind of input I find that automatic correction can produce 98% or 99% accuracy on most volumes. (This is what I am now able to achieve on early-19c volumes, even with the long s.) For many scholarly purposes, this accuracy will still not be sufficient; so, over the long term, we will need to rely on crowdsourcing supported by TypeWright. But for some kinds of search and text-mining, 98% accuracy is already very good indeed, so it will probably make sense to generate a tentatively-corrected version of the collection for immediate research use.

I will make my code and 18c lexical information available, and help you adapt them to support the OCR process (to identify passages for rescanning or crowdsourcing). I would also be glad to help you develop a tentatively-corrected version of the resulting collection for immediate research use.

Yours,

Ted Underwood

Associate Professor of English, University of Illinois

Brief Biographies of the Main Grant Participants:

Dr. Apostolos Antonacopoulos is director of PRImA Labs, Pattern Recognition and Image Analysis Research Laboratory at the University of Salford, Manchester, UK. He developed Aletheia, a tool for the semi-automated layout analysis of documents. He is Senior Lecturer and Head of the Pattern Recognition and Image Analysis (PRImA - www.primaresearch.org) Laboratory in the School of Computing, Science and Engineering at the University of Salford. He received his PhD from the University of Manchester Institute of Science and Technology (UMIST), UK in 1995. He received the International Association for Pattern Recognition (IAPR)/ICDAR Young Investigator Award in 2005 for his "Outstanding service in the field of Document Analysis and Recognition and his innovative research on the Analysis of Historical Documents". Apostolos has worked and published extensively on various problems in Document Analysis and in Pattern Recognition and applications. He is a member of the Editorial Boards of the International Journal on Document Analysis and Recognition (IJDAR) and the Electronic Letters on Computer Vision and Image Analysis (ELCVIA) journal, and Chair or member of a number of high-profile International Committees. He is a member of program committees of most conferences in the field of Document Analysis and Recognition and he has co-edited the first special issue on the Analysis of Historical Documents in IJDAR. He has significant experience in leading and participating in national, European and industry-sponsored projects. Notable EU projects are the current CIP-ICT-PSP *European Newspapers* and the FP7 IP *IMPACT* projects, and the completed FP5 *MEMORIAL* project.

Loretta Auvil, Boris Capitanu, and the SEASR Services project. Another Mellon-funded project, the Software Environment for the Advancement of Scholarly Research (SEASR), created the Meandre Workbench, the platform upon which the MONK project for data-mining novels was built. This team developed post-processing analysis of OCR text in order to clean up the Google N-grams data for faculty use.

Loretta Auvil works at the Illinois Informatics Institute (I3) at the University of Illinois at Urbana Champaign. She received a MS in Computer Science from Virginia Tech and a BS in Applied Mathematics and Computer Science from Alderson-Broaddus College. She has worked with a diverse set of application drivers to integrate machine learning and information visualization techniques to solve the needs of research partners. She has led software development and research projects for many years. Prior to working for I3, she spent many years at NCSA on machine learning and information visualization projects and several years creating tools for visualizing performance data of parallel computer programs at Rome Laboratory and Oak Ridge National Laboratory.

Hildelies Balk, Clemens Neudecker, and the IMPACT team (Improving Access to Texts) of the National Library of the Netherlands have worked for the last ten years on OCR'ing European texts. They have created the Center for Competence as a way of offering and sustaining services to libraries who have OCR needs.

Boris Capitanu is a research programmer working in the Illinois Informatics Institute (I3) at the University of Illinois at Urbana Champaign. Boris holds a B.S. and M.S. in Computer Science from University of Illinois at Urbana-Champaign. His research interests include data mining and educational technologies. Boris is currently working on the SEASR project creating software platforms for the advancement of scholarly research.

Anton duPlessis, MA, Curator of the Colonial Mexican Collection at Cushing Library and Director of the Primeros Libros Project, and principle developer of the Cobré tool.

Co-PI: **Richard Furuta** is a faculty member at Texas A&M University where he is a Professor in the Department of Computer Science, Director of the Center for the Study of Digital Libraries, and Director of the Hypermedia Research Laboratory. He received the B.A. degree from Reed College in 1974, the M.S. degree in Computer Science from the University of Oregon in 1978, and the Ph.D. degree in Computer Science from the University of Washington in 1986. Dr. Furuta's current areas of research include digital libraries, digital humanities, hypermedia systems and models, structured documents, and document engineering. He also has studied applications in computer supported cooperative work, software engineering, visual programming, document structure recognition from bitmapped sources, and management systems for three-dimensional-gesture-based user interfaces. In the area of Digital Libraries, he was one of the founders of the 1994 and 1995 Digital Libraries Conferences, which subsequently became the ACM Digital Libraries series, and later merged with the IEEE-CS series to form the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL). He was program chair for JCDL 2009 and ACM Digital Libraries 2000. He currently serves on the Steering Committee for ACM/IEEE-CS JCDL and was its chair from 2001-2005. He also is an Editor-in-Chief of the International Journal of Digital Libraries. Many of Dr. Furuta's current research projects are highly interdisciplinary, especially those in the area of Digital Humanities. These current projects include the Cervantes Project, centered on the iconic author of Don Quixote, the Picasso Project, which is creating a digital reasoned catalog that already contains more than 10,000 of Picasso's art works, and the Nautical Archaeology Digital Library, in conjunction with the campus' Institute for Nautical Archaeology. In other technical activities, Dr. Furuta has been co-program chair for ACM Document Engineering 2002, co-program chair for ACM Hypertext '93, program chair for Electronic Publishing '90, co-program chair of the 1991 DC ACM Chapter annual symposium, Chair of ACM SIGLINK from 1993–1995, member of the ACM SIG Board/SIG Governing Board Executive Committee from 1997–2001, and has served on many other program committees, conference committees, steering committees, and editorial boards.

Co-PI: **Ricardo Gutierrez-Osuna** is an associate professor in the Department of Computer Science at Texas A&M University. A native of Spain, he received a BS in Electrical Engineering from the Polytechnic University of Madrid in 1992, and MS and PhD degrees in Computer Engineering from North Carolina State University in 1995 and 1998, respectively. He has diverse research interests that revolve around machine learning, signal processing, and intelligent sensors. His current research projects are in speech processing for foreign-accent conversion, active perception with chemical sensor arrays, stress monitoring with wearable sensors, and perception of 3D facial caricatures (by humans). His previous research projects involved biologically inspired algorithms for artificial olfaction, audiovisual speech processing, and probabilistic navigation for mobile robotics.

Dr. Jacob Heil is currently a Postdoctoral Associate at Texas A&M University. He has been working with the IDHMC staff to create a database of the watermarks in the printed works of John Donne. This project may become part of the Digital Donne (http://digitaldonne.tamu.edu/).

PI: **Laura Mandell**, Professor of English Literature and Director of the Initiative for Digital Humanities, Media, and Culture at Texas A&M University, has published *Misogynous Economies: The Business of Literature in Eighteenth-Century Britain* (1999), a Longman Cultural Edition of *The Castle of Otranto* and *Man of Feeling*, and numerous articles primarily about eighteenth-century women writers. Her recent article in *New Literary History*, “What Is the Matter? What Literary History Neither Hears Nor Sees,” describes how digital work can be used to conduct research into conceptions informing the writing and printing of eighteenth-century poetry. She is Editor of the Poetess Archive, an online scholarly edition and database of women poets, 1750-1900 (http://www.poetessarchive.org), and Director of 18thConnect (http://www.18thConnect.org). Her current research involves developing new methods for visualizing poetry, developing software that will allow all scholars to deep-code documents for data mining, and improving OCR software for early modern and eighteenth-century texts via high performance and cluster computing. She developed the TypeWright crowd-sourced correction tool through Mellon support to Miami University. Mandell is currently a member of the Text Encoding Initiative Executive Board, and her IDHMC will be hosting the next member's meeting in November 2012.

R. Manmatha is a research associate professor in the Dept. of Computer Science at the University of Massachusetts, Amherst and is also part of the Center for Intelligent Information Retrieval. His research interests are in the areas of image and video retrieval and on the recognition and retrieval of printed and handwritten books. He has previously worked on the automatic annotation and retrieval of images and he and his students developed the first automatic search engine for historical handwritten scanned documents—specifically those of George Washington. He is currently an associate editor of PAMI and Pattern Recognition Letters and has served on several conference committees. He is a co-PI on a large NSF project investigating new techniques for text search, mining and OCR on a large collection of scanned books in collaboration with the Internet Archive and the Tufts University. He spent summer 2005 at Google as a visiting research scientist working on Google Books. He co-founded SnapTell, a mobile image search company which was acquired by Amazon. He is currently a consultant to A9/Amazon.

Dr. James Mosley has written extensively on variations in the font-face used on 80% of all eighteenth-century documents and one of our foremost historians of typography.[1] He worked for many years as the Director of the St. Bride Library and Institute, one of the world's great repositories of typographical history: http://www.stbride.org/. Their special collections contain remarkable holdings of physical material and archives related to England's typographical past. Dr. Moseley will advise Todd Samuelson and postdoctoral fellow Jacob Heil how to trace font movements and build a database about them that feeds into maps.

Performant Software Solutions LLC specializes in developing web applications and interactive user experiences. The multi-disciplinary team has years of experience supporting digital humanities research, and has developed projects funded by the Andrew W. Mellon Foundation and the National Endowment for the Humanities. Performant uses agile software development practices and open-source technologies that are maintainable and cost effective. Performant's software development process is focused on effective communication, transparency, and results.

Dr. **Todd Samuelson** is Curator of Manuscripts and Rare Books at Cushing Library. He directs and teaches at the Book History Workshop at Texas A&M University.

[1] http://typefoundry.blogspot.com/2009/01/recasting-caslon-old-face.html

http://typefoundry.blogspot.com/2009/07/lost-caslon-type-long-primer-no-1.html

This contract--a modified form of the NINES contract with contributing proprietary collections--promises to let people have texts that they correct (see news release, after Appendix)

TEXT AND METADATA SHARING AGREEMENT

THIS TEXT AND METADATA SHARING AGREEMENT (this "Agreement") is entered into by and between GALE-CENGAGE LEARNING, and NINES (Networked Interface for Nineteenth-century Electronic Scholarship) ("NINES") / 18THCONNECT as of the Contract Date specified below (the "Effective Date").

WHEREAS, GALE-CENGAGE LEARNING's mission is to help the scholarly community take advantage of advances in information technologies;

WHEREAS, NINES/18THCONNECT seeks to enable work in digital scholarly media to be produced, vetted, published, and recognized by the discipline and to provide scholars with the ability to conduct research within a federated digital environment of such work; and

NOW, THEREFORE, in consideration of the premises and the mutual promises contained herein and other good and valuable consideration, the receipt and sufficiency of which are hereby acknowledged, the parties agree as follows:

1. DEFINITIONS

For purposes of this Agreement, the following definitions shall apply:

1.1 "Typed Plain Text" shall mean the text that has been manually re-keyed and reviewed by the Text Creation Partnership (TCP).

1.2 "OCR Plain Text" shall mean the text files derived from page images (1.6) by 18THCONNECT using open-source Optical Character Recognition Programs (1.8), and any corrections to that text provided by 18THCONNECT's membership.

1.3 "ECCO" shall mean the database published by GALE-CENGAGE LEARNING called Eighteenth-Century Collections Online and Eighteenth Century Collections Online, Part II (ECCO).

1.4 "Metadata" shall mean bibliographic information describing the objects contained in the resource. Metadata has been provided by the English Short Title Catalogue (ESTC).

1.5 "Project" shall mean NINES's/18THCONNECT's use of the Typed Plain Text as given to 18THCONNECT by the TCP, OCR Plain Text as created by 18THCONNECT, and Metadata as given to NINES/18THCONNECT by the British Library for indexing and enabling scholars to search and locate content in the resource in connection with NINES's/18THCONNECT's integrated environment for aggregated, peer-reviewed research and online scholarship.

1.6 "Page Images" shall mean the digitized page images of the texts in the ECCO Collection.

1.7 "OCR Results" shall refer to the OCR xml (as described in Appendix 1) and word coordinates generated by Gamera and OCRopus. These "OCR Results" shall be returned to GALE-CENGAGE by NINES/18THCONNECT in the following xml format:

- <p> tag used to separate paragraphs within the text
- <wd pos="X1,Y1,X2,Y2">OCR_WORD<wd>
    - X1,Y1 = is the pixel coordinate of the upper left hand corner of the bounding box containing the word
    - X2,Y2 = is the pixel coordinate of the bottom right hand corner of the bounding box containing the word

- OCR_WORD is the OCR'd word from the page

Example:

```
-----------------------------------------------------------------------------------
<p>
        <wd pos="393,172,762,214">REPORTS</wd>
</p>
<p>
        <wd pos="542,342,566,354">ON</wd>
        <wd pos="579,341,620,354">THE</wd>
        <wd pos="142,468,350,518">FISHES,</wd>
        <wd pos="388,469,658,509">REPTILES</wd>
        <wd pos="697,468,816,508">AND</wd>
        <wd pos="852,468,1017,508">BIRDS</wd>
</p>
----------------------------------------------------------------------------------
```

1.8 "Open-source Optical Character Recognition Programs" shall refer to the programs used by the 18THCONNECT team to generate "OCR Plain Text," currently but not limited to Gamera and OCRopus.

1.9 "Open Source Libraries" shall refer to the modifications made to open-source Optical Character Recognition Programs by the 18THCONNECT team as it attempts to derive better results from the Optical Character Recognition Process.

2. GRANT OF LICENSE

2.1 GALE-CENGAGE LEARNING hereby grants NINES/18THCONNECT a limited, non-exclusive, royalty-free right to use the Typed Plain Text, OCR Plain Text and Metadata solely in connection with the Project. NINES/18THCONNECT may store Metadata and OCR Plain Text on its servers for the purpose of indexing and data mining (data mining to be restricted to GALE-CENGAGE LEARNING'S ECCO customers only) and shall display them to users only in short, query-dependent texts or data ("Snippets"). Snippets may include Metadata as well as a few lines of query-dependent Full Text and are not likely to exceed three (3) lines at a reasonable font size. NINES/18THCONNECT will also display as part of the Project a link to a page on the GALE-CENGAGE LEARNING website for ECCO, your website here_____, where GALE-CENGAGE LEARNING will make the content referenced by the Snippet available to users authorized by subscription to access the ECCO archive. For avoidance of doubt, this Grant of License does not permit any person or party to access any Gale product (for example, ECCO) unless such person or party is an authorized user as determined by appropriate license and purchase agreements.

2.2 NINES/18THCONNECT shall not under any circumstances redistribute the Typed Plain Text, OCR Plain Text or Metadata to any third party, agent, or affiliate except as set forth in Article 2.1 herein and integrated into NINES's/18THCONNECT's search-and-browsing interface that is part of the Project, unless such redistribution is explicitly in writing authorized by GALE-CENGAGE LEARNING.

2.3 NINES/18THCONNECT will use the ECCO Metadata as supplied by the ESTC, and 18THCONNECT will extract OCR Plain Text from page images delivered to the NCSA (National Center for Supercomputer Applications) by GALE-CENGAGE for 18THCONNECT following the execution of this Agreement.

2.4 NINES/18THCONNECT will return the OCR Results to GALE-CENGAGE LEARNING, per an agreed-upon schedule but not less than annually, for use within the ECCO database at no cost/obligation to GALE-CENGAGE LEARNING, should GALE-CENGAGE LEARNING choose to use the OCR Results.

2.5 NINES/18THCONNECT and GALE-CENGAGE LEARNING agree to handling text correction by scholars of OCR Results that will create OCR Plain Text per the attached Appendix 1.

2.6 GALE-CENGAGE LEARNING may request removal of the Typed Plain Text, the OCR Plain Text or Metadata or any part thereof from the Project at any time and for any reason whatsoever, by sending a request to the NINES/18THCONNECT Notice Contact including the specific URLs to be removed. On receipt of such a request, NINES/18THCONNECT will make reasonable efforts to remove and cease the display of the OCR Plain Text pertaining to the URLs within sixty (60) days. Notwithstanding the foregoing, GALE-CENGAGE LEARNING may make an emergency request to NINES/18THCONNECT for the removal of OCR Plain Text or any part thereof from the Project. In the event of an emergency request, NINES/18THCONNECT shall make reasonable efforts to remove and discontinue the display of the OCR Plain Text pertaining to the URL's specified within five (5) business days.

2.7 During the term of this Agreement, representatives of GALE-CENGAGE LEARNING and NINES/18THCONNECT shall cooperate in discussing the progress of the Project on an as-needed basis.

3. PAYMENTS AND FEES

Nothing in this Agreement shall be construed as providing for compensation to either party for its services under this Agreement, and each party shall bear the costs of performing its obligations hereunder.

4. NON-DISCLOSURE

Except as set forth in Section 2.1 herein, in connection with the use of OCR Plain Text and Metadata for purposes of conducting the Project, NINES/18THCONNECT shall not, during the term of this Agreement or any time thereafter, disclose, or permit any of their employees, agents, or assignees to disclose, to any other person or entity any Confidential Information (as defined below) of the other party. Confidential information does not include information which (a) was or becomes generally available to the public other than as a result of a disclosure by the receiving party or its representatives or (b) was or becomes available to the receiving party on a non-confidential basis from a source other than the disclosing party or its advisers, provided that such source was not known by the receiving party to be bound by any agreement to keep such information confidential, or otherwise prohibited from transmitting the information to the receiving party by a contractual, legal or fiduciary obligation.

5. PUBLICITY

Either party may issue a press release regarding the existence of this Agreement, but not without prior written authorization from the other party. Such authorization shall not be unreasonably withheld, and the parties shall make best efforts to reply to such requests within ten business days.

6. INTELLECTUAL PROPERTY

Subject to the intellectual property rights of third parties, GALE-CENGAGE LEARNING shall retain all rights to the OCR Results. Further, GALE-CENGAGE shall have the right to determine the use of the Typed Plain Text, the OCR Plain Text and Metadata, and will have unlimited access to the Open Source Libraries. Except as expressly granted herein, under no circumstances shall anything in this Agreement be construed as granting, by implication, estoppel, or otherwise, to one party a right or license to the other party's name, logo, design marks, trade names, or service marks (collectively, the "Trademarks"), issued patents and patent applications, copyrights and copyrights registration and applications, rights in ideas, designs, works of authorship, derivative works, or any other rights or license relating to the other party's intellectual property or services.

7. TERM AND TERMINATION

7.1 This Agreement shall continue in effect for the length of the Project, unless otherwise terminated in writing by a party or by the parties mutually, with or without cause.

7.2 Upon termination of this Agreement, all rights granted hereunder shall be terminated, except that the rights obligations set forth in Sections 2.5, 4, 5, and 9 shall survive termination, as shall any other provisions which by their terms are intended to survive termination. On termination NINES/18THCONNECT immediately shall remove or deactivate all links within its control built using the Metadata, cease using GALE-CENGAGE LEARNING Trademarks in association with NINES's/18THCONNECT's product, and forthwith provide a written statement that the aforementioned deactivation has taken place.

8. USE OF TRADEMARK

NINES/18THCONNECT and GALE-CENGAGE LEARNING agree to adhere to standards consistent with the high level of quality associated with one another's Trademarks. Neither party shall use any Trademarks or other database content to make it appear that the other party is endorsing, sponsoring, or recommending the information, goods, or services in its web site. Neither party shall suggest the other party is affiliated with it, its advertisers, or other entities to which it is linked, and neither party will otherwise misrepresent its relationship with the other party or present false or misleading information about the other party's products or services. Neither party shall use the Trademarks, or other materials, in a manner that is likely to cause confusion with, dilute, or damage the reputation of the other party or its database. Neither party shall place the other party's web pages in a "frame" within that party's web site, use other techniques that alter or obstruct the visual or other presentation of the other party's database in whole or in part, or otherwise impose editorial comment or commercial material or any other type of identification on or in proximity to content displayed in the other party's database without written permission from an authorized representative of the other party. Neither party shall use the name or logo of the other party, its participants, or content in any "metatag" without the other party's express written permission.

9. NOTICES

All notices given pursuant to this Agreement shall be in writing and sent to the Notice Address specified below. These addresses also are applicable for general contact between the parties. Notices may be delivered by hand, by overnight carrier, or shall be deemed to be received within five (5) business days after mailing if sent by registered or certified mail, postage prepaid. If any notice is sent by facsimile, then confirmation copies must be sent as specified above. Either party may from time to time change its Notice Address by written notice to the other party.

If to GALE-CENGAGE LEARNING:

Your contact info here
__Jim Draper_________________
__VP and Publisher____________
__Gale – Cengage Learning_____
__27500 Drake Rd.___________
__Farmington Hills, MI 48331___
(tel): 248-699-8519.__________
(fax): _______________
(email): jim.draper@cengage.com

If to NINES/18THCONNECT:

Laura Mandell
Associate Director of NINES / Director of 18THCONNECT
_Department of English_______

_356 Bachelor Hall___________
_Miami University___________
Oxford, OH 45056 USA_______
(tel): *(513) 529-5276*_________
(fax): *(513) 529-1392*_________
(email): mandellc@muohio.edu__

10. REPRESENTATIONS AND WARRANTIES

10.1 Each party hereby represents and warrants that it is duly organized and validly subsisting and has full authority to enter into this Agreement and to bind the party to the terms and conditions herein. Each party further represents and warrants that it has caused this Agreement to be executed by a duly authorized representative.

10.2 **OTHER THAN THE EXPRESS WARRANTIES STATED IN THIS SECTION 9, USE OF THE OCR PLAIN TEXT AND METADATA ARE PROVIDED ON AN "AS IS" BASIS, AND GALE-CENGAGE LEARNING DISCLAIMS ANY AND ALL OTHER WARRANTIES, CONDITIONS, OR REPRESENTATIONS (EXPRESS, IMPLIED, ORAL OR WRITTEN), RELATING TO THE FULL TEXT AND METADATA OR ANY PART THEREOF, INCLUDING, WITHOUT LIMITATION, ANY AND ALL IMPLIED WARRANTIES OF QUALITY, PERFORMANCE, COMPATIBILITY, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. NEITHER PARTY MAKES ANY WARRANTIES RESPECTING IMPROPER USAGE OF ITS RESPECTIVE DATABASE BY USERS OR ANY HARM THAT MAY BE CAUSED BY THE TRANSMISSION OF A COMPUTER VIRUS, WORM, TIME BOMB, LOGIC BOMB OR OTHER SUCH COMPUTER PROGRAM.**

10.3 GALE-CENGAGE LEARNING shall not be liable for any loss, injury, claim, liability or damage of any kind resulting from the unavailability of its database, interruption of the services provided hereunder, or arising out of or in connection with the its linking activities. Neither party shall be liable for any indirect, special, incidental, punitive or consequential damages, including but not limited to loss of data, business interruption, or loss of profits, even if advised of the possibility of a claim. Neither party shall have any responsibility or liability for any content appearing in the other party's database or for any link to or from third party web sites.

11. MISCELLANEOUS

11.1 This Agreement constitutes the entire agreement of the parties and supersedes all prior communications, understandings and agreements relating to the subject matter hereof, whether oral or written. No modification or claimed waiver of any provision of this Agreement shall be valid except by written amendment signed by authorized representatives of GALE-CENGAGE LEARNING and Licensee.

11.2 This Agreement and any amendments may be executed in one or more counterparts, each of which shall be deemed an original, but all of which together shall constitute one agreement.

11.3 Nothing contained herein shall be deemed to create an agency, joint venture, or partnership relationship between the parties. Neither party shall use or attempt to exercise apparent authority concerning the other party in its dealings with third parties.

11.4 Neither party shall be liable in damages or have the right to cancel this Agreement for any delay or default in performing hereunder if such delay or default is caused by conditions beyond its control including, but not limited to Acts of God, Government restrictions (including the denial or cancellation of any export or other necessary license), wars, insurrections, strikes, fires, floods, work stoppages, unavailability of materials, carriers or communications facilities, and/or any other cause beyond the reasonable control of the party whose performance is affected.

11.5 Waiver of any provision herein shall not be deemed to be a waiver of any other provision herein, nor shall waiver of any breach of this Agreement be construed as a continuing waiver of other breaches of the same or other provisions of this Agreement.

11.6 If any provision or provisions of this Agreement shall be held to be invalid, illegal, unenforceable or in conflict with the law of any jurisdiction, the validity, legality and enforceability of the remaining provisions shall not in any way be affected or impaired thereby.

11.7 Neither party may assign, directly or indirectly, all or part of its rights or obligations under this Agreement without the prior written consent of the other party, which consent shall not be unreasonably withheld or delayed, except however that GALE-CENGAGE LEARNING may make such assignment to one or more of its affiliated entities.

NINES / 18THCONNECT

BY: Laura Mandell

NAME: Laura C. Mandell

TITLE: Prof. / Assoc. Dir. / Dir.

DATE: 22 June 2010

GALE-CENGAGE LEARNING

BY:

NAME: Jim Draper

TITLE: Vice President and Publisher

DATE: 6 23 2010

Appendix 1.

A.1 DEFINITIONS

A.1.1 "Crowd-sourced Editing Platform" shall mean any open-source interface designed to allow users to see a word unreadable in OCR Results in the context of 7-9 words on either side, currently but not limited to AnnoLex.

A.1.2 "Editor" shall mean anyone appointed by NINES/18THCONNECT to approve and accept text corrections proposed by users of NINES/18THCONNECT through its Crowd-Sourced Editing Platform.

A.1.3 "Users" shall mean any registered user of NINES/18THCONNECT.

A.1.4. "Full Text" shall mean the text of a complete work or portion of a work in the ECCO Catalogue that has been OCR'd and corrected by users of NINES/18THCONNECT.

A.2 GRANT OF LICENSE

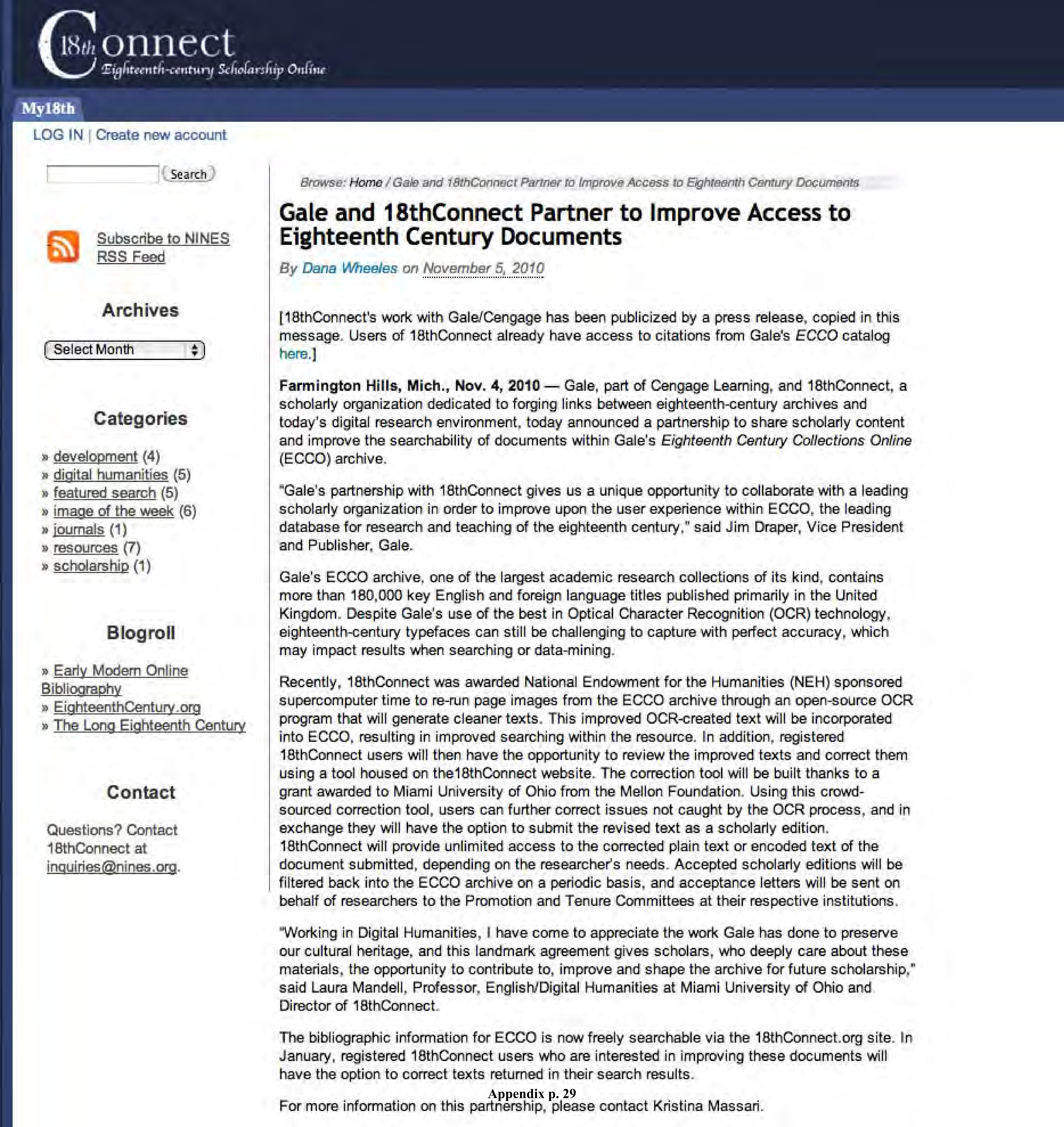
A.2.1 GALE-CENGAGE LEARNING allows NINES/18THCONNECT to display OCR Results along with the mis-scanned word image in its "Crowd-sourced Editing Platform."

A.2.2 Text corrected in the "Crowd-sourced Editing Platform" and approved by an Editor will be returned by NINES/18THCONNECT to GALE-CENGAGE LEARNING per an agreed-upon schedule but not less than annually, for use within the ECCO database at no cost/obligation to GALE-CENGAGE LEARNING, should GALE-CENGAGE LEARNING choose to use the OCR Plain Text (OCR Results + text-correction by users).

A.2.3 GALE-CENGAGE LEARNING may grant to specified users the right to correct OCR results of texts in exchange for Full Text of agreed-upon documents.

A.2.4. Users who have corrected and received Full Text from Gale may publish electronic versions of those Full Texts by submitting them to peer review at NINES/18THCONNECT.

A.2.4 GALE-CENGAGE LEARNING shall have first right of refusal to publish in print-on-demand or electronic form the Full Texts peer-reviewed by NINES/18THCONNECT.



18th Connect
Eighteenth-century Scholarship Online
My18th
LOG IN | Create new account
Search

Subscribe to NINES RSS Feed

Archives

Categories

- » development (4)
- » digital humanities (5)
- » featured search (5)
- » image of the week (6)
- » journals (1)
- » resources (7)
- » scholarship (1)

Blogroll

- » Early Modern Online Bibliography
- » EighteenthCentury.org
- » The Long Eighteenth Century

Contact

Questions? Contact 18thConnect at inquiries@nines.org.

Browse: Home / Gale and 18thConnect Partner to Improve Access to Eighteenth Century Documents

Gale and 18thConnect Partner to Improve Access to Eighteenth Century Documents

By Dana Wheeles on November 5, 2010

[18thConnect's work with Gale/Cengage has been publicized by a press release, copied in this message. Users of 18thConnect already have access to citations from Gale's *ECCO* catalog here.]

Farmington Hills, Mich., Nov. 4, 2010 — Gale, part of Cengage Learning, and 18thConnect, a scholarly organization dedicated to forging links between eighteenth-century archives and today's digital research environment, today announced a partnership to share scholarly content and improve the searchability of documents within Gale's *Eighteenth Century Collections Online* (ECCO) archive.

"Gale's partnership with 18thConnect gives us a unique opportunity to collaborate with a leading scholarly organization in order to improve upon the user experience within ECCO, the leading database for research and teaching of the eighteenth century," said Jim Draper, Vice President and Publisher, Gale.

Gale's ECCO archive, one of the largest academic research collections of its kind, contains more than 180,000 key English and foreign language titles published primarily in the United Kingdom. Despite Gale's use of the best in Optical Character Recognition (OCR) technology, eighteenth-century typefaces can still be challenging to capture with perfect accuracy, which may impact results when searching or data-mining.

Recently, 18thConnect was awarded National Endowment for the Humanities (NEH) sponsored supercomputer time to re-run page images from the ECCO archive through an open-source OCR program that will generate cleaner texts. This improved OCR-created text will be incorporated into ECCO, resulting in improved searching within the resource. In addition, registered 18thConnect users will then have the opportunity to review the improved texts and correct them using a tool housed on the18thConnect website. The correction tool will be built thanks to a grant awarded to Miami University of Ohio from the Mellon Foundation. Using this crowd-sourced correction tool, users can further correct issues not caught by the OCR process, and in exchange they will have the option to submit the revised text as a scholarly edition. 18thConnect will provide unlimited access to the corrected plain text or encoded text of the document submitted, depending on the researcher's needs. Accepted scholarly editions will be filtered back into the ECCO archive on a periodic basis, and acceptance letters will be sent on behalf of researchers to the Promotion and Tenure Committees at their respective institutions.

"Working in Digital Humanities, I have come to appreciate the work Gale has done to preserve our cultural heritage, and this landmark agreement gives scholars, who deeply care about these materials, the opportunity to contribute to, improve and shape the archive for future scholarship," said Laura Mandell, Professor, English/Digital Humanities at Miami University of Ohio and Director of 18thConnect.

The bibliographic information for ECCO is now freely searchable via the 18thConnect.org site. In January, registered 18thConnect users who are interested in improving these documents will have the option to correct texts returned in their search results.

For more information on this partnership, please contact Kristina Massari.

ProQuest®
Start here.

May 4, 2012

Donald J. Waters, Program Officer
Helen Cullyer, Associate Program Officer
The Andrew W. Mellon Foundation
140 East 62nd Street
New York, NY 10065

Dear Mr. Waters and Ms. Cullyer,

I am writing this letter of support on behalf of the *OCR'ing Early Modern Texts* project being led by Dr. Laura Mandell, Director of the Initiative for Digital Humanities, Media and Culture at Texas A&M University.

ProQuest has digitized several early modern collections, the most prominent being *Early English Books Online*. The main challenge that we face in developing these projects is the inability to create reliable searchable ASCII text using OCR technology. Searchable text is extremely valuable to scholars in the discovery of materials from this period for their research.

ProQuest is willing to collaborate with Dr. Mandell to build tools like TypeWright which offer Early Modern Texts to the crowd for the sake of having its OCR corrected by human hands. Working with REKn on this endeavor at the start will enable them to properly insure the security of our investment in these and similar texts because we will be able to approve the designs for document security before they are implemented in the tools and suggest any changes needed to make the tool's security system meet with our approval. We are also willing in principle to contract with REKn so that crowd-sourced correction tools can be used by any user who comes to the REKn site, beyond mere EBBO subscribers, because their tools will allow only for reading texts a few lines at a time, not for any direct exporting. Finally, ProQuest is willing in principle to allow REKn and JISC Historic Collections to each contribute and share corrections. The combined stream of corrected OCR will be returned to ProQuest and the Text Creation Partnership in the form required by each party, as determined by technical staff at ProQuest and the University of Michigan.

ProQuest is fully supportive of the work that is being proposed as a part of this project and believe that the results will provide great value to the scholarly community. ProQuest is willing to provide access to the digital images from *Early English Books Online* for use in this project provided the document security as outlined above is adhered to.

We look forward to participating in this project!

Best Regards,

Mary Sauer Games

Mary Sauer-Games
Vice President, Humanities Publishing
ProQuest

Subject: RE: clarification
Date: Friday, March 9, 2012 2:58:29 PM CT
From: Sauer-Games, Mary
To: Mandell, Laura

The earlier understanding, now superseded by the outcome of the April 24 meeting.

Laura,

I am imagining that the crowd would come from EEBO institutions, but not necessarily TCP institutions. EEBO institutions would be a much larger group than just TCP members and is more international in scope. We might consider allowing access to non-affiliated users if they contributed some base amount of work to the project.

Looking forward to the 24th.

mary

-----Original Message-----
From: Mandell, Laura [mailto:mandell@tamu.edu]
Sent: Friday, March 09, 2012 12:56 PM
To: Sauer-Games, Mary
Subject: clarification

Dear Mary:

Good news about our meeting: it is set to be hosted by Northwestern Libraries on April 24. We'll send a list of participants soon, but Rebecca Welzenbach has confirmed that she can be there along with me and Martin Mueller.

I have a clarification concerning your most recent letter for the Mellon grant, concerning the meaning of the following sentence:
At a minimum, ProQuest is willing to provide access to the digital images from Early English Books Online for use in this project, as well as any further development that happens post this grant, to the extent that Texas A&M and their partners have already purchased access to these materials.
This is very generous, but I just need to make certain: the "crowd" who will be using the tools to correct documents, seeing the page images that we have mounted in them -in a way that makes them inextricable - the members of that crowd might not be affiliated with any university at all. Does your sentence above allow for that, or are you imagining that the members of the crowd who make corrections must come from A&M and other TCP Partner institutions?

Thanks in advance.
Best, Laura
--
Laura Mandell
Professor of English
Director, Initiative for Digital Humanities, Media, and Culture
Texas A&M University
237 Blocker, MS 4227
College Station, TX 77843-4227
(979) 845-8345
mandell@tamu.edu

This email arranges a meeting of April 24 during which we demonstrated successfully that ProQuest would only benefit from allowing all members of the crowd to correct their texts and then, like Gale, giving a text to the person who corrects it.
--see next email.

Subject: Re: more information
Date: Thursday, May 3, 2012 2:01:46 PM CT
From: Laura Mandell
To: Sauer-Games, Mary

Mary Sauer-Games is running the Gale Contract through the ProQuest legal department so that we can sign a contract with them like Gale's by June 1.

Dear Mary:

Thanks so much for this message -- and for trying to make the June 1 timeframe. I'm attaching a zip of the original contract. There is a "Phase II" addition to Gale's contract but it involved Gale's OCR (you won't be giving us OCR, so it is not relevant). Also, I'm attaching our executive summary in case you need it. Thanks again, and so glad to hear you are home safe and sound.

Best, Laura

--
Laura Mandell
Professor of English
Director, Initiative for Digital Humanities, Media, and Culture
Texas A&M University
237 Blocker, MS 4227
College Station, TX 77843-4227
(979) 845-8345
FAX: (979) 826-2292
mandell@tamu.edu

On 5/3/12 12:25 PM, "Sauer-Games, Mary" <Mary.Sauer-Games@proquest.com> wrote:

> Laura,
>
> I am sorry for being a bit late in responding, but I was in the UK this week. I had a very good meeting with Caren Milloy at the JISC about this project and look forward to further discussions and participation.
>
> I definitely can start work on a contract and have a meeting set up next week with our legal staff. If you could re-send the Gale agreement, that would be helpful. I can't seem to find a copy of it on my laptop and I may have deleted it as I recall it was quite large. If June 1 looks to be an issue, I will let you know.
>
> All the Best,
>
> Mary
>
> -----Original Message-----

OCR’ing Early Modern Texts

Nick Laiacona and Kristin Jensen, Performant Software Solutions LLC
May 2, 2012 Document Version 1.4

The Project

OCR’ing Early Modern Texts is a multi-institutional project led by Laura Mandell and the Initiative for Digital Humanities, Media, and Culture (IDHMC) at Texas A&M University (TAMU). The project will combine the efforts of several universities, research groups, and other commercial and non-profit entities to address a problem facing scholars and librarians who work with early modern texts: the inadequacy of machine-readable transcriptions currently produced by optical character recognition (OCR). Performant Software Solutions LLC will work with IDHMC to support the following aims:

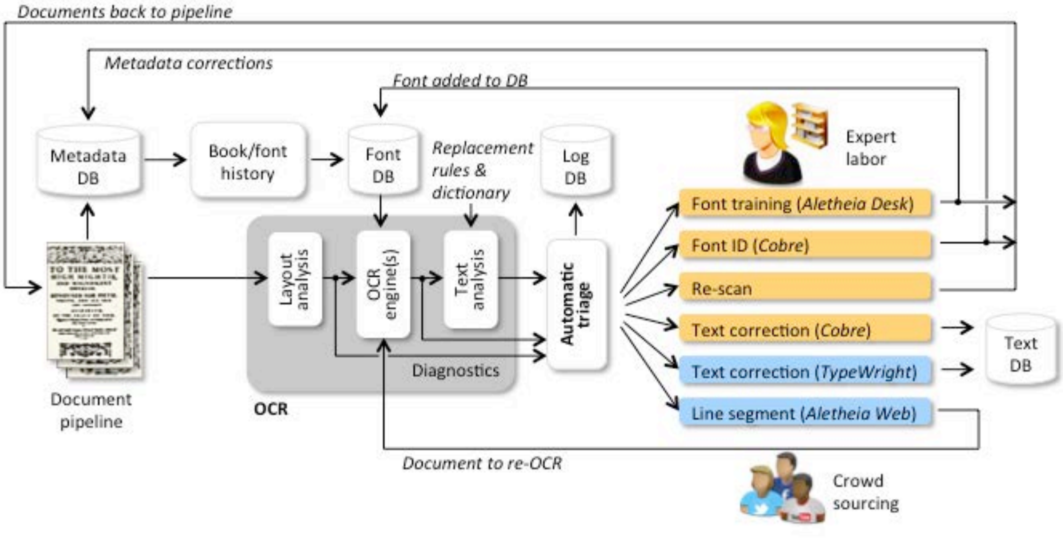
- **Optimized mixed-initiative workflow**. We will develop a document-processing workflow to make efficient use of limited expert knowledge, crowd-sourcing, and computational power by leveraging and adapting existing software tools for OCR, text analysis, font training and line segmentation.

- **Automated document evaluation**. We will develop texting mining and supervised learning techniques to evaluate the correctness of OCR outputs (without ground-truth) and to identify the specific needs of each document (e.g., re-scanning, manual alignment, minor text corrections) by analyzing the text output and numeric diagnostics provided by the OCR engines.

- **Case study with EEBO and ECCO.** We will deploy these techniques on EEBO and ECCO to produce materials that can be crowd-source corrected.

1. Optimized Mixed-Initiative Workflow

Performant will develop a document-processing workflow to make efficient use of limited expert knowledge, crowd-sourcing, and computational power by leveraging and adapting existing software tools for OCR, text analysis, font training and line segmentation.



Processing Pipeline for OCR'ing Early Modern Texts

In the course of creating the processing workflow, the following work will be performed:

- The workflow will be implemented using a Service Oriented Archicture (SOA). Individual tools will be made available as web services with a Web Service Description Language (WSDL) description. This will allow each tool to be discovered and manipulated using the Taverna scientific workflow management system.
- The workflow will be deployed on TAMU's Brazos Supercomputing Cluster.
- The following tools will be made available as web services: Gamera, Tessaract, Aletheia Web, Typewright, JuxtaWS, SVN, and Apache Solr. We will also need to provide web services for the layout analysis, text analysis, and evaluation metric components.
- Juxta WS, an open-source text collation web service, will be installed to facilitate comparing the texts produced by the OCR engines with each other and with the corresponding ground truth.
- Adapt the OCR engines to accept already-segmented texts from Aletheia and to use the font libraries recommended by Cobre.
- Determine optimal time-outs so that an OCR engine's inability to find lines in a particular image or document will not consume too many computing resources.
- Index texts for search. This index will be analogous to the current ARC index and will, like the ARC index, be hosted on the IDHMC's servers at TAMU. Performant will also provide an appropriate interface giving authorized project participants access to the OCR output data.
- When OCR output is evaluated at unacceptable level of errors, the page images will be sent to Aletheia for line segmentation, to Cobre for font recognition, or to both. The page images will then be fed back into the OCR engines for a second

attempt at OCR processing. Performant will develop software for passing page images through this re-processing loop.

- If it turns out not to be possible to determine automatically whether OCR output has achieved an acceptable level of errors (and which kind of remediation is needed), Performant will create software for sending page images to Aletheia and Cobre based upon human determinations rather than automatic indications.
- Develop software for passing OCR output evaluated at acceptable levels to Typewright for crowdsource correction of texts and to post-processing correlation n-gram analyzer provided via the Meandre workbench by SEASR.

2. Automated Document Evaluation

Performant Software will develop of evaluation processes and metrics based on research preformed at TAMU by Ricardo Gutierrez-Osuna, at Illinois by SEASR, and at the IDHMC by Mandell. The evaluation metrics are a series of tests that identify OCRing errors. These tests will determine whether the output from OCR'ing a book is passed directly into post-processing or diverted to a re-processing loop. It will also attempt to determine what kind of intervention (font recognition, line segmentation, or both) is required before the page images are OCR'd again. Performant will send up to three representatives to the OCR evaluation conference to be convened at the University of Wisconsin.

3. Case Study with EEBO and ECCO

EEBO and ECCO combined contain upwards of 300,000 documents of varying sizes and from various periods of print technology. Our goal is to support IDHMC in correcting between 30,000 to 70,000 documents during grant tenure, and set up two systems by its end: one in which corrections continue to be made, and another workflow system that can be adopted by museums, collections, and libraries.

In the course of this case study, the following work will be performed:

- Re-index textual data in the SOLR indexer by page rather than by document
- Add features to TypeWright (uncertain words, drop-downs, italics, line adjustments)
- Implement feedback from usability studies on crowd-sourced correction tools
- Submit metadata and OCR outputs to REKn and 18thConnect TypeWright, Cobre, and Aletheia installations
- Set up Taverna workflow in conjunction with IMPACT
- Create automated export procedures for returning corrected texts to Gale, ProQuest, and TCP
- Make code for all tools and workflows available on Github.

What We Will Provide

In our partnership with the IDHMC, Performant Software Solutions will be responsible for all software development tasks, including:

- Software development
- User interface design and user experience
- Functional testing and regression testing
- Cross-browser compatibility testing
- Automated unit tests and code coverage
- User stories, wireframes, design documents
- Software project plans and schedules
- Development site hosting

We understand that OCR’ing Early Modern Texts is a collaborative, multi-institutional project and we commit ourselves to working effectively with the other participants towards a common goal.

What the Client Will Provide

The IDHMC and Texas A&M University, working with other project participants, will provide:

- Overall project direction and coordination
- Ground truth textual data
- Hosting for the Tesseract, Gamera and other OCR engines on TAMU’s Brazos Supercomputing Cluster
- OCR output data
- Hosting for the OCR output database, Solr index, and interface software on IDHMC’s servers at TAMU
- All raw data necessary for developing transcription correctness metrics
- Hosting of, and access to, the post-processing software provided via the Meandre workbench by SEASR

Project Schedule and Budget

The chart below shows the estimated number of hours and dollars spent per month.

Year 1:

Role	**Rate**	**Month 1**	**M2**	**M3**	**M4**	**M5**	**Total Hours**	**Total Cost**

****PAGE**DELETED****

****PAGE**DELETED****

****PAGE**DELETED****

The University of Illinois
at Urbana-Champaign

Proposal Entitled: OCR'ing Early Modern Texts

University of Illinois Principal Investigator: Loretta Auvil

Submitted by and make contract to:
THE BOARD OF TRUSTEES OF THE
UNIVERSITY OF ILLINOIS
c/o Office of Sponsored Programs and Research Administration
1901 South First Street, Suite A
Champaign, Illinois 61820-7406

Authorizing Official

Name: D. Dutta

Title: Chair, Research Board

Signature: D Dutta

Date: 6/8/12

Administrative Contact

Name: Linda Learned

Title: Co-Interim Director, Office of Sponsored
Programs and Research Administration

Signature: Linda Learned

Date: 6/8/12

Telephone: (217) 333-2187

Fax: (217) 239-6830

Chartered 1867

****PAGE**DELETED****

****PAGE**DELETED****

Statement of Work OCR’ing Early Modern Texts

The Illinois team will perform post-processing spell checking to correct for known OCR errors. We will work with experts to determine the dictionaries to be used for spell checking that contain the period spellings and other normalization to be used. We will work with Ted Underwood, local expert, to generate dictionaries of eighteenth-century proper nouns, common foreign words, and period spellings, including syncopated forms ("silv'ry") and period hyphenations ("to-day") that will be used as part of the spell checking. We will leverage the spell checking capabilities that have been developed to work with the Google Ngrams data for processing the collections. The algorithm allows a known set of common OCR transformations to be applied to misspelled words in an attempt to get a correctly spelled word. This will create a set of replacement rules for correcting the OCR errors that can be applied to the collections. We will also research additional methods for determining OCR errors. We have done some experiments in studying 2-character and 3- character ngrams occurrences with correct and incorrect (OCR error) words.

We will also be helping to determine metrics that can be used to evaluate the correctness of an OCR text. These metrics can include items like total number of tokens per page, and total number or percentage of tokens found in dictionary. These metrics and others can be used to create predictive models that can determine which part of the workflow the text will proceed for further processing.

UNIVERSITY OF MASSACHUSETTS
AMHERST

Research Administration Building
70 Butterfield Terrace
Amherst, MA 01003-9242

Office of Grant and
Contract Administration

voice: 413.545.0698
fax: 413.545.1202

February 10, 2012

Texas A and M University
Dr. Gerianne M. Alexander
Psychology Dept.
4223 TAMU
College Station, TX 77843-4223

RE: Proposed Subcontract to University of Massachusetts
UM Reference No. 112-1169
Entitled: OCRing Early Modern Texts

Dear Dr. Alexander:

Attached is the subject proposal submitted on behalf of Professor Raghavan Manmatha of the Computer Science Department.

It is our understanding that this proposal will be included in a prime proposal which you are submitting to the Andrew Mellon Foundation.

The University is looking forward to participating in this project, subject to the execution of a mutually acceptable subcontract.

If you have questions on the technical aspects of the proposal, please contact Professor Manmatha at (413) 545-3623. Administrative concerns may be directed to Marcia Day, CRA, Grant and Contract Administrator, at (413) 545-0698. In all future correspondence about this proposal, please refer to UMass Proposal No. 112-1169.

Sincerely,

Carol P. Sprague
Director

CS/as
attachment
cc: Manmatha, P.I.

Scope of Work

The work will be done by the PI (R. Manmatha) and a graduate student working under his supervision over the course of a year.

The aim is to do automatic text alignment of multiple text sequences to do both OCR evaluation and also allow the automatic alignment of two OCR engine (text) outputs.

Specifically, we will create a fast algorithm to take the OCR output of one book and the groundtruth (or a groundtruth transcript) and automatically align the two sequences. This can be used to produce an estimate of the OCR error. The technique will work with raw text (for example Tesseract only produces raw text) and will handle multiple languages.

A different version of this alignment algorithm will be used to rapidly align two different OCR outputs (eg Tesseract and Gamera) so that the aligned output can be used for downstream adjudication and OCR improvement. This tool will also handle multiple languages.

The source code will be released as open source and provided to IDHMC.

Timeline:

a) Sep'12 – Jan'13. Create tool for fast OCR evaluation. Evaluate tool with test data.

b) Feb'13 – Jul'13. Create tool for fast alignment of two OCR sequences. Evaluate tool with test data.

c) Jul’13 -Aug'13 – Provide code to IDHMC and

Data: **Groundtruth and OCR outputs will be provided by IDHMC.

**Please note: this is a very small
subset of the ordinary, expensive kind of ground truth,
not the kind that Manmatha will be creating for us.

****PAGE**DELETED****

****PAGE**DELETED****

****PAGE**DELETED****

****PAGE**DELETED****

****PAGE**DELETED****

****PAGE**DELETED****

****PAGE**DELETED****

****PAGE**DELETED****

IMPACT 7

Koninklijke Bibliotheek
Nationale bibliotheek van Nederland

Statement of Work

Automated text recognition, carried out by Optical Character Recognition (OCR) engines, does in many cases not produce satisfying results for historical documents. Recognition rates are often poor or even useless, thereby drastically reducing the usefulness of any digitized historical collection for scholars in the humanities. No free and open OCR engine is currently able to cope satisfactorily with the historical fonts found in printed materials published between the Gutenberg age and the start of the industrial production of books in the middle of the 19th century.

The project *OCR'ing Early Modern Texts* is set out to significantly enhance the usability of such collections by adapting two open OCR engines, Tesseract and Gamera, with capability for processing a large collection of digitized early modern texts from EEBO and ECCO by improving line segmentation and recognition of historical European fonts, and adopting crowd-sourcing of post-correction of OCR results.

The specific challenges the Koninklijke Bibliotheek (KB) – National Library of the Netherlands wishes to address in the project *OCR'ing Early Modern Texts* as a partner are:

1.) To review progress reports and discuss project planning with the principal investigator, Laura Mandell, and to integrate the experience gained during the lead of the IMPACT project to ensure high quality and aptitude of project deliverables with (a total of) 57 hours from Hildelies Balk-Pennington De Jongh, and

2.) To assist in the setting up of in-depth scenario-driven evaluation of OCR results with the help of state-of-the-art technologies and expertise developed during the IMPACT project, such as evaluation under consideration of reading order, alignment of OCR and ground truth, and the treatment of special (historical) characters, with (a total of) 57 hours from Clemens Neudecker, and

3.) To support technical staff in the integration of project tools into an automated executable workflow that can be provided to other stakeholders, using the open source Taverna tool and Interoperability Framework developed in IMPACT, with (a total of) 114 hours from Clemens Neudecker, and

4.) To participate in one project meeting in College Station, TX, to identify optimal evaluation and training strategies, with Hildelies Balk-Pennington De Jongh and Clemens Neudecker.

The KB will also act as the liaison to the IMPACT Centre of Competence for Digitisation (http://www.digitisation.eu/), a not for profit organisation with the mission to make the digitization of historical printed text "better, faster, cheaper" by offering expertise and tools for all parts of the digitisation workflow.

University of Salford MANCHESTER

School of Computing, Science & Engineering
PRImA – Pattern Recognition and Image Analysis Lab

Newton Building, Salford, Greater Manchester, M5 4WT, United Kingdom
www.primaresearch.org

OCR’ing Early Modern Texts

Funding proposal to the Mellon Foundation

Description of Proposed Work

As stated in the proposal, the work undertaken by PRImA will focus on identifying the most relevant (to the crowdsourcing task and target environment) functionality of *Aletheia* and creating a web-based implementation of this subset.

More specifically, the work can be broken down to the following tasks:

M1-6 (6PM)

1. Develop web-based viewing functionality to display PAGE information (region outlines, text etc.) overlaid on the image of the document page.

2. Develop web-based functionality to edit region outlines represented in PAGE, resulting from the application of OCR.

M7-12 (6PM)

3. Develop web-based functionality to enter text (including a virtual keyboard with special characters).

4. Develop web-based region tagging functionality for text regions (in terms of their semantic label – e.g. headline, page number, body text etc.) and also for graphics regions (e.g. illustrations, decorative borders etc.)

5. Identify possible ways to interface the above functionalities with TypeWright and implement most suitable option.

6. Collect feedback from users and make necessary revisions.

M13-24 (3PM)

7. Support use of software, collect use data and make necessary updates.

****PAGE**DELETED****

PRImA 3

University of Salford
MANCHESTER

Dr Apostolos Antonacopoulos
Senior Lecturer in Pattern Recognition

School of Computing, Science and Engineering
Director, PRImA Lab
The University of Salford
Newton Building
Salford, Greater Manchester
M5 4WT. United Kingdom

a.antonacopoulos@primaresearch.org

www.primaresearch.org

Prof Laura Mandell
Director
Initiative for Digital Humanities, Media and Culture
Texas A&M University
USA

This letter promises to make "Web Aletheia" freely available as a tool but also in its code as well.

23 February 2012

Dear Laura,

Re: OCR'ing Early Modern Texts

We at the PRImA Lab are very pleased to participate in the above proposal to the Mellon Foundation. The goals and consortium are exciting and very much aligned with our research interests and capabilities.

The *Aletheia* tool we created over the past 4 years during the IMPACT project has been very well received by researchers and service providers around the world and the natural next step will be to port some of its functionality onto a web-based application. The freely available PAGE format we designed for IMPACT has also been very successfully used to represent document analysis and OCR results, for the large datasets of IMPACT (over 50,000 pages with full text and layout description created using Aletheia), a number of international competitions and an increasing number of other research projects in different countries.

As stated in the proposal, we will create a web application that will port a suitable (for running on the web) subset of the functionality of Aletheia. This application will interface with the image and text repository at TAMU, via references to page images and PAGE file descriptions. The software will be capable of viewing page images, overlaid with PAGE information (region outlines and text), and of editing the region outlines (and text). We will also explore the best way to interface with TypeWright or run as a standalone application.

The above software will be made available under the Apache License, Version 2.0 as requested.

Best regards,

Search Store

Questions? Need Advice? Call 1-800-800-2775

Help | Account | Cart

Quote for Desktops

Equipment for graduate and undergraduate student use at the IDHMC

The Apple Store

Thank you.

Your proposal has been sent.

Omniture

Please call The Apple Store at **800-800-2775 (Education) or 800-GO-APPLE (Government)** if you have questions. Your proposal is shown below for your reference.

Web Proposal Number : **W72395154**

Items you have selected	Part No.	Qty	Unit Price	Ext. Price
Apple Keyboard with Numeric Keypad	MB110LL/B	1	$49.00	$49.00
iMac 27-inch, Quad-Core	Z0M7	2	$2,079.00	$4,158.00
AMD Radeon HD 6970M 1GB GDDR5	065-0257			
8GB 1333MHz DDR3 SDRAM - 2x4GB	065-0559			
Accessory kit	065-8995			
1TB Serial ATA Drive	065-0255			
Apple Mouse	065-0343			
Apple Keyboard with Numeric Keypad (English) + User's Guide	065-0349			
3.1GHz Quad-Core Intel Core i5	065-0249			
AppleCare Protection Plan for iMac - Auto-enroll	S3128LL/A	2	$119.00	$238.00

Print this page for your records.

Subtotal Please note that your subtotal does not include sales tax or rebates. **$4,445.00**

Proposer information:
Trey Dockendorf
Texas A&M University
treydock@tamu.edu
(979)4212830

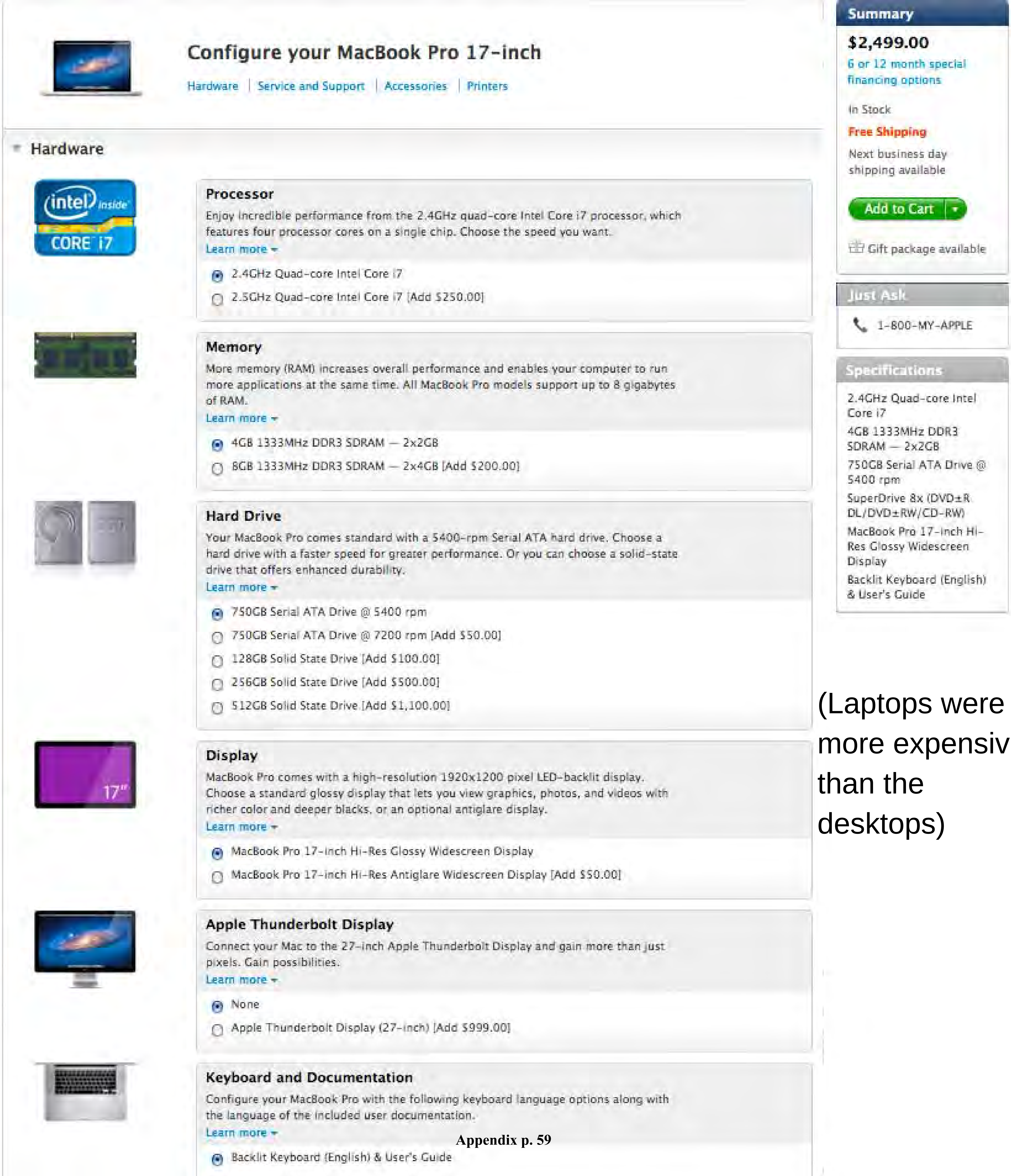
Proposer comments:
Additional iMacs for IDHMC

Continue shopping

Order Status

| Terms of Use | Privacy Policy | Help | Sales and Refunds | Contact Us

For more information about Apple products and programs,
please call 800-800-2775 (Education) or 800-GO-APPLE (Government).



Configure your MacBook Pro 17-inch

Hardware

Processor

Enjoy incredible performance from the 2.4GHz quad-core Intel Core i7 processor, which features four processor cores on a single chip. Choose the speed you want.

Learn more

● 2.4GHz Quad-core Intel Core i7

○ 2.5GHz Quad-core Intel Core i7 [Add $250.00]

Memory

More memory (RAM) increases overall performance and enables your computer to run more applications at the same time. All MacBook Pro models support up to 8 gigabytes of RAM.

Learn more

● 4GB 1333MHz DDR3 SDRAM — 2x2GB

○ 8GB 1333MHz DDR3 SDRAM — 2x4GB [Add $200.00]

Hard Drive

Your MacBook Pro comes standard with a 5400-rpm Serial ATA hard drive. Choose a hard drive with a faster speed for greater performance. Or you can choose a solid-state drive that offers enhanced durability.

Learn more

● 750GB Serial ATA Drive @ 5400 rpm

○ 750GB Serial ATA Drive @ 7200 rpm [Add $50.00]

○ 128GB Solid State Drive [Add $100.00]

○ 256GB Solid State Drive [Add $500.00]

○ 512GB Solid State Drive [Add $1,100.00]

Display

MacBook Pro comes with a high-resolution 1920x1200 pixel LED-backlit display. Choose a standard glossy display that lets you view graphics, photos, and videos with richer color and deeper blacks, or an optional antiglare display.

Learn more

● MacBook Pro 17-inch Hi-Res Glossy Widescreen Display

○ MacBook Pro 17-inch Hi-Res Antiglare Widescreen Display [Add $50.00]

Apple Thunderbolt Display

Connect your Mac to the 27-inch Apple Thunderbolt Display and gain more than just pixels. Gain possibilities.

Learn more

● None

○ Apple Thunderbolt Display (27-inch) [Add $999.00]

Keyboard and Documentation

Configure your MacBook Pro with the following keyboard language options along with the language of the included user documentation.

Learn more

● Backlit Keyboard (English) & User's Guide

Summary

$2,499.00

6 or 12 month special financing options

In Stock

Free Shipping

Next business day shipping available

Gift package available

Just Ask

☎ 1-800-MY-APPLE

Specifications

2.4GHz Quad-core Intel Core i7

4GB 1333MHz DDR3 SDRAM — 2x2GB

750GB Serial ATA Drive @ 5400 rpm

SuperDrive 8x (DVD±R DL/DVD±RW/CD-RW)

MacBook Pro 17-inch Hi-Res Glossy Widescreen Display

Backlit Keyboard (English) & User's Guide

(Laptops were more expensive than the desktops)

CORAID

Name/Company Dockendorf, Trey / Texas A&M University
Phone 979-845-5712
Email treydock@tamu.edu
Address 3139 TAMU College Station, TX 77843

EtherDrive SAN	**Description**	**Quantity**
SRX3200-G	Six GigE Ports, 24 disk holders for 3.5" SATA, SAS, or SSD Drives	1
Software	**Description**	
CorOS™	CorOS™ Scale-out Distributed Storage Operating System	ALL
Disks	**Description**	
Coraid 2TB SATA HDD	Coraid 2TB SATA HDD - 3.5”, 7.2K RPM	6
	Product Totals	**$21,289.00**

Support		**Extended Total**
CORAID E-Support (Business Day) for 1-Year	CORAID Email Only Support for Business Day with Standard RMA	$2,128.90
CORAID E-Support (Business Day) for 3-Years	CORAID Email Only Support for Business Day with Standard RMA	$5,811.90
Extended Warranty		**Extended Total**
Standard 36 Months	CORAID Standard 36 Months	$0.00

*SRX chassis must be filled at least 75% with disks

*All SFP+ SRXs, VSXs, SP's, and HBAs DO NOT include Fiber Optic Transceivers

*This quote does not include cost of shipping, handling, or taxes (if applicable)

	Extended Total
Overall Total with 1-Year of Support	**$23,417.90**
Overall Total with 3-Years of Support	**$27,100.90**

LAURA C. MANDELL
CURRICULUM VITAE

Professional Address: Department of English, Texas A&M University, MS 4227, College Station, TX, 77843-4227. 513-560-7860; mandell@tamu.edu

EMPLOYMENT:
2011 to present: Professor, English, Director, Initiative for Digital Humanities, Media, and Culture, Texas A&M University
2008 to 2011: Professor, English, Miami University
1999 to 2008: Associate Professor, Miami University of Ohio
2004-present: Affiliate, Women Studies;
2007-present: Affiliate, Interactive Media Studies
1993 to 1999: Assistant Professor, Miami University of Ohio, Department of English.
1992-1993: Teacher (11th and 12th grades), Albuquerque Academy

OTHER APPOINTMENTS:
Director, ARC (Applied Research Consortium), 2011 to present
Director, 18thConnect (http://www.18thConnect.org), 2009 to present
Technological Editor, Romantic Circles (http://www.rc.umd.edu), 2009 to present
Chair, MLA Committee on Information Technology, 2009 to 2011
Associate Director, NINES (http://www.nines.org), 2007 to March, 2012
Director, Digital Humanities, (2008 to 2010), and Director of Research Initiatives (2006 to 2008) Interactive Media Studies Program, Miami University

INTERNET RESOURCES (peer-reviewed)
General Editor, The Poetess Archive (http://unixgen.muohio.edu/~poetess) and The Poetess Archive Journal (http://unixgen.muohio.edu/~poetess/PAJournal/); accepted after peer review by the Modern Language Association (MLA), the Networked Interface for Nineteenth-century Electronic Scholarship (NINES; http://www.nines.org), and Romantic Circles (http://www.rc.umd.edu)

BOOKS:
Misogynous Economies: The Business of Literature in Eighteenth-Century Britain, Univ. of Kentucky Press, 1999.
"Feeling Real: Melancholy, Romantic Poetry, and Print," book manuscript

BOOKS IN PROGRESS
"Breaking the Book," under contract for Blackwell Manifesto Series.
"XSLT for Humanists," with Brian Zillig, Syd Bauman.

ARTICLES (recent):
"Disciplining the Real: John Haslam, Johanna Southcott, and the Emergence of Modern Disciplines" (forthcoming from Eighteenth Century Studies)
"Evaluating Digital Scholarship," with Stephen Olsen, Susan Schreibman, Profession 2011 (123-135). Also available online:

http://www.mlajournals.org/toc/prof/2011/1

"Brave New World: A Look at 18thConnect," *Age of Johnson* 21 (January 2012). Print.

"Non-Consuming Relevance: the Grub Street Project," The Shape of Things, ed. Jerome McGann (Rice Univ. Press, 2010), and online: http://shapeofthings.org/papers/

"Histories of Print, Histories of Emotion," Introduction to a special issue, "Technologies of Emotion," ed. Laura Mandell, The Eighteenth Century: Theory and Interpretation 50.2-3 (2010).

"Special Issue: 'Scholarly Editing in the Twenty-First Century' – A Conclusion," Literature Compass 7.2 (2010): 120-133.

"The Poetess Archive Database" (poster), Digital Humanities Quarterly 3.3 (2009) http://www.digitalhumanities.org/dhq/vol/3/3/000059.html

"Hymn, Prayer, Action: Anna Barbauld and the Public Worship Controversy," Studies in Eighteenth-Century Culture 38 (2009): 117-142.

"Bad Marriages, Bad Novels: The Jacobin 'Philosophical Romance,'" Recognizing the Romantic Novel, ed. Charlotte Sussman, Jill Heydt-Stevenson (Liverpool Press, 2008), pp. 49-76.

"Encoding Matter," by Invitation, Special Forum: "Digitisation and Materiality," 19: Interdisciplinary Studies in the Long Nineteenth Century 6 (April 2008), http://www.19.bbk.ac.uk/issue6/digital%20forum/mandelldigitalforum.pdf

"What Is the Matter? What Literary Theory Neither Hears Nor Sees," New Literary History 38 (2007): 757-778.

"Imaging Interiority: Photography, Psychology, Lyrical Poetry," Victorian Studies 49.2 (2007): 218-227.

"Putting Contents on the Table: The Disciplinary Anthology and the Discipline of Literature," Poetess Archive Journal 1.1 (12 April 2007) <http://unixgen.muohio.edu/~poetess/PAJournal/index.html> [peer-reviewed]

AWARDS AND SCHOLARSHIPS (recent):

"Humanities Visualization Space," with Patrick Burkart (Communications Dept.) and Philip Galanter (Visualization Dept.), Tier One Program (TOP) Activity 2, Texas A&M, $110,000, 2012-2014.

"Assessment in the Humanities: A National Symposium," with Cecilia Shore and Paul Anderson, $9,000, awarded by the Teagle Foundation, December 2010

"18thConnect and Open-Access Full-Text," Mellon Officer's Grant for $41,000 awarded July 14, 2010.

NCSA (National Center for Supercomputing Applications / I-CHASS (Illinois Center for Computing in the Humanities, Arts, and Science), 200,000 hours of supercomputer time for 18thConnect for OCR development, 2009-2010.

President's Academic Enrichment Award (PAEA), Miami University, with Kerry Powell, Wiestse de Boer, and Charles Ganelin: $250,000 to start Miami's Humanities Center (2008)

CACR (Center for Academic Computing Research, Miami Univ.) Grant, 2007-2008 ($4,000)

Assigned Research Appointment for The Poetess Archive – Paid Leave for Fall 2005.

Ohio Learning Network, Principal Investigator for "Bringing Knowledge Closer through Web Interactivity," 2004-2005 ($20,000); CETE Grant for 2004-2005 ($5,000)

Loretta Auvil
Senior Project Coordinator
Illinois Informatics Institute
1205 W. Clark St, Urbana, IL 61801
E-mail: lauvil@illinois.edu

Professional Preparation

Alderson-Broaddus College, BS, Applied Mathematics and CS, May 1990, Summa Cum Laude.
Virginia Polytechnic Institute and State University, MS CS, May 1992.
University of Illinois at Urbana-Champaign, PhD CS, ABD.

Appointments

20011-present	Senior Project Coordinator, Illinois Informatics Institute, U. of Illinois
2003-2011	Senior Project Coordinator, NCSA, U. of Illinois, Champaign, IL
1997-2003	Visualization Programmer, NCSA, U. of Illinois, Champaign, IL
1997	Graduate Research Assistant, NCSA, U. of Illinois, Champaign, IL
1993-1997	Graduate Student, Rome Laboratory/C3CB, Rome, NY
1992-1993	Computer Scientist, Rome Laboratory/C3CB, Rome, NY
1990-1992	Graduate Student, Rome Laboratory/C3CB, Rome, NY
1990	Math Laboratory Assistant, Applied Math & Computer Science, Alderson Broaddus College, Philippi, WV
1990	Student Researcher, Oak Ridge National Laboratory, Oak Ridge, TN
1989	Student Researcher, Oak Ridge National Laboratory, Oak Ridge, TN

Publications Most Closely Related To This Proposal

- Ács, Bernie, Xavier Llorà, Loretta Auvil, Boris Capitanu, David Tcheng, Mike Haberman, Limin Dong, Tim Wentling, and Michael Welge. 2010. A general approach to data-intensive computing using the Meandre component-based framework. In *Proceedings of the 1st International Workshop on Workflow Approaches to New Data-centric Science*, 1-12. Indianapolis, Indiana: ACM. doi:10.1145/1833398.1833406.
- Ács, Bernie, Xavier Llorà, Loretta Auvil, Boris Capitanu, David Tcheng, Mike Haberman, Tim Wentling, and Michael Welge. 2009. Flowing Homogeneously: Meandre Transparent Integration of Data-Intensive Computing and Third Party Services. In 5th IEEE International Conference on e-Science 2009, University of Oxford.
- Capitanu, Boris, Xavier Llorà, Loretta Auvil, Michael Welge, and Bernie Acs. 2009. SEASR Integrates with Zotero to Provide Analytical Environment for Mashing up Other Analytical Tools presented at the Digital Humanities 2009 Poster Session, University of Maryland.
- Auvil, Loretta, Eugene Grois, X. Llorà, G. Pape, Vered Goren, Barry Sanders, Bernie Acs, and Robert McGrath. 2007. A Flexible System for Text Analysis with Semantic Network. In Digital Humanities 2007, 17-20. Champaign-Urbana, IL.
- Clement, T., Loretta Auvil, C. Plaisant, G. Pape, and Vered Goren. 2007. "Something that is interesting is interesting them": Using text mining and visualizations to aid interpreting repetition in Gertrude Stein's The Making of Americans. In Digital Humanities 2007, 40-44. Champaign-Urbana, IL.
- Don, Anthony, Elena Zheleva, Machon Gregory, Sureyya Tarkan, Loretta Auvil, Tanya Clement, Ben Shneiderman, and Catherine Plaisant. 2007. Discovering interesting usage patterns in text collections: integrating text mining with visualization. In Lisboa, Portugal.
- Llorà, X, B. Ács, Loretta Auvil, B. Capitanu, M. E. Welge, and David E. Goldberg. 2008. Meandre: Semantic-Driven Data-Intensive Flows in the Clouds. In *Proceedings of 4th IEEE International Conference on eScience*, 238-245. IEEE Press.
- Cai, Y. D., D. Clutter, G. Pape, J. Han, M. Welge, and Loretta Auvil. 2004. MAIDS: Mining Alarming Incidents from Data Streams. In *Proceedings of the 2004 ACM SIGMOD*, 919-920. Paris, France.

Other Significant Publications

- Larkin, D. M, G. Pape, R. Donthu, L. Auvil, M. Welge, and H. A Lewin. 2009. "Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories". Genome Research 19, no. 5: 770.
- W. Murphy et al., "Dynamics of Mammalian Chromosome Evolution Inferred from Multispecies Comparative Maps," American Association for the Advancement of Science, vol. 309, 2005, pp. 613-617.
- W. H. Hsu, L. S. Auvil, T. Redman, D. Tcheng, and M. Welge. High-Performance Knowledge Discovery and Data Mining Systems Using Workstation Clusters (Extended Abstract). Presented at National Conference on High Performance Networking and Computing (SC99), Portland, OR, November 1999.

Synergistic Activities

- Collaboration with Harris Lewin, Denis Larkin and Jian Ma on the development of Evolution Highway visualization tool for comparative genomic analysis.
- Collaboration with John Unsworth and others on the Nora and Monk projects to analyze 18th and 19th century literature leveraging data and text mining techniques deployed in the D2K and SEASR environment.
- Lead instructor of "SEASR Analytics" Course at Digital Humanities Summer Institute at University of Victoria in June 2009, June 2010, and June 2011
- Member of the Automated Learning Group at NCSA
- Team member on the development of Data to Knowledge (D2K)
- Team member on the development of SEASR
- IEEE Visualization Conference Finance Chair (1998-present) and IEEE Visualization Conference Tutorial Chair (2002-2003)
- IEEE 2001 Certificate of Appreciation

Collaborators

Bernie Acs, U. of Illinois
Catherine Blake, U. of Illinois
Rachael Brady, Duke U.
Tanya Clement, U. of Texas
Timothy Cole, U. of Illinois
J. Stephen Downie, U. of Illinois
David Enstrom, U. of Illinois
David E. Goldberg, U. of Illinois
Kevin Franklin, U. of Illinois
Jiawei Han, U. of Illinois
Brant Houston, U. of Illinois
William Hsu, KSU
Matthew Jockers, Stanford U.
Denis Larkin, U. of Illinois
Loren Leonard, PSU
Harris Lewin, U of CA, Davis
Jian Ma, U. of Illinois
Carole Palmer, U. of Illinois
Catherine Plaisant, U. of Maryland
Marshall Scott Poole, U. of Illinois
Ray Siemens, U. of Victoria
Natasha Smith, U. of North Carolina
David Tcheng, U. of Illinois
John Unsworth, U. of Illinois
Mike Ward, U. of Illinois
Michael Welge, U. of Illinois

Graduate and Postdoctoral Advisors

Calvin J. Ribbens, Virginia Tech, and Michael Heath, U. of Illinois.

Student Supervisor

Supervised students and graduate students that work with the Automated Learning Group.

Richard Furuta
Professor
Department of Computer Science
Texas A&M University
College Station, Texas 77843-3112
E-mail: furuta@cs.tamu.edu

PROFESSIONAL PREPARATION:

- B.A., Biology, Reed College, 1974.
- M.A., Computer Science, University of Oregon, 1978.
- Ph.D., Computer Science, University of Washington, 1986.

APPOINTMENTS:

- Professor, Department of Computer Science, Texas A&M University, 2001-present; Associate Professor, 1993-2001.
- Director, Center for the Study of Digital Libraries, Texas Engineering Experiment Station, 2004-present; Associate Director 1995-2004.
- Director, Hypermedia Research Laboratory, Department of Computer Science, Texas A&M University, 1995-present; Co-Director, 1993-1995.
- Assistant Professor, Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, 1986-1993.

PUBLICATIONS:

- F. Shipman, R. Furuta, D. Brenner, C. Chung and H. Hsieh. "Guided Paths through Web-Based Collections: Design, Experiences, and Adaptations," *Journal of the American Society of Information Sciences (JASIS)*, **51(3)**, March 2000, pp. 260-272.
- P. Stotts, R. Furuta, and J. Ruiz Cabarrus, Hyperdocuments as automata: verification of trace-based browsing properties by model checking, *ACM Trans. Inf. Syt.,* **16(1)**, January 1998, 1-30.
- R. Furuta, J. Scofield, and A. Shaw, Document formatting systems: Survey, concepts and issues, *Computing Surveys*, **14(3)**, 417-472.
- R. Furuta, S. S. Kalasapur, R. Kochumman, E. Urbina, and R. Vivancos-Pérez, "The Cervantes Project: Steps to a Customizable and Interlinked On-line Electronic Varorium Edition Supporting Scholarship," *Research and Advanced Technology for Digital Libraries: 5th European Conference, ECDL 2001*, September 2001, pp. 71-82.
- R. Furuta and J-C. Na, "Applying caT's Programmable Browsing Semantics to Specify World-Wide Web Documents that Reflect Place, Time, Reader, and Community, Proceedings of the ACM Symposium on Document Engineering, November 2002, pp. 10-17.
- U.P. Karadkar, A. Kerne, R. Furuta, L. Francisco-Revilla, F. Shipman, and J. Wang. "Connecting Interface Metaphors to Support Creation of Path-Based Collections," *Research and Advanced Technology for Digital Libraries: 7th European Conference, ECDL 2003 Proceedings*, Lecture Notes in Computer Science LNCS 2769, Springer, 2003, pp. 338-349.
- P. Dave, U.P. Karadkar, R. Furuta, L. Francisco-Revilla, F. Shipman, S. Dash, and Z. Dalal. "Browsing Intricately Interconnected Paths," *Hypertext 03: The fourteenth ACM conference on hypertext and hypermedia*, ACM Press, 2003, pp. 95-103.
- U. P. Karadkar, R. Furuta, S. Ustun, Y.J. Park, J.C Na, V. Gupta, T. Ciftci, and Y. Park. "Display-agnostic Hypermedia." *Hypertext 2005: Proceedings of the Fifteenth ACM Conference on Hypertext and Hypermedia*, ACM Press, 2004, pp. 58-67.
- E. Urbina, R. Furuta, S. E. Smith, N. Audenaert, J. Deng, and C. Monroy. "Visual Knowledge: Textual Iconography of the *Quixote*, a Hypertextual Archive." *Literar and Linguistic Computing*, 2006.
- U. Karadkar, M. Nordt, R. Furuta, C. Lee, and C. Quick. "An Exploration of Space-Time Constraints on Contextual Information in Image-based Testing Interfaces." *European Conference on Digital Libraries, ECDL 2006, Proceedings*. 2006

Synergistic Activities

- Current digital library building project: Nautical Archaeology Digital Library. Focuses particularly on wooden ships and ship reconstruction. Current support from the NSF.
- Current digital library building project: Cervantes Project. Provides a resource on the works and life of Miguel de Cervantes Saavedra (1547-1616), as well as a regularly-updated bibliography of scholarly works discussing his writings. One current project is an "electronic virtual variorum edition" of his best-known work, *Don Quixote*, using images of the editions published during his lifetime. Previous support from the NSF (ITR/DLI). Another is a digital presentation of the textual iconography of the *Quixote*—based an extensive set of editions obtained specifically for the project (funded by the National Endowment for the Humanities).
- Organization of Digital Libraries conference and community activities. Conference Chair, Digital Libraries 94. Program Chair, Digital Libraries 95. Program Chair, Digital Libraries 2000. Program chair the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) 2009. Member of the steering committee for JCDL (2001 and following; chair 2001-2005). Tutorials co-chair, European Conference on Digital Libraries 2001. Workshops Chair, ACM/IEEE-CS JCDL 2002 and JCDL 2003. Member of the program committee for others in these conferences' series as well as Asian Digital Libraries conferences. Editor-in-chief *International Journal of Digital Libraries*. Chair of the IEEE-CS Technical Committee on Digital Libraries.
- Work supporting a Digital Humanities Initiative on the Texas A&M University campus. Member of the steering board directing this initiative. Chair of the working group drawn from the campus. Was a member of the selection committee for the Director if the Institute for Digital Humanities, Media, and Culture.
- Active in the Hypertext and Document Engineering research fields. Program co-chair, ACM Hypertext 93. Mentoring chair, ACM Hypertext 2001. Editorial board, *New Review of Hypermedia* and the *Journal on Digital Information*. Program co-chair, ACM Document Engineering 2002. Member of steering committee, ACM Document Engineering, 2002-2008.

Collaborators and co-editors

Nabil Adam, Rutgers University; James Caverlee, Texas A&M University; Lillian Cassel, Villanova University; Lauren Cifuentes, Texas A&M University; Luis Vieira de Castro, Texas A&M University; Lois Delcambre, Portland State University; Edward Fox, Virginia Tech; Luis Francisco-Revilla, University of Texas; Greg Hislop, Drexel University, Hao-Wei Hsieh, University of Iowa; Unmil Karadkar, Texas A&M University; John J. Leggett, Texas A&M University; Du Li, Nokia; Enrique Mallen, Sam Houston State University; Catherine C. Marshall, Microsoft Research; Erich Neuhold, University of Vienna; Marlo Nordt, Chevron; Frank M. Shipman III, Texas A&M University; Gary Stringer, Texas A&M University; Eduardo Urbina, Texas A&M University

Graduate (Ph.D.) Advisor:

Alan Shaw, University of Washington

Postdoctoral Advisor:

None.

Thesis Advisor and Postgraduate-Scholar Sponsor:

Shueh-Cheng Hu, consultant
David Kingery, Boeing
Vijay Kumar, consulting in Austin
Marlo Nordt, Chevron
Jie Deng, CGG Veritas
Sheiyao Augustine Su
Jin-Cheon Na, Nanyang Technological University
Unmil Karadkar, University of Texas

Carlos Monroy, Texas A&M University
Yungah Park, Texas A&M University
Youngjoo Park, Texas A&M University
Tolga Ciftci, Texas A&M University
Neal Audenaert, Texas A&M University
Paul Bogen, Texas A&M University
Omar Alvarez, Texas A&M University
Hamed Alhoori, Texas A&M University
Luis Davi Meneses Macchiavello, Texas A&M University

Total number of graduate students advised is 44 (28 Masters awarded, 12 PhD awarded).
Total number of postdoctoral scholars sponsored is 0.

Ricardo Gutierrez-Osuna

Department of Computer Science and Engineering
Texas A&M University
3112 TAMU, College Station, TX 77843-3112

Tel: (979) 845-2942
Email: rgutier@cse.tamu.edu
URL: http://research.cse.tamu.edu/prism/

Education

Polytechnic University of Madrid (Spain)	Electrical Engineering	BS, 1992
North Carolina State University	Computer Engineering	MS, 1995
North Carolina State University	Computer Engineering	PhD, 1998

Appointments

Associate Professor	Computer Science, TAMU	09/06-present
Research Fellow	Center for Speech Technology Research, University of Edinburgh	07/09-12/09
Assistant Professor	Computer Science, TAMU	07/02-08/06
Assistant Professor	Computer Science and Engineering Wright State University, Dayton, OH 45435	09/98 -06/02

Honors and awards

- Association of Former Students Teaching Award, TAMU College of Engineering 2009
- Barbara and Ralph Cox '53 Faculty Fellow, TAMU College of Engineering 2009
- Tenneco Meritorious Teaching Award, TAMU College of Engineering 2009
- Graduate Faculty Teaching Excellence Award, TAMU Computer Science 2007
- NSF Faculty Early Career Development (CAREER) Award 2000-2004

Five publications most closely related to the proposed project

- D. Felps* and R. Gutierrez-Osuna, "Developing objective measures of foreign-accent conversion," *IEEE Trans Audio, Speech & Language Proc*, 18(5), 1030-1040, 2010.
- D. Felps*, H. Bortfeld and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Commun*, 51, 920-932, 2009.
- D. Felps*, C. Geng, M. Berger, K. Richmond, and R. Gutierrez-Osuna, "Relying on critical articulators to estimate vocal tract spectra in an articulatory-acoustic database," in *Proc. Interspeech,* 2010, 1990-1993.
- J. Rodríguez*, H. Bortfeld, I. Rudomín, B. Hernández and R. Gutiérrez-Osuna, "The Reverse-caricature Effect Revisited: Familiarization with Frontal Facial Caricatures Improves Veridical Face Recognition, *Appl Cog Psych*, 23(5), 733-742, 2009.
- S. Fu*, R. Gutierrez-Osuna, A. Esposito, K. Praveen and O. N. Garcia, "Audio/Visual Mapping with Cross-Modal Hidden Markov Models," *IEEE Trans Multimedia*, 7(2), 243-252, 2005.

Five other significant publications

- J. Choi*, B. Ahmed, and R. Gutierrez-Osuna, "Development and evaluation of an ambulatory stress monitor based on wearable sensors," *IEEE Trans Information Technology in BioMedicine*, in press.
- J. Choi* and R. Gutierrez-Osuna, "Removal of respiratory influences from heart rate variability in stress monitoring," *IEEE Sensors Journal*, 11, 2649-2656, 2011.
- N.-Y. Yu, T. Yamauchi, H.-F. Yang; Y.-L. Chen and R. Gutierrez-Osuna, "Feature selection for inductive generalization," *Cognitive Science*, 34, 1574–1593, 2010.
- P. Kakumanu*, A. Esposito, O. N. Garcia and R. Gutierrez-Osuna, "A comparison of acoustic coding models for speech-driven facial animation," *Speech Communication*, 48(6), pp. 598-615, 2006.
- R. Gutierrez-Osuna, P. K. Kakumanu*, A. Esposito, O. N. Garcia, A. Bojorquez, J. L. Castillo and I. J. Rudomin, "Speech-driven Facial Animation with Realistic Dynamics," *IEEE Transactions on Multimedia*, 7(1), pp. 33-42, 2005.

* denotes students under my direct supervision

Synergistic Activities

Panel reviewer	NSF, NIH, European Commission
Paper reviewer	PNAS, IEEE Trans Computers, IEEE Trans Neural Networks, IEEE Sensors J, IEEE Trans Systems, Man and Cybernetics, IEEE Trans Biomed Circuits Sys, Sensors and Actuators B, Neurocomputing
Curriculum development	New courses: Speech Processing, Intelligent Sensor Systems, Pattern Analysis. Undergraduate research: Supervision of Capstone Design Projects since 1999 (40+ projects, 125+ students to date)
Minorities	Moderator. NSF 9910768 Cyber-conference: "Research Foundations on Successful Participation of Underrepresented Minorities in IT"
Editorial board / program committee	Associate Editor: IEEE Sensors J, ICRA (2009). Program Chair: ISOEN 2011. Program Committee: 2010 IEEE MFI, IJCNN (2008, 2009), IEEE Sensors Conf (2007, 2008), Robotics: Systems and Science (2006).

Collaborators and other affiliations

i. *Collaborators*: D Song, R Stoleru, A Ames, R Murphy, S Smith, F Shipman, SM Smith (TAMU), A Hierlemann (ETH-Zurich), M Carreira-Perpinan (UC Merced), H Bortfeld (U Conn), S Semancik (NIST).

ii. *Graduate Advisors*: HT Nagle (NC State), RC Luo (Chung Cheng Univ, Taiwan)

iii. *Doctoral and Postdoctoral Advisees*: A Perera-Lluna (UPC, Spain), T Yamanaka (Sophia University, Japan), A Gutierrez-Galvez (Univ Barcelona, Spain), B Raman (WUSTL), D. Felps (NGIA), J. Rodriguez (Polytechnic University of Puerto Rico), J. Choi (Intel).

Total number of advisees: 2 Post-Doctoral, 5 PhD (6 more in progress), 11 MS.

Vita for R. Manmatha

Raghavan Manmatha
Department of Computer Science,
University of Massachusetts, Amherst, Massachusetts 01003-4610.
(413)-545-3623
manmatha@cs.umass.edu
http://www.cs.umass.edu/ manmatha

a. Professional Preparation

- Phd in Computer Science, University of Massachusetts at Amherst 1997.
- M.S. in Electrical Engineering, University of Hawaii, Manoa, 1986.
- B.Tech. in Electrical Engineering, Indian Institute of Technology, Kanpur, India 1983.

b. Appointments

- Sept 2006 - Current: Research Associate Professor, Computer Science, University of Massachusetts.
- Sept 1998 - Aug 2006: Research Assistant Professor, Computer Science, University of Massachusetts.
- Oct 2009 - Current: Consultant A9/Amazon, Palo-Alto.
- May 2006 - Jun 2009: Co-Founder/Advisor Snaptell, Inc.
- Jan 2006 - Jun 2006: Consultant Google.
- Jun 2005 - Aug 2005: Visiting Research Scientist Google.
- Jul 1997 - Aug 1998: Post-doctoral Research Associate, Computer Science,
- Feb 1995 - Aug 1998: Lead researcher in the multimedia indexing and retrieval group, Center for Intelligent Information Retrieval. Supervised 4 graduate and 2 undergraduate students.
- Apr 1994 - Jan 1995: Worked for ACSIOM as Project Manager in charge of developing algorithms to read stock certificates. Project funded by the Chicago Stock Exchange.
- Jul 1993 - Jan 1994: Visited IBM Almaden Research Labs, San Jose.

c. Some Related Publications

- Yalniz, I. Z. and Manmatha, R., Finding Translations in Scanned Book Collections, *Accepted to Proc. ACM SIGIR 2012*.
- Frinken, V., Fischer A., Manmatha, R. and Bunke, H., A Novel Word Spotting Method Based on Recurrent Neural Networks. IEEE Trans. PAMI 34(2): 211-224 (2012)
- Yalniz, I. Z., Can, E. F. and Manmatha, R., Partial Duplicate Detection for Large Book Collections, *Proc. ACM CIKM 2011, pg 469–474*.
- Yalniz, I. Z. and Manmatha, R., A Fast Alignment Scheme for Automatic OCR Evaluation of Books, *Proc ICDAR 2011: 754-758*

- Rasagna, V. Kumar, A., Jawahar, C. V. and Manmatha, R., Robust Recognition of Documents by Fusing Results of Word Clusters *ICDAR 2009: 566-570*.
- S. L. Feng and R. Manmatha, A Hierarchical HMM Based Automatic Evaluation of OCR Accuracy for a Digital Library of Books, *Proc. JCDL'06, pg 109–118*.
- Manmatha, R.and Rothfeder, J. A Scale Space Approach for Automatically Segmenting Words from Historical Handwritten Documents *IEEE PAMI, 27(8): 1212-1225, 2005*.
- T. Rath, R. Manmatha and V. Lavrenko, "A Search Engine for Historical Manuscript Images", *Proc. ACM SIGIR'04, pp. 369-376*.
- T. Rath and R. Manmatha, "Word Image Matching Using Dynamic Time Warping"'*Proc. IEEE CVPR'03, vol. 2, pp. 521-527*.
- R. Manmatha, T. Rath and and F. Feng, Modeling Score Distributions for Combining the Outputs of Search Engines, *Proc. ACM SIGIR'01, pg 267–275*.

d. Synergistic Activities

- Associate Editor IEEE PAMI Aug 2011 - Current, Pattern Recognition Letters 2004 - Current. ACM TOIS 2000-2002.
- Area Chair for ACM SIGIR conf. in '01, '03-05,'09, '11, Senior Program Committee CIKM'12, Area Chair for CVPR'11. Program Co-Chair for ICFHR'14.
- Program Committee member for a number of conferences including CVPR'04-10,12, ICCV'06-08, CIVR'04-09, ICDAR'05-11, ICFHR'08, DAS'06-12, ECCV'06-10.

e. Collaborators and Other Affiliations

1. **Collaborators.**
   James Allan, Bruce Croft, Deepak Ganesan, Allen Hanson, David Smith, Nicholas Howe - Smith College. Brewster Kahle - Internet Archive, Greg Crane - Tufts. C. V. Jawahar, Pramod Kumar, V. Anand - IIIT Hyderabad. V. Frinken, A. Fisher, H. Bunke - U. Bern.
2. **Graduate and Postdoctoral Advisors.**
   Ph.D. advisors were Professors Allen Hanson and Edward Riseman at the University of Massachusetts, Amherst and Dr. John Oliensis at Stevens Institute.
   Postdoctoral advisor was Professor Bruce Croft at the University of Massachusetts.
3. **Thesis Advisor and Postdoctoral-Scholar Sponsors.**
   I. Z. Yalniz, E. Can, V. N. Murthy - University of Massachusetts, S. Feng - Siemens Corporate Research, J. Jeon, N. Mohanty, T. Rath - Google, J. Rothfeder - IBM, M. Korn - Amazon, S. Ravela, - MIT, M. Das - Kodak Research, N. Srimal, Y. Ren, V. Wu, C. Han, R. Liang, Vikrant Kobla - Industry, 16 graduate students and 1 postdoctoral scholar.

Key

Q= Quarter *University = Texas A&M

L.M. = Laura Mandell (IDHMC/ARC)*

C.L. = Cushing Rare Books Library*

P.S. = Performant Software

P.R. = PRImA (University of Salford)

R.G. = Ricardo Gutierrez-Osuna*

R.F. = Rick Furuta*

S.Z. = SEASR (University of Illinois)

R.M. = R. Manmatha (University of Massachusetts)

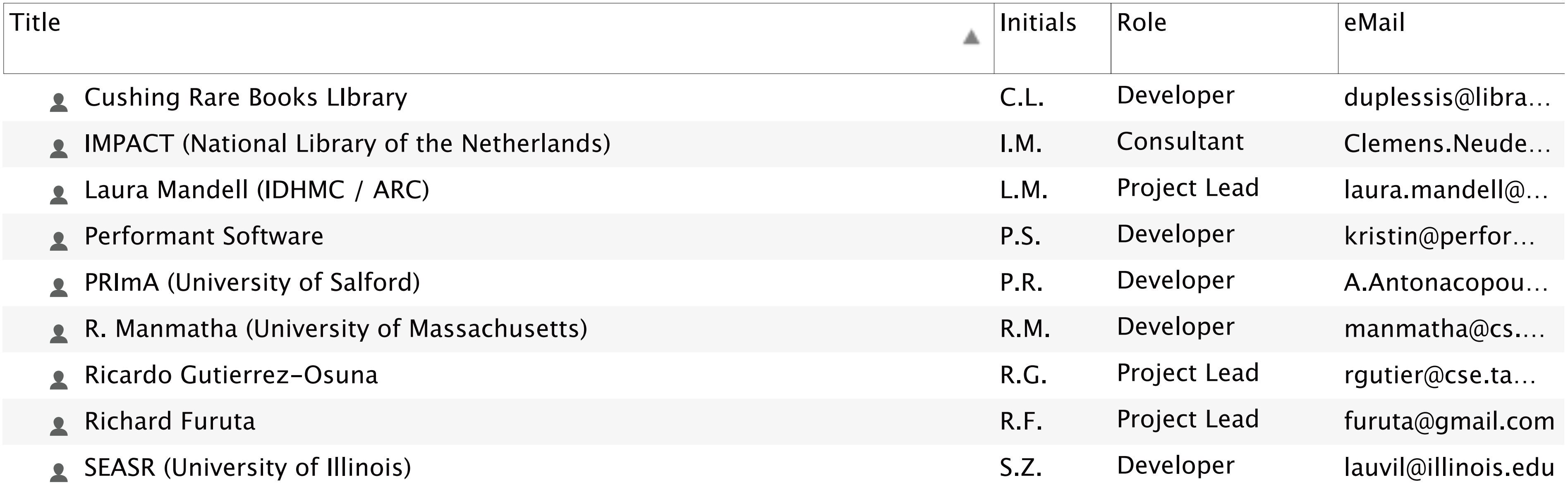
I.M. = IMPACT (National Library of the Netherlands)

Appendix: Detailed Schedule of Tasks

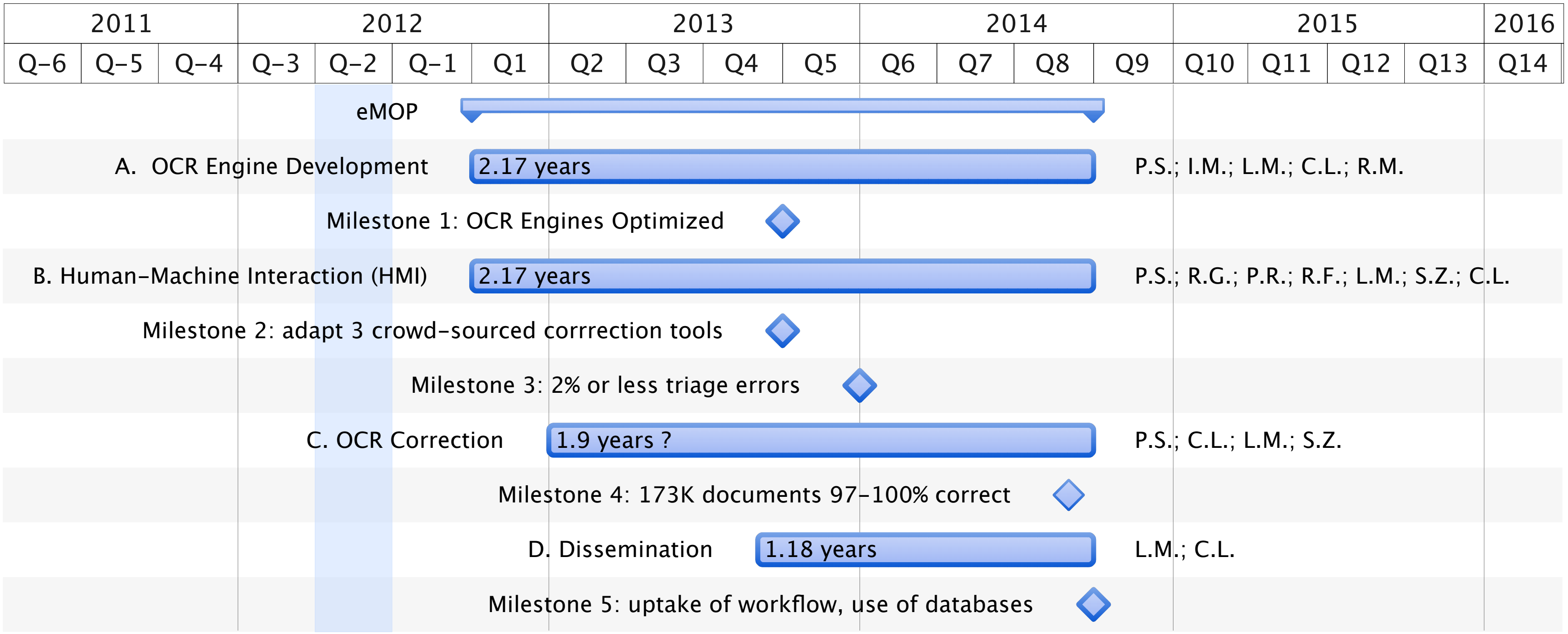
Unit	Goal	Tasks	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**
			10/12–12/12	1/13–3/13	4/13–6/13	7/13–9/13	10/13-12/13	1/14-3/14	4/14-6/14	7/14-9/14
A. OCR Engine Development	1. Engines	Optimize what goes in: find optimal image settings using ImageMagick	**L.M.**							
		Optimize what happens inside: put Tesseract's line segmentation procedure into the Gamera Toolkit	**P.S.**							
		Optimize what comes out: create and tweak XSLT transforms that a) put xml outputs (hOCR, Gamera's xml) into the xml form required by Gale, ProQuest; b) create TEI-A; c) use whitespace to mark up paragraphs	**L.M.**	■	■	■				
	2. Fonts	Select documents containing representative fonts & run them to see results, creating typed versions to test them against	■							
		Create a font importation database	**C.L.**	■	■	■				
		Scan samples of fonts from Cushing, Antwerp, St. Bride's	■	■	■	■				
		Train engines in fonts from EEBO/ECCO; train engines in and transcribe samples of font images from Cushing, Antwerp, St. Bride's	**L.M.**	■	■	■	■	■	■	■
	CHECKPOINT 1: make sure font database needed		**Nov. 2012**							
	3. Testing	Add x-y coordinates for each line of the test data set, indicating place on the page image, making font documents usable to calibrate	■							
		Calibrate the algorithm that compares OCR outputs with hand-typed text;	**R.M.**							
		Modify algorithm to compare OCR outputs with hand-typed text	■	■	■	■				
		Create API for sending us (and making available to all) early modern test set & comparison algorithm, and then use it to test all OCR engine tweaks				**R.M.**				
	4. OCR'ing EEBO and ECCO page images	Set up Taverna workflow to run OCR process			**I.M.**	■				
		After getting best results, 93% accuracy or higher, run 260,000 documents through engines on HPC at 10 seconds per page			**P.S.**	■	**L.M.**	■	■	■
	CHECPOINT 2: make sure test set can be made automatically			**Jan. 2013**						
Milestone 1: we now know that 23.7 million pages can and will be 93% correct and are running through the engines – Sept. 2013										
B. Human-machine interaction	5. Crowdsourcing a) Cobre	Launch Django server with instance of Cobre backed by D-Space allowing all 18thConnect and REKn members to create and save Frankenbooks	**C.L.**							
		Add features to Cobre that allow automated creation of structure that allows for filmstrip presentation, metadata-editing, font identification, and transcription	**C.L.**							
		Load page images of "unreadable" documents into Cobre along with other editions of the same title		**L.M.**						
		Conduct usability studies by consultants who are book history and early modern experts (Raven, Hume, and Mosley)			**R.F.**					
		Re-design tool and					**C.L.**			
		Re-work the interface based upon usability studies						**P.S.**		

	Goal	Tasks	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
	b) Aletheia Web	Create web version of Aletheia	P.R.	■	■	■				
		Design interface and stand up Aletheia in 18thConnect and REKn, Ruby on Rails				P.S.				
		Conduct usability studies on graduate and undergraduate students, adjusting the design and interface				R.F. P.R.				
	c) Type-Wright	Add capacity for adjusting lines	P.S.	■	■					
		Add other features to tool, including red squiggly underline feature to thumbnail for indicating probable errors as indicated in post-processing output	P.S.	■	■					
	d) all tools	Conduct wide-ranging usability studies and measure effectiveness of all tools					R.F.	■	■	■
		Re-work both the triage system based upon document evaluation (see last item of goal 6, immediately below) and the interfaces based upon usability and effectiveness studies							■ P.S.	■ ■
	Milestone 2: Release Tools – Sept. 2013									
	6. Document Evaluation a) Check coordinates produced by OCR engine	Run clustering algorithm on word coordinates to isolate documents with too many letter sizes per letter		R.G.	■	■				
		Run clustering algorithm on line coordinates to isolate pages with inconsistently ordered lines		■	■	■				
	b) Check N-grams and words	Count number of words that are unique and that contain internal punctuation other than hyphen	S.Z.	■	■	■				
		Count number of impossible n-grams in three or four languages	■	■	■	■				
		Count number of unique words in the dictionary with 0, 1, 2, and 3 editing distances	■	■	■	■				
		Count number of replacement rules that apply	■	■	■	■				
	c) Find Document Signature	Select among 47,000 keyed texts the documents with OCR results that fail because of font id, line segmentation, and page-image inadequacy		R.G.	■	■				
		Measure these known failures using clustering and counting 6a. and b., immediately above, and correlate ranges of measures obtained into document signatures corresponding to specific engine failures (font, lines, bad images) if possible		■	■	■				
	d) Use signals	Correlate typical n-gram errors in three languages with need for font training		■	■	■				
		Count number of single-and-double character words in document		■	■	■				
	e) Draw conclusions	Determine document signatures and signals that indicate what went wrong in OCR process, whether it was font misidentification, unknown layout, or unknown problems		■	■	■				
	7. Optimize OCR Output with Human Assistance (Optimize HMI)	Set up automated triage system: font mis-id and unknown go to Cobre; layout indeterminacy goes to Aletheia Web			P.S.	■				
		Select subset of documents in each tool to monitor			L.M.	■				
		Based on usability studies and human-made improvements in document subset, determine how to optimize human / machine intervention (i.e., tool tweaking; adding automated processes for tasks that are too repetitive; not sending specific problems to the tools, or allowing agents to forward problems to 18thConnect / REKn directors					R.F.	■	■	■
		Adjust measures that indicate where document needs to be sent based on degrees to which crowd is able and willing to help (first item in 7, immediately above)					■	■	■	■
	CHECKPOINT 3: confirm time/correctness correlation		**Apr. 2013**							
	8. Launch Crowd Tools	TypeWright and Cobre demo at pre-conference workshop at MLA, January 2013 (dhCommons, already scheduled)		L.M.						
		TypeWright/Cobre paper, REKN announcement, on Restoration and 18thC division panel, with James Raven, MLA, January 2013 (chair Catherine Ingrassia; already scheduled)		■						

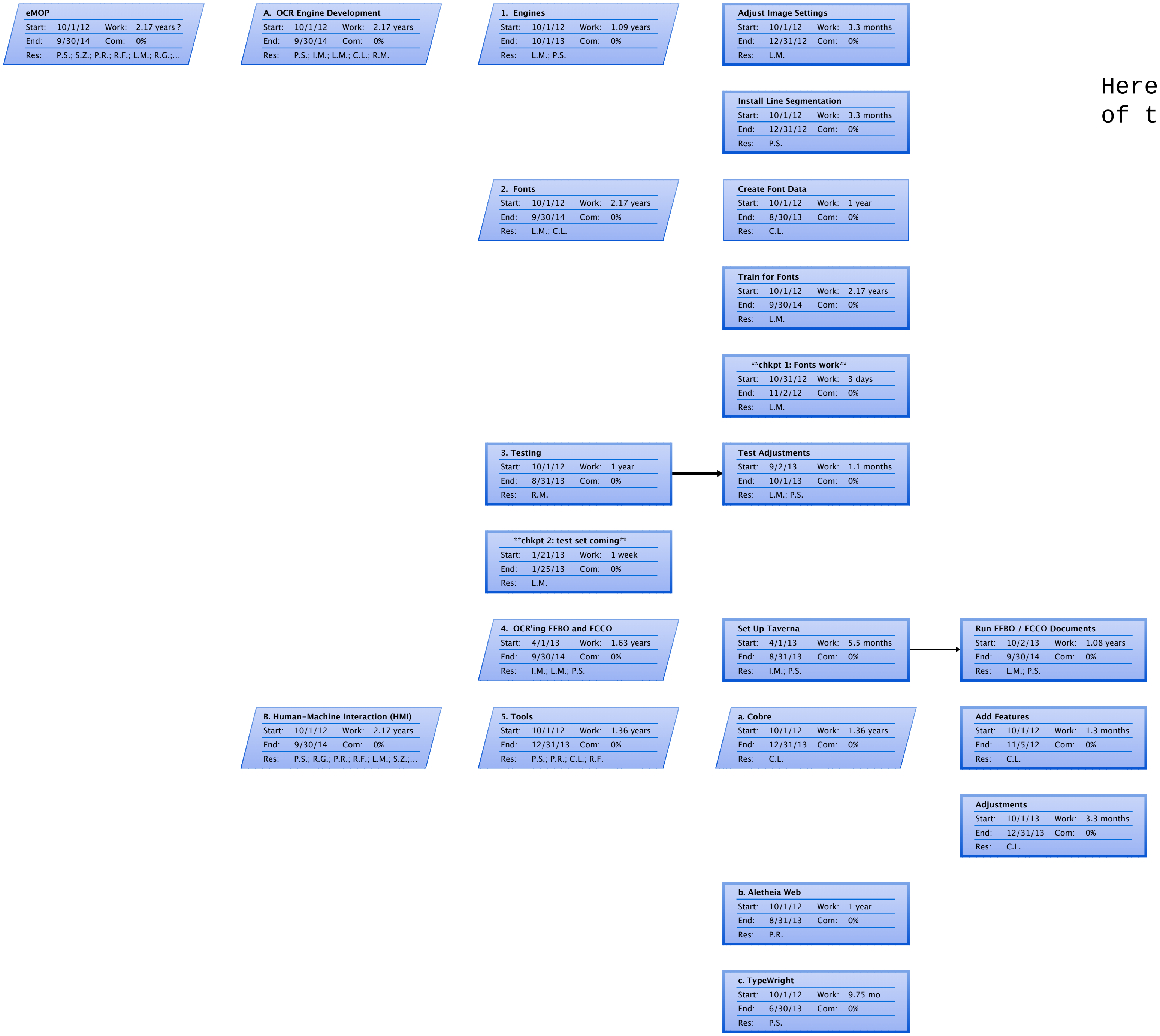
	Goals	Tasks	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
		Day-long pre-conference workshop on how to use TypeWright, Aletheia Desktop, and Cobre workshop, REKN announcement, at ASECS national meeting, 2013 (already scheduled)			L.M.					
		Set up editing groups to work on Cobre documents (Defoe Society, History of Science etc.)			L.M.	■	■	■	■	■
		REKn Launch (Ray Siemens, Richard Cunningham): will apply by 1 March 2013/2014 for paper session to be held at SAA (Shakespeare Association of America) meeting in St. Louis and Vancouver								
Milestone 3: Document Evaluation Working – December 2013										
C. OCR Correction	9. Manually Correct the OCR Output	Set up and run "voting algorithm" to compare the outputs of the three engines and choose the reading that has the most votes				R.M.	L.M.	■	■	■
		Create n-gram analysis and replacement rules		S.Z.	■	■				
		Create dictionary lookups by Levenshtein editing distance		■	■	■				
		Develop parameters for replacement rules of name and place gazetteers		■	■	■				
		Install Gazetteers from Underwood in Taverna					P.S.			
	10. Engage Humans in the Correction Process	Crowds work in Cobre, Aletheia Web, and TypeWright					L.M.	■	■	■
		Re-run documents after people have identified fonts or diagramed the page layouts, then send documents to TypeWright					P.S.	■	■	■
		Send all corrected documents to TypeWright as set up in 18thConnect and REKn					■	■	■	■
		Forward texts corrected in TypeWright by users (deemed reliable) to Gale, ProQuest, and the TCP, and index them in the ARC Catalog.								■
	CHECKPOINT 4: mechanical correction improves by 60%							March 2014		
	11. Save the Data	Give corrected texts to the people who corrected them					L.M.	■	■	P.S.
		Help correctors create library-quality electronic editions					■	■	■	■
		Export metadata corrections to the English Short Title Catalog for review								■
		Save correction histories to create a crowd-sourced correction data set in Institutional Repository (IR)					C.L.			P.S.
		Extract font identifications from Cobre Frankenbooks into Font History database, correlating ESTC number with typeface								
Milestone 4: 23.7 million pages now 97% correct, 99.9% once through TypeWright – Sept. 2014										
D. Dissemination	12. Release of Tools, OCR Workflow, and ESTC Databases to Improve knowledge	Release History of Font Importation Database				C.L.				L.M.
		Release database of documents needing rescanning by ESTC number								
		Submit for publication in REKn all revisions made in Cobre and saved by author (corrector) in the Texas A&M D-Space								L.M.
		Release the tools and Taverna workflow for download on Github and in IMPACT Competency Center								■
	13. Strengthen and sustain Crowd intervene-tion process	Create a plan for strengthening ARC support of NINES, 18thConnect, and REKn								L.M.
		Enlist Professors among special interest groups to lead (as editor/promoters) users of the tools for correcting and assisting OCR								L.M.
		Formulate a plan for how to record, monitor, and pool corrections made in tool instances worldwide								■
	14. Publish Results	Publish History of Fonts Database and Rescanning Database in Institutional Repository (IR)								■
		Publish Report on OCR'ing Early Modern Texts in IR and submit to CLIR								■
		Submit paper on optimizing Human-Computer Interaction								R.F.
Milestone 5: Tools and Workflow are being used worldwide – Dec. 2014										



Title	Initials	Role	eMail
Cushing Rare Books LIbrary	C.L.	Developer	duplessis@libra...
IMPACT (National Library of the Netherlands)	I.M.	Consultant	Clemens.Neude...
Laura Mandell (IDHMC / ARC)	L.M.	Project Lead	laura.mandell@...
Performant Software	P.S.	Developer	kristin@perfor...
PRImA (University of Salford)	P.R.	Developer	A.Antonacopou...
R. Manmatha (University of Massachusetts)	R.M.	Developer	manmatha@cs....
Ricardo Gutierrez-Osuna	R.G.	Project Lead	rgutier@cse.ta...
Richard Furuta	R.F.	Project Lead	furuta@gmail.com
SEASR (University of Illinois)	S.Z.	Developer	lauvil@illinois.edu


2011
2012
2013
2014
2015
2016
Q–6
Q–5
Q–4
Q–3
Q–2
Q–1
Q1
Q2
Q3
Q4
Q5
Q6
Q7
Q8
Q9
Q10
Q11
Q12
Q13
Q14
eMOP
A. OCR Engine Development
2.17 years
P.S.; I.M.; L.M.; C.L.; R.M.
Milestone 1: OCR Engines Optimized
B. Human–Machine Interaction (HMI)
2.17 years
P.S.; R.G.; P.R.; R.F.; L.M.; S.Z.; C.L.
Milestone 2: adapt 3 crowd–sourced corrrection tools
Milestone 3: 2% or less triage errors
C. OCR Correction
1.9 years ?
P.S.; C.L.; L.M.; S.Z.
Milestone 4: 173K documents 97–100% correct
D. Dissemination
1.18 years
L.M.; C.L.
Milestone 5: uptake of workflow, use of databases

Here we offer a diagram of the Task Table above.



eMOP
Start: 10/1/12 Work: 2.17 years ?
End: 9/30/14 Com: 0%
Res: P.S.; S.Z.; P.R.; R.F.; L.M.; R.G.;...

A. OCR Engine Development
Start: 10/1/12 Work: 2.17 years
End: 9/30/14 Com: 0%
Res: P.S.; I.M.; L.M.; C.L.; R.M.

1. Engines
Start: 10/1/12 Work: 1.09 years
End: 10/1/13 Com: 0%
Res: L.M.; P.S.

Adjust Image Settings
Start: 10/1/12 Work: 3.3 months
End: 12/31/12 Com: 0%
Res: L.M.

Install Line Segmentation
Start: 10/1/12 Work: 3.3 months
End: 12/31/12 Com: 0%
Res: P.S.

2. Fonts
Start: 10/1/12 Work: 2.17 years
End: 9/30/14 Com: 0%
Res: L.M.; C.L.

Create Font Data
Start: 10/1/12 Work: 1 year
End: 8/30/13 Com: 0%
Res: C.L.

Train for Fonts
Start: 10/1/12 Work: 2.17 years
End: 9/30/14 Com: 0%
Res: L.M.

chkpt 1: Fonts work
Start: 10/31/12 Work: 3 days
End: 11/2/12 Com: 0%
Res: L.M.

3. Testing
Start: 10/1/12 Work: 1 year
End: 8/31/13 Com: 0%
Res: R.M.

Test Adjustments
Start: 9/2/13 Work: 1.1 months
End: 10/1/13 Com: 0%
Res: L.M.; P.S.

chkpt 2: test set coming
Start: 1/21/13 Work: 1 week
End: 1/25/13 Com: 0%
Res: L.M.

4. OCR'ing EEBO and ECCO
Start: 4/1/13 Work: 1.63 years
End: 9/30/14 Com: 0%
Res: I.M.; L.M.; P.S.

Set Up Taverna
Start: 4/1/13 Work: 5.5 months
End: 8/31/13 Com: 0%
Res: I.M.; P.S.

Run EEBO / ECCO Documents
Start: 10/2/13 Work: 1.08 years
End: 9/30/14 Com: 0%
Res: L.M.; P.S.

B. Human-Machine Interaction (HMI)
Start: 10/1/12 Work: 2.17 years
End: 9/30/14 Com: 0%
Res: P.S.; R.G.; P.R.; R.F.; L.M.; S.Z.;...

5. Tools
Start: 10/1/12 Work: 1.36 years
End: 12/31/13 Com: 0%
Res: P.S.; P.R.; C.L.; R.F.

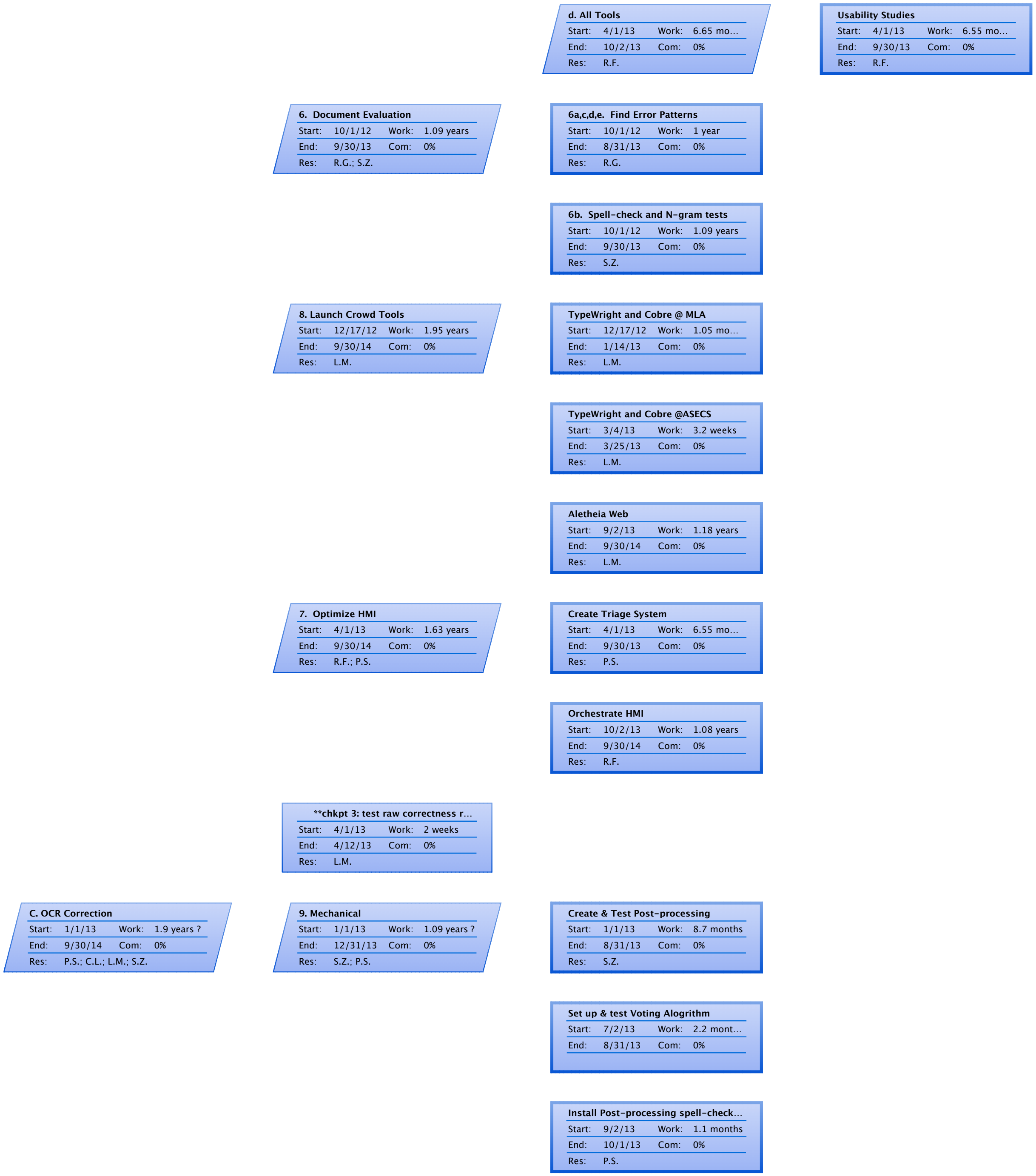
a. Cobre
Start: 10/1/12 Work: 1.36 years
End: 12/31/13 Com: 0%
Res: C.L.

Add Features
Start: 10/1/12 Work: 1.3 months
End: 11/5/12 Com: 0%
Res: C.L.

Adjustments
Start: 10/1/13 Work: 3.3 months
End: 12/31/13 Com: 0%
Res: C.L.

b. Aletheia Web
Start: 10/1/12 Work: 1 year
End: 8/31/13 Com: 0%
Res: P.R.

c. TypeWright
Start: 10/1/12 Work: 9.75 mo...
End: 6/30/13 Com: 0%
Res: P.S.



d. All Tools
Start: 4/1/13 Work: 6.65 mo…
End: 10/2/13 Com: 0%
Res: R.F.

Usability Studies
Start: 4/1/13 Work: 6.55 mo…
End: 9/30/13 Com: 0%
Res: R.F.

6. Document Evaluation
Start: 10/1/12 Work: 1.09 years
End: 9/30/13 Com: 0%
Res: R.G.; S.Z.

6a,c,d,e. Find Error Patterns
Start: 10/1/12 Work: 1 year
End: 8/31/13 Com: 0%
Res: R.G.

6b. Spell-check and N-gram tests
Start: 10/1/12 Work: 1.09 years
End: 9/30/13 Com: 0%
Res: S.Z.

8. Launch Crowd Tools
Start: 12/17/12 Work: 1.95 years
End: 9/30/14 Com: 0%
Res: L.M.

TypeWright and Cobre @ MLA
Start: 12/17/12 Work: 1.05 mo…
End: 1/14/13 Com: 0%
Res: L.M.

TypeWright and Cobre @ASECS
Start: 3/4/13 Work: 3.2 weeks
End: 3/25/13 Com: 0%
Res: L.M.

Aletheia Web
Start: 9/2/13 Work: 1.18 years
End: 9/30/14 Com: 0%
Res: L.M.

7. Optimize HMI
Start: 4/1/13 Work: 1.63 years
End: 9/30/14 Com: 0%
Res: R.F.; P.S.

Create Triage System
Start: 4/1/13 Work: 6.55 mo…
End: 9/30/13 Com: 0%
Res: P.S.

Orchestrate HMI
Start: 10/2/13 Work: 1.08 years
End: 9/30/14 Com: 0%
Res: R.F.

****chkpt 3: test raw correctness r…**
Start: 4/1/13 Work: 2 weeks
End: 4/12/13 Com: 0%
Res: L.M.

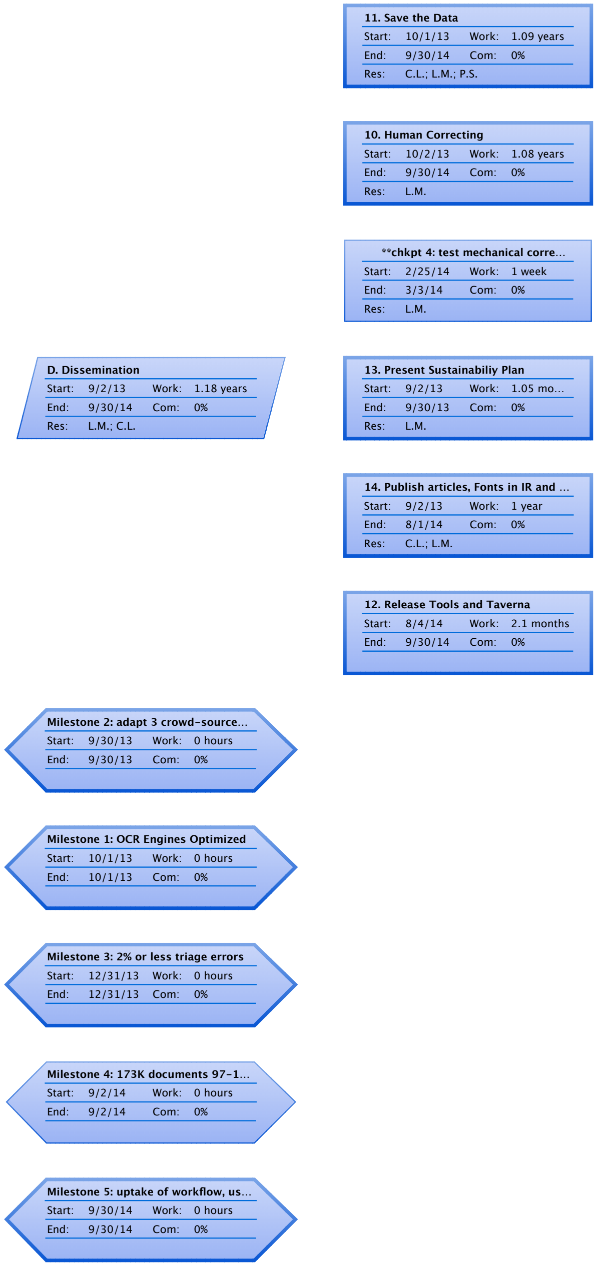
C. OCR Correction
Start: 1/1/13 Work: 1.9 years ?
End: 9/30/14 Com: 0%
Res: P.S.; C.L.; L.M.; S.Z.

9. Mechanical
Start: 1/1/13 Work: 1.09 years ?
End: 12/31/13 Com: 0%
Res: S.Z.; P.S.

Create & Test Post-processing
Start: 1/1/13 Work: 8.7 months
End: 8/31/13 Com: 0%
Res: S.Z.

Set up & test Voting Alogrithm
Start: 7/2/13 Work: 2.2 mont…
End: 8/31/13 Com: 0%

Install Post-processing spell-check…
Start: 9/2/13 Work: 1.1 months
End: 10/1/13 Com: 0%
Res: P.S.



11. Save the Data
Start: 10/1/13 Work: 1.09 years
End: 9/30/14 Com: 0%
Res: C.L.; L.M.; P.S.

10. Human Correcting
Start: 10/2/13 Work: 1.08 years
End: 9/30/14 Com: 0%
Res: L.M.

**chkpt 4: test mechanical corre...
Start: 2/25/14 Work: 1 week
End: 3/3/14 Com: 0%
Res: L.M.

D. Dissemination
Start: 9/2/13 Work: 1.18 years
End: 9/30/14 Com: 0%
Res: L.M.; C.L.

13. Present Sustainabiliy Plan
Start: 9/2/13 Work: 1.05 mo...
End: 9/30/13 Com: 0%
Res: L.M.

14. Publish articles, Fonts in IR and ...
Start: 9/2/13 Work: 1 year
End: 8/1/14 Com: 0%
Res: C.L.; L.M.

12. Release Tools and Taverna
Start: 8/4/14 Work: 2.1 months
End: 9/30/14 Com: 0%

Milestone 2: adapt 3 crowd-source...
Start: 9/30/13 Work: 0 hours
End: 9/30/13 Com: 0%

Milestone 1: OCR Engines Optimized
Start: 10/1/13 Work: 0 hours
End: 10/1/13 Com: 0%

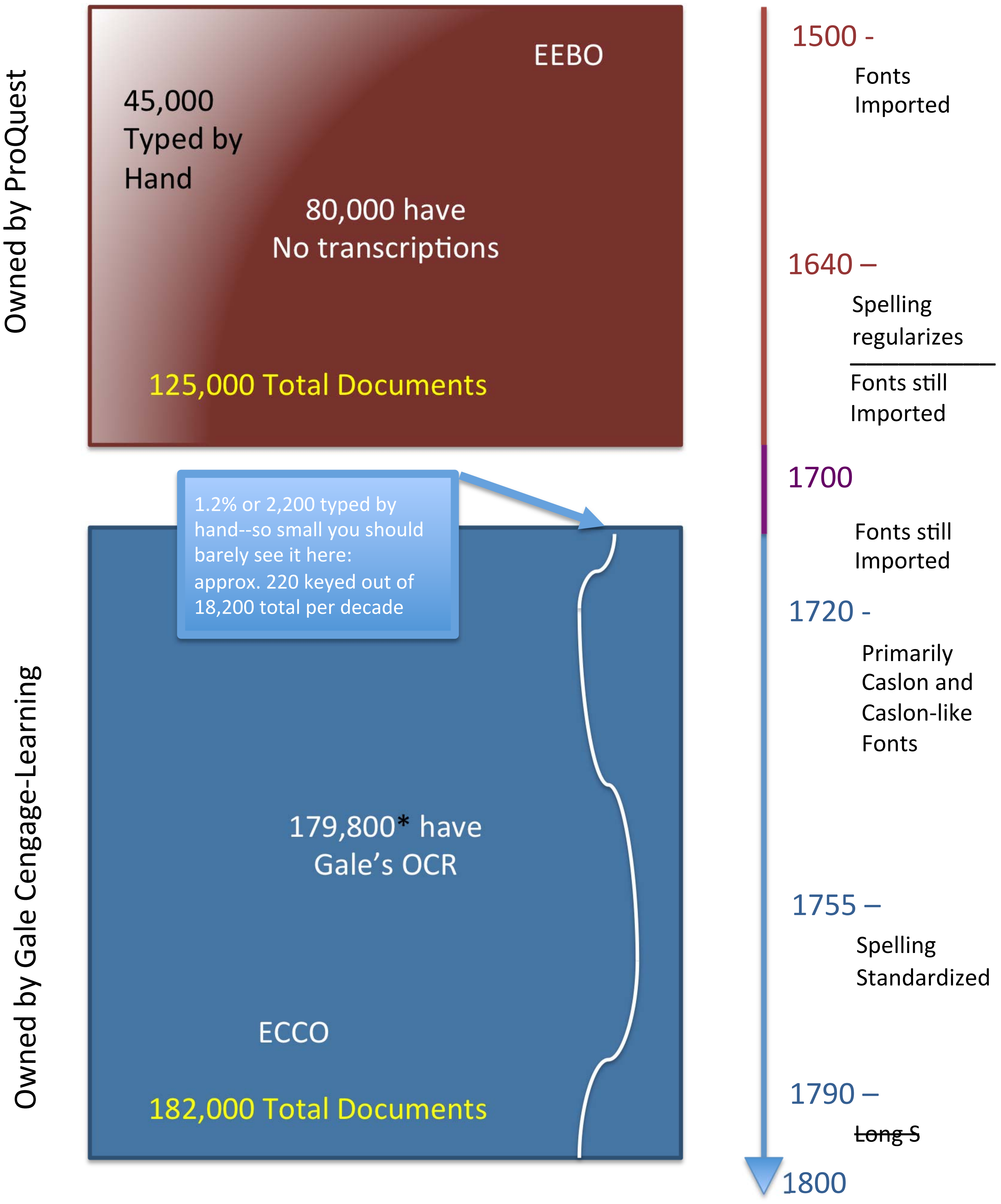
Milestone 3: 2% or less triage errors
Start: 12/31/13 Work: 0 hours
End: 12/31/13 Com: 0%

Milestone 4: 173K documents 97-1...
Start: 9/2/14 Work: 0 hours
End: 9/2/14 Com: 0%

Milestone 5: uptake of workflow, us...
Start: 9/30/14 Work: 0 hours
End: 9/30/14 Com: 0%

The Early Modern Data Set

Time



Owned by ProQuest

EEBO

45,000 Typed by Hand

80,000 have No transcriptions

125,000 Total Documents

1.2% or 2,200 typed by hand--so small you should barely see it here: approx. 220 keyed out of 18,200 total per decade

Owned by Gale Cengage-Learning

179,800* have Gale's OCR

ECCO

182,000 Total Documents

1500 - Fonts Imported

1640 – Spelling regularizes

Fonts still Imported

1700

Fonts still Imported

1720 - Primarily Caslon and Caslon-like Fonts

1755 – Spelling Standardized

1790 – ~~Long S~~

1800

*We rounded off to 2,000 typed / 180,000 OCR'd in the grant narrative

This early prospectus has been included because some of the commitment letters and subaward contracts refer to the Milestones included here.

Laura Mandell
Prospectus for an Application
November 30, 2011

OCR'ing Early Modern Texts

Attempts to preserve and make more accessible Anglo-American documents of literary and historical value began in the last third of the twentieth century with a major microfilming effort. Almost all the books, pamphlets, journals, and printed ephemera that had been catalogued by nineteenth- and twentieth-century antiquarians was microfilmed, and those images are now preserved digitally as well, primarily in the EEBO and ECCO datasets. To make those digital images of documents fully findable and to make it possible for new research questions to be asked based on them, the digital surrogates of print materials dating from 1500 to 1900 must be transformed into textual data—that is, into letters encoded in the way that machines can read them.

There are two ways to accomplish this transformation from page images into typed text:

1. Type the page images by hand; or
2. Read the page images mechanically using software programs, called Optical Character Recognition engines (hereafter OCR), to type them automatically.

The first alternative is very expensive and of course slow, rather like copying manuscripts by hand instead of producing copies using a printing press. In this case, we will have to make a selection of what to save, as has the Text Creation Partnership at the University of Michigan. If literary historians have learned anything from the advent of new historicism and cultural studies, movements in criticism that have shifted attention from a select canon of works onto everything from cheap popular newspaper stories to pamphlets and ephemera, it is to save everything.[1]

Digital tools will make it possible to ask new questions and learn new things from what could in fact become the entirety of our cultural heritage. But adequate tool functioning requires that mechanical typing work well, as close as possible to the 99%-correct figure formulated as a requirement by the US Government Printing Office's *Report on the [2nd] Meeting of Experts on Digital Preservation* (2004).[2]

However, OCR programs have not worked very well in mechanically typing texts from page images of books that were printed before 1800. For example, I have seen demonstrated a tool that could replace the Oxford English Dictionary. The OED was created by myriad contributors reporting upon instances of words that they encountered in their reading. Imagine now, with a properly digitized archive, being able to search all texts printed between 1500 and 1900 in order to determine where words first appeared in the English language, their shifts in meaning, and the heyday of their popular usage? The results returned by this tool, however, were flat-out wrong. "This word was never used

[1] David Simpson, "Is Literary History the History of Everything?" *SubStance* 88 (1999): 5-16.

[2] http://www.gpo.gov/pdfs/fdsys-info/documents/WhitePaper-OptimizingOCRAccuracy.pdf

before 1750," the demo insisted, and it was a word I had read recently in Jonathan Swift's *Gulliver's Travels* published in 1726. The tool itself is sound: the data going into it is making it look foolish. At stake is not just a digital OED; at stake is faith in digital technology on the part of humanists, librarians, and historians—all scholarship that relies upon philological data. It is precisely in order to avoid the problem of OCR's functioning upon early modern and eighteenth-century texts that the builders of the Google n-gram viewer used only documents printed after 1800 as part of their dataset.[3]

The reasons that OCR engines have trouble mechanically typing texts printed before roughly 1830 fall into two categories:

1. Poor quality page images (microfilmed images are "bitonal"—black and white only—making it very difficult to distinguish wayward marks from intentional imprints);
2. Irregularities in early modern and eighteenth-century printing practices.

We simply cannot rescan 300,000 documents, some of which are 1500 pages long, some 7 volumes, almost all housed in rare books rooms. Irregularities in print pose at least two problems for OCR engines:

a) It is difficult for these programs to properly segment any given page image into lines of text;

b) It is difficult to recognize images as discrete letters.

Without solving the first problem, it is impossible to solve the second: that is, without being able to "see" each line of text as a line, OCR engines cannot correctly identify images of letters as letters in a typed version of text and then transform them to their properly encoded alphabetic equivalent.

As R. Manmatha puts it, there is widespread belief among computer scientists that the problem of OCR was solved in the 1990s. Image expert and IMPACT member Apostolos A agrees: OCR is considered "good enough" for modern documents, presenting only engineering problems to be solved by companies commercializing the OCR systems. There are limited theoretical advances to be made in machine printed OCR of modern documents. It is because they are only worried about the modern that research-funding bodies such as the NSF do not see early modern OCR as a growth area and researchers do not engage in it.

However, we know that OCR of printed historical documents is far from being pushed to its limits. In this case, though, research will spring from implementing what we already know about OCR. For instance, given that some text-images or portions of them cannot be read by machines, how can we maximize human-computer interaction? How can we create an architecture sufficient for such triage that not only sorts machines from humans but selects the most qualified people for specific tasks, distinguishing good solid correcting ability from the capacity to actually train the OCR engines or create metadata necessary for productive re-scanning? And it turns out that expert knowledge in the field of book history may help us to maximize OCR output in ways not yet tried. If

[3] Jean-Baptiste Michel, et. al., "Quantitative Analysis of Culture Using Millions of Digitized Books," *Sciencexpress* < http://www.sciencemag.org/content/early/2010/12/15/science.1199644.full.pdf>; Erez Lieberman-Aiden and Jean-Baptiste Michel, "Culturonomics," *DH2011 Conference Abstracts*, p. 8.

training OCR engines improves their accuracy, couldn't tracing the importation of European fonts help? We could then create training sets to run on specific documents printed with these specific fonts. In a fictional example, knowing that all the documents printed by a London printer between 1683 and 1694 used an early Dutch font could allow us to use one set of training materials on those documents in particular.

At the OCR Summit meeting held at Texas A&M University this Fall, leaders in the field of OCR from Google's Ranjith Unnikrishnan to IMPACT's Clement Neudecker met to discuss how to move forward in producing better OCR results of early modern and eighteenth-century texts. We all agreed at the outset that indeed, as NSF suspects, there has been enough *innovation* in the field and not enough *implementation* of innovative ideas. Partly this is because the only archives that have enough resources to attempt implementation also have enough resources to pay for keying, and they do so because keying is sure fire.

What we—a subset from the group of people who met here in October (http://idhmc.tamu.edu/commentpress/participants/)—propose to do is to prove, as part of our full application to the Mellon Foundation for funding, that applying innovations will in fact pay off when OCR'ing Early Modern Texts: in the full application, we will submit a sample to show what can be done. If awarded a grant, we will use Mellon funding to enhance that payoff in greater accuracy by refining multiple open-access OCR engines and creating human-computer interaction processes, all informed by those computer-scientists' innovations. Next, we will run through those engines and crowd-sourcing procedures over 300,000 texts, many of them multivolume, collected in the ECCO and EEBO catalogues—texts printed from 1450 to 1800. In addition, we will use our engines on the substantial collections at Texas A&M Libraries of early modern materials that have been collected for the Cervantes Project and Digital Donne. Finally, we'll conclude by using our engines on the 65,000 nineteenth-century texts in JISC Historic Collections.

Who Owns What?

Before describing what we intend to do, proprietary issues must be discussed concerning 1) OCR engines and enhancements, 2) early modern texts, and 3) textual corrections.

1) Out of two of the best OCR engines, Google's Tesseract and ABBYY FineReader, one is proprietary (the latter). According to Tesseract developers, there could be some question as to who owns the OCR output generated by commercial engines. In any case, though ABBYY FineReader currently has better results, it seems to us possible with less financial investment than ABBYY has made to get better results using two open-access engines: Gamera and Tesseract itself. All training and improvements of these engines funded by the Mellon Foundation will be freely available.

2) The history of holding libraries such as the British Library is such that ownership of microfilmed copies of their collections has passed out of their hands. That means that we must deal with proprietary texts unless one plans to re-scan all 300,000 titles—again, many of them multivolume.

Not only would costs be exorbitant, but rare books rooms at many US libraries have contracts with the owners of the microfilms, ProQuest and Gale Cengage Learning, prohibiting them from rescanning for open access. Additionally, JISC Collections, a commercial venture, was founded precisely to make ECCO and EEBO affordable to all UK Universities. In short, the lines between proprietors and users are not simply drawn.

The contracts that 18thConnect, NINES, and ARC have been able to negotiate with the companies who own these early modern and eighteenth-century texts are good ones, as these contracts will protect the libraries' and the companies' investments while curtailing the effects of a digital divide between scholars working at universities who can afford to purchase these full-text collections and those who cannot.

3) All textual corrections produced on funding provided by Mellon, whether on data owned for distribution exclusively in the UK or by the companies with whom we are partnering, Gale Cengage Learning and ProQuest, will be sharable with anyone and open access.

Ultimately, procuring the best possible texts—machine-readable and fully searchable—matters tremendously beyond who actually owns the data. Professors have in fact long been supporting presses by working on editions, commercial presses as well as university presses, the editors themselves reaping little extra remuneration beyond their salaries. Now, under the new dispensation, faculty salaries and student educational experiences can both contribute toward creating scholarly editions of many, many texts beyond the canonical—Gale Cengage Learning and ProQuest merely being the equivalent of Oxford World Classics in terms of commercial involvement. More scholars and many more students have to be involved in creating this massive digital archive adequate to preserving our cultural heritage, texts of historical interest as well as literary merit. As Martin Mueller says, however, it is time for scholars now to do our own dishes. But this new task imparts more than the dishwashing metaphor implies. If creating reading editions of texts has been something that students and professors have not had to do in general, the results of this lacuna have been intellectual poverty, bias, and learning activities too inconsequential to become passionate about. We are entering "the age of editing," Jerome McGann says, and that is all to the good.

What We Propose to Do:

We propose to undertake a program that is very clear-cut and limited in its goals, but nonetheless will have a powerful impact upon the future of scholarship: create a better set of OCR programs for early modern texts; run through OCR engines the 300,000 digitized early modern texts (billions of pages, digitized from microfilm) in EEBO and ECCO plus 65,000 19th-century texts; create multiple interacting venues for crowd-sourced correction and further OCR training; create methods for sorting based on communities.

Here follow the work-packages or milestones that we hope to achieve, listing participants involved and tasks to be worked on simultaneously, with roughly estimated costs that would be spelled out completely in a full proposal

I. Train Tesseract, open-access OCR program

Milestone 1. Train Tesseract, which, based on the research done by Ray Smith, has solved the line segmentation problem, as thoroughly as we have Gamera on the long 's' to produce results that are better than the OCR produced for Gale by Prime. This milestone will be accomplished before we apply for Mellon funding as it seems necessary to prove that this project warrants further investment. Results will be reported as part of the Final Report for the Officer's Grant.

II. Train Gamera, open-access OCR program

Milestone 2. Create better OCR engines ($177,000):

1. The Initiative for Digital Humanities, Media, and Culture (hereafter IDHMC) will create a ground truth using Aletheia that has been developed by the Pattern Recognition and Image Analysis research lab (hereafter PRImA) and text that has been double-keyed by the Text Creation Partnership (hereafter TCP). $66,000
2. The IDHMC and Google will work together on improving Tesseract based on evaluations using our Ground Truth. $0
3. The IDHMC will work together on improving Gamera using the Ground Truth created in #1 and programming Gamera for better line segmentation, copying Tesseract's strategies; R. Manmatha, working for the IDHMC, will adapt Ray Smith's line segmentation rules built for Tesseract to Gamera. We need to develop engines (plural) in order to help us with estimating errors of documents that have not been keyed and whose spelling variants appear nowhere. $81,000
4. The IDHMC will work with Performant Software to run our data through Tesseract and Gamera. $30,000

III. Sort texts by fonts and use appropriately trained engines.

Milestone 3. Sorting Texts with Fonts ($80,000)
Relying on the data set provided by Ian Maxted (the basis of the British Book Trade Index) as well as Stationer's Registry, IDHMC is creating a database-driven mapping of the movement of paper and fonts through London's booksellers and printers. Crucial before 1720 when fonts were imported from Europe, this mapping will help us to determine: a) which fonts to train extensively our OCR engines to read; b) which libraries to use with the engines when running sets of texts—it may be that all texts printed from 1681 to 1693 by a printer who worked for 13 booksellers could be run using one training set. Jon Orwant of Google has said about getting good OCR results: "training, training, training." We proved that mantra to be correct with Gamera's training in recognizing the long 's'. Breaking up the 300,000 ECCO and EEBO texts into smaller subsets according

to their fonts and thus focusing the kinds of training needed for each subset will improve OCR results remarkably.

1. Hire a postdoc to complete Maxted's database. $65,000
2. Help training two OCR engines, Tesseract and Gamera, in European fonts, as they have done for ABBYY: 180 hours from Clemens Neudecker and Hildelies Balk. $15,000

IV. Choose which output is most correct to use.

Milestone 4. Evaluation ($37,500)

1. Performant Software, working for the IDHMC, will create another Solr indexer, identical to the current ARC indexer, to hold the output of the OCR engines. Hosting will of this index will be on the IDHMC servers that also host the current ARC Solr index and search engine. $17,500
2. Performant Software, working for the IDHMC, and SEASR will develop correctness metrics for automatically determining how well a page has been "read"—not a test of the OCR engines themselves using ground truthing, but instead a test of how well an engine has performed on a specific document. Optimal time-outs need to be set so that an engine's inability to find lines in a particular image or document will not consume too many computing resources. Number of unique n-grams per document is a rough measure of how well a document has been read, and percentage correctness can be passed forward to the triage procedures developed in Milestone 4. $20,000

V & VI. Build and adapt tools for correcting / postprocessing / more training

Milestone 5. Post-Processing ($70,000)

1. Performant Software, working for the IDHMC, will hook up the OCR output that is evaluated at acceptable levels of correctness (85% or above) to the post-processing correlation n-gram analyzer provided via the Meandre workbench by SEASR. $10,000
2. SEASR will perform post-processing dictionary lookups. The dictionary of variant spellings that this tool will use will come from Martin Mueller of Northwestern University who has developed this from the TCP ECCO data, from Ted Underwood of the University of Illinois who has created variant spelling lists from the MONK data, and from the IDHMC who is digitizing early modern dictionaries held by Larry Mitchell, director of the Cushing Special Collections Library at Texas A&M University. The analyzer based on these historically precise dictionaries and Martin's usage tool will live in Bamboo Corpora Space and will contain dictionaries that can be added and subtracted from the tool's analysis. $60,000

Milestone 6. Tools ($145,000)

The development of these tools will be overseen by the IDHMC: Aletheia will be adapted by PRImA Labs; Active OCR is a work under development by Travis Brown of Bamboo Corpora Space; and Texas A&M libraries will work with Brian Geiger and Carl Stahmer of the English Short Title Catalog (hereafter ESTC) Database Project.

1. Aletheia needs to be simplified and developed into a web-based tool: $65,000
2. CONCERT, a web-based training tool for expert correctors that has been developed by IBM, could be used with permission or at a certain cost. If not, Active OCR System needs to be developed into a web-based tool. $50,000 Bamboo Corpora Space
3. Add metadata forms to Texas A&M's Cobré tool, used to compare North American Incunables in the Primeros Libros Project, for gathering more extensive metadata based on the comparison of multiple title pages: $20,000 IDHMC; $10,000 ESTC

VII. Triage: create an automated way to send texts to appropriate correction tools and qualified correctors:

Milestone 7. Triage ($75,000)

Richard Furuta ($25,000) plus a graduate student ($50,000) at the Center for the Study of Digital Libraries (hereafter CSDL) will create a system for sorting out OCR results in order to funnel texts and pages images to specific user communities:

1. Advanced experts will work with documents printed in unusual or heretofore unidentified fonts training the OCR engine using an interactive training tool (IBM's CONCERT or MITH's Active OCR system);
2. Advanced experts capable of extracting bibliographical data that will be forwarded to the ESTC;
3. Users capable of adjusting line segmentation using Aletheia, the resulting document being resubmitted to the engines for re-OCRing;
4. The "Crowd" capable of making typographical corrections: these documents will be sent on to post-processing and then to the TypeWright crowd-sourced correction tool.
5. Set up versioning so that data corrections do not overwrite each other but work together as "votes" for the most correct reading.

In the US, the tools and crowds will live and meet via Bamboo Corpora Space; the data will be locked down in the SOLR indexer hosted by ARC and accessible via the ARC web service. For the UK, see Milestone 7.

VIII. Coordinate global corrections.

Milestone 8: Collect User Corrections Globally ($17,500)

JISC Collections is applying for funding from the JISC funding agency to install versions of the TypeWright tool in the UK JISC Historical Collections Platform. Gale-Cengage Learning has agreed to let the UK share its corrected ECCO data with us, despite the fact that JISC Collections has purchased ECCO for use only in the UK. Our goal will be to set up a way of sharing those corrections back and forth, using TEI-L for version control.

Total amount: $602,000

Laura Mandell <laura.mandell@gmail.com>

Experiments with Tesseract

4 messages

Ranjith Unnikrishnan <ranjith@google.com> **Thu, Feb 9, 2012 at 5:50 PM**
To: Laura Mandell <laura.mandell@gmail.com>

Hi Laura,

I ran several experiments using the labeled data that you sent me. Unfortunately the results are somewhat inconclusive. Here are some observations, roughly ordered from bad to good news:

First, adding the Caslon font and the manually labeled images of characters to the Tesseract training process does not seem to affect its overall accuracy. This is based on evaluation with the images that you sent as well as on the few books in our test set. There are a couple of possibilities for why the addition of real data is not making an immediate difference:

- the Caslon font as well as the manually labelled data are not very visually different from characters in the Wyld font that we already train Tesseract on.

- the images you sent me are binarized (black and white) images and have relatively poor quality compared the images we usually work with. Normally Tesseract takes as input a grayscale image and then does its own custom binarization operation before doing the recognition. It is possible that the already binarized images that you sent result in Tesseract computing poor features which reduces its overall accuracy. If you expect all your images to be of the quality of the pages that you sent me, it might be necessary to do some preprocessing on them to "clean" them up. For example, I was able to extract slightly better text by changing Tesseract's configuration to do more aggressive noise removal on the page. But doing that on a relatively clean page could result in loss of genuine characters that have small size, like punctuation marks. So in short, we'd either need to know beforehand what the image quality will be like so we can set the right configuration mode for Tesseract, or invest time to develop image cleanup algorithms, or preferably work on the original grayscale images from which the images you sent me were derived.

- The type of errors Tesseract is making is not the kind that will go away easily by using more labeled training data. For example, the bulk of the errors occur on italicized words, which are particularly challenging in older books because of how different and stylized the italic fonts are in comparison to regular upright fonts from the same period. The problem of accuracy on italic words is something that we're aware of and are working on, but is not something that will be easily resolved by training on more labeled data. So while it shouldn't hurt, I'm doubtful from the nature of the errors that significant accuracy gains can be had simply by acquiring more labeled training data.

All that said, I do believe that Tesseract either as is or trained with the labeled data you provided will do a good job of (i) textline finding, and (ii) recognizing historic ligature forms particularly on upright text when the images are of good quality. I've attached some examples of images and OCR output to illustrate:

trial_lines.enm.png is a page from one of our scanned books that contains the same content as the book whose pages that you sent me but is of better image quality. (I was hoping to find the exact same book in our corpus but couldn't find it). The image shows in blue boxes the text lines detected by Tesseract's page layout analysis. The page also contains characters with several historic ligatures and italic words, and although the OCR is not perfect you'll see in the accompanying **trial.txt** file that it gets most of them correctly.

wright_lines.enm.png is a page from another book. It contains Old English, but I include it here because it has (i) a more complex layout despite which we are still able to find correct textlines as well as correctly order and group them into paragraphs, and (ii) several words where some of the letters are in a different font that other letters of the same word, as well as some that contain historic symbols like the 'Thorn' symbol that looks remarkably like a lowercase 'p'. The OCR output for the page is in **wright.txt**.

Let me know what you think, and if you could use some more example images.
~Ranjith

4 attachments

THE LATE

Trial and Conviction

OF

Count TARIFF.

[illegible]

trial_lines.enm.png
78K

wright_lines.enm.png
77K

trial.txt
1K

wright.txt
2K

Laura Mandell <laura.mandell@gmail.com> **Fri, Feb 10, 2012 at 9:32 AM**
To: Ranjith Unnikrishnan <ranjith@google.com>, m-farrington@tamu.edu

Ranjith:

First, THANK YOU SO MUCH. It is bad news in one way, but not in another: it's exactly the same result we had with Gamera, which means that we were not doing something wrong. The line segmentation is beautiful, isn't it? Tesseract is clearly very powerful. So all that is to the good, and we will send you what we make of it in the next few days to see if our proposal about how to move forward makes sense to you.

Thanks.
Best, Laura

[Quoted text hidden]

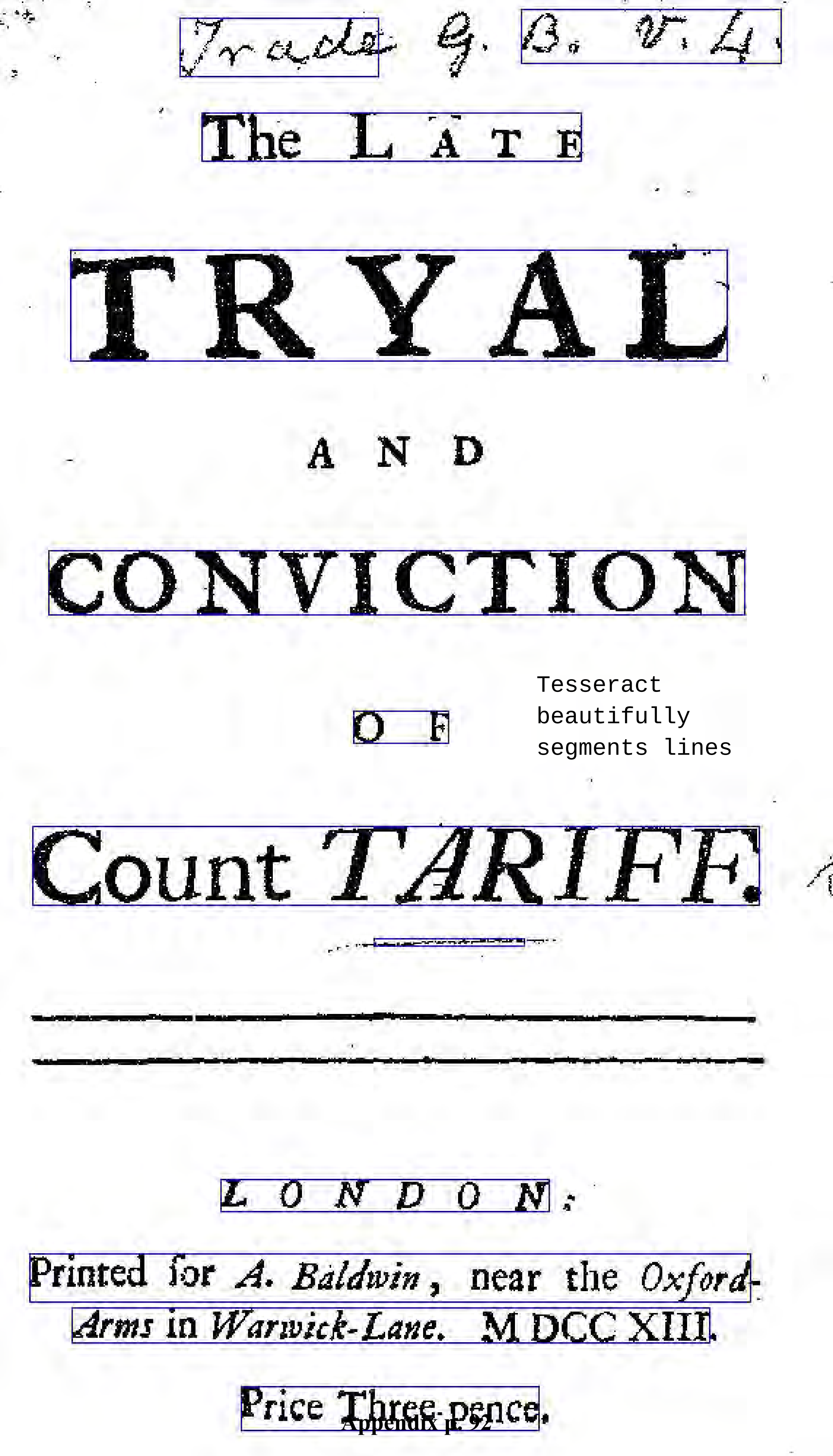
--
Laura Mandell
Director, Initiative for Digital Humanities, Media, and Culture
Professor, English
Texas A&M University
p: 979-845-8345
e: mandell@tamu.edu
@mandellc
http://idhmc.tamu.edu

Laura Mandell <laura.mandell@gmail.com> **Fri, Feb 10, 2012 at 9:32 AM**
To: m-farrington@tamu.edu

[Quoted text hidden]

[Quoted text hidden]

4 attachments



Trade G. B. V. 4.

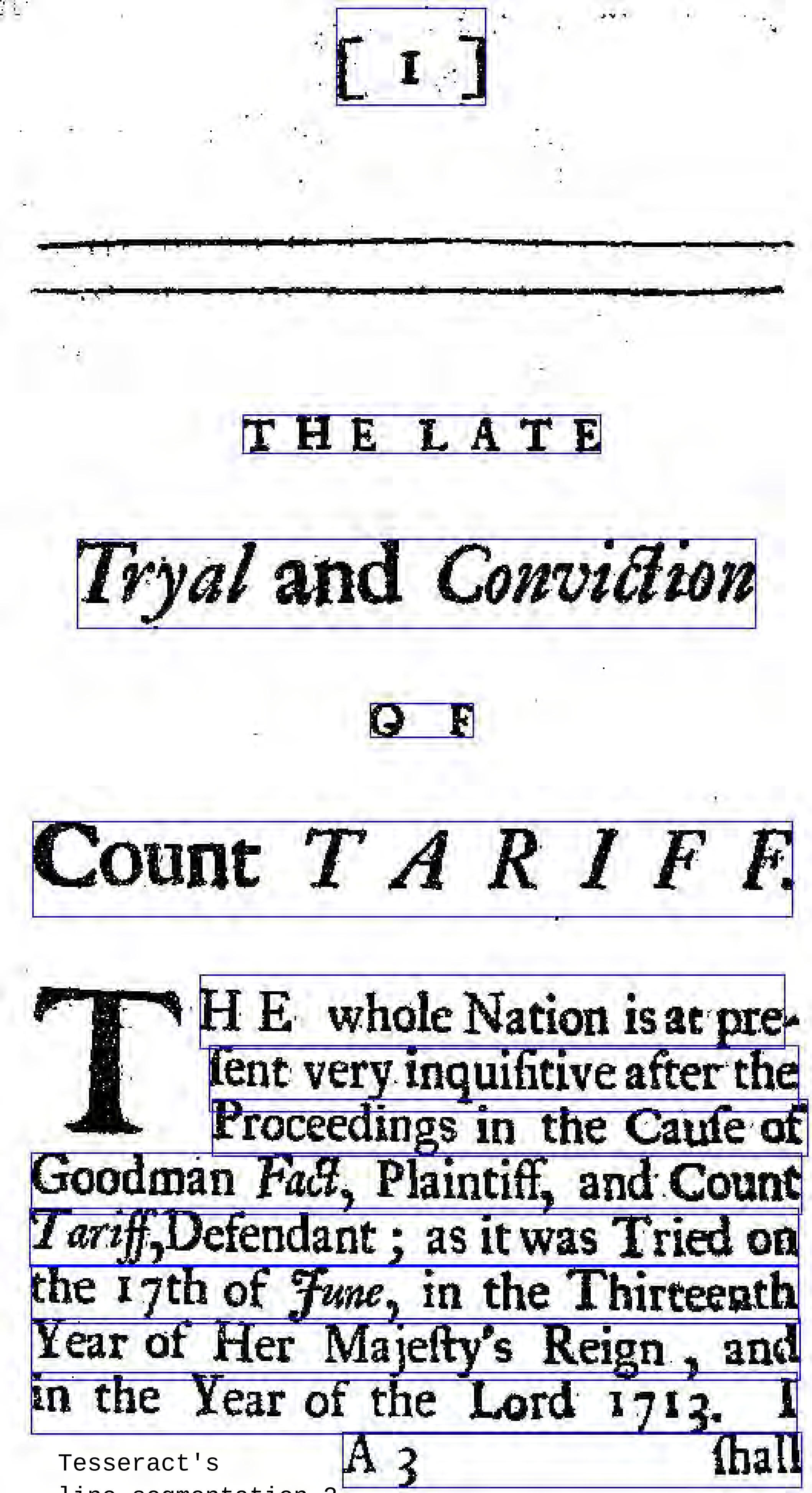
The LATE TRYAL AND CONVICTION OF Count *TARIFF.*

LONDON:

Printed for *A. Baldwin*, near the *Oxford-Arms* in *Warwick-Lane*. MDCCXIII.

Price Three-pence.

Tesseract beautifully segments lines



[1]

THE LATE

Tryal and *Conviction*

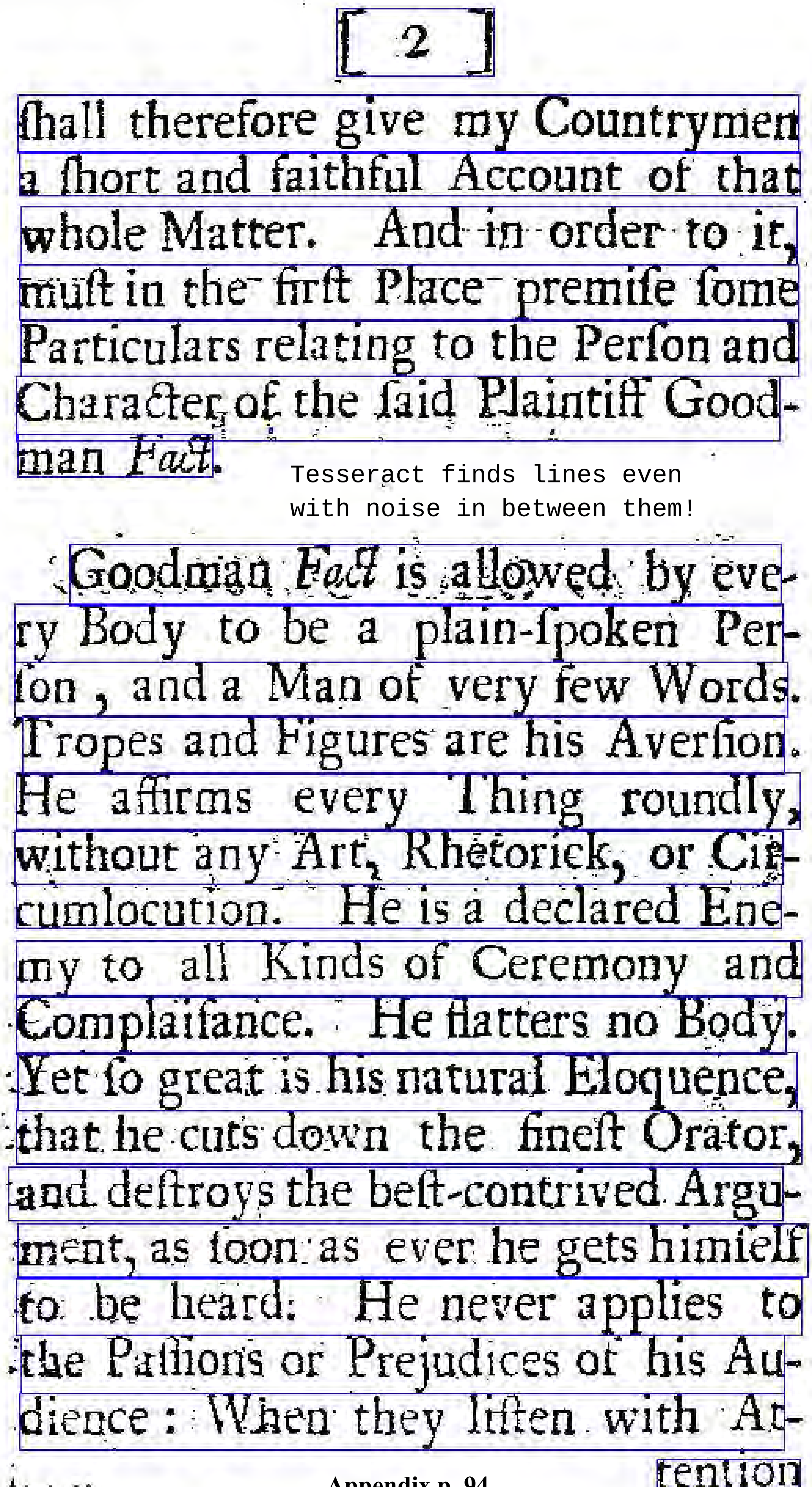
OF

Count *TARIFF*.

THE whole Nation is at preſent very inquiſitive after the Proceedings in the Cauſe of Goodman *Fact*, Plaintiff, and Count *Tariff*, Defendant; as it was Tried on the 17th of *June*, in the Thirteenth Year of Her Majeſty's Reign, and in the Year of the Lord 1713. I

A 3 ſhall

Tesseract's line segmentation 2



[2]

ſhall therefore give my Countrymen a ſhort and faithful Account of that whole Matter. And in order to it, muſt in the firſt Place premiſe ſome Particulars relating to the Perſon and Character of the ſaid Plaintiff Goodman *Fact*.

Tesseract finds lines even with noise in between them!

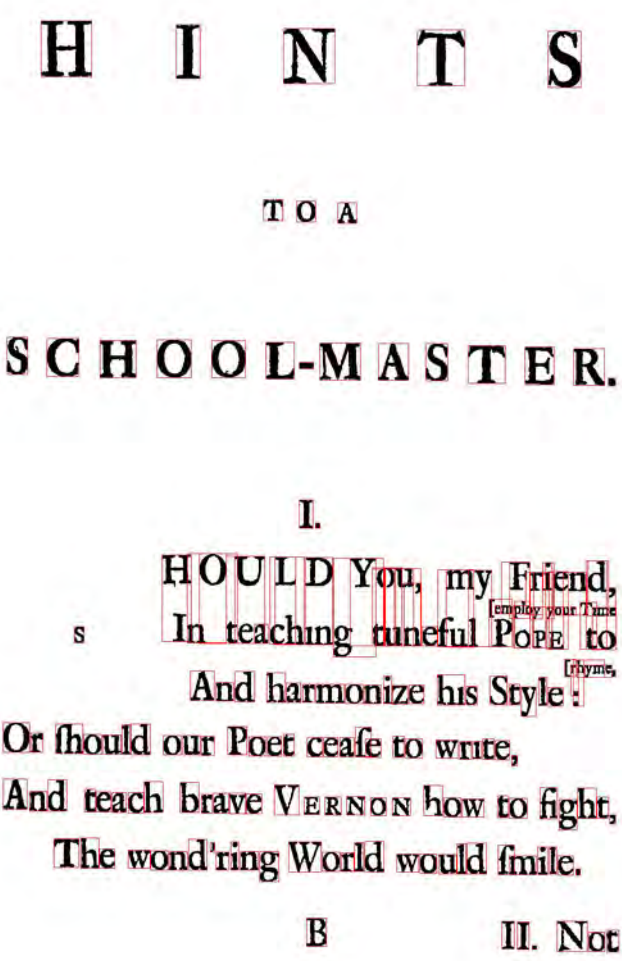
Goodman *Fact* is allowed by every Body to be a plain-ſpoken Perſon, and a Man of very few Words. Tropes and Figures are his Averſion. He affirms every Thing roundly, without any Art, Rhetorick, or Circumlocution. He is a declared Enemy to all Kinds of Ceremony and Complaiſance. He flatters no Body. Yet ſo great is his natural Eloquence, that he cuts down the fineſt Orator, and deſtroys the beſt-contrived Argument, as ſoon as ever he gets himſelf to be heard. He never applies to the Paſſions or Prejudices of his Audience: When they liſten with At-

tention

Tesseract's Box File Format

```
s 129 1976 157 2026
h 149 1975 177 2025
a 183 1976 205 2007
l 229 1977 240 2026
l 212 1977 224 2025
t 273 1976 290 2013
h 293 1977 322 2026
e 330 1977 352 2009
r 358 1977 379 2008
e 381 1977 405 2009
f 409 1979 434 2027
o 427 1979 456 2011
r 461 1980 481 2011
e 484 1981 509 2013
g 535 1963 566 2012
i 573 1982 584 2032
v 591 1982 617 2011
e 621 1981 646 2012
m 685 1980 730 2012
y 738 1960 767 2012
C 789 1979 831 2031
o 834 1980 863 2012
u 867 1979 894 2012
n 900 1979 928 2010
t 935 1980 950 2018
  957 1978 978 2012
y 979 1958 1009 2011
m 1015 1977 1060 2014
e 1066 1977 1089 2010
n 1093 1975 1124 2013
a 132 1905 157 1934
s 185 1906 210 1956
h 201 1907 230 1953
o 236 1907 263 1936
r 267 1907 287 1936
t 292 1907 311 1941
a 333 1907 356 1936
n 364 1907 391 1936
d 395 1907 425 1956
f 452 1910 475 1956
a 472 1909 494 1939
i 501 1911 515 1960
t 520 1911 538 1947
h 542 1911 572 1956
f 577 1911 600 1956
u 597 1908 625 1939
l 628 1911 644 1959
A 678 1911 722 1957
c 727 1911 751 1941
c 753 1910 776 1941
o 781 1910 809 1940
u 814 1909 840 1941
n 846 1910 874 1940
t 880 1911 898 1947
o 934 1909 963 1940
f 967 1909 992 1952
t 1020 1909 1037 1943
```

The line segmentation algorithm in the Gamera Toolkit is too primitive to use with early modern texts:



HINTS

TO A

SCHOOL-MASTER.

I.

SHOULD You, my Friend, [employ your Time
In teaching tuneful POPE to [rhyme,
And harmonize his Style:
Or ſhould our Poet ceaſe to write,
And teach brave VERNON how to fight,
The wond'ring World would ſmile.

B II. Not

Here, two lines are being read as one

Here, lines are not
being recognized.

Earl Oſmond's jeſter had fled the country. I exerted my knavery for the laſt time in ſtealing the fugitive's caſt coat, was accepted in his place by the Earl, and now gain an honeſt livelihood by perſuading my neighbours that I'm a greater fool than themſelves.

PERCY. And your change is for the better?

MOTL. Infinitely; indeed your fool is univerſally preferred to your knave—and for this reaſon; your fool is cheated, your knave cheats: Now every-body had rather cheat, than be cheated.

PERCY. Some truth in that.

MOTL. And now, ſir, may I aſk, what brings you to Wales?

PERCY. A woman, whom I adore.

MOTL. Yes, I gueſſed that the buſineſs was about a petticoat. And this woman is——

PERCY. The orphan ward of a villager, without friends, without family, without fortune!

MOTL. Great points in her favour, I muſt confeſs. And which of theſe excellent qualities won your heart?

PERCY. I hope I had better reaſons for beſtowing it on her. No, Gilbert; I loved her for a perſon beautiful without art, and graceful without affectation—for an heart tender without weakneſs, and noble without pride. I ſaw her at once beloved and reverenced by her village companions: they looked on her as a being of a ſuperior order; and I felt, that ſhe who gave ſuch dignity to the cottage-maid, muſt needs add new luſtre to the coronet of the Percies.

MOTL. From which I am to underſtand that you mean to marry this ruſtic.

PERCY. Could I mean otherwiſe, I ſhould bluſh for myſelf.

MOTL.

Notes on using Gamera with 18th Connect texts and Brazos

by David Woods, Miami University

4 May 2012

Tools being used are Gamera (http://gamera.informatik.hsnr.de/) and the OCR Toolkit Gamera addon (http://gamera.informatik.hsnr.de/addons/ocr4gamera/index.html). Both tools are written in Python and C++. According to the documentation, Gamera will process images in TIF or PNG format, but all testing was done with TIF images.

There are two main phases for using Gamera. One is to run the gamera_gui program to open an image. Gamera then identifies all of the instances of black pixels completly surrounded by white pixels (glyphs). The user then trains Gamera by specifying what character each glyph represents. Training also allow combining glyphs, etc. The result is training data stored in an XML file. Training for the 18th Connect images was done on a sub-set of images by Mike Behrens at Illinois. This should be done on a local computer because the need for a graphical interface.

Here is an example of the XML training data for a glyph:

```
<glyphs>
  <glyph uly="136" ulx="385" nrows="56" ncols="62">
    <ids state="MANUAL">
      <id name="latin.capital.letter.a" confidence="1.000000"/>
    </ids>
    <data>
      30 3 58 6 55 8 54 8 54 9 52 10 52 10 51 12 50 12 49 14 47 15 47 4 1
      10 46 5 2 9 46 4 4 9 45 4 4 9 45 4 5 8 44 4 6 9 42 5 7 8 42 4 8 9 41
      4 9 8 40 4 10 8 40 3 11 9 38 4 11 9 38 3 12 9 37 3 14 9 36 3 14 9 34
      4 15 8 35 3 16 9 33 4 16 9 33 5 15 10 31 19 1 11 31 32 29 33 29 4 3 2
      2 8 4 10 29 3 22 9 27 4 22 9 27 3 23 10 26 3 23 10 26 2 25 9 25 2 26
      10 23 3 27 9 23 3 27 10 21 4 27 10 20 5 28 10 19 5 28 10 18 6 28 11
      17 6 29 10 17 6 29 11 15 7 29 11 14 8 28 13 12 10 27 14 10 12 24 19 5
      17 18 25 1 18 18 26 1 15 20 26 42 19 1 0
    </data>
  </glyph>
```

The “<id name=”latin.capital.letter.a …” shows the name assigned during the training process. The names should follow Unicode format (http://www.unicode.org/charts/PDF/U0000.pdf) when possible, but other names can be used.

The other phase is to do the actual OCR on an image. This is implemented in the OCR toolkit, which uses training data and an additional file to process images. The images are split into glyphs and processed with line and word segmentation routines. The training data is used to find the best match for each glyph, and then processed to provide text output. The names assigned in the training data are used to extract the text character for the output. These can be overridden using an “extra characters” CSV file to map the “id name” from the training data to a specific character. The “extra characters” file is currently used for ligatures, italics characters and some joined capitals.

Both Gamera and the OCR toolkit are installed in my user area on Brazos - /home/ext-woods/tools/gamera. Note that the Gamera and OCR Toolkits installed into separate directories under gamera/lib64 - python and python2.4 - and I manually collected the code under the python sub-directory.

The installation on Brazos contains some modifications:

- OCR Toolkit code has been modified to add functionality to produce XML output in an 18th Connect specific format.
- Gamera code has been modified to comment out Python import statements for wx (graphics library) which is not available on Brazos compute nodes. This code is only needed for the Gamera GUI.

The current training data and extra characters CSV file for 18th Connect are in /home/ext-woods/training on Brazos. The training data file is named “Baskerville_library.xml” since the training focused on the Baskerville font. It can be used with any images, but will work best with works using the Baskerville font.

18th Connect XML output format

The XML files produced for 18th Connect are actually fragments since they do not contain headers, etc. The XML starts with a <page> tag that contains the name and path to the image file. Within the page are <line> elements, which contain <wd> (word) elements. The word elements have position info for a bounding box around the word, and the recognized text.

Running Gamera on Brazos

Setup :

- Add /home/ext-woods/tools/gamera/bin to PATH
- Add /home/ext-woods/tools/gamera/lib64/python to PYTHONPATH

The command that is run for OCR processing is “ocr4gamera.py”, which has lots of command options. The most interesting options are:

- -v n – sets the verbosity level. The default is 0 (silent). A setting of 1 adds comments about progress to the output. A setting of 2 will produce images showing the original image with overlaid red boxes showing the character, word, and line segmentation. The files are debug_chars.png, debug_words.png and debug_lines.png
- -d – this will attempt to deskew the image by rotating it before doing the OCR processing.

- -f – this applies a very basic filter to the glyphs before doing the OCR processing. I think it calculates the average number of pixels in a glyph and then eliminates any glyphs with more than 10 times this average or less than 1/10th of the average. This can be good for eliminating illustrations on a page, but can have very bad results if the image is dirty.
- -a – this applies a routine that tries to group glyphs before the OCR processing. This can be helpful when for broken characters and no harmful effects were observed.
- -x filename – specifies the name of the training library
- -o filename – file where recognized text is written
- -c filename – name of CSV file containing “extra characters” name conversions
- -18 (not documented in the help) – create 18th Connect XML and write this to the output file
- -k n (not documented in the help) – despeckle the image before doing OCR. Glyphs with less than n pixels will be eliminated. Useful for dirty images, but at higher values of n periods, dots on letters, etc. are removed.
- -h – help
- There are additional options. These were either found to be not useful or not even explored.

The command to process one image, with the image, training data, and extra characters CSV file in the current directory would be:

```
ocr4gamera.py –x Baskerville_library.xml –a –f –k 50 –c extra_chars.csv -18 –o file.xml file.tif
```

See /home/ext-woods/test/ocr-K022188.000.job on Brazos for an example of a batch job that does serial processing on a set of images.

Work on the Abe cluster at NCSA showed an average processing time of one minute per images, but images that processed for several hours were also seen. Processing time depends on the number of glyphs recognized, and experience showed that images that processed for more than 10 – 20 minutes would not produce usable output. Typically the image was very dirty or contained an illustration.

Parallel processing with Gamera

Processing images with Gamera is an embarrassingly parallel process since processing of each image can be done independently. However, to make the best use of the system setup on the Abe cluster at NCSA, a parallel tool was developed. This tool is a very preliminary stage of development.

The tool takes a list of images to process, and farms them out to the CPU cores assigned to a batch job. It does a check on the XML output for each file to check for the closing </page> tag. I think it removes the XML file if this tag is not found. The tool also enforces a time limit on processing for each image by killing processes after a period of time (this is currently fixed in the code – look for “pidsvec = waitANDQuery(1200,i);” which specified 1200 seconds). The specific arguments for the ocr4gamera.py command are also hard coded in the parallel dispatcher.

The mpiDispatcher code can be found in /home/ext-woods/tools/mpiDispatcher/. To compile it on Brazos, do:

```
module load openmpi/gcc
```

```
mpiCC mpiDispatcher.cc –o mpiDispatcher
```

The mpiDispatcher takes four arguments:

```
mpiDispatcher  imageList n1 n2 n3
```

Where:

imageList is the full path of a file with the list of images to process

n1 – number of lines in the image file

n2 – number of images to process – should be a multiple of 8 – it's assumed the code is running on an 8-core system

n3 – starting location in the image file – this allows you to submit multiple jobs that start at different points in the image file

To go with this code, is a simple Python script - /home/ext-woods/tools/utils/listTIFs.py - that will generate a list of all the images found under a specified directory. It does assume the image files have a .TIF extension. It takes one argument which is the directory to start at.

For an example of a parallel batch job using the mpiDispatcher code see /home/ext-woods/test/ocr-mpi.job

NOTES added from Email Dated 25 April 2012:

At a minute per image, the time does add up quickly. The 6 month estimate came from dividing the total time required by the number of processing cores on Abe (9600). Brazos doesn't have as many cores – only about 2400 – but processing images in parallel will reduce the time.

The attached document has some notes about Gamera on Brazos and Abe. We were just starting to figure out the best way to process images on Abe when it was shut down, so there may be a better approach. Also, this approach was developed after discussions with the NCSA support consultants who suggested that batch jobs that used all of the cores on a node was the best approach based on their scheduling policies.

One key thing to keep in mind is that Gamera will keep working on an image until it finishes – I found some images where Gamera ran for a couple of hours. However, I found that if Gamera ran for more than 10 minutes or so, that indicated that Gamera wasn't going to produce useful output, so our parallel processing approach has a time limit. In some cases, re-processing with the addition of de-speckling (removing glyphs with a small number of pixels) did produce useful output.

This approach to parallelizing Gamera processing uses C++ code since that is what my colleague who wrote the code was most familiar with, but if someone on the team is familiar with Python, that might be a better tool since Gamera is written in Python. Another possible advantage to using Python might be to reduce the processing time by reducing the amount of work is starting/ending the Gamera code – the current approach starts up a separate instance of Gamera for each image.

David Woods, Ph. D
Assistant Director for Research Computing
310K Laws Hall
Miami University

(513) 529-1857

Response from Rick Furuta:

The estimates of how long it would take that I saw in email seem to be consistent, at least within what we would call a constant factor. (Admittedly the constant can make a big difference here in terms of practicality.) The six month estimate might have been optimistic in terms of assuming that the problem can be completely parallelized, but it also seems like it is quite possible to limit Gamera's resource use since the observation is that it spends the longest on problems it will never solve. If Brazos has half the number of cores, that would then be around 12 months times three OCR engines or 36 months total. This suggests that the strategy is to get enough documents done to tune the procedures and get the workflow in place and then to approach other supercomputer centers for time allocations. I think it does reassure, however, that a significant chunk of the work will be able to be done during the grant period.

Cobre Summary, by Anton DuPlessis

Cobre (pronounced Cobré, short for comparative book reader and "copper" in Spanish) is a web application developed by the Texas A&M University Libraries that offers a suite of tools to aid interaction with the corpus that forms *Los Primeros Libros* Project. The *Primeros Libros* Project is an international collaboration to digitize and provide web access to the first books produced in the New World (1539 – 1601).

The interface leverages animated scrolling in the filmstrip views and tiled thumbnails with quick previews to the viewing of the pages in context. However, magnification and comparison tools within Cobre facilitate detailed examination and create opportunities for academic investigation and instruction, especially in the examination of variations in print, missing / obstructed text, missing / damaged pages, fire marks, marginalia and other copy specific attributes.

Because of its user-driven design, Cobre incorporates functionality to view and compare multiple exemplars of the same title that would be impossible with the physical books. Some of the functions allow for the alignment of one book with another and synchronous page views truly allowing the examination of state, emission, edition, etc. different examples of a title via parallel comparison.

Advanced editing features allow for annotations, via adding to the existing metadata, flexible enough to accommodate identifiable aspects of a book across all exemplars, like a table of contents, or nuanced markings of specific pages of inquiry, such as engravings, devices or firemarks, or exemplars. Another feature permits a scholar to collate pages from different books or exemplars into a custom exemplar, termed a Frankenbook.

The Cobre interface is broken into two sets of tools:

- Reading Tools
    - Book View
    - Reading View
    - Detailed View
    - Repository View
    - Comparison View
        - Quick Comparison View
- Editing Tools
    - Basic
    - Canonical
    - Frankenbook
    - Annotations
        - Structural
        - Non-structural

The reading tools are available to all visitors and support both casual reading and scholarly interaction with the books. The editing tools are available only to authorized users to create and modify the book structures and annotations.

Reading Tools. Cobre provides searching and browsing interfaces for book discovery. Book metadata are indexed for the search interface. The browsing interface is hierarchical by repository and collection with books ordered by title.

Book Overview. The book overview provides all of the available metadata describing the book along with icons providing navigation to the reading view, detailed view, and repository view. Icons are also available for downloading a PDF copy and adding the book to a list for the comparison view.

Reading View. The reading view starts with the front cover and progresses through the interior of the book, two pages at a time, to replicate the behavior that is available when reading a physical book. Users can either use the arrow keys on their keyboard or click on the pages to navigate through the book. Additionally, each page has an icon linking to the detailed view of that page.

Repository View. The books are hosted in a DSpace repository. The repository view is a custom DSpace/Manakin theme that displays tiled page thumbnails. Each tile contains links to the other views available in Cobre, as well as a quick link to a larger preview of the page.

Detailed View. The detailed view provides a "zoomable" interface for each page and a filmstrip view of a run of pages. The page can be dragged around the image viewer pane and zoomed using mouse clicks or a scroll wheel. Movement and zoom controls are overlaid on the viewer. The overlaid open book icon links back to the reading view of the book opened to the current page, while the printer icon links to a high-resolution image that is suitable for printing purposes. A scrollable and playable filmstrip of the pages is displayed beneath the detailed image viewer. The filmstrip centers the current page being viewed and provides context before and after this page. Clicking on a thumbnail in the filmstrip will load that page in the image viewer pane. The filmstrip provides three thumbnail sizes and can be played at six different speeds.

Comparison View. The comparison view displays parallel, scrollable and playable filmstrips of selected books. Quick comparison view opens a floating window showing a larger image of each page selected.

As in the detailed view, the filmstrips provide three thumbnail sizes and can be played at six different speeds. This allows for quick and easy comparison of page images from multiple books.

The browse, search and view interfaces provide methods to add books to a comparison list. When two or more books have been added to the list, the comparison view is accessible by clicking on the "compare" button. Locking the filmstrips together and scrolling or playing through the pages allows a quick comparison of the pages of the books. Different copies of the same book that have been individually scrolled can easily be synchronized with a particular filmstrip (book) by selecting its sync icon. Individual pages from each book being compared can be inspected at the same time in a collateral quick zoom view

Editing Tools. Various features of the books can be edited by authorized users (curators, scholars, etc.) by using the integrated editing tools. The basic page editor allows for the reordering of pages, as well as adding and removing pages as necessary, via a drag and drop interface. Augmentation of the basic page editor adds the functionality to create Frankenbooks and to align copies of a book to the canonical version. The annotation editor provides for adding, editing, and removing annotations or copying the annotations from one copy of a book to another. The annotations discussed here are descriptions of structure, special markings, etc., as one may find in a critical bibliography.

The book that is being edited can be compared with another book. Comparing books while editing is useful for aligning books with the canonical version. To aid alignment, pages of the canonical version display the annotations on the page when the page is clicked. The user adds blank pages into the book that is being edited as necessary to account for pages that are missing with respect to the canonical version.

Frankenbook Editor. The basic page editor functionality is extended when editing a Frankenbook. In addition to the standard operations, pages from the compared books can be dragged and dropped onto pages in the Frankenbook. When there are multiple copies of a book, the user can select which copy has the most complete version of a page, on a page by page basis, or in the case of unique attributes, which copy of a page has the most desirable attributes (for example, the most interesting marginalia).

Annotation Editor. An annotation editor is also available that allows for adding, editing, and removing annotations on existing copies, canonical copies, and Frankenbooks. The user inputs the starting page and length of the annotation, whether the annotation is part of a structural hierarchy, and text of the annotation itself. The annotations can be input in the context of an existing copy and then duplicated in the abstract canonical copy, or may be input directly in the context of the canonical copy if they are available from some external source, such as a critical bibliography. Structural annotations provide support for describing sections and chapters in books. Non-structural annotations are used for annotating special markings, marginalia and other items of interest.

Why a “canonical” book?

Since many of these books are missing pages in arbitrary locations, it is very difficult to align one against the other, and in our case we need to align an arbitrary number of books. A canonical book (an abstract construct that permits alignment of different exemplars and witnesses of the same work by leveraging the structural metadata) allows each individual book to be aligned with it separately. An arbitrary number of books can then easily be aligned for comparison

Frankenbook

A Frankenbook is a canonical book that has actual page images, drawn as desired from any existing copy of a book, replacing the abstract pages of the canonical book. A Frankenbook makes alignment and recognition and reduces the need for multiple copies. While there is generally a need for only one canonical book, there may be several

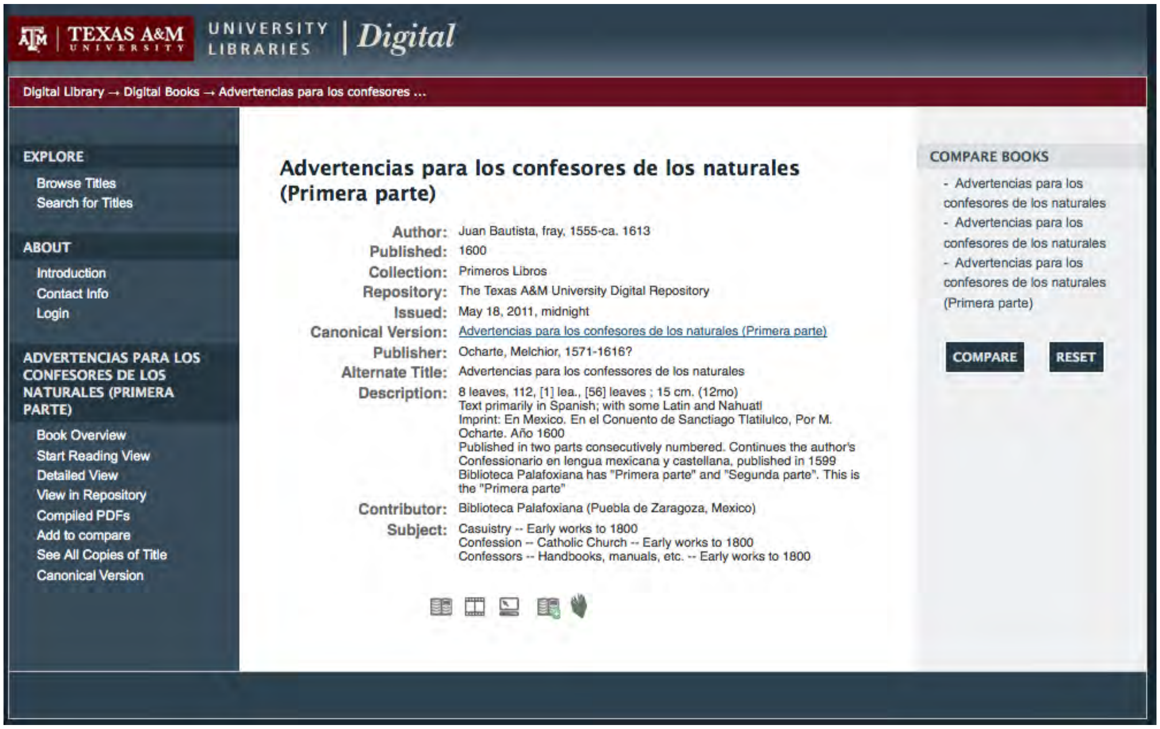
Frankenbooks per title. We have found different full-page engravings in different copies of a Bulla and variation on the number of licenses among several exemplars of *Adevertencias para los confessores 1ª parte* (1600).

One question that arises is: When is a book a copy (with variation) of another book and when is it an edition of the book? Bibliographers generally agree that if fifty percent or more of a book has been re-typeset, then the book should be considered a new edition. In our case, we do have books that have been re-typeset. Type was a precious commodity and one could not afford to store type for future printings. This would argue for multiple canonical books, one for each edition. But one can also argue for alignment of all copies, exemplars and witnesses, to one canonical book. At the current time, we generate one canonical book and align all copies to this standard. This requires that we add "pages" containing the phrase "for synchronization only" in the appropriate places.

Screenshots:

The Metadata View:

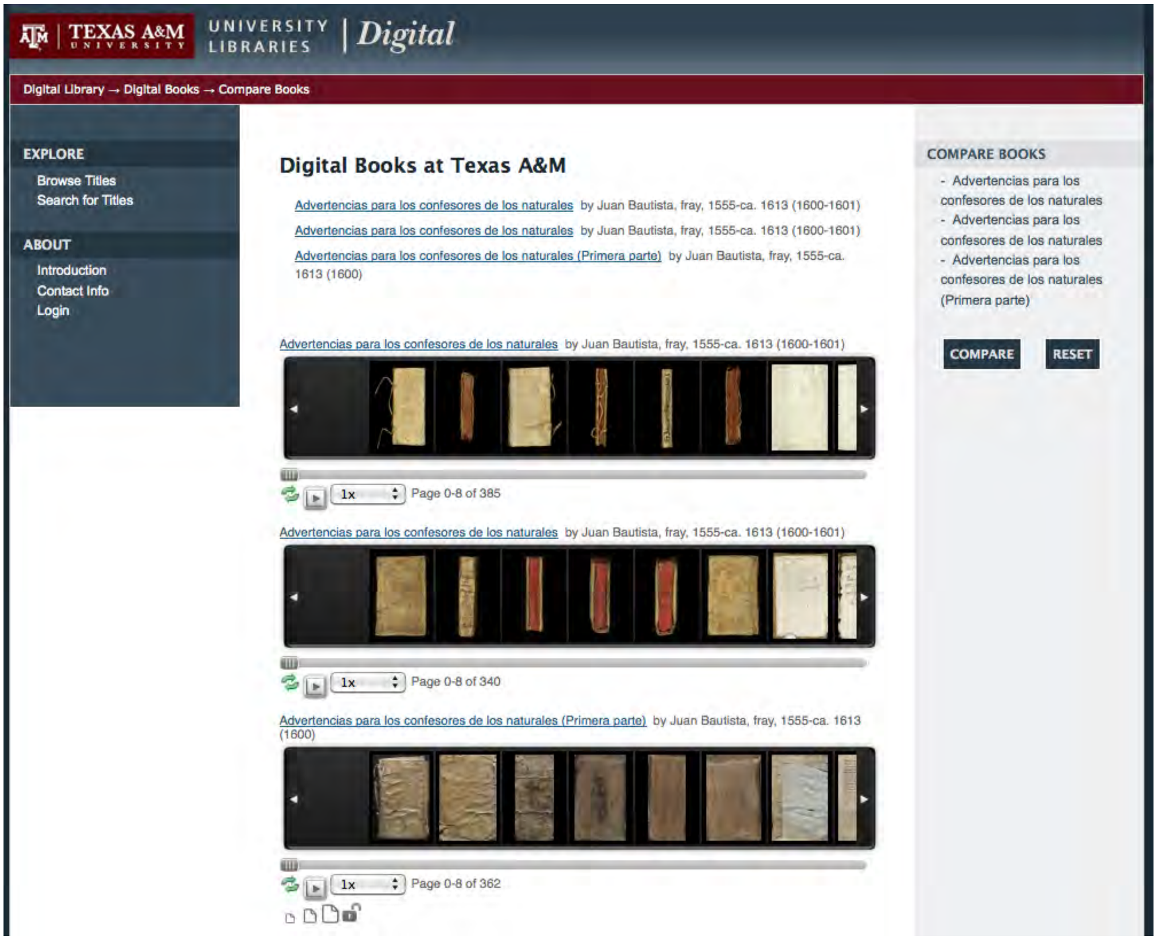
The Book Overview page


TEXAS A&M UNIVERSITY
UNIVERSITY LIBRARIES
Digital
Digital Library → Digital Books → Advertencias para los confesores ...
EXPLORE
Browse Titles
Search for Titles
ABOUT
Introduction
Contact Info
Login
ADVERTENCIAS PARA LOS CONFESORES DE LOS NATURALES (PRIMERA PARTE)
Book Overview
Start Reading View
Detailed View
View in Repository
Compiled PDFs
Add to compare
See All Copies of Title
Canonical Version
COMPARE BOOKS
- Advertencias para los confesores de los naturales
- Advertencias para los confesores de los naturales
- Advertencias para los confesores de los naturales (Primera parte)
COMPARE
RESET

Advertencias para los confesores de los naturales (Primera parte)

Author: Juan Bautista, fray, 1555-ca. 1613
Published: 1600
Collection: Primeros Libros
Repository: The Texas A&M University Digital Repository
Issued: May 18, 2011, midnight
Canonical Version: Advertencias para los confesores de los naturales (Primera parte)
Publisher: Ocharte, Melchior, 1571-1616?
Alternate Title: Advertencias para los confessores de los naturales
Description: 8 leaves, 112, [1] lea., [56] leaves ; 15 cm. (12mo)
Text primarily in Spanish; with some Latin and Nahuatl
Imprint: En Mexico. En el Conuento de Sanctiago Tlatilulco, Por M. Ocharte. Año 1600
Published in two parts consecutively numbered. Continues the author's Confessionario en lengua mexicana y castellana, published in 1599
Biblioteca Palafoxiana has "Primera parte" and "Segunda parte". This is the "Primera parte"
Contributor: Biblioteca Palafoxiana (Puebla de Zaragoza, Mexico)
Subject: Casuistry -- Early works to 1800
Confession -- Catholic Church -- Early works to 1800
Confessors -- Handbooks, manuals, etc. -- Early works to 1800

Different Exemplars and Witnesses:



Digital Books at Texas A&M

Advertencias para los confesores de los naturales by Juan Bautista, fray, 1555-ca. 1613 (1600-1601)

Advertencias para los confesores de los naturales by Juan Bautista, fray, 1555-ca. 1613 (1600-1601)

Advertencias para los confesores de los naturales (Primera parte) by Juan Bautista, fray, 1555-ca. 1613 (1600)

- Advertencias para los confesores de los naturales
- Advertencias para los confesores de los naturales
- Advertencias para los confesores de los naturales (Primera parte)

Advertencias para los confesores de los naturales by Juan Bautista, fray, 1555-ca. 1613 (1600-1601)

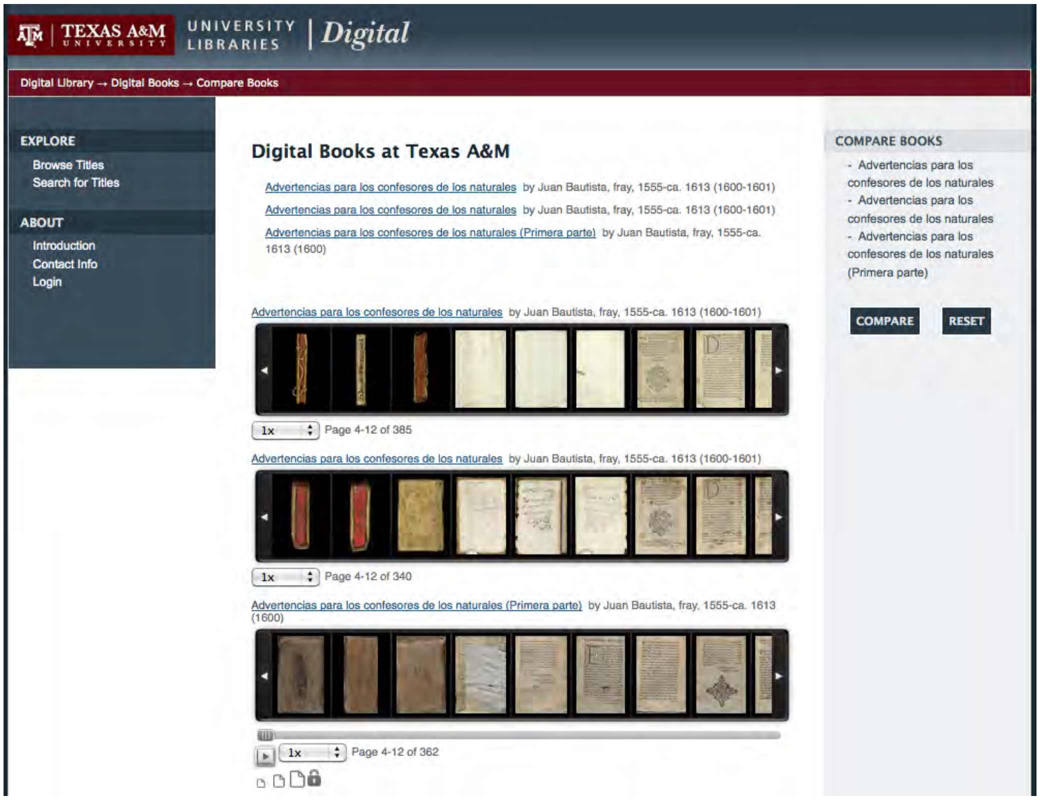
1x Page 0-8 of 385

Advertencias para los confesores de los naturales by Juan Bautista, fray, 1555-ca. 1613 (1600-1601)

1x Page 0-8 of 340

Advertencias para los confesores de los naturales (Primera parte) by Juan Bautista, fray, 1555-ca. 1613 (1600)

1x Page 0-8 of 362



Digital Books at Texas A&M

Advertencias para los confesores de los naturales by Juan Bautista, fray, 1555-ca. 1613 (1600-1601)

Advertencias para los confesores de los naturales by Juan Bautista, fray, 1555-ca. 1613 (1600-1601)

Advertencias para los confesores de los naturales (Primera parte) by Juan Bautista, fray, 1555-ca. 1613 (1600)

COMPARE BOOKS

- Advertencias para los confesores de los naturales
- Advertencias para los confesores de los naturales
- Advertencias para los confesores de los naturales (Primera parte)

Advertencias para los confesores de los naturales by Juan Bautista, fray, 1555-ca. 1613 (1600-1601)

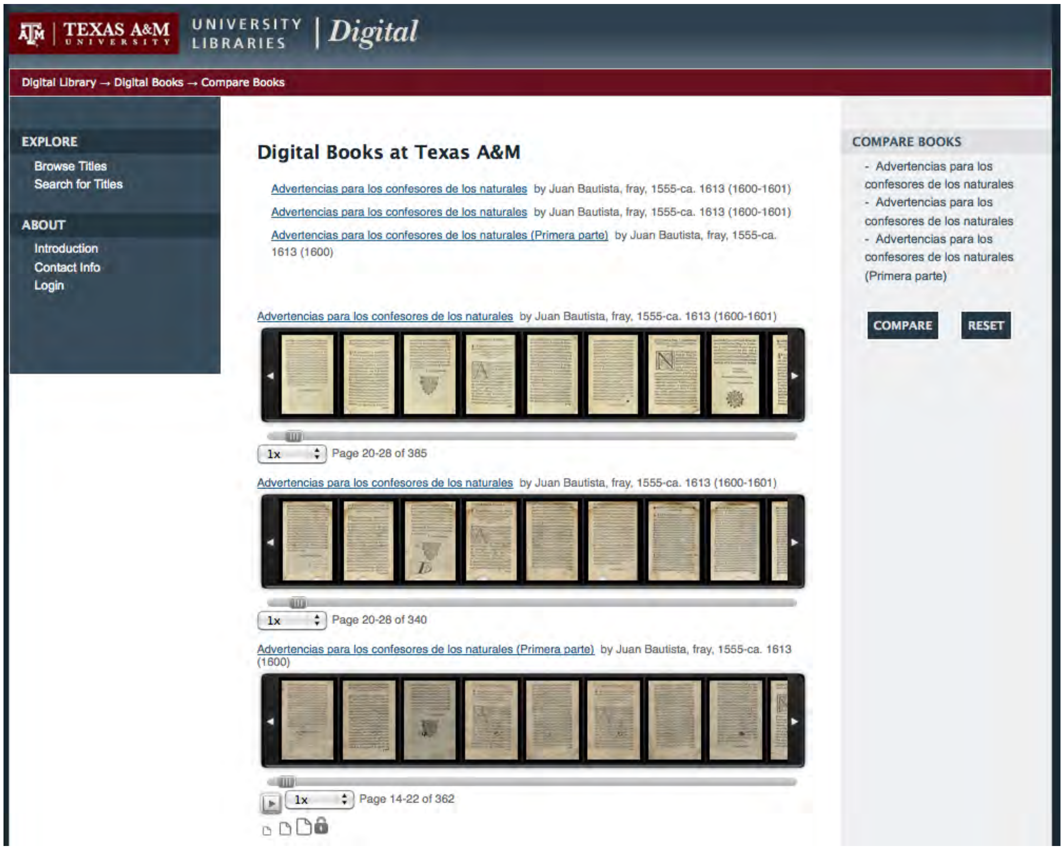
Page 4-12 of 385

Advertencias para los confesores de los naturales by Juan Bautista, fray, 1555-ca. 1613 (1600-1601)

Page 4-12 of 340

Advertencias para los confesores de los naturales (Primera parte) by Juan Bautista, fray, 1555-ca. 1613 (1600)

Page 4-12 of 362



Digital Books at Texas A&M

Advertencias para los confesores de los naturales by Juan Bautista, fray, 1555-ca. 1613 (1600-1601)

Advertencias para los confesores de los naturales by Juan Bautista, fray, 1555-ca. 1613 (1600-1601)

Advertencias para los confesores de los naturales (Primera parte) by Juan Bautista, fray, 1555-ca. 1613 (1600)

COMPARE BOOKS

- Advertencias para los confesores de los naturales
- Advertencias para los confesores de los naturales
- Advertencias para los confesores de los naturales (Primera parte)

Advertencias para los confesores de los naturales by Juan Bautista, fray, 1555-ca. 1613 (1600-1601)

Page 20-28 of 385

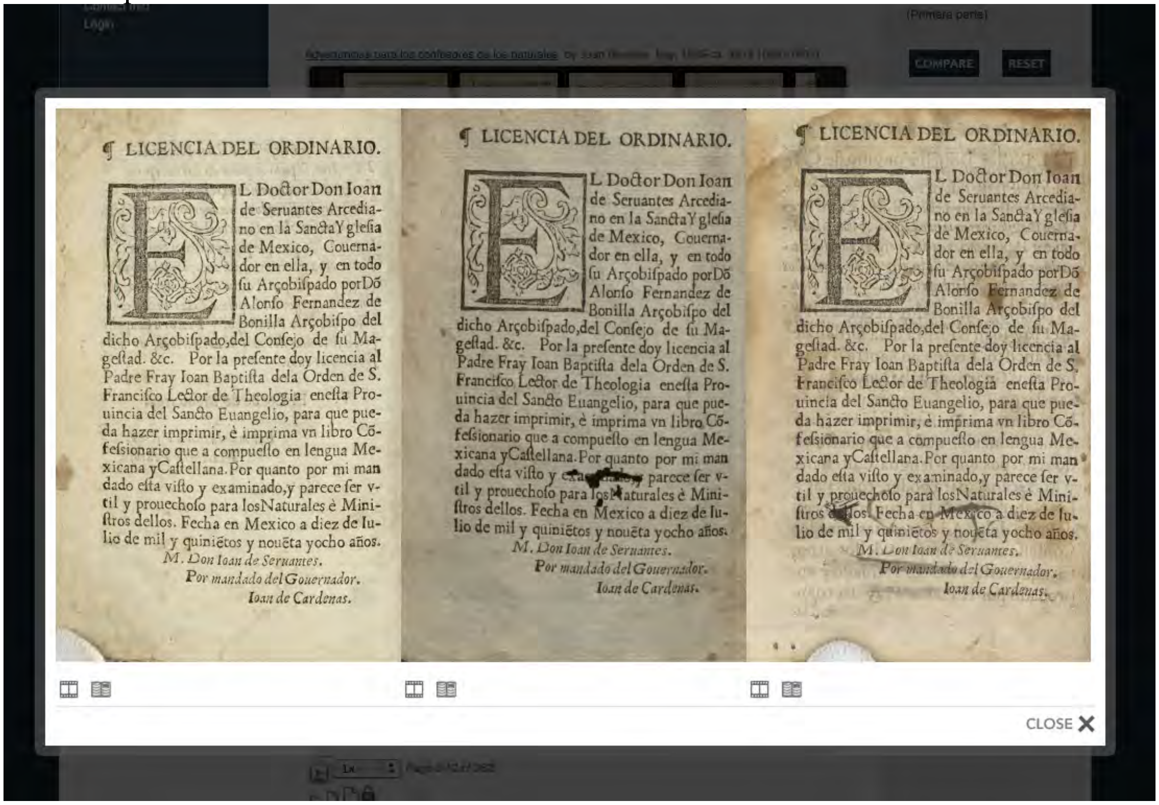
Advertencias para los confesores de los naturales by Juan Bautista, fray, 1555-ca. 1613 (1600-1601)

Page 20-28 of 340

Advertencias para los confesores de los naturales (Primera parte) by Juan Bautista, fray, 1555-ca. 1613 (1600)

Page 14-22 of 362

The Comparative View:



¶ LICENCIA DEL ORDINARIO.

EL Doctor Don Ioan de Seruantes Arcediano en la SanctaYgleſia de Mexico, Gouernador en ella, y en todo ſu Arçobiſpado porDõ Alonſo Fernandez de Bonilla Arçobiſpo del dicho Arçobiſpado,del Conſejo de ſu Mageſtad. &c. Por la preſente doy licencia al Padre Fray Ioan Baptiſta dela Orden de S. Franciſco Lector de Theologia eneſta Prouincia del Sancto Euangelio, para que pueda hazer imprimir, è imprima vn libro Cõfeſsionario que a compueſto en lengua Mexicana yCaſtellana.Por quanto por mi mandado eſta viſto y examinado,y parece ſer vtil y prouechoſo para losNaturales è Miniſtros dellos. Fecha en Mexico a diez de Iulio de mil y quiniẽtos y nouẽta yocho años.

M. Don Ioan de Seruantes.

Por mandado del Gouernador.

Ioan de Cardenas.

¶ LICENCIA DEL ORDINARIO.

EL Doctor Don Ioan de Seruantes Arcediano en la SanctaYgleſia de Mexico, Gouernador en ella, y en todo ſu Arçobiſpado porDõ Alonſo Fernandez de Bonilla Arçobiſpo del dicho Arçobiſpado,del Conſejo de ſu Mageſtad. &c. Por la preſente doy licencia al Padre Fray Ioan Baptiſta dela Orden de S. Franciſco Lector de Theologia eneſta Prouincia del Sancto Euangelio, para que pueda hazer imprimir, è imprima vn libro Cõfeſsionario que a compueſto en lengua Mexicana yCaſtellana.Por quanto por mi mandado eſta viſto y examinado,y parece ſer vtil y prouechoſo para losNaturales è Miniſtros dellos. Fecha en Mexico a diez de Iulio de mil y quiniẽtos y nouẽta yocho años.

M. Don Ioan de Seruantes.

Por mandado del Gouernador.

Ioan de Cardenas.

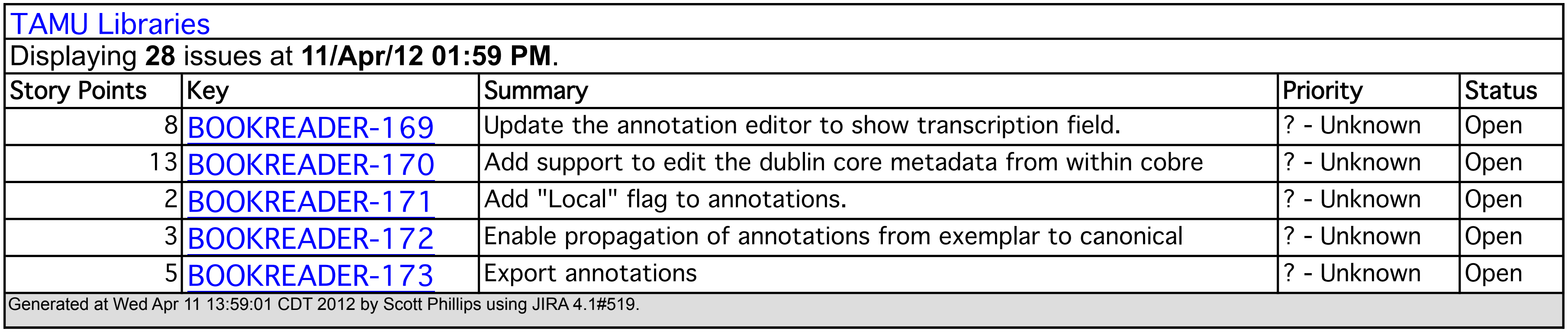
¶ LICENCIA DEL ORDINARIO.

EL Doctor Don Ioan de Seruantes Arcediano en la SanctaYgleſia de Mexico, Gouernador en ella, y en todo ſu Arçobiſpado porDõ Alonſo Fernandez de Bonilla Arçobiſpo del dicho Arçobiſpado,del Conſejo de ſu Mageſtad. &c. Por la preſente doy licencia al Padre Fray Ioan Baptiſta dela Orden de S. Franciſco Lector de Theologia eneſta Prouincia del Sancto Euangelio, para que pueda hazer imprimir, è imprima vn libro Cõfeſsionario que a compueſto en lengua Mexicana yCaſtellana.Por quanto por mi mandado eſta viſto y examinado,y parece ſer vtil y prouechoſo para losNaturales è Miniſtros dellos. Fecha en Mexico a diez de Iulio de mil y quiniẽtos y nouẽta yocho años.

M. Don Ioan de Seruantes.

Por mandado del Gouernador.

Ioan de Cardenas.



TAMU Libraries

Displaying **28** issues at **11/Apr/12 01:59 PM**.

Story Points	Key	Summary	Priority	Status
8	BOOKREADER-169	Update the annotation editor to show transcription field.	? - Unknown	Open
13	BOOKREADER-170	Add support to edit the dublin core metadata from within cobre	? - Unknown	Open
2	BOOKREADER-171	Add "Local" flag to annotations.	? - Unknown	Open
3	BOOKREADER-172	Enable propagation of annotations from exemplar to canonical	? - Unknown	Open
5	BOOKREADER-173	Export annotations	? - Unknown	Open

Generated at Wed Apr 11 13:59:01 CDT 2012 by Scott Phillips using JIRA 4.1#519.

Here we have put our requests into the Cobre development schedule, and the "story points" estimate the amount of development hours necessary for working on the tools. These changes will be implemented in the first development cycle (2012); the second (2013) will be devoted to working on the tool after usability studies by expert users.

Subject: Fwd: Mellon
Date: Monday, April 30, 2012 1:58:45 PM ET
From: Anton DuPlessis
To: Todd Samuelson, Laura Mandell

Lots of stuff from James on DC and typefaces. This is the sort of message that leaves wanting to print it out and really read it...

Sent from my iPod

Begin forwarded message:

In these two emails, the Cobre development team works with Rare Books Librarians in order to determine the Dublin Core Metadata fields that we will need to add to get the typeface information we need in order to better OCR the Early Modern Corpus.

From: James Creel <jcreel@library.tamu.edu>
Date: April 30, 2012 12:11:12 PM CDT
To: Anton DuPlessis <duplessis@library.tamu.edu>
Cc: Alexey Maslov <ak_maslov@library.tamu.edu>, Scott Phillips <scott.phillips@tamu.edu>, Douglas Hahn <dhahn@library.tamu.edu>, Micah Cooper <jmicah@library.tamu.edu>
Subject: Re: Mellon

Hi Anton,

Thanks for the information about the typeface metadata analysis needed for the Mellon grant application. I had a few thoughts about how Dr. Samuelson's suggestions could be manifested in Dublin-Core-style metadata.

According to a decision made by a committee long ago, the dc.format element is reserved in our DSpace repository for information about digital formatting: MIME type, file size, etc. If this rule is respected, then this leaves dc.description fields as the only suitable place for recording typeface metadata. In either case, we would probably use qualified fields.

> Type family (which could be general, such as roman, italic, textura, or batarde, or could be more specific: Bembo Roman, Jenson Roman, &c)

We might use something like dc.description.typefamily for these values. To facilitate sorting on these values, we might suggest something like "Roman - Bembo" instead of "Bembo Roman" so that it would appear near "Roman" in a lexicographic ordering.

> Type name (in the rare case where bibliographers have affixed particular names to the types of notable printers: e.g. Caxton 1-7)

A qualifier like "typename" might not be appropriately unambiguous to the uninitiated, so we might go with something like dc.description.typefacename.

> Punchcutter (the most significant piece of information, according to Harry Carter, but not always ascertainable)

Seems pretty unambiguous, so why not dc.description.punchcutter, eh?

> Body Size - Name (point size or the old names for type sizes, an approximation before regularization: long primer, pica, English, &c)

To avoid ambiguity, we might preface the qualifier with just "type" as in dc.description.typebodysizename . It does get a bit verbose at that point though, and I feel it could more reasonably be shortened if we were using the dc.format element.

> Body Size - Measurement (in millimeters)

dc.description.typebodysizemeasurement

Better yet, though, those last two seem like they could easily go in one field with some well-defined formatting. Something like "dc.description.typebody: long primer, 5mm"

> Earliest Date (the date of the first known type specimen or book utilizing the type)
>
> Other information appearing in a scholarly work like H. D. L. Vervliet's French Renaissance Printing Types: A Conspectus, completed in 2010, includes the names of the books in which the types were first seen; names and holdings of any type specimens known; any printing materials (punches, matrices, and moulds) which have been preserved, and their location; any secondary literature specifically addressing the typeface; and notes about the development of the type (including any revisions in the form of adaptations or recuttings of the punches).

These sorts of information probably belong in dc.description fields as well, but may not be amenable to machine readability without some very rigorous formatting and vocabulary control.

The use of Dublin-Core-style fields for this purpose sort of presupposes that there is only one typeface for the document. If multiple values were given for each of these fields, there is no good way to see which of them go together (for example, which one of several typebody fields would go with which one of several typefacename fields). This problem could be avoided by putting all the information in a single field value (an extension of the suggestion of merging the type body information). For example: "dc.description.typeface: Bembo Roman; Caxton 1. Bob the punchcutter. long primer, 5mm. Earliest known date: 1800." However, as certain values may often be missing, it would be tricky to devise a consistent, machine readable or even human-readable structure that could any and all such information. It also prevents sorting in DSpace on any of the specific constituent values.

In any case, Dr. Mandell may be unconcerned with sorting issues and choice of dc elements. Since her use-case for the repository is as a back-end to Cobre and there might be copyright issues on the documents, her data would probably best be hosted in a private repository separate from the IR.

The metadata experts in Scholarly Communications are the local authorities on these matters, but we appreciate you hearing our thoughts as well. Your development team is at your service if you would like any more information.

Thanks,

James et al.

On Apr 28, 2012, at 152, Anton DuPlessis wrote:

> Hi James,
>
> This is the message from Todd to Laura about typefaces that I mentioned yesterday.
>
> Best,
>
> Anton
>
> Sent from my iPod
>
> Begin forwarded message:
>
>> **From:** Todd Samuelson <toddsamuelson@library.tamu.edu>
>> **Date:** April 13, 2012 9:49:31 AM CDT
>> **To:** "Mandell, Laura" <mandell@tamu.edu>
>> **Cc:** Anton DuPlessis <duplessis@library.tamu.edu>, "Jacob Heil (jah8p@neo.tamu.edu)" <jah8p@neo.tamu.edu>
>> **Subject: RE: Mellon**
>>
>> Dear Laura,
>>
>> I hope that all has been going well with the Mellon planning. I've been working with Jake to develop a document about the roles of the postdoc and my position, and we should have the file for you soon. Yesterday, I had a long conversation with Anton about the Cobre reader and some issues with the metadata schema, and I wanted to check in with you to discuss some of the implications this may have for our typographical research. This can wait, of course, but I think that determining an approach to some of the structural issues may be relevant further along.
>>
>> I've been working on some background regarding type in England (both domestic and imported), and have constructed a rough timeline of English typography. (Essentially, a great deal of the type used in English printing were Flemish, French, and Dutch at various stages and in various type styles, mainly imported but also designed by a few significant native figures.) One thing that has become clear is that the identification of type is a fairly complex issue, requiring more than the name of an associated printer to determine confidently. Of course, it may not be necessary to identify the types for the project, but since the postdoc will be spending so much time researching secondary material, it may make sense to build in metadata fields to accommodate the sorts of information which will be found.
>>
>> My impression is that since the project is going to move back to the

earliest printing in England, we should anticipate the more difficult period of typefaces (before the typefounding trade had developed and even before type bodies were made standard). In the most recent texts describing types of the fifteenth and sixteenth centuries, as many as eight or nine fields are used to identify the faces. Not all of these are, in my opinion, relevant to our project, but I would suggest that we consider the following:

Type family (which could be general, such as roman, italic, textura, or batarde, or could be more specific: Bembo Roman, Jenson Roman, &c)
Type name (in the rare case where bibliographers have affixed particular names to the types of notable printers: e.g. Caxton 1-7)
Punchcutter (the most significant piece of information, according to Harry Carter, but not always ascertainable)
Body Size - Name (point size or the old names for type sizes, an approximation before regularization: long primer, pica, English, &c)
Body Size - Measurement (in millimeters)
Earliest Date (the date of the first known type specimen or book utilizing the type)

Other information appearing in a scholarly work like H. D. L. Vervliet's French Renaissance Printing Types: A Conspectus, completed in 2010, includes the names of the books in which the types were first seen; names and holdings of any type specimens known; any printing materials (punches, matrices, and moulds) which have been preserved, and their location; any secondary literature specifically addressing the typeface; and notes about the development of the type (including any revisions in the form of adaptations or recuttings of the punches).

I'm sorry to send this lengthy letter, but my feeling in talking with Anton yesterday is that the process of building up the reader is moving forward, and making some of these decisions may be necessary. Since the postdoc will be compiling so much of this information, being able to place it in the project in a manner that can be searched and utilized seems sensible to me, though of course you'll need to decide how to reckon with it. If you'd like to meet, I'd look forward to it; Jake and I will continue our dialogue otherwise. I'll also continue compiling a bibliography of critical material about English typography as a starting point.

Many thanks,
Todd.

Bibliography of English Typography

Ball, Johnson. *William Caslon 1693-1766: the ancestry, life and connections of England's foremost letter-engraver and type-founder*. Kineton, 1973.

Berry, W. T. and Johnson, A.F. *Catalogue of specimens of printing types by English and Scottish printers and founders 1665-1830.* London: Oxford University Press, 1935.

Blades, William. *The Life and Typography of William Caxton.* 2 vols. London: Joseph Lilly, 1861-3.

Blayney, Peter W. M.. *The First Folio of Shakespeare.* Washington, D.C.: Folger Library Publications, 1991.

Bowman, John H. *Greek printing types in Britain from the late eighteenth century to the early twentieth century.* Thessaloniki: Typophilia, 1998.

---. *Greek printing types in Britain in the nineteenth century, a catalogue.* Oxford: Oxford Bibliographical Society, 1992.

Carter, Harry. *A history of the Oxford University Press: volume 1, to 1780.* Oxford, 1975.

---. "A List of Type Specimens." *The Library* 4th series 22.4 (1941-2): 185-204.

---. *A View of Early Typography Up to About 1600. The Lyell Lectures 1968.* Oxford: Clarendon, 1969.

Caslon, William. *A specimen of printing types,* London, 1766. Facsimile, ed. James Mosley, *Journal of the printing Historical Society,* 16, 1981-2.

Clement, Richard W. "The beginnings of printing in Anglo-Saxon." *Papers of the Bibliographical Society of America* 91 (1997): 192-244.

De Ricci, [Seymour]. *A Census of Caxtons.* Illustrated Monographs, No. XV. Oxford: Bibliographical Society, 1909. "Caxton's eight types are reproduced in facsimile, printed on rough paper, and showing full pages of the types. . . . A list of Caxton's books in chronological order, showing the types used in each year, is appended" (Updike 120 note 1).

Dreyfus, John. "The Baskerville punches 1750-1950." In John Dreyfus, *Into Print: Selected Writing on Printing History, Typography and Book Production.* London, 1994, pp. 13-36. First published in *The Library* third series 5.1 (1950): 26-48.

Dreyfus, John. "Baskerville's methods of printing." *Signature* new series 12 (1941): 44-9.

Duff, [Edward] Gordon. *A Century of the English Book Trade. Short notices of all printers, stationers, book-binders, and others connected with it from the issue of the first dated book in 1457 to the incorporation of the company of stationers in 1557.* London: Bibliographical Society, 1905.

--. *Early English Printing: a Series of Facsimiles of All the Types Used in England During the 15th Century.* New York: Burt Franklin, 1970.

--. *Fifteenth Century English Books* (Bibliographical Society's Monographs, No. XVIII, 1917).

--. *Life of Caxton.* (Caxton Club of Chicago)

[Fell Types]. *Specimen of the Several Sorts of Letter given to the University by Dr. John Fell, later Lord Bishop of Oxford.* 1693 (other editions in 1695, 1706, 1768, 1787, 1794, etc). Also London: James Tregaskis, 1928.

Ferguson, W. Craig. *Pica roman type in Elizabethan England.* Aldershot, 1989.

Gaskell, Philip. *A bibliography of John Baskerville.* 2nd ed. Chicheley, 1973. First published Cambridge, 1959. Includes a facsimile of Baskerville's type specimen of 1775.

---. *A bibliography of the Foulis Press.* London, 1964.

---. *A New Introduction to Bibliography.* New Castle: Oak Knoll, 1995.

---. "Photographic enlargements of type forms." *Journal of the Printing Historical Society* 7 (1971): 51-3. [Enlarged details of types of Jenson, Aldus, Garamond, Baskerville, Wilson, etc.]

Hart, Horace. *Notes on a century of typography at the University Press, Oxford, 1693-1794. Oxford, 1900.* Reprinted, with an introduction and additional notes by Harry Carter, 1970.

Howes, Justin. "Caslon's Punches and Matrices." *Matrix* 20 (2000): 1-7. [With an insert, 'Caslon Old Face: an inventory', pp. i-v888, listing surviving punches and matrices.]

Howes, Justin. "Caslon's Patagonian." *Matrix* 24 (2005): 61-71.

---. *English and Scottish Printing types, 1501-35, 1508-41.* Oxford, 1930.

---. *English and Scottish Printing types, 1535-58, 1552-58.* Oxford, 1932.

---. *English printers' types of the sixteenth century.* London, 1936.

---. *William Bulmer: the fine printer in context, 1757-1830.* London, 1993.

Marrot, H. V. *William Bulmer, Thomas Bensley: a study in transition.* London, 1930.

McGuinne, Dermot. *Irish type design: a history of printing types in the Irish character.* Dublin, 1992.

McKenzie, D. F. *The Cambridge University Press 1696-1712, a bibliographical study.* Cambridge, 1966.

Mores, Edward Rowe. *A dissertation upon English typographical founders and foundries, 1778; edited, with an introduction and notes, by Harry Carter and Christopher Ricks.* London, 1961 (corrected reprint, 1963).

Morison, Stanley. *John Bell.* London, 1930.

---. *John Fell, the University Press, and the 'Fell' types, by Stanley Morison, with the assistance of harry Carter.* Oxford 1967.

Mosley, James. "The early career of William Caslon." *Journal of the Printing Historical Society* 3 (1967): 66-81.

---. *British type specimens before 1831: a hand-list.* Oxford, 1984.

---. "English Vernacular: a Study in Traditional Letter Forms." *Motif* 11 (1963): 3-55.

---. "English Vernacular Revisited." *Forum: Journal of Letter Exchange* 13 (April 2007): 8-11.

---. *Handmade Type: Thoughts on the Preservation of Typographic Materials.* Oldham: Incline, 2007.

---. *The Nymph and the Grot: the Revival of the Sanserif Letter.* London, 1999.

---. *Typefoundry: Documents for the History of Type and Letterforms.* [online blog].
< http://typefoundry.blogspot.com/>

---. "The Typefoundry of Vincent Figgins 1792-1836." *Motif* 1 (1958): 29-36.

Nuttall, Derek. English printers 1600-1700 and their supra-text roman and italic types. D. Phil. Thesis, University of Reading, 1985.

Pardoe, Frank. *John Baskerville of Birmingham.* London, 1975.

Reed, Talbot Baines. *A history of the old English letter foundries.* New ed. Rev. and enlarged by A. F. Johnson. London, 1952.

Stephenson, S. & C. *A specimen of printing types and various ornaments, 1796, reproduced together with the Sale catalogue of the British Letter Foundry, 1797, with an introduction by James Mosley.* London, 1990.

Straus, Ralph. *John Baskerville, a memoir, by Ralphs Straus and Robert K. Dent.* London, 1907.

Tanselle, G. Thomas. *Bibliographical Analysis: A Historical Introduction.* Cambridge: Cambridge UP, 2009.

Treadwell, Michael. "The Grover typefoundry." *Journal of the Printing Historical Society* 15 (1980/81): 36-53.

Updike, D. B. *Printing Types: Their History, Forms, and Use.* 2 vols. Cambridge, Mass.: Belknap, 1966.

Mellon Proposal

Rationale for and description of the Postdoctoral researcher.

The OCR project will require a postdoctoral research position (hereinafter postdoc) whose job it will be to build a database of fonts used in English print shops. This database will provide the newly-developed OCR engines with the criterion needed to discern, through integration with the metadata of scanned texts, the most appropriate subset of font to apply to the reading of a given text. The postdoc will compile this database by, first, researching the dissemination of fonts and matrices from the continent into England and, second, identifying and utilizing best practices for a system of nomenclature that can be integrated into the project's OCR engines.

It should be noted that, in addition to furthering the goals of the governing OCR project, these "best practices" and this "system of nomenclature" will fill a gap in the field of analytical bibliography long ago lamented by Harry Carter in his Lyell Lectures at Oxford in 1968 (published as *A View of Early Typography up to About 1600* by Clarendon in 1969). While One of Carter's truisms is that type identification is more an art than a science,[1] he nonetheless offers some approaches to a systematic and mechanical method for citing types of the early modern period: by measuring the body, linking it to the conventional historical names, and then linking it, as well as possible, to the name of the punchcutter.[2] Carter posits that,

> "Concentrated in that [the identity of the source] is all manner of information as to place and time, circumstances and relationships on which a history can be built: the knowledge of who cut it enables one not only to describe a face of type, it makes it worth describing – fits it into the whole scheme of things. To associate a type with Jenson or Caslon is to give it a quality which many sentences would be needed to express in any other way if, indeed, words could do it." (Carter 23)

It will be the duty of the postdoc to build a history of such things as "place and time, circumstances and relationships." In doing so the postdoc should be able to discern the probability of a given subset of typefaces having migrated into the particular shop in which a

[1] "[I]t is evident that in considering the face of a fount of type we are in a world of art, styles, difficulty of saying what styles, inherited forms, human hands; a humble art it maybe, but not a mechanical proceeding or anything susceptible of scientific treatment" (Carter 24).

[2] In response to the question of why it should be difficult to identify types used in early printing, Carter writes that "if you measure it, and find that 20 lines set in it take up, say, 85 mm., you restrict it to a class of a particular body – a property of a typefounder's mould. It remains to describe the face, which might be cast on a variety of bodies. I had rather name typefaces for size by the conventional body that would best fit the, Pica, Long Primer, Minion, and such, than by numbers, qualifying these terms if necessary by adding 'large' or 'small.' Some time in the early part or the middle of the sixteenth century these names acquired fixed meanings. Until it becomes appropriate to use them it is safest to measure the face of a fount, which you can do if you have a powerful magnifying-glass and a fine scale and measure from the top of b to the bottom of p of the extent of an Italic *f*. This, called the gauge of the face, cannot vary. Much the best indication of the character of a face of type is the name of the person who cut it" (Carter 23).

given early modern book was printed and, thus, provide a dataset that the OCR engines might then apply to the reading of that scanned document.

In sum, the position will require at least one year of research and database construction, the former leading, as a matter of course, toward the latter. The postdoc will begin with the guided construction of the knowledge necessary to discern the kinds of questions to ask of fonts and foundries. In the course of this research phase the postdoc will begin to apply the information acquired to early database structures that will be revised over time. The research phase will culminate with a research trip that will bridge the research and the database construction. Once the data are collected, he or she will work with affiliated entities to aggregate and manipulate the data into the form most useful to the newly-designed OCR engines.

What follows is a detailed, phased timeline of the postdoc's duties with expected outcomes.

1) Phase One: Initial research period (3 months, OCT-DEC)
    a. Working with Dr. Todd Samuelson, Rare Books & Manuscripts Curator in Texas A&M University's Cushing Library, the postdoc will use the resources in local and other domestic libraries to build a knowledge base (a) of the development of typefaces and (b) of the flow of fonts in the early modern period between the continent and England. This research will take an antechronological course of major type categories (modern, transitional, old style, etc.) with emphasis upon significant English type designers before reaching the (exclusively) imported types. This course will allow the research to begin with focused control groups, as it were, and then to expand back in time and outward toward the continent.
    b. Establish contacts in St. Bride Library in London and the Plantin-Moretus Museum in Antwerp; research trips to these locations will mark the culmination of the research phase of the postdoc's portion of the project (detailed in Phase Five, below). Such contacts will help to establish the kinds of questions the postdoc should explore in thinking about ways to discriminate between fonts. Furthermore, such early contact may develop a network of scholars who can additionally supplement the research phase.
    c. Begin drafting a co-authored, article-length document that will contribute to the field of analytical bibliography.
    d. Plan for a visit and consultation with James Mosley, Visiting Professor in the Department of Typography & Graphic Communication, at the University of Reading. (Details in Phase Three, below.)

2) Phase Two: Initial steps toward database construction (1 month, JAN)
    a. Working closely with IDHMC-affiliated database specialists, the postdoc will develop a general structure for the database. This early version will be based on the results of the initial research phase of the project. Of necessity, this will be tentative and it will have a highly portable, manipulable interface so as to

accommodate further design (in Phases Four and Six, below) and the importation of data during the research trips (Phase Five).

b. Additionally, during this phase the postdoc will work with the Dr. Samuelson and the network of contacts in the European libraries established in Phase One to determine a specific plan for the research travel.

3) Phase Three: Capstone to domestic research efforts, James Mosley. (1week, FEB)

   a. Host James Mosley at Texas A&M. Mosley has taught at the Univeristy of Virginia's Rare Book School, was the founding editor of the Journal of the Printing Historical Society, and, in 2003, his contributions to printing history earned him the American Printing History Association's annual award.
   b. Design and participate in workshops for Prof. Mosley and interested parties.
   c. Arrange for Prof. Mosley to lecture in either an open forum or in Book History classes, thus contributing to the academic environment of the university.
   d. Consult with Prof. Mosley on the findings of the research undertaken in Phase One and the early framework for the database designed in Phase Two.
      i. Brainstorm and troubleshoot plans for nomenclature.
      ii. Troubleshoot database categories.
      iii. Refine the set of questions and avenues for exploration to be undertaken during the research trip in Phase Five.

4) Phase Four: Solidfy database framework (3 weeks, FEB)

   a. Using the information gathered in Phase Three, revise the database scheme initially developed in Phase Two.
   b. Work with IDHMC-affiliated database specialists to finalize a robust, portable database interface that can travel in Phase Five.

5) Phase Five: European Research Trip (6 Weeks, MAR-APR)

   a. The postdoc and Dr. Samuelson will travel to St. Bride Library and the Plantin-Moretus Museum to further develop an understanding of the flow of fonts and to collect data. In the interest of time and intellectual productivity, both researchers will need to travel: the trip has focused on these libraries because of the wealth of materials available therein, and the sheer quantity will require a division of the research labors.
   b. St. Bride Library, London (3 weeks)
      i. Build on the history of transmission developed in Phase One by supplementing it with resources from the collection, e.g.:[3]

[3] Information about the collections at St. Bride provided here and in point "ii. Collect data…" is compiled from the description of St. Bride's collections on their website. <http://stbride.org/library/collections>

1. **British Printing Industries Federation**, several hundred minute books and other records of local master printers' associations and their alliances.
2. **Trade documents**, two substantial collections of early printing-trade ephemeral documents mainly concerning industrial relations.
3. **Trade literature**, extensive collections of literature on trade exhibitions, conferences and competitions; printing ink trade literature; bookcloth trade literature; printers' and auxiliary services' trade literature; printing; printing machinery and suppliers trade literature.

ii. Collect data using the materials onsite, e.g.:

1. **Type specimens**, a major collection of nearly 10,000 items, books, pamphlets and single-sheet specimens. The collection of British type specimens is the largest in existence.
2. **Oxford University Press**, 170 cases and 950 packets of printing type from the 17th to 20th centuries, mostly cast at Oxford for printing western and oriental languages, including those known as the 'Fell' types.
3. **Caslon Foundry**, 1,051 boxes of punches from the typefoundry begun by William Caslon with original inventories, albums of smoke proofs etc.

c. Plantin-Moretus Museum, Antwerp (3 weeks)

i. Expand knowledge of the workings of the seventeenth-century printing house and of the trade in general

1. Plantin-Moretus houses the two oldest printing presses in the world, the only full typographic collection from the early modern period, as well as a number of dies and molds of the most famous type designers of the 16th and 17th century.[4]

a. Citing Graham Pollard, G. Thomas Tanselle wrote that "[w]hen books by individual printers are gathered together and all this evidence examined, Pollard says, 'the knowledge which may be gained of the inner working of a printing office is often surprising.' This remark is important for suggesting that bibliographical analysis is a form of historical recovery and that the story of how a book was produced (or part of it) can be learned from the book itself" (Tanselle 14).

[4] Information about the Plantin-Moretus is compiled, in large part, from their website. <http://www.museumplantinmoretus.be/eCache/MFN/30/05/437.html>

b. This research will take Tanselle's observation further by exploring the inner working of a printing office first hand and supplementing this knowledge with the archival information about the Plantin business.

2. Plantin will offer the opportunity for a close study of punchcutters from the sixteenth century, information that will allow the postdoc to buttress conclusions about the flow of type from the continent into England.
    a. "Punchcutters in the late sixteenth century were so few that to generalize about them is difficult. Plantin dealt with the best of them in Paris and the Netherlands, and there is little sign of the production of new typefaces elsewhere at that time. . . . Punchcutters were always scarce" (Carter 11).
    b. This scarcity facilitates the study of a closed control group that will confirm or helpfully complicate the conclusions reached in the earlier phases of the research project.

ii. Collect data using materials onsite
    1. See above (5.c.1) regarding the typographic specimens available, as well as the "closed control group" of continental punchcutters discussed in (5.c.2).

6) Phase Six: Build databases for OCR engines (2.5 months, APR-JUN)
    a. Aggregate, analyze, and organize data gathered on research trip in Phase Five.
    b. Work with IDHMC programmers to manipulate data into a form that will be (a) most beneficial to the OCR engines and (b) robust enough to withstand various further manipulations.

7) Phase Seven: Drafting co-authored articles (3 months, JUL-SEP)
    a. Using information gathered in Phases Three and Five, supplement the draft of the article that was begun in Phase One that attends to the gap in analytical bibliography noted by Harry Carter in the opening of this section.
        i. Create a narrative/history of the flow of fonts, the primary type casters and foundries, and the distinguishing characteristics, as best as the postdoc and Dr. Samuelson can determine, of various fonts that find their way into England.
        ii. Appendices may include some version of the data collected over the course of the research project.
    b. Draft an article (with Dr. Laura Mandell) on the unique experience of collecting and parsing these data for the context of creating OCR engines.

c. These articles will contribute substantially to the body of knowledge in the humanities not only the field of analytical bibliography but also by exhibiting the ways in which data might be collected, disseminated, and used.

Timeline of English Typography

1476-1490: Beginning of Printing in England
Flemish typefaces imported by Caxton and others

"Caxton returned [from Bruges] to England in 1476, and set up a press in the Abbey precincts at Westminster – at the Sign of the Red Pale – and also brought with him some type and equipment. The remaining seven fonts of type that Caxton used fall into two classes: *batarde* types of the Burgundian school; and *lettres de forme* more on the model of pointed gothic types of the Mainz school" (Updike 116-7). "In the days before 1490 Caxton, De Worde, Lettou, and Machlinia had typefaces of Flemish cut, some of which are found in use in Flanders too" (Carter 63). A consideration: "Besides the books that were printed in England, a great many volumes were printed abroad for the English market – some at Antwerp, Paris, and Rouen, others at Basle, Louvain, and Cologne. These do not concern us in our study of English type-forms, though in passing it may be said that they were more finished than the books printed in England" (Updike 123).

1490-1540
French Black Letter

"No doubt the textura as cut in France was the letter that the English regarded as their own and took to calling 'English' before the time of Moxon in the late seventeenth century" (Carter 64). Similarly, Carter writes that "talking about the French Black Letter we in this country are on familiar ground. From 1490 until 1540 it was our national idiom in type. . . . I do not mean that London printers followed a French fashion in the early years: they bought French matrices" (63). Updike's view of the source is rather different: he writes that "in Caxton's day, gothic letter was in vogue for all English printing. Later, this gothic crystallized into an English pointed black-letter character, similar to some of the black-letter of the Netherlands, from which, tempered perhaps by French influences, it was derived. It was the characteristic type of England, and we find it in the English workrooms of De Worde, who greatly perfected it, at the beginning of the sixteenth century, as well as in use by Pynson and Berthelet. This character was commonly employed throughout the sixteenth century, and until the end of the seventeenth century, and even in the eighteenth century it was still used for law-books, proclamations, licenses, etc." (Updike 88-9). This question of the connection between genre and typeface is treated by Carter, too, who says that the "French *bâtarde* and Italian rotunda had their parts to play in London printing, the *bâtarde* mainly for law texts in Norman French, and the rotunda on small bodies for compendious volumes in Latin; but it was a humble role as compared with the scope given them abroad" (Carter 64).

1540-1650
Rise of Roman, Decline of Black Letter

Speaking of the French textura which dominated English printing in the previous half-century, Carter writes that its influence continued, though "in the half-century after 1540 we used mainly Flemish or Dutch imitations of it" (63). Carter points to the beginnings of roman types in England before this period: "In England, Richard Pynson was the first to print in Roman. A sermon of Savonarola of 1509 from his press is the earliest example. Thereafter this kind of type was commonly used for Latin if it savored of the New Learning. In the 1540s it occurs quite often in short passages of English, chapter summaries, poems, and prefaces" (92). He dates the first book entirely printed in English in roman types to 1555; Updike places his landmark of printing full volumes in roman types in England at 1518, though these are works in Latin. In describing the growing acceptance of the roman letter in Britain, Carter writes "the Geneva Bible of 1560 must have accustomed a great many Englishmen and Scots to the new letter. The officially-sponsored Bishop's Bible and its successor the Authorized Version of 1611 being Black Letter books, it is probable that Roman type had in British minds associations with puritanical Calvinism, or perhaps rather with dissent, for the Romanists favored it too. As late as 1637 Archbishop Laud insisted on the Book of Common Prayer for Scotland being set mainly in Black Letter. The last of our Black Letter Bibles was of 1640" (Carter 92). Many of the roman faces used by London printers were French, and Carter discusses the appearances of Garamond's types in 1538-42, but states that they were fairly uncommon, "probably because later in the century, when Roman and Italic types became usual, a French typefounder, Jerome Haultin, worked in London supplying founts from matrices which he bought from his uncle, Pierre Haultin, in France. What look like Garamond's faces in our Elizabethan books often turn out on closer acquaintance to be Haultin's" (Carter 86).

Of indigenous punchcutting in England, Updike centers his attention upon John Day, "the London printer . . . [who] left the most distinct mark on early sixteenth century English typography" (90). Active from the 1560s-1580s, Day cut a face in Saxon characters, and Updike writes that "his roman and italic used in the volume are of extreme importance in the history of early English type-founding. . . . Day was one of the first English printers to cut roman and italic letters on uniform bodies. Before that time, roman and italic types had been considered characters without mechanical interrelation; as examination of books in which they are both employed too plainly shows" (90-2). Day did purchase Dutch types as well, as Carter notes, from Henric Pieterszoon de Lettersnijder of Rotterdam. "The smallest, a Pica Black, occurs in London in the printing of Day and of Waldegrave" (105-6).

During this period, Updike writes of a general decline in English typography due in part to "the general falling off which began as soon as the restraining traditions of the manuscript volumes had passed away" (92) and in part to State restrictions on printing and typefounding. One implication of this is, as Updike quotes Talbot Reed, that “this transitional period of decline is due to the fact that so many earlier printers had used purchased types, and that after Day’s time, the restrictions of the State required printers to become letterfounders, a state which caused a reduction in quality. (93) As of 1557, no presses could be founded outside London aside from one each at Oxford and Cambridge, and the Star Chamber decree of 1637 restricted punchcutters sanctioned in the country to four positions selected by a commission. Even when the restriction was lifted in 1693, a lack of practiced letterfounders to fill the demand led to most type being imported (Updike 94).

1650-1720
Ascendancy of Dutch typefaces

The press at Oxford, which was founded in 1585, was revitalized “between 1667 and 1672 [as] the press received some fine types imported from Holland by Dr. John Fell, Dean of Christ Church and later Bishop of Oxford. . . . Fell employed Marshall, afterwards Dean of Gloucester, to buy some of these types in Holland, and Marshall’s negotiations for their purchase (between 1670 and 1672) were chiefly with Abraham van Dyck, son of Christoffel, the celebrated type-cutter, and Dirk Voskens. . . . Dr. Fell also imported a Dutch letter-cutter, Peter Walpergen, to direct the Oxford foundry” (Updike 95-7). Cambridge University also imported most of its press from Dutch foundries (Updike 96 note 2). Ultimately, Updike writes that “the types of most seventeenth century English books were probably Dutch. For this there were several reasons. One was the success of the Elzevirs, then the prominent publishers and printers of Europe, whose types were Dutch. Then there was the influence of fashion, for ‘the caprices of the court have always been to some extent responsible for the evolution of taste;’ and court taste was to some degree Dutch. Moreover, with the Revolution, English restrictions on the importation of types were removed, and the use of Dutch fonts came about partly because, on account of previous hampering governmental regulations, there were not enough trained letter-cutters left in England to produce good types. That was the most potent reason of all for the general English use of the Dutch letter” (99).

1720-1800
Caslon and His Descendants

Although Talbot Reed states that “there was probably more Dutch type in England between 1700 and 1720 than there was English,” Updike notes that “the rise of William Caslon, the greatest of English letter-founders, stopped the importation of Dutch types; and so changed the history of English type-cutting, that after his appearance the types used in England were most of them cut by Caslon himself, or else fonts modeled on the style which he made popular. . . . Caslon’s various specimens will show the English style. These, with Baskerville’s specimens, are the chief sources for the study of eighteenth century English type-forms” (100).

Eighteenth Century Collections Online

Unprecedented insight into a vital era

All illustrations reproduced with kind permission of the British Library and the British Museum.

> *"The content, scope and accessibility of* Eighteenth Century Collections Online *are astonishing. Enthusiastically recommended for all academic, public and research libraries serving serious literary scholarship."*
>
> — Library Journal

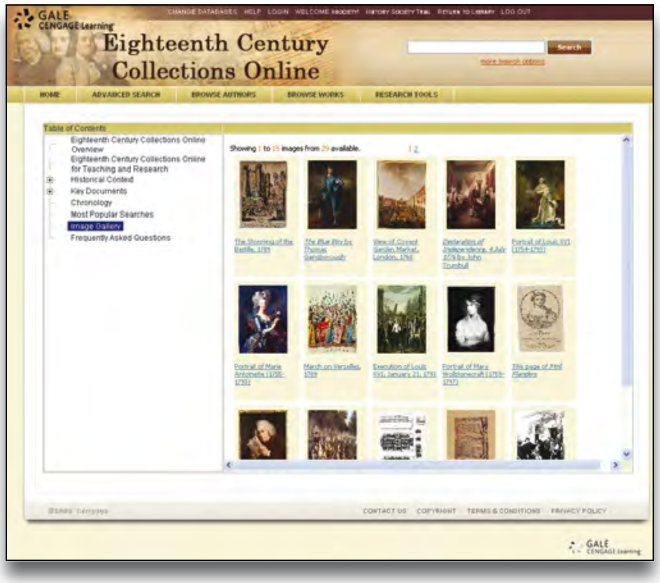
On its debut, *Eighteenth Century Collections Online* represented the most ambitious single digitization project ever undertaken – more than 26 million pages of text based on the English Short Title Catalogue. It delivered into users' hands more than 136,000 key English- and foreign-language titles printed in Great Britain during the 18th century including thousands of works from the Americas.

Today, this collection – along with its recent *New Editions* module of nearly seven million new pages (see reverse) – remains the premier online resource for scholars, faculty and researchers worldwide:

- More than 26 million pages of text from more than 136,000 titles (33 million pages and 180,000 titles when New Editions is added)
- In-depth coverage of well-known and lesser-known authors
- Canonical titles of the period as well as contemporary works that analyze and debate those titles
- Source institutions include the British Library, Oxford University, Harvard University, Cambridge University, National Library of Scotland, National Library of Ireland, the Library of Congress

Powerful technology, pinpoint results

A new interface and even more search options and research tools has enhanced *Eighteenth Century Collections Online* (see reverse). Users can easily search every word of every page of the entire collection by a word or a phrase, and view digitized page images with their search term highlighted.



GALE CENGAGE Learning
CHANGE DATABASES
HELP
LOGIN
WELCOME
History Society Trial
Return to Library
LOG OUT
Eighteenth Century Collections Online
Search
more search options
HOME
ADVANCED SEARCH
BROWSE AUTHORS
BROWSE WORKS
RESEARCH TOOLS
Table of Contents
Eighteenth Century Collections Online Overview
Eighteenth Century Collections Online for Teaching and Research
Historical Context
Key Documents
Chronology
Most Popular Searches
Image Gallery
Frequently Asked Questions
Showing 1 to 15 images from 29 available.
1 2
The Storming of the Bastille, 1789
The Blue Boy by Thomas Gainsborough
View of Covent Garden Market, London, 1760
Declaration of Independence, 4 July 1776 by John Trumbull
Portrait of Louis XVI (1754-1793)
Portrait of Marie Antoinette (1755-1793)
March on Versailles, 1789
Execution of Louis XVI, January 21, 1793
Portrait of Mary Wollstonecraft (1759-1797)
Title page of Moll Flanders
CONTACT US
COPYRIGHT
TERMS & CONDITIONS
PRIVACY POLICY
GALE CENGAGE Learning

▲ Users will find a growing collection of images to complement millions of pages of text

Try it today

For details on *Eighteenth Century Collections Online*, contact your Gale Representative or visit gale.com/gdctrial to register for a free trial.

» Gale Digital Collections

Eighteenth Century Collections Online, Part II: New Editions

Complement your holdings with millions of new pages

> *"A university that does not have* ECCO *is not a serious player in eighteenth-century British and American studies – in literature or anything else."*
>
> — Rob Hume, Pennsylvania State University

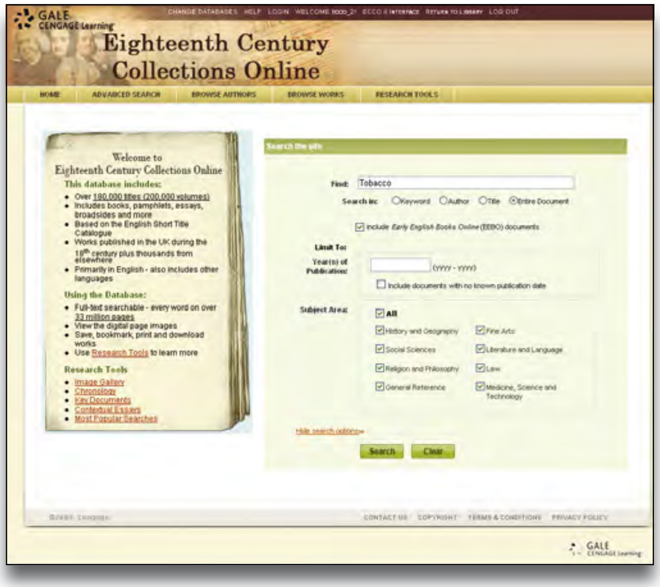
The English Short Title Catalogue continues to uncover both new works and new holdings of previously unavailable titles. The result is *Eighteenth Century Collections Online, Part II: New Editions*, an addition to the original collection that will be welcomed by serious researchers as an example of your commitment to their studies.

Engage and empower researchers with new content

Created specifically to supplement the original, *New Editions* delivers the online ease and depth of content that made the *Eighteenth Century Collections Online* a must-have resource for scholarly libraries worldwide.

Enhancements include:

- Nearly 50,000 new titles of previously unavailable or inaccessible materials
- Cross-searchable content from Early English Books Online from ProQuest
- Image Gallery, Most Popular Searches and Key Documents sections
- Contextual essays and chronology aimed at novice researchers
- Citation generator and export functionality
- Expanded download and e-mail features
- Keyword in Context feature from results list
- Updated user interface, and much more


GALE CENGAGE Learning
CHANGE DATABASES HELP LOGIN WELCOME ecco_21 ECCO II INTERFACE RETURN TO LIBRARY LOG OUT
Eighteenth Century Collections Online
HOME ADVANCED SEARCH BROWSE AUTHORS BROWSE WORKS RESEARCH TOOLS
Search the site
Hide search options»
Search
Clear
CONTACT US COPYRIGHT TERMS & CONDITIONS PRIVACY POLICY
GALE CENGAGE Learning

Welcome to Eighteenth Century Collections Online

This database includes:

- Over 180,000 titles (200,000 volumes)
- Includes books, pamphlets, essays, broadsides and more
- Based on the English Short Title Catalogue
- Works published in the UK during the 18th century plus thousands from elsewhere
- Primarily in English - also includes other languages

Using the Database:

- Full-text searchable - every word on over 33 million pages
- View the digital page images
- Save, bookmark, print and download works
- Use Research Tools to learn more

Research Tools

- Image Gallery
- Chronology
- Key Documents
- Contextual Essays
- Most Popular Searches

Find: Tobacco

Search in: ○ Keyword ○ Author ○ Title ● Entire Document

☑ Include *Early English Books Online* (EEBO) documents

Limit To:

Year(s) of Publication: (yyyy - yyyy)

□ Include documents with no known publication date

Subject Area: ☑ All

☑ History and Geography ☑ Fine Arts

☑ Social Sciences ☑ Literature and Language

☑ Religion and Philosophy ☑ Law

☑ General Reference ☑ Medicine, Science and Technology

▲ A new interface and millions of new pages enhance the acclaimed original *Eighteenth Century Collections Online*

Try it today

For details on *Eighteenth Century Collections Online, Part II: New Editions* contact your Gale Representative or visit gale.com/gdctrial to register for a free trial.

FREE TRIAL — For more information or to request a free trial of this unique collection, please contact your Gale Sales Representative or visit www.gale.com/GDCTrial.

» Gale Digital Collections

EEBO (Early English Books Online)							
	1473-1639	1640-1661	1662-1700	Total			
	33,807	39,223	52,530	**125,560**			
		percentage total published 1640 or after:					
				73.08%			
				TCP typed**			
				45,500			
				**by 2012			
		Total for which no typed text exists					
				80,060			
Work suggesting that Donne Collection Font Samples will be Represented in EEBC							
Samples from the Donne Collection					**Publication Cities, EEBO**	**Date Ranges**	
Printers		Location					
M. Fletcher (MF)		UK					
TN		UK					
W Stansby		UK					
Thomas Harper		UK					
John Dawson		UK					
Cambridge UP Printers: Buck and Daniel							
Thomas Buck		UK					
Roger Daniel		UK					
Ja. (J.) Flesher		UK					
					Aberdeen	1640-1663	
					Amsterdam	1640-1650	
C. Plantin (Elsevir Family)		Antwerp			Antwerp	1640-1650	

Samples from the Donne Collection					Publication Cities, EEBO	Date Ranges	
Ja. (J.) Flesher		Basel					
					Bristol	1640-1661	
					Bruges	1670-1681	
					Brussels	1670-1683	
					Caen	1640-1666	
					Cambridge	1640-1650	
					Canterbury	1550-1560	
					Chemsford	1670-1685	
					Colchester	1640-1670	
Ja. (J.) Flesher		Cologne			Cologne	1550-1560	
Ja. (J.) Flesher		Cologne					
					Cork	1640-1667	
					Coulogne	1670-1688	
					Deif	1640-1659	
					Dieppe	1670-1689	
					Dordrect	1640-1656	
					Douai	1640-1652	
					Dublin	1550-1560	
					Edinburgh	1550-1560	
					Emden/Basil	1550-1560	
					Exeter	1640-1668	
					Falmouth (Cornwall)	1670-1687	
Typographia Vaticana		Franeker					
Ja. (J.) Flesher		Frankfurt			Frankfurt	1550-1560	
					Gant	1670-1684	
Ja. (J.) Flesher		Geneva			Geneva	1550-1560	
					Glasgow	1640-1651	
					Gothenburg	1640-1671	
					Grenwich	1550-1560	
					Hamburgh	1640-1655	

Samples from the Donne Collection					Publication Cities, EEBO	Date Ranges	
					the Hague	1640-1654	
Typographia Vaticana		Ingolstadt					
					Kilkenny	1640-1669	
					Leiden	1640-1650	
					Lisbon	1640-1672	
					London	1550-1560	
					Louvain	1640-1662	
Ja. (J.) Flesher		Lyons			Lyons	1550-1560	
					Middelburg	1640-1664	
					Newcastle	1640-1665	
					Norwich	1670-1690	
					Orleans	1670-1686	
					Oxford	1640-1650	
Ja. (J.) Flesher		Paris			Paris	1550-1560	
Typographia Vaticana		Paris					
Typographia Vaticana		Pont-a-Mousson					
					Regensburg	1670-1682	
Typographia Vaticana		Rome					
					Rotterdam	1640-1653	
					Rouen	1550-1560	
					Saint Andrews	1550-1560	
					Saint Omer	1640-1657	
					Salisbury	1670-1680	
					Shrewsburty	1640-1660	
					Strasburgh	1550-1560	
Typographia Vaticana		Valenciennes					
Ja. (J.) Flesher		Venice			Venice	1550-1560	
					Waterford	1550-1560	
					Wesel	1550-1560	
					Worcester	1550-1560	

Samples from the Donne Collection					**Publication Cities, EEBO**	**Date Ranges**	
Ja. (J.) Flesher		Wurzberg					
					York	1640-1658	
					Zurich	1550-1560	
ECCO	Phase I	Phase II					
Total Documents Phase I	136000	182,898					
1700-1720 Phase I	21284						
Percentage after 1720	0.1565	180,000	(less 2k typed)				
	16.00%	151,200					
Applying 16% to Phase II total give you this figure	16% of 180k	Checking Column B: 180k minus 151.2 k=					
	28800	28,800	most difficult for OCR to type				

LAURA MANDELL

"Brave New World:
A Look at *18thConnect*"

for

The Age of Johnson

A Scholarly Annual

Volume 21
Vol. 21 ISBN-10: 0-404-62771-4
Vol. 21 ISBN-13: 978-0-404-62771-3
Published December 2011

Editor
Jack Lynch

Book Review Editor
J. T. Scanlan

Contact: editorial@amspressinc.com

AMS Press, Inc.
New York
amspressinc.com

BRAVE NEW WORLD: A LOOK AT *18THCONNECT*

LAURA MANDELL

The rapidly changing scholarly digital universe requires both new tools and new practices to meet the challenges it presents. Because the skill sets required to design the digital universe vary, scholars, librarians, computer scientists, and even administrators must work together to create platforms for locating, searching, and accessing the data now stored in electronic media. If scholars fail to join librarians in this task, many valuable digital materials will effectively vanish, their traces washed away in the data deluge. Similarly, institutions must adjust their practices to meet the realities of the new digital world, especially with regard to approaches to tenure and promotion. These institutional practices require a vetting process equivalent to juried publication for assessing the quality of novel forms of scholarship such as databases, electronic editions, software, finding aids, or digital tools. What is needed is a collaborative effort that permits young scholars with savvy ideas for digital projects to develop those projects, confident that their work will receive the peer review necessary for keeping their jobs. Finally, there is a larger systemic issue that Jerome McGann notes: our cultural heritage is going to be digitized, and scholars need to be at the table when people are deciding when, where, and how.[1]

To address these needs, I founded and currently co-direct *18thConnect* (http://www.18thConnect.org), an exciting new community, electronic platform, and research portal currently under construction and supported by Miami University of Ohio, the University of Virginia, an NEH grant from the Institute for Computing in the Humanities, Arts, and Social Sciences (I-CHASS) at the University of Illinois, and a grant from the Andrew W. Mellon Foundation awarded to Miami University, my home institution. Brad Pasanek of the University of Virginia co-directs *18thConnect*, which takes its inspiration from *NINES* (*Nineteenth-Century Scholarship Online*) founded by Jerome McGann and directed by Andrew Stauffer, also at the University of Virginia. Like *NINES*, *18thConnect* will offer scholars access to electronic resources by allowing for an aggregated integration of those resources. In other words, *18thConnect* will serve as a single platform through which multiple resources can be searched. Freely available scholarly resources, such as the *English Short-Title Catalogue* (*ESTC*), the *Old Bailey Online*, and the 2,180 *ECCO* texts that have been transcribed by Michigan's Text Creation Partnership, can be consulted with a simple click on a link returned in a search; while proprietary resources can be fully accessed only by those whose institutions subscribe to them, all searches will return bibliographic data into the *18thConnect* portal.[2] The editorial board and steering committee, moreover, will be actively soliciting submissions from additional digital resources in the fields of art history, literature, history, and philosophy.

18thConnect peer-reviews each database and electronic scholarly edition in its collection. In fact, one of *18thConnect*'s tasks will be to serve as a vetting agent, helping to legitimize worthy electronic scholarship for promotion and tenure committees. The editorial board members of *18thConnect* are as illustrious as those for any major press. When *18thConnect* accepts an electronic scholarly edition, it will provide letters of acceptance designed to help interpret an accepted project's value for promotion and tenure committee members for whom such work may be new.[3] *18thConnect* follows in the tradition established by Jerome McGann's purpose in establishing *NINES* for nineteenth-century scholarship (http://www.nines.org). Both projects ensure that high-quality scholarship is distinguished from the mass of materials available online; both ensure that promising young scholars will receive the credit they need and deserve for high quality work.[4] *NINES* has received an NEH Summer Institute Grant to conduct two institutes on this very topic: representatives from the ACLS and MLA participating in this *NINES* Summer Institute will work together with chairs, administrators, and digital humanists in order to come up with criteria both for judging work and for communicating their value to promotion and tenure committees.

It should be emphasized that *18thConnect* serves only as an *aggregator* for online projects; the databases, editions, and electronic scholarship to which it offers access remain in the hands of their creators and will not be owned, controlled, or operated by *18thConnect*. Open-access sites thus retain their creative freedom and can be submitted for peer-review while still undergoing major development. In fact, it is ideal for creators or editors to submit their work in its developmental stage: *18thConnect* and *NINES* together offer instruction via summer workshops and institutes on how to create library-quality, state-of-the-art digital resources, projects worthy of becoming, for instance, MLA Electronic Scholarly Editions. Though *18thConnect* only points to rather than ingests digital scholarship, proprietary collections and journals such as *ECCO, E-C Journals*, and *JSTOR* are searchable through *18thConnect*. Full text can be searched when it has been made available by the owners or creators of the resource. Users are returned text snippets, and then can click on the title of the work to see the full article, digital facsimile, electronic edition, or data strand. As mentioned above, freely available texts are completely accessible through *18thConnect*, whereas access to the texts in proprietary collections such as *ECCO* or the *E-C Journals* portal are accessible to users only if their home institutions subscribe. Users with access to *ECCO* at work or home but using a proxy server can immediately access an *ECCO* text by searching *18thConnect*. Once *18thConnect* is fully functional, a link to that primary text will be returned, accompanied by links to all the reliable scholarly information available on the web related to the search term or title—from scholarly articles to versions, editions, images, and the like.

What of those whose institutions do not subscribe to *ECCO*, the major eighteenth-century resource for primary materials, containing literally millions of page images of texts in the fields of language, literature, law, social science, medicine, etc.? *18thConnect* has worked out what may prove to be a historic agreement with Gale–Cengage that will provide search access and more to users whose institutions cannot afford to subscribe.

Gale's *ECCO* contains page images for over 182,000 texts, some of them multi-volume texts as lengthy as *Clarissa*. Creating such a set of images has taken decades of work, but some of the page images are not sufficiently readable to be transformed into typed texts by computer programs designed for this work. Earlier OCR (Optical Character Recognition) software was unequipped to handle the variability of eighteenth-century typography, often misreading words and thereby compromising text searches. *ECCO*'s digitized images, moreover, were made from microfilmed copies of the original titles, and this also sometimes compromises the legibility of the scanned page. To address this problem, new, more precise OCR software is needed. Grants from the Mellon Foundation and NEH (through the National Center for Supercomputer Applications [NCSA] and I-CHASS) are funding Miami University's development of a new, open-source software program for mechanically transforming images into typed texts. *18thConnect* will re-run *ECCO* page images through this new program in order to generate cleaner text than Gale has been able to produce so far. Next, *18thConnect* will provide a window for users—anyone who wishes to register with an e-mail address—to correct the typing of these texts.

The new OCR software we will use, Gamera, is an open-source program originally developed by Professor Ichiro Fujinaga of McGill University and published by Johns Hopkins University (http://gamera.informatik.hsnr.de/). Because Gamera was originally created for recognizing musical characters, it is less dependent than other OCR software on recognizing characters only if they occur on the same line as others. This feature is valuable for scanning texts produced before 1820 because the characters in those texts are often not evenly aligned on any given line, but instead can fall above and below, the result of the punch not being situated in the matrix with mathematical precision when the type was made. We have already been able to train Gamera to distinguish between the long s and the lowercase *f*, something that was previously possible only through dictionary look-up. There are, however, some things Gamera may not do as well as Gale's OCR, so we are further developing automated correction, and the centerpiece of our process: a crowd-sourced correction tool. It is time, Martin Mueller has said, for scholars to wash their own dishes: the more scholars help us correct texts, the better scholarly searches will be, and ultimately—because everything will be sent back to Gale—the better the archive will be for future scholars.[5] That these texts be correctly typed is crucial for searching and data-mining; only by providing texts that can be accurately read by machines will we make them locatable, usable, and comprehensible to future generations.

Contributing to the future robustness and integrity of the archive, users of *18thConnect* can search for documents in the *ECCO* collection and (1) correct errors found in the snippets, or (2) register in order to see and read portions of a text in exchange for correcting it. If someone has registered as a user at *18thConnect*, which requires only a username and e-mail address, that person can save documents that he or she would like to correct into a personal account on the "My18" page. Additionally, if a user decides to correct a whole text, once the corrections are completed through the online correction tool accessible through the "My18" page, *18thConnect* will immediately give the correctly typed version of that document to the person who corrected it *to use as he or she likes,* and we will do so in several forms: both plain text and text encoded in TEI, which the MLA requires for electronic scholarly editions.[6] We will also provide guidance so that users can create a library-quality digital edition and use all the newest tools on their documents such as JuXta and the Versioning Machine, which enable one to compare various editions, as well as exciting new visualization tools such as TAPor, Voyeur: Reveal Your Texts, and TokenX.[7] Many of these tools will be made accessible through *18thConnect* in 2011. If sufficiently researched and annotated, editions built with our help can be submitted to *18thConnect* for peer review. Furthermore, library-quality scholarly editions are eligible to become MLA Electronic Scholarly Editions. Positively reviewed editions are first accepted into the *18thConnect* online finding aid. If a scholar's edition has been accepted (positively peer-reviewed), Gale–Cengage may choose to publish the edition along with the page images as a print-on-demand edition, or other print-on-demand publishers can be enlisted to print the transcribed, annotated edition.[8] Thus, in exchange for correcting texts and improving the archive, users will have produced a scholarly resource that will count toward tenure and that is both digital edition and printed book.

18thConnect will give coded documents, instructions, and support so that scholarship produced by users—ideally professors of eighteenth-century literature, culture, history, and art—is of the highest digital quality. The TEI coding that we provide ensures that these books can be produced in any number of forms—not just web pages but E-books, for instance, and whatever forms become necessary as new technologies emerge. In other words, one's scholarly work will be preserved for future use in a way that book publishing no longer guarantees. Though print copies will continue to exist, the standard means for discovering and searching scholarly work will be electronic; only by making one's work fully searchable by word will it rise to the surface of searches.

Let me give one example, situating the problem in recent debates about the scholarly uses of digital resources. John Guillory has taken issue with N. Katherine Hayles's praise of multitasking forms of attention, insisting on the value of the sustained attention required for close reading.[9] The former is associated with the digital, as is the "distant reading" proposed by Franco Moretti in his *Graphs, Maps, and Trees,* as well as elsewhere in his arguments with Katie Trumpener.[10] The activity of distant reading involves looking at visual representations of searching and data-mining many texts, and it can be alternated by close reading in and around texts discovered to be most interesting. But as is plain in reading an explicit account of the reading practices of scholars at our moment, neither concept—neither distant nor close—fully captures what we do.[11] Scholars always have, and ideally always will, act as "filters" of information based on minds developed through discipline and intensive study that cannot be duplicated by machines. While that will not change, what will change in the new digital universe is the advent of new and powerful search functions that will become indispensable, both for the cognitive filtering traditional scholarship demands and for the proper ordering and cataloguing of new digital archives, making findable the information they hold.

A search that I conducted in *18thConnect* further illustrates this point. I have been interested in historically changing notions of factuality and began searching for a phrase I had encountered in my reading that I thought might be a forerunner to the legal notion of "circumstantial evidence." My goal was to investigate whether the evidence of circumstance was as devalued during

THE TEXAS A&M

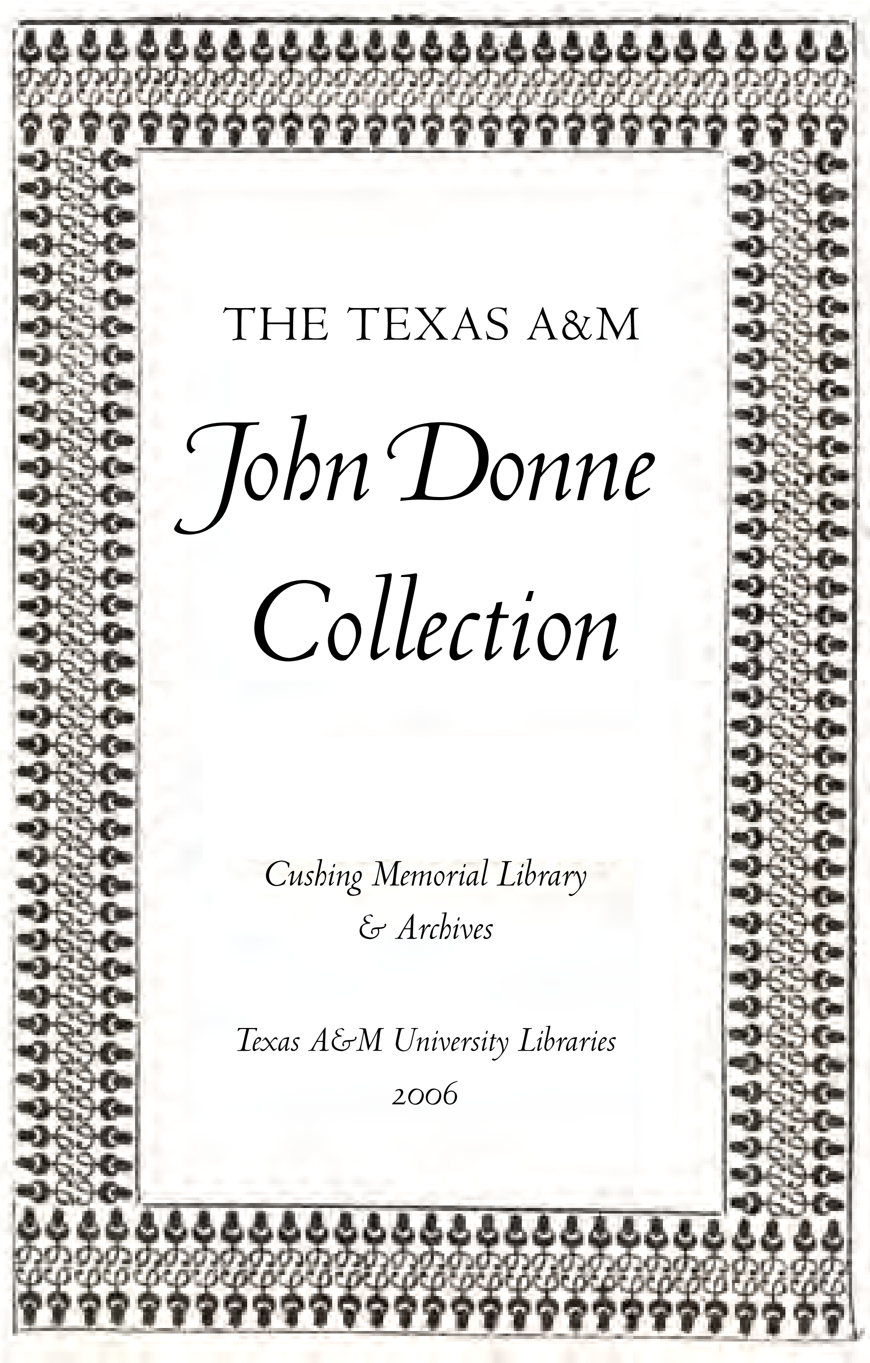
John Donne Collection

CUSHING MEMORIAL LIBRARY & ARCHIVES

Texas A&M University Libraries
2006

The John Donne Collection

TEXAS A&M UNIVERSITY



THE TEXAS A&M

John Donne Collection

Cushing Memorial Library
& Archives

Texas A&M University Libraries
2006

Acknowledgements and Sponsors

Preface: Steven Escar Smith
Introduction: Gary A. Stringer

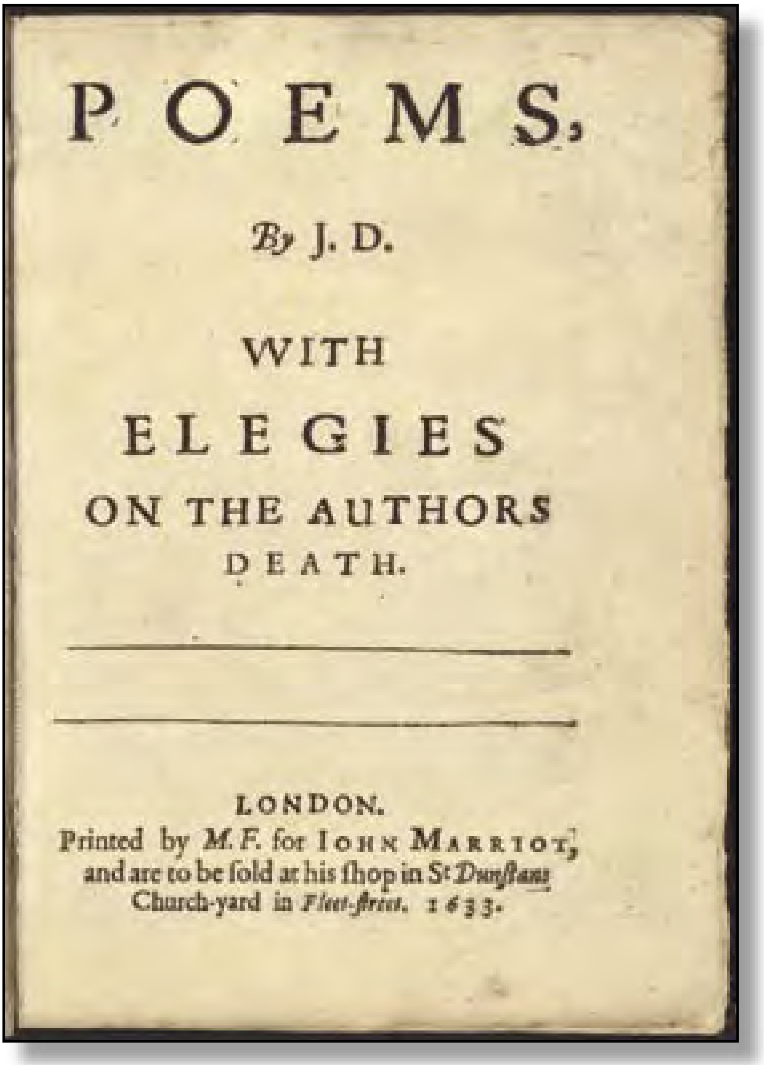
Editor: Christopher L. Morrow
Assistant Editor: Lisa Richter
Editorial Support: Gary A. Stringer, Paul A. Parrish, Steven Escar Smith, Tina Budzise-Weaver, Kristin Hill, and Hillary Little.
Technical Support: Stephanie Elmquist and Timothy A. Weaver
Administrative Support: Teri Czajkowski, Jo Youngblood, Dalaiah Eiland, Melyssa S. Henderson and Marian Eide

Catalogue Design: Mary Ann Jacob

Exhibit Design and Production: Christopher L. Morrow, Lisa Richter, and Lynn Zynda

Editorial Note: Annotations in this catalogue are heavily indebted to an unpublished catalogue of the Donne Library editions by Maggs Booksellers, the catalogue for the 16 December 2004 Sotheby's auction, and Sir Geoffrey Keynes's *A Bibliography of Dr. John Donne* (Oxford: Clarendon Press, 1973).

Made possible through the generous support of:
Melbern G. Glasscock Center for Humanities Research
Friends of the Sterling C. Evans Library
College of Liberal Arts
Department of English
Texas A&M University Libraries
John H. Hinton Cushing Memorial Library Endowment



POEMS,

By J. D.

WITH

ELEGIES

ON THE AUTHORS

DEATH.

LONDON.
Printed by *M.F.* for IOHN MARRIOT,
and are to be ſold at his ſhop in St *Dunſtans*
Church-yard in *Fleet-ſtreet*. 1633.

Poems, 1633

Preface

STEVEN ESCAR SMITH

SOMETIMES A LIBRARY buys something on faith and a hunch. In the early 1980s, Paul Parrish, professor of English at Texas A&M, approached Jay Poole, Assistant Director for the Sterling C. Evans Library, with a recommendation that we buy two books—the 1633 edition of John Donne's *Poems, by J. D. With Elegies on the Author's Death* and Donne's 1669 *Poems, &c.* The first was on offer from the Zeitlin & Ver Brugge Booksellers and the second from Bennett & Marshall Rare Books & Manuscripts, both of Los Angeles. Of course, the Library had always collected modern reprints of Donne and such secondary literature as was appropriate for the teaching and research of English and literary studies on the undergraduate and graduate levels. A few nineteenth-century editions had also worked their way into the collections over the years. But not until the purchase of these two items were any rare or first editions of Donne a part of the collection. Parrish, a Donne scholar and an editor on the *Donne Variorum* project, no doubt made a compelling case for these two books based on their inherent research value and interest. But for the Library to make such an investment, he would have also had to make the case that Donne studies in particular and Renaissance or Early Modern Studies in general were areas of growth and long-term commitment for the English department. Though the details of his argument, now over two decades old, have long since faded from memory, the presence of these two volumes is evidence of his persuasiveness and the willingness of the librarian, undoubtedly balancing this request against limited resources and many other demands, to listen and take a chance.

So the books were bought and added to the Special Collections Department where they resided happily for many years. Since then they have been used for research and teaching. They have also appeared in exhibits from time to time as examples of "high spots" in English literature or in support of other themes. Meanwhile, both the English Department and the Texas A&M University Libraries have enjoyed years of remarkable growth and development. The old Department of Special Collections and the University Archives are now the Cushing Memorial Library and Archives and occupy an historic and renovated building all their own. The special collections have increased in both breadth and depth in tandem with A&M's continued development as a large, research intensive, comprehensive educational institution. Cushing's collections and collecting areas include, but are not limited to, military history, science fiction, western Americana, ninteenth-century American prints and illustrators, modern politics, Texana, natural history, Africana, Hispanic studies, ornithology, nautical archaeology, eighteenth-century French history and culture, Mexican colonial history, and the history of books and printing. The literary material has developed and grown along with the rest, the most comprehensive of its author collections being Miguel de Cervantes, Rudyard Kipling, Somerset Maugham, the Powys family, Christina Rossetti, and Walt Whitman, to name only a few. The English Department has achieved distinction as a center for Early Modern Studies and in many other areas, with strengths across the range of English and American Literature as well as Rhetoric and Composition, Linguistics, Creative Writing, Discourse Studies, and Women's Studies. The Department is now or recently has been home to and supportive of publications and publishing enterprises such as the *World Shakespeare Bibliography*, the *South Central Review, Journal of American Folklore, Studia Mystica, Seventeenth-Century News, The Powys Journal*, and *Callaloo;* it is a participant in scholarly and professional initiatives, such as the Carnegie Initiative on the Doctorate; and it is a member and supporter of cultural organizations and institutions, such as the Folger Institute. Most relevant for this exhibit, however, the Department is now home to *The Variorum Edition of the Poetry of John Donne,* with Gary Stringer, the edition's founder and a co-founder of the John Donne Society of America, as General Editor and Visiting Professor of English.

The arrival of the *Donne Variorum* project at A&M made the acquisition of those two Donne volumes, which have been well-enough used over the years, an even better and more far-sighted investment. They were never considered ends in themselves, but now they took on added value as perhaps the beginning of a larger Donne collection in support of the English Department's growing programs. The fact that we have been adding other material of relevance to Early Modern Studies (most notably in relation to the *Don Quixote,* Shakespeare, French literature and history, and the history of books and printing) and that the Library is partnering to give electronic expression to the Donne project made acquiring more Donne material an argument for adding strength to strength.

Thus, in the fall of 2004, when Gary Stringer brought the Sotheby's auction of the I. A. Shapiro library, which contained copies of all the seventeenth-century Donne editions, to the Library's attention, it seemed like an opportunity worth pursuing. The subsequent sale of books represented in Donne's personal library, also from the Shapiro collection, by Maggs was yet another good opportunity. I. A. Shapiro (1904-2004), long-time Professor at the University of Birmingham, was known for his extraordinary knowledge of Elizabethan and Jacobean literature and history. He also held the contract with Oxford University Press for the collected edition of Donne's letters for 65 years, a project he never completed.

An ad hoc committee consisting of Parrish and Stringer along

Notes by and correspondence to I. A. Shapiro with Don Dickson and Larry Mitchell (also of the English Department) formed to look more closely at the Sotheby's sale. Members of the group also held discussions with Colleen Cook, Dean of Libraries, and Ben Crouch, Executive Associate Dean of the College of Liberal Arts. With encouragement from these administrators, approval from the College of Liberal Arts Library Enhancement Fund committee, expert advice from the ad hoc group, help from Julian Rota of Bertram Rota Booksellers (who bid on our behalf at the Sotheby's auction), and a few additional purchases to fill in gaps here and there, those two books, bought on faith and a hunch, have blossomed into one of the largest and most distinctive Donne collections in the United States. As is evident from the above, the collection's provenance is mostly I. A. Shapiro, though the two books acquired in the 1980s, *Biathanatos* (1644), the disbound "Printer to the Understanders" (from the 1633 *Poems*), and a few other items have come from other sources. I refer the reader to Gary Stringer's engaging introduction to this catalog for a more complete story of the acquisition of this collection.

In their acquisition decisions, librarians try very hard to anticipate future use and weigh very carefully cost against value. In their research, scholars hope that certain volumes or collections will answer particular questions or lead them in one direction or another. But at some point most collecting decisions, especially in regard to primary material, require a leap of faith, just as most acts of scholarship require a willingness to follow new paths and confront unanticipated questions. Using the A&M Donne collection, one scholar has already identified unknown variants in the 1633 edition, a discovery that would not have been possible without two copies of this edition (the one purchased back in the 1980s and the other from the Shapiro library) to compare side-by-side. The O'Flahertie copy of Donne's *Letters* (1654), packed with tipped-in notes, inter-leavings, and marginalia, is a treasure-trove of textual data waiting to be mined. The seventy or so books gathered by Shapiro in his effort to recreate Donne's working library offer an unequalled chance on this side of the Atlantic for exploring the author's influences. Other opportunities abound. From this vantage point, the question is no longer if the books will be used, but how? Watching the answer to that question unfold is one of the chief rewards of scholarship and librarianship.

Steven Escar Smith
Associate Dean and Director
C. Clifford Wendler Professor
Cushing Memorial Library and Archives
Texas A&M University Libraries

Table of Contents

17th-century editions of John Donne

Introduction

GARY A. STRINGER
General Editor, *The Variorum Edition of the Poetry of John Donne*
Visiting Professor
Department of English
Texas A&M University

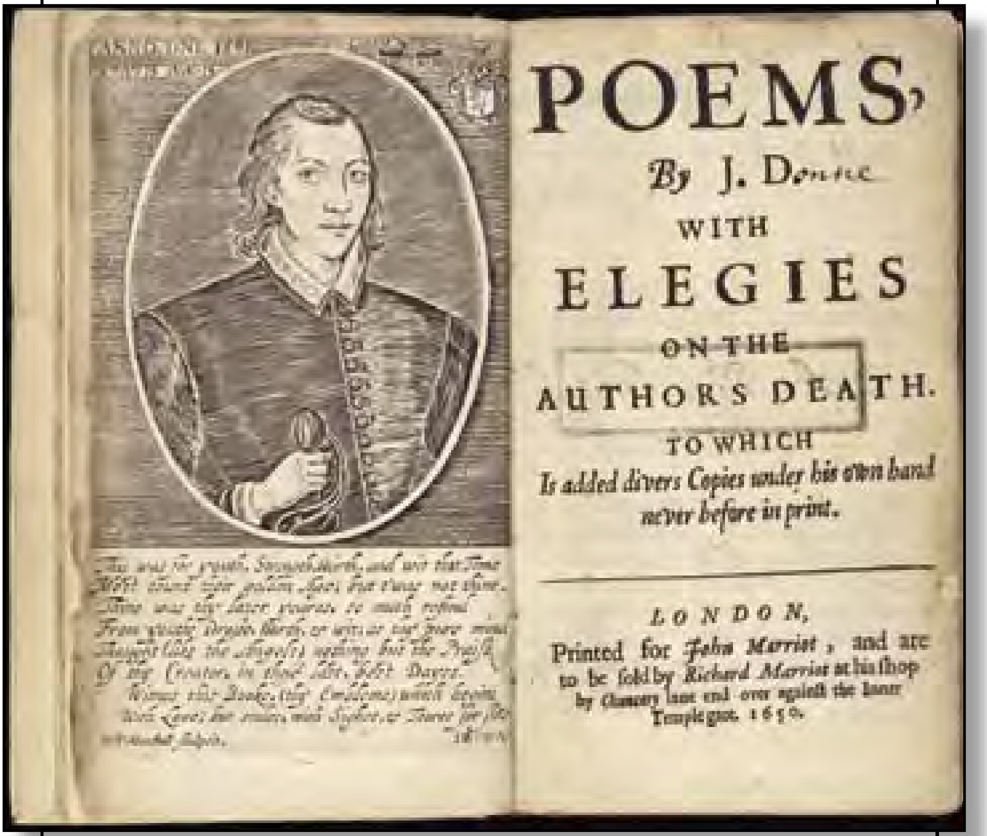
AT TEXAS A&M UNIVERSITY, interest in John Donne is not a new thing. Earlier generations of students routinely read A&M Professor Stanley Archer's "Meditation and the Structure of Donne's 'Holy Sonnets,'" an essay first published in 1961, but catapulted to prominence by its inclusion in 1966 in *John Donne's Poetry,* a Norton Critical Edition widely used as a classroom textbook.[1] Among those who cut their critical teeth on the work of Archer and his contemporaries was Paul A. Parrish, who joined Archer on the A&M faculty in 1974 and in 1981 became one of the founding members of the Advisory Board for *The Variorum Edition of the Poetry of John Donne,*[2] a project he has served in the additional capacities of Volume Commentary Editor for *The Anniversaries and the Epicedes and Obsequies* (1995), Volume Commentary Editor for the recently published *Holy Sonnets* (2005), and—since 1996—Chief Editor of the Commentary for the edition as a whole. And there is more. Among those whom Parrish recruited as a Contributing Editor to the *Variorum's* volume on the *Anniversaries* was his (then) young colleague Donald R. Dickson, who had joined the A&M faculty in 1981. Having completed this assignment, Dickson withdrew from the Donne scene to pursue other interests for a number of years, but reemerged in 2003, when he contracted to reedit Donne's poetry for a new issue of the Norton Critical Edition —the same work in which Stanley Archer's essay had appeared 37 years earlier. Dickson's volume is slated for publication in

December of this year, and he has recently rejoined the *Variorum* project as a textual editor.

In the fall of 2004, after retiring from the University of Southern Mississippi the previous spring, I joined the English faculty at A&M as a Visiting Professor and moved the *Variorum* headquarters to College Station, relocating the project into an institutional environment that seemed to optimize its chances of flourishing and reaching completion. The negotiations that brought me here involved my pledge to aid in the development of an electronic archive of Donne materials that would not only support work on the *Variorum,* but also enhance the university's growing emphasis on digital humanities scholarship as manifested in its online Picasso project and in the *Electronic Variorum Edition of Don Quixote.*[3] This effort was to draw on the resources of the Center for the Study of Digital Libraries in the Department of Computer Science and of the University Libraries' Digital Initiatives program, and we intended to begin with my collection of microfilms of the manuscripts and early editions of Donne's poetry. Before we could get this project up and running, however, word came through friends of a rare opportunity that would cause us to realign our priorities: in March of 2004, at the age of 100, the British scholar I. A. Shapiro had died in Birmingham, England, and many of his books—including a number of seventeenth-century editions of the works of John Donne—were to be auctioned off at Sotheby's in December of that year. The descriptions in the pre-auction list were eye-popping, so a group of English department faculty members decided to approach the administration about the possibility of putting together a fund with which to bid on these books.[4] It was not a hard case to make, but that we succeeded so painlessly is perhaps less a tribute to our powers of persuasion than to the understanding and vision of the three administrators who instantly agreed to collaborate in this enterprise: Dr. Ben M. Crouch, Executive Associate Dean of the College of Liberal Arts, who pledged major support from CLA library enhancement funds; Dr. Steven Escar Smith, Associate Dean for Advancement of the University Libraries and Director of the Cushing Memorial Library and Archives, who pledged supplemental support from his acquisitions budget; and Dr. Colleen Cook, Dean of the University Libraries, who supported the initiative throughout and played a major role in certain post-auction purchases. Future generations of scholars and book lovers will join us in applauding their foresight.

Once the funds were authorized, I rescheduled a previously postponed research trip to London to coincide with the date of the auction. Since I had never been to such an auction before, Steven Smith engaged Julian Rota (of the Bertram Rota Antiquarian Booksellers firm in London) to serve as our purchasing agent and do the actual bidding. I agreed to undertake a pre-auction examination of the books, confer with Julian, and make sure all the volumes conformed to their published descriptions. Together with friends, I first saw the books at a pre-auction open house at Sotheby's, and they were breathtaking. A couple of days later, just before the sale, Julian and I spent about two hours in a cubbyhole at Sotheby's leafing through the books and making notes on the condition of the various volumes. By this time, of course, we had copies of the published sale catalogue, which listed the books as a numbered series of "lots," described each lot, and provided an estimated price range for each. What we saw in the catalogue presented us with a problem of strategy—our funds were insufficient to cover the purchase of all the Donne items, even if each one sold at the low end of the estimate, and our list of priorities did not coincide with the order in which the items were scheduled to be sold. We wanted, above all else, volumes of the collected *Poems* to fill in the gaps between the 1633 and 1669 editions that the Cushing Library already owned and, if necessary, were prepared to spend everything we had to get them. Secondarily, we wanted copies of both the 1651 and the 1654 prose letters and, after that, whatever volumes of sermons and other prose works we might have funds for. The various editions of the poems were listed first in the catalogue, but the prose letters (our second priority) weren't scheduled for sale until after various volumes of sermons and other items had been sold. We thus had to hypothesize a what-if scenario in which our bidding would vary according to whatever amount of money we had left at given points in the auction, and my wife Mary Ann—much better with numbers than I am—was a great help in this. She drew up a chart showing what the cost of each item would be if it sold at the highest estimate, at twice the highest, and at thrice the highest, and then calculated what we would have left at each point in the proceedings under those three scenarios. In the event, all but two of the items we bought sold above the upper end of the pre-auction estimate, and some items—such as a rare copy of the 1650 *Poems*—went for more than twice the highest price the Sotheby's experts had projected.

Julian, Mary Ann, and I planned to work as a team at the actual



This was for youth, Strength, Mirth, and wit that Time
Most count their golden Age; but t'was not thine.
Thine was thy later yeares, so much refind
From youths Drosse, Mirth, & wit; as thy pure mind
Thought (like the Angels) nothing but the Praise
Of thy Creator, in those last, best Dayes.
Witness this Booke, (thy Embleme) which begins
With Love; but endes, with Sighes, & Teares for sins.
Will: Marshall. Sculpsit. IW.

POEMS,
By J. Donne
WITH
ELEGIES
ON THE
AUTHORS DEATH.
TO WHICH
Is added divers Copies under his own hand
never before in print.

LONDON,
Printed for John Marriot, and are
to be sold by Richard Marriot at his shop
by Chancery lane end over against the Inner
Temple gate. 1650.

Poems, 1650

auction, with him bidding and us feeding him information. Mary Ann was to sit on one side keeping a visible declining balance as the sale progressed; I was to sit on the other nudging Julian whether to keep raising the bid or not. On the day of the auction we walked in and took seats in the front row—which prevented our seeing what others were doing behind us, but also perhaps signaled that we didn't much care. There were probably 150-200 people in the room, including Sotheby's employees, other bidders, and spectators interested in the proceedings. At the front of the room was the auction block, a kind of round pulpit such as you sometimes see in older churches, and on tables along one side of the room was a bank of phones from which Sotheby's employees could stay in moment-to-moment contact with absentee bidders. The December 16 auction was divided into two separate sessions, the first (which included the Shapiro books) beginning at 10:30 A.M., the second at 2:30 P.M. Two-hundred-and-sixty-eight lots were to be auctioned in the first session—or slightly more than 1 per minute. This meant that things were expected to move quickly, and they did.

At the appointed hour the auctioneer, a clean-cut young man in a gray suit, walked up into the block, explained the rules, and began with the first lot. For each item he would first announce the starting price, based on bids previously submitted by mail or email, and then open the bidding from the floor. He treated each bidder with great courtesy—there was no badgering, and none of the rapid-fire, pressure-packed patter that one usually associates with auctions. The auctioneer just clearly announced each bid, surveyed the room to see whether anyone wished to raise it, and—when he could sense that we were "all done"—would strike the rail of the block with his hammer, a circular device more like a cookie cutter than a regular hammer, and that would be that. People bid by nodding, holding up a little paddle, or raising a finger. Phone bidders communicated with the employees engaged for that purpose, who then announced the bids as they came in. In most cases, the number of bidders was fairly quickly reduced to a very few contenders, and once the auctioneer sensed that the competition was going to be fierce, he would declare a minimum raise for the next bid to prevent the process from being bogged down in nickel-and-dime niggling. A 1916 broadside poster proclaiming the independence of the Irish republic, for instance, which had been estimated to go for £50,000-60,000, jumped to £75,000 in just a few seconds, whereupon the auctioneer declared that subsequent bidding would proceed in £10,000 increments. This item finally topped out at £130,000, I believe, and he sold it in just about a minute.

About 40 minutes into the auction we got to the Shapiro items, and within 12 or 13 dizzying minutes, A&M had acquired the books listed below. After the volumes of poetry had sold, we had obtained everything we had hoped for, and still had over $21,000 left, which allowed us to bid on some prose items we never expected to have money for. The only misstep, if you can call it that, came with lot 72—a lot comprising both the 1640 *LXXX Sermons* and the 1649 *L Sermons.* This lot was a priority for us, but the bidding escalated so rapidly that Julian and I decided to drop out and save our remaining money for subsequent items. We then took lot 73 (*LXXX Sermons* only) for less than the lowest estimate, and abstained from bidding on lot 74 (a second two-volume set of the sermons) because we wanted to save our money for the upcoming volumes of prose letters (lots 77 and 79). Fortunately —as I shall mention below—lot 74 did not sell, and we were able to acquire it later at £100 less than the low-end estimate. One of the most interesting items we acquired was an extensively annotated copy of the 1654 prose *Letters* that had once belonged to the Reverend T. R. O'Flahertie, a great nineteenth-century collector and afficianado of Donne, and I later learned that the phone bidder who had been driving up the price on this item was my friend and *Donne Variorum* colleague Tom Hester, who wanted the copy to consult for a modern edition of the letters that he, Dennis Flynn, and Ernest Sullivan are now preparing for Oxford University Press.

As is evidenced in the following pages of this catalog, the collection of volumes we obtained in this auction is truly astonishing. It includes the 1635 (second edition), 1639 (third edition), 1649 (fourth edition, first issue), and 1650 (fourth edition, second issue) editions of Donne's collected *Poems*; a first edition of *Pseudo-Martyr* (1610), Donne's first published work; a first edition of *Death's Duell* (1632), preached when Donne was so ill that a contemporary called it Donne's own funeral sermon; a first edition of *Six Sermons* (1634); a first edition of *LXXX Sermons* (1640); a first edition of *XXVI Sermons* (1660/61); a first edition of *Letters to Severall Persons of Honour* (1651); a first edition, second issue, of *Letters to Severall Persons of Honour* (1654); and a first edition of *A Collection of Letters made by Sr Tobie Matthews* (1660), which contains six Donne letters not contained in the other editions of his letters. Even were there no more, these volumes would constitute an invaluable cache of materials for critical and bibliographical research. Remarkably, however, there is more. A couple of days after the sale, through Sotheby's, I queried Shapiro's heirs as to the disposition of three lots that had not sold in the auction, and we subsequently made offers on and obtained four other very important items, one of which was the single most important item in the entire auction—Shapiro's copy of the first edition of the collected *Poems,* published in 1633. Also obtained in this transaction were a first edition of *L Sermons* (1649); a second copy of the 1640 *LXXX Sermons*; and a third edition of *Ignatius His Conclave* (1635), Donne's trenchant satire against the Jesuits. Still later, I received notice from Maggs Brothers Rare Books in London that they had purchased from Shapiro's library—and were offering to sell as a single lot—a collection of around 70 sixteenth- and seventeenth-century works, written in Latin and several modern European languages, concerning astronomy, biography, contemporary news and gossip, ecclesiastical polity, the epistolary arts, history, law, linguistics, liturgy, poetry, philosophy, politics, and theology that at one time or another Donne had alluded to, quoted from, or was known to have read. I passed this word on to Steven Smith, and in due course this collection of secondary materials, embodying much of the contemporary thought that helped to furnish Donne's mind, also arrived in the Cushing Library. It is described elsewhere in this catalogue and displayed in this celebratory exhibition. Also on exhibit is a recently purchased copy of the first edition of *Biathanatos* (1644), Donne's treatise on suicide, which Steven Smith obtained from the bookseller Phillip J. Pirages a few months ago.

In adding these recently acquired items to the 1633 and 1669 editions of the *Poems* that it already held, the Cushing Memorial Library and Archives—in one fell swoop, as it were—has assumed a noteworthy place among the world's repositories of primary Donne editions and contemporaneous background materials. That it has done so is a tribute to those individuals and agencies that have provided resources for the purchase of such items, as well as to those within the institution who have exhibited the courage and vision necessary to seize a once-in-a-lifetime opportunity. We hope to expand the collection as other items and resources become available. In the meantime, we are working diligently to digitize these newly acquired volumes and make images of them available on the internet, together with tools that will facilitate manipulation and analysis of their texts. We expect individual scholars to come to College Station to examine the actual books, but this web project will ensure that users the world over will have easy access to this truly remarkable resource.

NOTES

[1] A. L. Clements, ed., *John Donne's Poetry* (New York: Norton, 1966). Archer's essay originally appeared in the literary journal *ELH*.

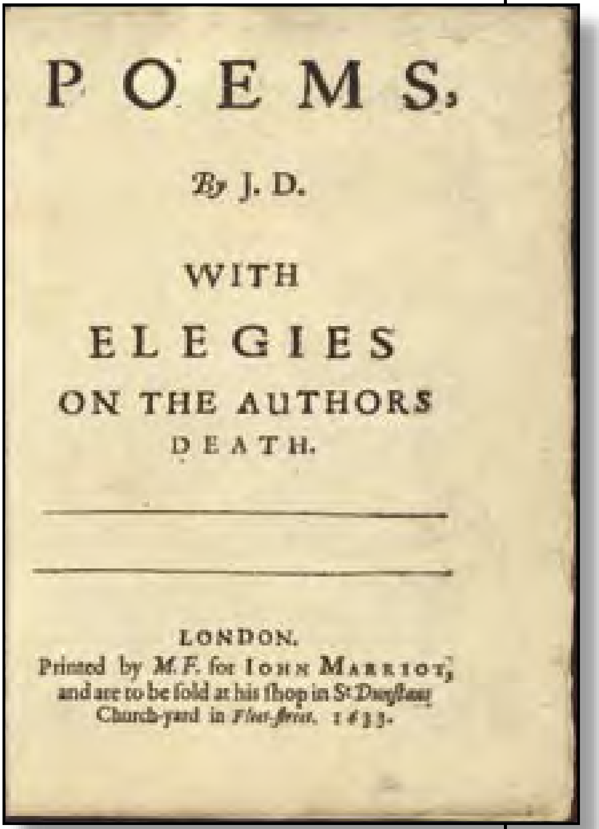
[2] Ed. Gary A. Stringer et. al., 8 vols. (Bloomington: Indiana University Press, 1995—)

[3] The director of this project is Dr. Eduardo Urbina. The URL for the *Quixote* edition is http://www.csdl.tamu.edu/cervantes/V2/variorum/index.htm

[4] Professor Donald R. Dickson wrote the formal proposal to Dean Ben Crouch, and the idea was enthusiastically endorsed by a number of English department faculty members, including Margaret J. M. Ezell, J. Lawrence Mitchell, Paul A. Parrish, and James L. Harner.

17th-Century Editions of John Donne

POEMS, *by J. D. with Elegies on the Authors Death.* London: M. F[letcher] for John Marriot, 1633. 4to. vi + 406 pp. Henry White copy.



POEMS,

By J. D.

WITH

ELEGIES

ON THE AUTHORS

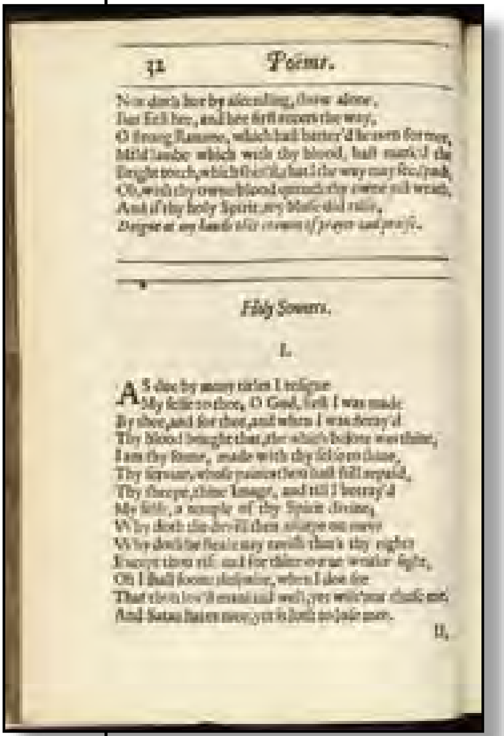
DEATH.

LONDON.
Printed by M.F. for IOHN MARRIOT,
and are to be ſold at his ſhop in St Dunſtans
Church-yard in *Fleet-ſtreet.* 1633.

Poems, 1633

First edition of Donne's collected poems, published two years after his death. This edition contains 154 authentic poems, 2 spurious poems, 11 Donne prose letters, and 13 elegies on Donne by various authors. The texts in this volume are based primarily on two manuscripts, but occasionally reflect the influence of at least two others. None of these manuscripts are written in Donne's hand. Either the publisher or another editor has eclectically constructed the wording of certain poems, regularized spelling to some extent, and smoothed the meter of many lines.

This copy is bound in contemporary calf with gilt-stamping on the front and back boards. The inside of the front board bears the following inscription: "Henry White. Cathedral Close Lichfield. Easter Eve & Lady Day. MDCCCXV." Reverend Henry White, M.A., was the Sacristan of the Lichfield Cathedral in the early decades of the 19th century. Lady Day, or the Feast of the Annunciation, falls on March 25th. Since Easter occurred on March 26th in 1815, Lady Day and Easter Eve both fell on March 25th of that year.



32 *Poems.*

Nor doth hee by aſcending, ſhow alone,
But firſt hee, and hee firſt enters the way.
O ſtrong Ramme, which haſt batter'd heaven for mee,
Mild Lambe, which with thy blood, haſt mark'd the path;
Bright torch, which ſhin'ſt, that I the way may ſee,
Oh, with thy owne blood quench thy owne juſt wrath,
And if thy holy Spirit, my Muſe did raiſe,
Deigne at my hands this crowne of prayer and praiſe.

Holy Sonnets.

I.

AS due by many titles I reſigne
My ſelfe to thee, O God, firſt I was made
By thee, and for thee, and when I was decay'd
Thy blood bought that, the which before was thine,
I am thy ſonne, made with thy ſelfe to ſhine,
Thy ſervant, whoſe paines thou haſt ſtill repaid,
Thy ſheepe, thine Image, and till I betray'd
My ſelfe, a temple of thy Spirit divine;
Why doth the devill then uſurpe on mee?
Why doth he ſteale, nay raviſh that's thy right?
Except thou riſe and for thine owne worke fight,
Oh I ſhall ſoone deſpaire, when I doe ſee
That thou lov'ſt mankind well, yet wilt not chuſe me,
And Satan hates mee, yet is loth to loſe mee.

II.

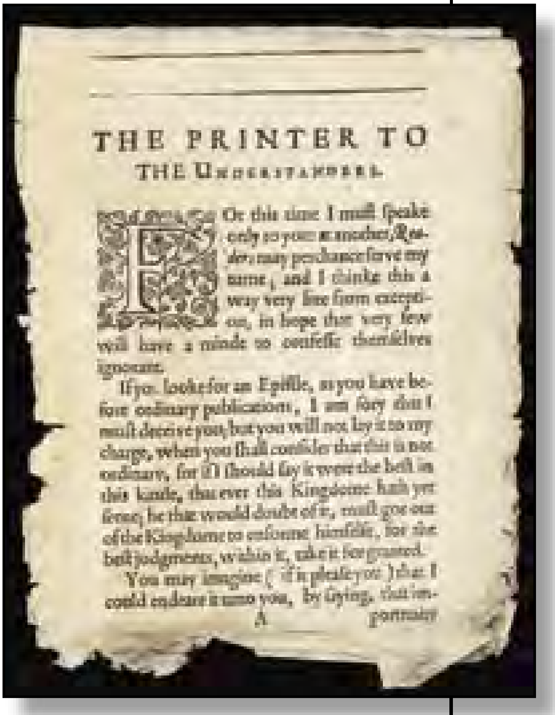
Poems, 1633

POEMS, *by J. D. with Elegies on the Authors Death.*
London: M. F[letcher] for John Marriot, 1633. 4to. vi + 406 pp.
I. A. Shapiro copy.

This copy is bound in contemporary calf with gilt-stamping on the front and back boards and "DONS POEMS" written in ink on the fore-edge. Contains ownership inscription by I. A. Shapiro on the front free endpaper with autograph and notes regarding the state (corrected or uncorrected) of some leaves.

"THE PRINTER TO THE UNDERSTANDERS."
From *Poems, by J. D. with Elegies on the Authors Death.*
London: M. F[letcher] for John Marriot, 1633. 4to. iv.

Unbound preface. Written by Miles Fletcher, this reader's preface was added to some copies of the first edition of Donne's collected poems and was likely an afterthought. Neither of Cushing Library's two copies of the 1633 *Poems* contains this rare preface.



THE PRINTER TO
THE UNDERSTANDERS.

FOr this time I muſt ſpeake only to you: at another, Readers may perchance ſerve my turne; and I thinke this a way very free from exception, in hope that very few will have a minde to confeſſe themſelves ignorant.

If you looke for an Epiſtle, as you have before ordinary publications, I am ſorry that I muſt deceive you; but you will not lay it to my charge, when you ſhall conſider that this is not ordinary, for if I ſhould ſay it were the beſt in this kinde, that ever this Kingdome hath yet ſeene; he that would doubt of it, muſt goe out of the Kingdome to enforme himſelfe, for the beſt judgements, within it, take it for granted.

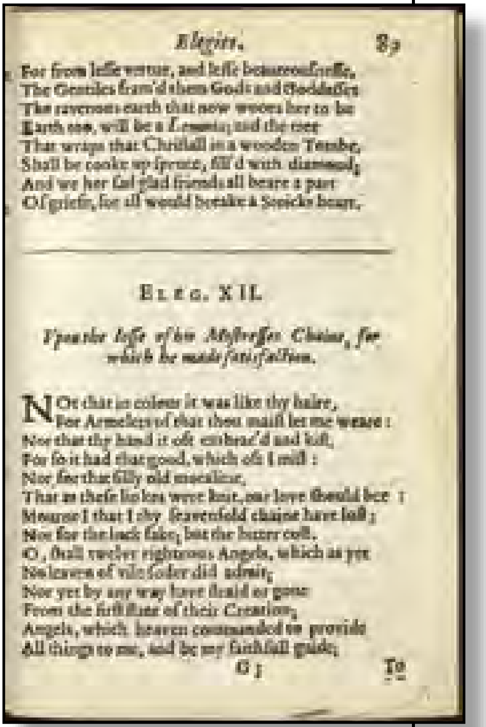
You may imagine (if it pleaſe you) that I could endeare it unto you, by ſaying, that im-

A portunity

Printer to the Understanders, 1633

POEMS, *by J. D. with Elegies on the Authors Death.* London: M. F[letcher] for John Marriot, 1635. 8vo. 419 pp. I. A. Shapiro copy.

Second edition of Donne's collected poems. This edition, in the smaller octavo format, adds 37 pieces to the text of the 1633 edition and omits Thomas Browne's elegy on Donne and Basse's *Epitaph upon Shakespeare*. Of the 37 additions, 17 are now accepted as Donne's. Together, the 1633 and 1635 editions essentially defined the Donne poetic canon until the 20th century. This copy contains ownership inscriptions by Arthur Kaye and I. A. Shapiro.



Elegies. 89

For from leſſe vertue, and leſſe beauteouſneſſe,
The Gentiles fram'd them Gods and Goddeſſes.
The ravenous earth that now wooes her to be
Earth too, will be a *Lemnia*; and the tree
That wraps that Chriſtall in a wooden Tombe,
Shall be tooke up ſpruce, fill'd with diamond;
And we her ſad glad friends all beare a part
Of griefe, for all would breake a Stoicks heart.

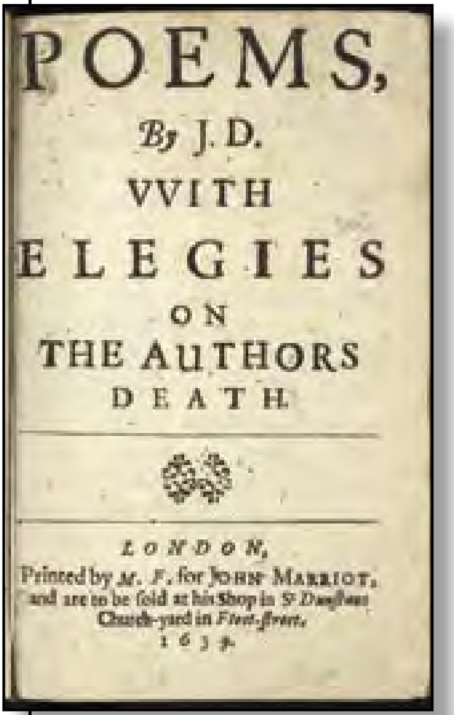
ELEG. XII.

Upon the loſſe of his Miſtreſſes Chaine, for which he made ſatisfaction.

NOt that in colour it was like thy haire,
For Armelets of that thou maiſt let me weare:
Nor that thy hand it oft embrac'd and kiſt,
For ſo it had that good, which oft I miſt:
Nor for that ſilly old moralitie,
That as theſe linkes were knit, our love ſhould bee:
Mourne I that I thy ſeavenfold chaine have loſt;
Nor for the luck ſake; but the bitter coſt.
O, ſhall twelve righteous Angels, which as yet
No leaven of vile ſoder did admit;
Nor yet by any way have ſtrayd or gone
From the firſt ſtate of their Creation;
Angels, which heaven commanded to provide
All things to me, and be my faithfull guide;

G 3 To

Poems, 1635



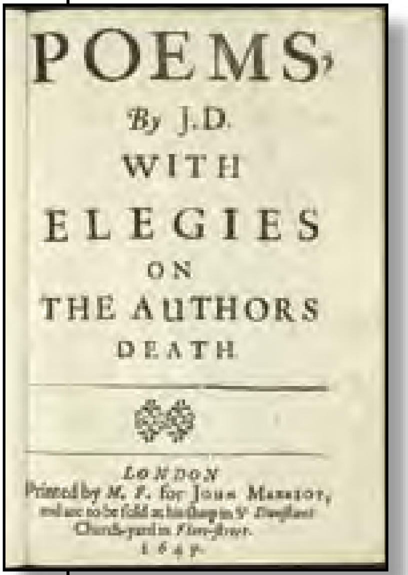
POEMS,
By J. D.
VVITH
ELEGIES
ON
THE AUTHORS
DEATH.

LONDON,
Printed by *M. F.* for JOHN MARRIOT,
and are to be ſold at his Shop in S^t *Dunſtans*
Church-yard in *Fleet-ſtreet.*
1639.

Poems, 1639

POEMS, *by J. D. with Elegies on the Authors Death.* London: M. F[letcher] for John Marriot, 1639. 8vo. 419 pp. I. A. Shapiro copy.

Third edition of Donne's collected poems. Essentially a page-for-page resetting of the 1635 edition, it contains a number of minor changes from the previous edition.



POEMS,
By J.D.
WITH
ELEGIES
ON
THE AUTHORS
DEATH.

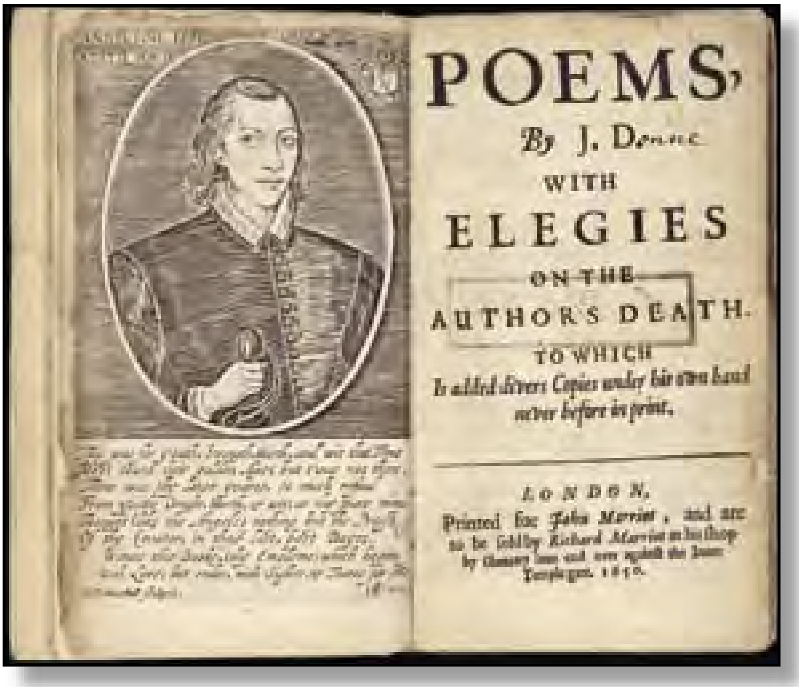
LONDON
Printed by M. F. for JOHN MARRIOT,
and are to be ſold at his ſhop in S^t Dunſtans
Church-yard in Fleet-ſtreet.
1649.

Poems, 1649

POEMS, ***by J. D. with Elegies on the Authors Death.***
London: M. F[letcher] for John Marriot, 1649. 8vo. vi + 368 pp.
I. A. Shapiro copy.

Fourth edition, first issue, of Donne's collected poems. This rare edition seems to have had a very low circulation and may not have actually been published. Most of the printed sheets were incorporated into the 1650 edition issued by the same publisher and edited by Donne's son, also John Donne. This edition also contains two additional poems, "Upon Mr. Thomas Coryat's Crudities" and "Sonnet. The Token."

POEMS, ***by J. D. with Elegies on the Authors Death. To Which Is added divers Copies under his own hand never before in print.***
London: For John Marriot, 1650. 8vo. vi + 392 pp.
I. A. Shapiro copy.

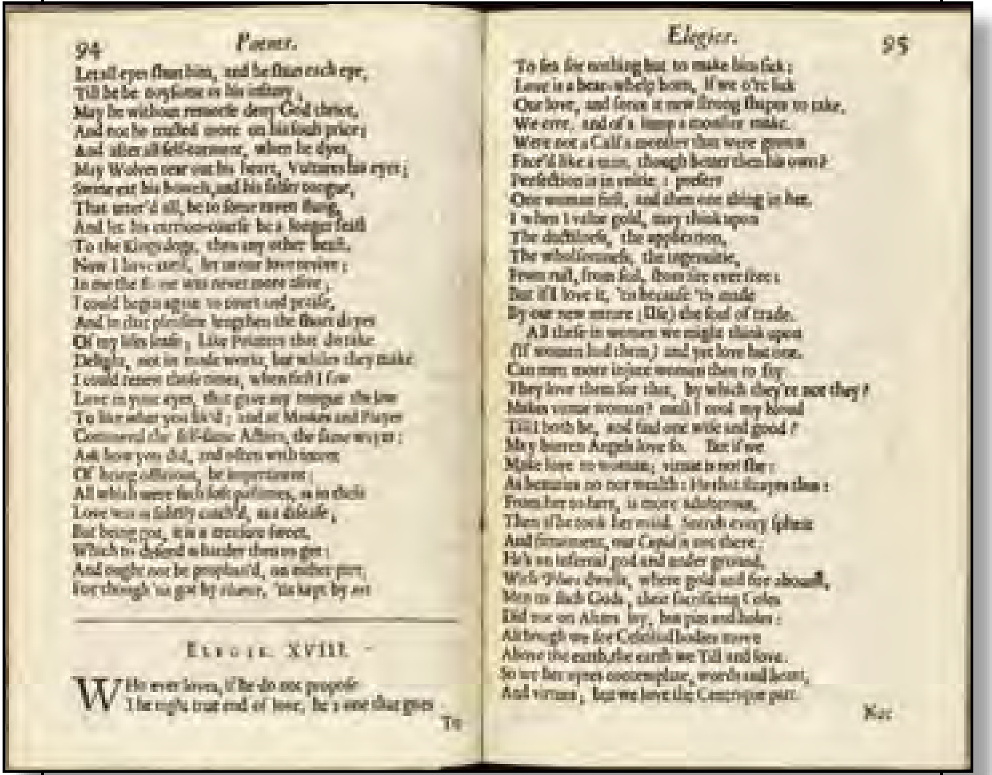


POEMS,
By J. Donne
WITH
ELEGIES
ON THE
AUTHORS DEATH.
TO WHICH
Is added divers Copies under his own hand
never before in print.

LONDON,
Printed for John Marriot, and are
to be ſold by Richard Marriot at his ſhop
by Chancery lane end over againſt the Inner
Temple gate. 1650.

Poems, 1650

Fourth edition, second issue of Donne's collected poems. This edition contains a frontispiece portrait of Donne at the age of 18 by William Marshall and adds 13 more poems and short prose pieces including a prefatory poem, "To John Donne," by Ben Jonson. The preface, "The Printer to the Understander," has been replaced with a dedication to Lord Craven by the younger John Donne.



94 *Poems.*

Let all eyes ſhun him, and he ſhun each eye,
Till he be noyſome as his infamy;
May he without remorſe deny God thrice,
And not be truſted more on his Souls price;
And after all ſelf-torment, when he dyes,
May Wolves tear out his heart, Vultures his eyes;
Swine eat his bowels, and his falſer tongue,
That utter'd all, be to ſome Raven flung,
And let his carrion-coarſe be a longer feaſt
To the Kings dogs, then any other beaſt.
Now I have curſt, let us our love revive;
In me the flame was never more alive;
I could begin again to court and praiſe,
And in that pleaſure lengthen the ſhort dayes
Of my lifes leaſe; like Painters that do take
Delight, not in made works, but whiles they make;
I could renew thoſe times, when firſt I ſaw
Love in your eyes, that gave my tongue the law
To like what you lik'd; and at Maskes and Playes
Commend the ſelf-ſame Actors, the ſame wayes;
Ask how you did, and often with intent
Of being officious, be impertinent;
All which were ſuch ſoft paſtimes, as in theſe
Love was as ſubtilly catch'd, as a diſeaſe;
But being got, it is a treaſure ſweet,
Which to defend is harder then to get:
And ought not be prophan'd, on either part,
For though 'tis got by chance, 'tis kept by art.

ELEGIE. XVIII.

WHo ever loves, if he do not propoſe
The right true end of love, he's one that goes
To

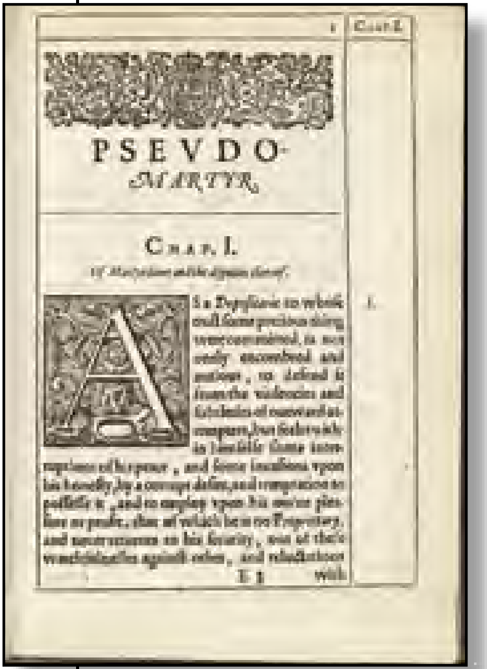
Elegies. 95

To ſea for nothing but to make him ſick;
Love is a bear-whelp born, if we o're lick
Our love, and force it new ſtrong ſhapes to take,
We erre, and of a lump a monſter make.
Were not a Calf a monſter that were grown
Fac'd like a man, though better then his own?
Perfection is in unitie: preferr
One woman firſt, and then one thing in her.
I when I value gold, may think upon
The ductilneſs, the application,
The wholſomneſs, the ingenuitie,
From ruſt, from ſoil, from fire ever free:
But if I love it, 'tis becauſe 'tis made
By our new nature (Uſe) the ſoul of trade.
All theſe in women we might think upon
(If women had them) and yet love but one.
Can men more injure women then to ſay
They love them for that, by which they're not they?
Makes virtue woman? muſt I cool my bloud
Till both be, and find one wiſe and good?
May barren Angels love ſo. But if we
Make love to woman; virtue is not ſhe:
As beauty is not nor wealth: He that ſtrayes thus
From her to hers, is more adulterous,
Then if he took her maid. Search every ſphear
And firmament, our Cupid is not there:
He's an infernal god and under ground,
With Pluto dwells, where gold and fire abound:
Men to ſuch Gods, their ſacrificing Coles
Did not on Altars lay, but pits and holes:
Although we ſee Celeſtial bodies move
Above the earth, the earth we Till and love:
So we her ayres contemplate, words and heart,
And virtues; but we love the Centrique part.
Nor

Poems, 1669

POEMS, ***&c. by John Donne, late Dean of St. Pauls. with Elegies on the Authors Death. To which is added Divers Copies under his own hand, never before printed.***
London: Printed by T. N. for Henry Herringman, 1669. 4to. vi + 414 pp.

Fifth edition of Donne's collected poems. The last 17th-century edition of Donne's poems, this edition adds five more poems, notably two important elegies – "Love's Progress" and "To his Mistresse going to bed"– which had not previously seen print.



1 Chap. I.

PSEVDO-MARTYR.

Chap. I.

Of Martyrdome and the dignitie thereof.

As a Depoſitarie to whom ſome precious thing were committed, is not onely encombred and anxious, to defend it from the violencies and ſubtleties of outward attempters, but feeles within himſelfe ſome interruptions of his peace, and ſome incurſions upon his honeſty, by a corrupt deſire, and temptation to poſſeſſe it, and to employ upon his owne pleaſure or profit, that of which he is no Proprietary, and never returnes to his ſecurity, out of theſe wraſtlings againſt others, and reluctations 1.

B 1 with

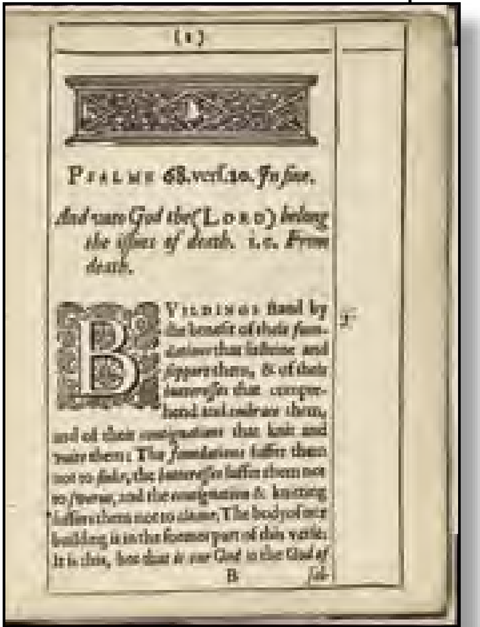
Pseudo-Martyr, 1610

PSEUDO-MARTYR. *Wherein Out of Certaine Propositions and Gradations, This Conclusion is evicted. That those which are of the Romane Religion in this Kingdome, may and ought to take the Oath of Allegeance.*
London: W. Stansby for Walter Burre, 1610. 4to. xxxviii + 392 pp.
I. A. Shapiro copy.

First edition of Donne's first published work. In this treatise, Donne defends the policy of James I regarding Roman Catholics, arguing that they should take the Oath of Allegiance and that those who chose death over compliance could not be genuine martyrs.

DEATH'S DUELL, *or, A Consolation to the Soule, against the dying Life, and living Death of the Body.*
London: Thomas Harper for Richard Redmer and Benjamin Fisher, 1632. 4to. ii + 47 pp.
I. A. Shapiro copy.

First edition, either first or second issue, lacking frontispiece portrait and title page. On the last Friday of Lent in 1631, Donne delivered this sermon on *Psalm* 68:20 at Whitehall before Charles I, King of England. Wracked with illness, this would be Donne's last sermon. He died five weeks later on March 31, 1631, and the sermon was first published the following year.



(1)

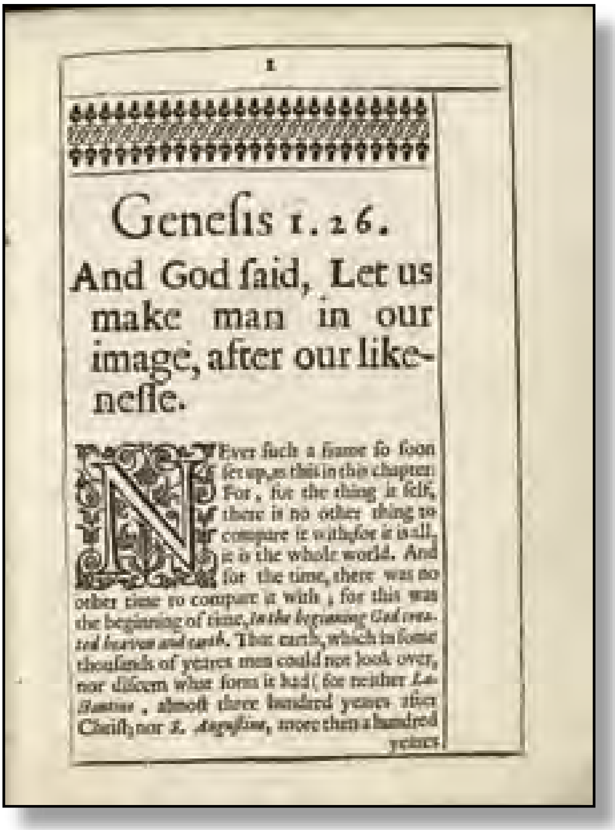
PSALME 68. verſ. 20. In fine.

And unto God the (LORD) *belong the iſſues of death. i.e. From death.*

BVILDINGS ſtand by the benefit of their foundations that ſuſteine and ſupport them, & of their butteresses that comprehend and embrace them, and of their contignations that knit and unite them: The *foundations* ſuffer them not to *ſinke*, the *butteresses* ſuffer them not to *ſwerve*, and the contignation & knitting ſuffers them not to cleave; The body of our building is in the former part of this verſe; It is this, hee that is our God is the God of 5.

B ſal-

Death's Duell, 1632



1

Geneſis 1.26.

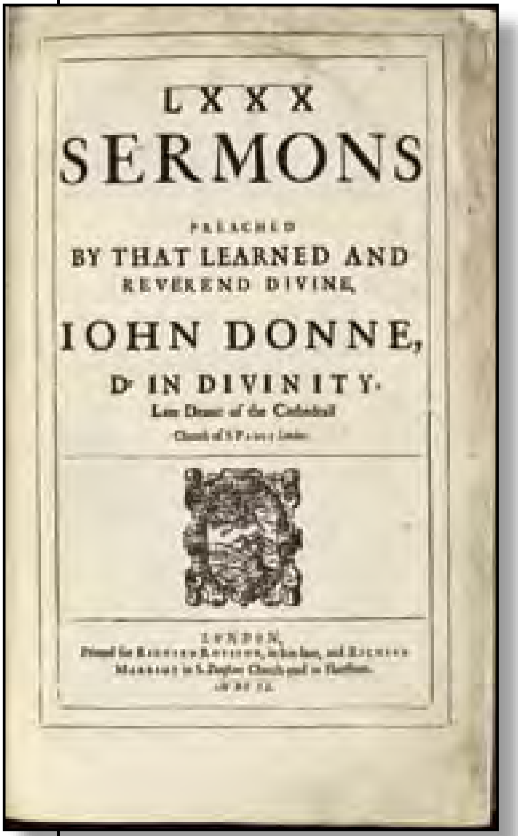
And God ſaid, Let us make man in our image, after our likeneſſe.

NEver ſuch a frame ſo ſoon ſet up, as this in this chapter: For, for the thing it ſelf, there is no other thing to compare it with, for it is all, it is the whole world. And for the time, there was no other time to compare it with; for this was the beginning of time, *In the beginning God created heaven and earth.* That earth, which in ſome thouſands of yeares men could not look over, nor diſcern what form it had (for neither *Lactantius*, almoſt three hundred yeares after Chriſt, nor *S. Auguſtine*, more then a hundred yeares

Six Sermons Upon Severall Occasions, 1634

SIX SERMONS UPON SEVERALL OCCASIONS, Preached before the King, and elsewhere: *By that late learned and reverend Divine John Donne, Doctour in divinitie, and Dean of S. Pauls, London.* Cambridge: the Printers to the University [Thomas Buck and Roger Daniel], 1634. 4to. iv + 92 leaves. I. A. Shapiro copy.

First edition. This collection contains six sermons, each with a separate title page featuring decorative borders. These six sermons include two sermons on *Genesis* 1:26 for Charles I, one on *Hosea* 2:19, one on *Matthew* 21:44, one on *John* 5:22, and one on *John* 8:15. A note by Shapiro asserts that the size of the margins and the presence of an ornamental leaf prior to the title page suggests that this copy is in its original calf binding. It was common for individual pages at the beginning or end of the book, such as the ornamental leaf, to become lost and for the entire volume to be trimmed during re-binding.



LXXX
SERMONS
PREACHED
BY THAT LEARNED AND
REVEREND DIVINE,
IOHN DONNE,
Dr IN DIVINITY.
Late Deane of the Cathedrall
Church of S. Pauls London.

LONDON,
Printed for Richard Royston, in Ivie lane, and Richard
Marriot in S. Dunstans Church-yard in Fleetstreet.
MDCXL.

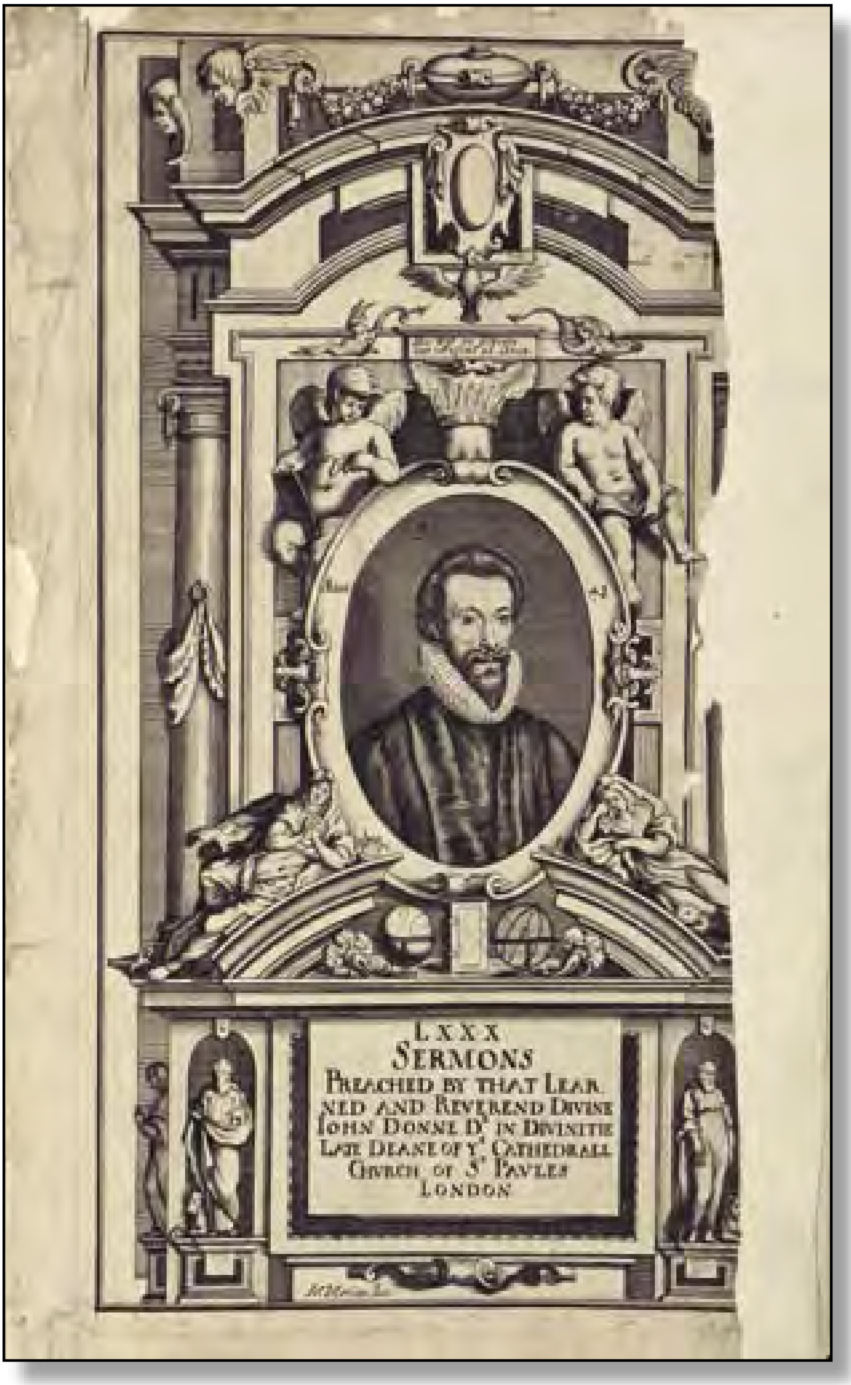
LXXX Sermons, 1640

LXXX SERMONS Preached by that Learned and Reverend Divine Iohn Donne, Dr. in Divinity, Late Deane of the Cathedrall Church of S. Pauls London.
London: For Richard Royston and Richard Marriot, 1640.
folio. xxxii + 826 pp.
I. A. Shapiro copy.

First edition. This large collection contains 80 sermons by Donne as well as "The Life and Death of the Dr. Donne, Late Deane of St. Paul's London" by Izaak Walton. The title page bears a device, illustrating Daniel at prayer, first used in 1597 by G. Simson. This copy is bound in contemporary speckled calf and contains bibliographic notes by Shapiro on the blank sheet facing the title page.

LXXX SERMONS Preached by that Learned and Reverend Divine Iohn Donne, Dr. in Divinity, Late Deane of the Cathedrall Church of S. Pauls London.
London: For Richard Royston and Richard Marriot, 1640.
folio. xxxii + 826 pp. I. A. Shapiro copy.

First edition. The frontispiece, title page, and dedication have all been crudely restored on their outer margins. The largest missing fragment is on the frontispiece but this does not affect the portrait or any of the text. This copy is bound in half-calf with marbled boards.

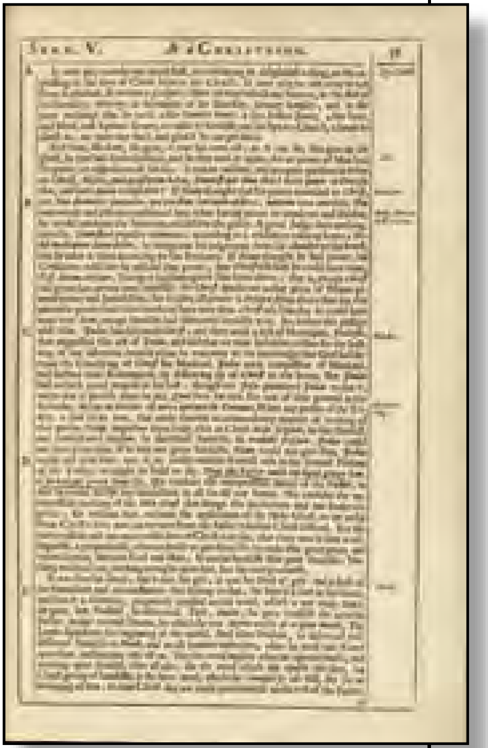


LXXX
SERMONS
PREACHED BY THAT LEAR
NED AND REVEREND DIVINE
IOHN DONNE D^{r} IN DIVINITIE
LATE DEANE OF Y^{e} CATHEDRALL
CHURCH OF S^{t} PAULES
LONDON

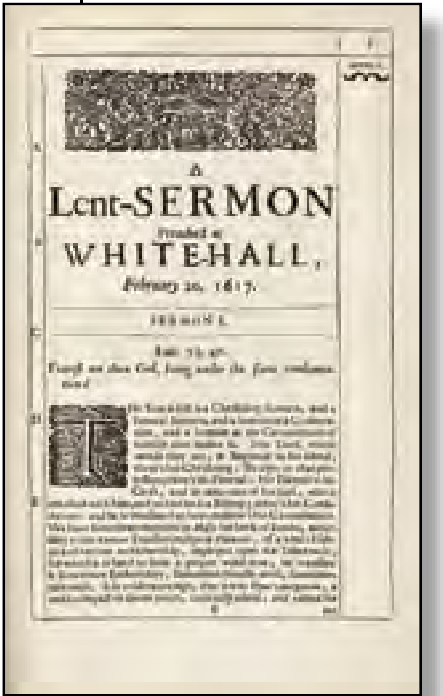
LXXX Sermons , 1640

FIFTY SERMONS Preached by that Learned and Reverend Divine John Donne, Dr. in Divinity, Late Deane of the Cathedrall Church of S. Pauls London.
The Second Volume. London: Ja. Flesher for M. F., J. Marriot and R. Royston, 1649. folio. viii + 474 pp. I. A. Shapiro copy.

First edition. It includes reprints of the six sermons from the 1634 collection. Bound in contemporary calf, this copy contains a few contemporary annotations and passages underlined in ink, providing an insight into contemporary interests and habits of reading.



Fifty Sermons, 1649



A
Lent-SERMON
Preached at
WHITEHALL,
February 20. 1617.

SERMON I.

XXVI Sermons, 1660

XXVI SERMONS Preached by that Learned and Reverend Divine John Donne, Doctor in Divinity, Late Dean of the Cathedral Church of St. PAULS, LONDON. The Third Volume. London: T. N. for James Magnes, 1660/61. folio. x + 411 pp. I. A. Shapiro copy.

First edition. This edition was carelessly edited and printed, including only 23 of the 26 sermons promised by the title (one sermon is omitted and two are repeated). According to the preface, by the younger John Donne, this edition was limited to 500 copies, which probably explains why this edition is the rarest of the three large collections of Donne's sermons. This copy is bound in 19th-century half-cloth and marbled boards.

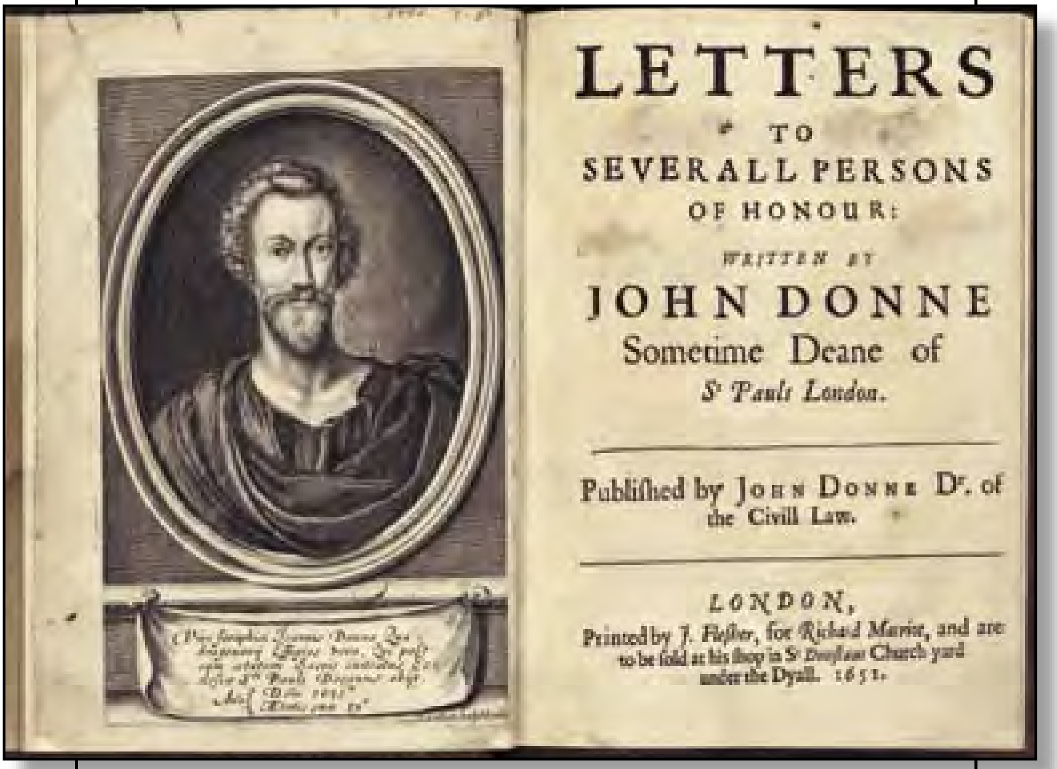
LETTERS TO SEVERALL PERSONS OF HONOUR.

Written by John Donne Sometime Deane of St. Pauls London.

London: J. Flesher for Richard Marriot, 1651. 4to. v + 318 pp.

I. A. Shapiro copy.

First edition, first issue, with an engraved frontispiece portrait of Donne at the age of 49 by Pieter Lombart. This printed collection contains 129 letters by Donne. This copy also includes handwritten marks and notes from early owners and readers. For example, one reader, perhaps objecting to the phrases, has blacked out "in good faith" in one letter and "by my troth" in two others.

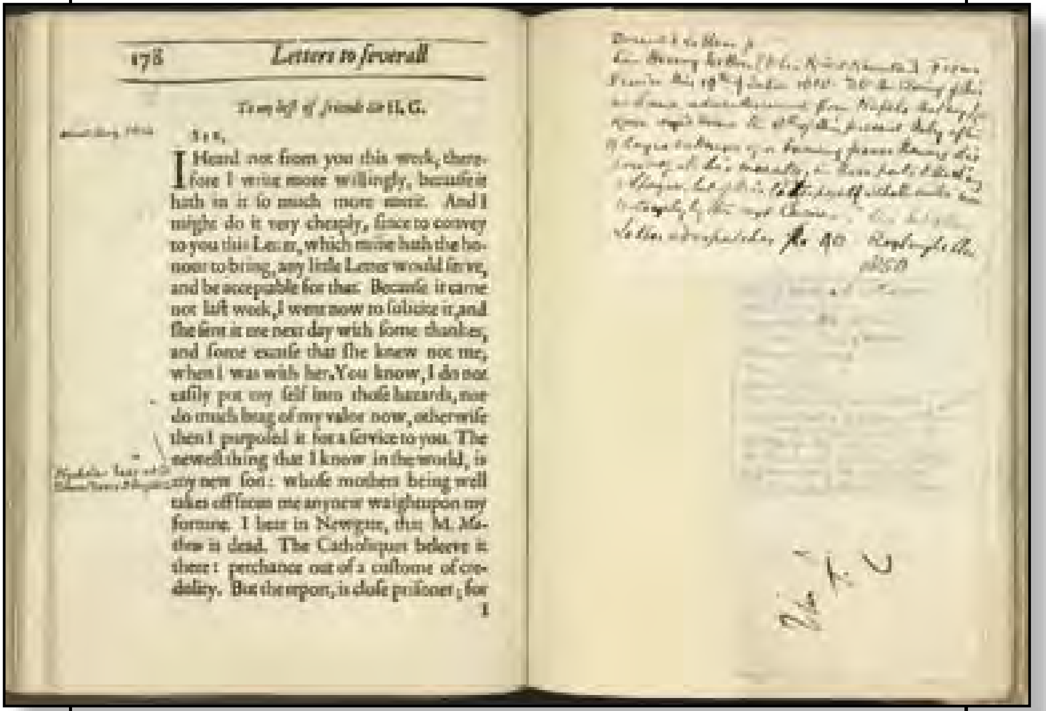


LETTERS
TO
SEVERALL PERSONS
OF HONOUR:
WRITTEN BY
JOHN DONNE
Sometime Deane of
St Pauls London.

Publiſhed by JOHN DONNE Dr. of
the Civill Law.

LONDON,
Printed by J. Fleſher, for Richard Marriot, and are
to be ſold at his ſhop in St Dunſtans Church yard
under the Dyall. 1651.

Letters to Severall Persons of Honour, 1651



178 *Letters to ſeverall*

To my beſt of friends Sir H. G.

Sir,

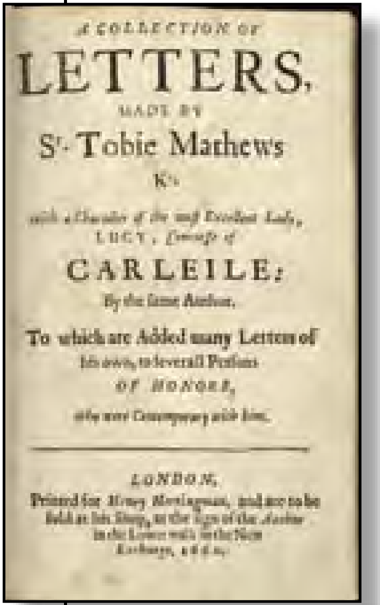
I Heard not from you this week, therefore I write more willingly, becauſe it hath in it ſo much more merit. And I might do it very cheaply, ſince to convey to you this Letter, which mine hath the honour to bring, any little Letter would ſerve, and be acceptable for that: Becauſe it came not laſt week, I went now to ſolicite it, and ſhe ſent it me next day with ſome thanks, and ſome excuſe that ſhe knew not me, when I was with her. You know, I do not eaſily put my ſelf into thoſe hazards, nor do much brag of my valor now, otherwiſe then I purpoſed it for a ſervice to you. The neweſt thing that I know in the world, is my new ſon: whoſe mothers being well takes off from me any new waight upon my fortune. I hear in Newgate, that M. Mathew is dead. The Catholiques believe it there: perchance out of a cuſtome of credulity. But the report, is cloſe priſoner; for

I

Letters to Severall Persons of Honour, 1654

LETTERS TO SEVERALL PERSONS OF HONOUR.
Written by John Donne Sometime Deane of St. Pauls London.
London: Printed by J. Flesher, and are to be sold by John Sweeting, 1654. 4to. vi + 318 pp.
I. A. Shapiro/T. R. O'Flahertie copy.

First edition, second issue. This copy also contains the signature of Henry Goodere, a close friend of Donne's, pasted on the front free endpaper. In the late 19th century, Reverend T. R. O'Flahertie, who collected both printed editions and manuscripts of Donne's works, owned this copy. When it was rebound in 1900, numerous notes by O'Flahertie, some on envelopes, were interleaved into this volume. The combination of O'Flahertie's and Shapiro's copious notes make this text a treasure-trove for Donne scholars and students interested in the poet's *Letters*.



A COLLECTION OF
LETTERS,
MADE BY
Sr. Tobie Mathews
Kt.
with a Character of the most Excellent Lady,
LUCY, Countesse of
CARLEILE:
By the same Author.
To which are Added many Letters of
his own, to severall Persons
OF HONOUR,
who were Contemporary with him.

LONDON,
Printed for Henry Herringman, and are to be
sold at his Shop, at the Sign of the Anchor
in the Lower walk in the New
Exchange, 1660.

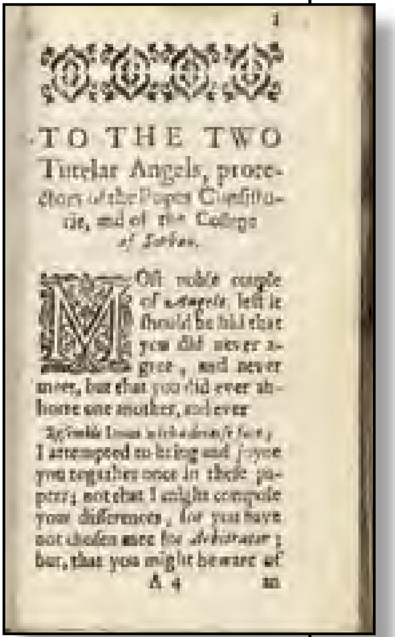
A Collection of Letters, 1660

Matthew, Sir Tobie. A COLLECTION OF LETTERS, Made by Sr. Tobie Mathews Kt. with a Character of the most Excellent Lady, Lucy, Countesse of Carleile: By the same Author. To which are Added many Letters of his own, to severall Persons of Honour, who were Contemporary with him. London: For Henry Herringman, 1660. 8vo. xviii + 356 pp. I. A. Shapiro copy.

First edition. Edited by the younger John Donne, this collection was intended to demonstrate the art of letter-writing and included letters to and from significant figures. In addition to six not included in the 1651 *Letters,* it contains 38 letters to and from Donne, notably a letter to Donne from the poet and playwright, Ben Jonson.

IGNATIUS HIS CONCLAVE*: or, His Inthronisation in a late Election in Hell: Wherein many things are mingled by way of Satyr. Concerning the Disposition of Jesuites, the Creation of a new Hell, the establishing of a Church in the Moone. There is also added an Apology for Jesuites. All dedicated to the two adversary Angels, which are protectors of the Papall Consistory and of the Colledge of Sorbon. By John Donne, Doctor of Divinitie, and late Deane of Saint Pauls.* London: For John Marriott, 1635. 12mo. vi + 138 pp. I. A. Shapiro copy.

Rare third edition, variant title page with date altered from 1634 to 1635. In *Ignatius his Conclave,* Donne attacks the Jesuits and the doctrines of Cardinal Bellarmine. Written in 1610, this work was first published in 1611 in two Latin editions. That same year it was translated into English, probably by Donne himself, but the English version lacks the "edge" of the Latin edition.



1

TO THE TWO
Tutelar Angels, prote-
ctors of the Popes Consistu-
rie, and of the College
of *Sorbon*.

MOst noble couple of *Angels*, lest it should be said that you did never agree, and never meet, but that you did ever abhorre one another, and ever

[illegible];

I attempted to bring and joyne you together once in these papers; not that I might compose your differences, for you have not chosen mee for *Arbitrator*; but, that you might be ware of

A 4 in

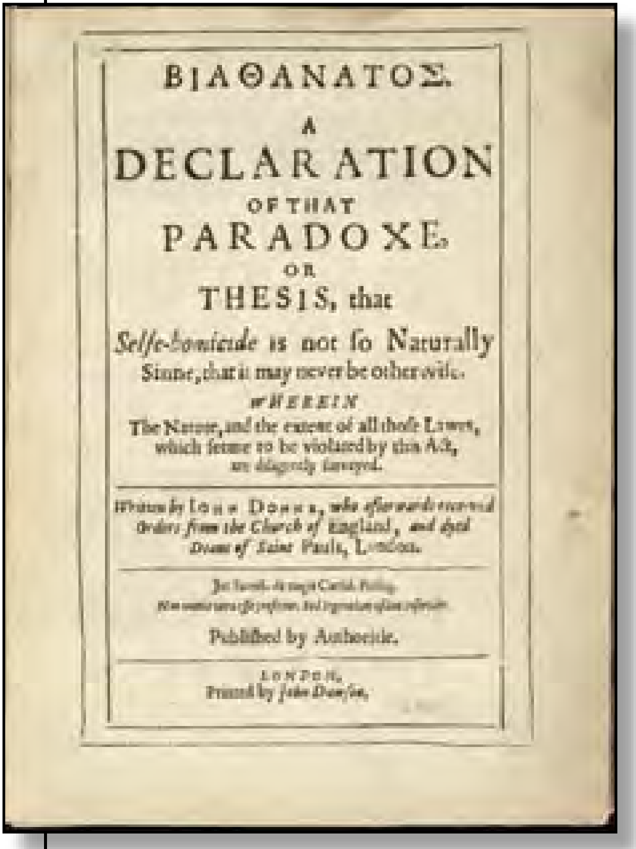
Ignatius his Conclave, 1635

IGNATIUS HIS CONCLAVE… Reproduced in facsimile from the edition of 1611. Introduction by Charles M. Coffin. New York: The Facsimile Text Society, Columbia University Press, 1941. 12mo. xxxii + 143 pp.

Modern facsimile of the first English edition of *Ignatius his Conclave.*

ΒΙΑΘΑΝΑΤΟΣ. *A Declaration of that Paradoxe, or Thesis, that Self-homicide is not so Naturally Sinne, that it may never be otherwise. Wherein the Nature, and the extent of all those Lawes, which seeme to be violated by this Act, are diligently surveyed.*
London: Printed by John Dawson, [1644]. 4to. xix +218 pp.

First edition, first issue. Though not printed until long after his death, this is Donne's earliest prose work. Written about 1608, *Biathanatos* argues that suicide is not always an act of despair or mortal sin and can be the equivalent of a noble martyrdom. It is believed to have been printed in 1644, but not actually published until as late as 1647. The text contains a preface by Donne's son, who notes that his father "forbid [it] both the Press and the Fire."



ΒΙΑΘΑΝΑΤΟΣ.
A
DECLARATION
OF THAT
PARADOXE,
OR
THESIS, that
Selfe-homicide is not ſo Naturally
Sinne, that it may never be otherwiſe.
WHEREIN
The Nature, and the extent of all thoſe Lawes,
which ſeeme to be violated by this Act,
are diligently ſurveyed.

Written by IOHN DONNE, *who afterwards received
Orders from the Church of* England, *and dyed
Deane of Saint* Pauls, London.

[illegible]
[illegible]

Publiſhed by Authoritie.

LONDON,
Printed by *John Dawſon*.

ΒΙΑΘΝΑΤΟΣ, 1644

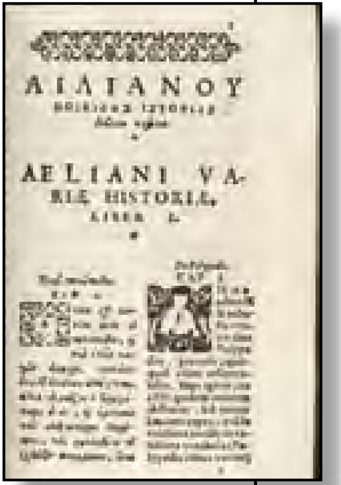
Donne's Reading:

A WORKING LIBRARY

AS IS EVIDENCED in both his poems and prose writings, Donne read broadly and deeply in many fields, and was especially learned in the areas of law and theology. In the twentieth century, scholars such as John Sparrow and Sir Geoffrey Keynes located in various British repositories many volumes that had once been in Donne's library; and in the 1973 edition of his *Bibliography of Dr. John Donne,* Keynes listed 213 of these copies, together with their present locations. Nearly 50 others have since come to light. While I. A. Shapiro did not own any of the actual copies that had once been Donne's, he did accumulate a scholar's working library of books Donne is known to have owned or read. The result is an important resource for researchers interested in examining the volumes he would have drawn on in producing his various texts. What follows is a list of the books collected by Shapiro and short notes where applicable.

Aelianus, Claudius, known as Aelian. ***Variae historiae libri XIIII. Cum latina interpretatione [of Justus Vulteius].*** **Geneva: J. de Tournes, 1613. 12mo. xviii + 460 pp.**

An important resource for anecdote, Aelian's *Varia Historia* was extremely popular in the 16th century. While this edition is printed in both Greek and Latin, Donne – like many contemporary readers – probably used primarily the Latin. He refers to Aelian in various places – specifically in *Biathanatos,* when he mentions the custom of killing and burying elderly parents.



ΑΙΛΙΑΝΟΥ

AELIANI VARIÆ HISTORIÆ, LIBER I.

Aelianus, 1613

Agrippa, Heinrich Cornelius. ***Opera.*** **Lyons: Fratres Beringi, [c. 1600.] 8vo. xxiv + 694 pp.**

Arcangelo [Pozzo] da Burgo Novo, OFM. ***Apologia pro defensione doctrinae Cabalae, contra Petrum Garziam…Mirandulam impugnantem… et conclusions cabalisticae…secundum opinionem… Mirandulae ex ipsius Hebraeorum sapientum fundamentis christianam religionem maxime declarantes.*** **Basel: S. Henripetri (1600). 8vo. xxxii + 568 pp.**

This copy contains a note by Shapiro on the front flyleaf which reads: "Referred to by Donne in *Essays in Divinity* and elsewhere."

INSTITVTIONVM MORALIVM

IN QVIBVS VNIVERSÆ QVÆSTIONES AD CONSCIENTIAM

IOANNE AZORIO LORCITANO

VENETIIS, MDCIII

Azor, 1690

Azor, Juan SJ. ***Institutionum moralium…Pars prima.*** **Venice: G. & N. Polo, 1603. 8vo. xii + 1690 pp.**

This influential religious text by Jesuit priest Juan Azor provides an interesting early 17th century window into moral issues. T. S. Healy describes Azor as one of Donne's "favorite moralists." This text is cited frequently in *Pseudomartyr* and *Biathanatos,* and Azor is cited in *Essays in Divinity, Sermons* and in the second part of *Ignatius His Conclave.*

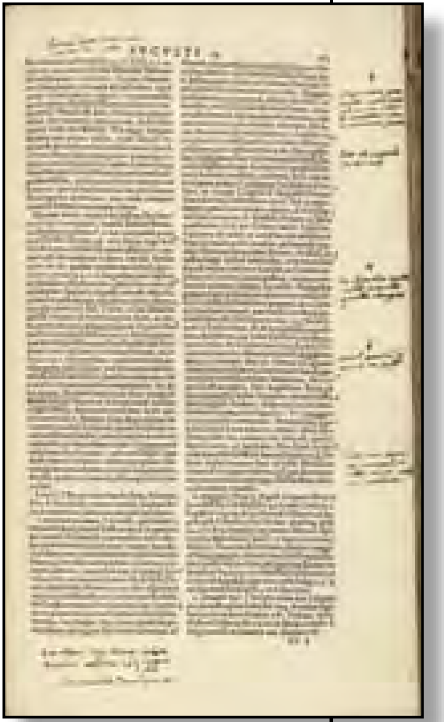
Azpilcueta, Martino. ***Enchiridion sive manuale confessariorum et poenitentium.*** **Venice: F. Ziletti & G. Ferrari, with the consent of the author, 1584. 4to. xvi + 1012 pp.**

Azpilcueta, Martino. *Manuale . . .cum allegationibus in margine repositis.* **Venice: heirs of D. de Farri, 1606. 4to. lxxxviii + 776 pp.**

Donne refers to Martino Azpilcueta's informal manual for confessors multiple times in *Pseudomartyr.*

Baronius, Caesar, Cardinal, ed. ***Martyrologium romanum... Accesserunt notations.*** **Paris: M. Sonnius, 1607. folio. xxii + 500 pp.**

This copy has been carefully read and heavily marked throughout. The book also bears an inscription indicating that the book was a gift from Samuel Walsall (1575-1626) to his father John Walsall on October 28, 1609. According to a xeroxed letter by Shapiro slipped into the book, Samuel probably wrote the majority of the annotations and his father wrote the rest.



Baronius, 1607

Baronius, 1610

Baronius, Caesar, Cardinal, ed. ***Sacrum martyrologium Romanum.*** **Cologne: J. Gymnich, 1610. 4to. xliv + 876 pp.**

The author, famous for *Annales Ecclesiasticae,* tried to suppress this early work after N. Fabius erroneously inserted Xynoris into the list of saints. Donne refers to this text, a dictionary of Catholic martyrs organized by calendar, several times in *Biathanatos.*

Brisson, Barnabé. *De formulis et sollemnibus populi romani verbis libri viii…Adiecti sunt rerum & verborum indices locupletissimi.* Frankfurt: J. Wechel & P. Fischer, 1592. 4to. ix + 750 pp.

A well-known moderate French jurist and author of a number of works on Roman law and Roman history, Brisson was hanged in 1591 by the "Seize," a group of insurgents who briefly captured Paris. Donne cites this text in *Ignatius His Conclave* and *Essays in Divinity*.

Bodin, Jean. *De republica libri sex… Editio sexta.* Frankfurt: H. Palthenius for the widow of Johann Rosa, 1622. 8vo. xvi + 1221 pp.

BOUND WITH:

Herpin, René. *Ad verius illustrandum J.B. de republica methodum… apologia.* Frankfurt: N. Hoffmann for J. Rosa, 1615. 8vo. 115 pp.

Donne refers to Bodin by name in *Biathanatos* where he writes "yea by a law of Venice, though Bodin say, it were better the town were sonke, than ever there should be any Example, or precedent therein."

Bosquier, Philippe. *Monomachia Iesu Christi, ex Luciferi, incruenta; seu concionum xi de tentationibus Christi in deserto, notae…editio ultima…auctior.* Cologne: J. Crithius, 1611. 8vo. xvi + 303 pp.

This collection of sermons by well-known preacher, Phillippe Bosquier, focuses on the temptation of Christ in the desert. Donne cited this collection frequently in *Pseudomartyr* and *Biathanatos*. In the former, Donne cites the sixth sermon in this collection: "One of your owne Authors related, that *Anastasius* a Monke, had a hundred Divels appointed to vexe and tempt him for four yeares, and after he

MONOMA-
CHIA IESV CHRISTI, ET LV-
CIFERI, INCRVENTA:
SEV
CONCIONVM XL. DE TEN-
TATIONIBVS CHRISTI IN DESERTO,
NOTAE.
Auctore F. P. PHILIPPO BOSQVIERO [illegible]
[illegible]
[illegible]
[illegible]
COLONIAE AGRIPPINAE,
Apud IOANNEM CRITHIVM, sub signo Galli
ANNO M. D. CXI.

Bosquier, 1611

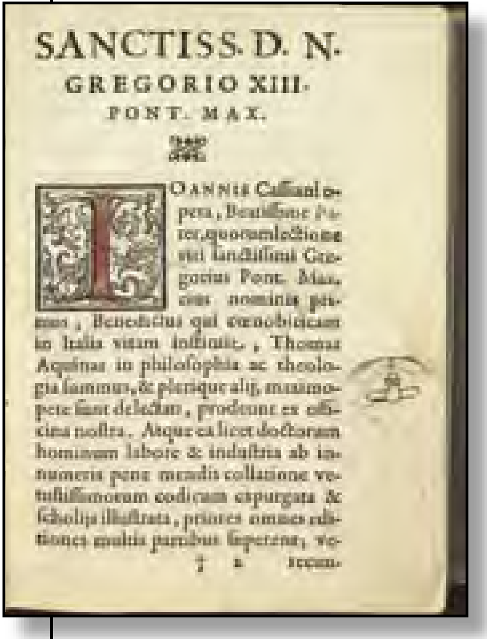
had overcome that trouble, and tamed them, he set them on work to build him a great Monastery, & to bring Aqueducts, and other conveniencies thereunto, for his temporal provision. . . .”

Cano, Melchior, Bishop of the Canaries. *Locorum theologicorum libri duodecim.* Cologne: officina Birckmannica f. A. Mylius, (Wurzburg: exc. H. Aquensis) 1585. 8vo. 921 pp.

One of the representatives at the Council of Trent, Cano viewed Jesuits, Calvinists, and Lutherans as precursors of the Antichrist. Donne cites Cano in *Pseudomartyr* : “To which opinion [that Purgatory seems to be ‘but the Mythologie of the Romane Church, and a morall application of pious and useful fables’] *Canus* expresses himselfe to have an inclination, when he sayes, ‘That men otherwise very grave, have gathered up rumours, and transmitted them to posterity…and that Noble Authors have been content to think, that that was the true law of History, to write those things which the common people thought to be true. . . .’”

Cassian, John. *De institutes renuntiantium libri XII. Collationes sanctorum patrum XXIV, [etc.]*. Rome: Typographia Vaticana, 1588. 8vo. xvi + 749 pp.

The anonymous editor of this volume indicates that Cassian was aided by Cardinal Antonio Caraffa and the Spanish scholar Pedro Chacon. This detailed glossary of Greek and other unfamiliar words and phrases includes “biathanatos.” John Donne refers to Cassian in both *Biathanatos* and *Pseudomartyr* and to Chacon in *Ignatius his Conclave,* and, according to a note by Shapiro, cites the 1606 Lyons edition, a reprint of this edition.



SANCTISS. D. N.
GREGORIO XIII.
PONT. MAX.

IOANNIS Caſſiani opera, Beatiſſime Pater, quorum lectione viri ſanctiſſimi Gregorius Pont. Max. cuius nominis primus, Benedictus qui cœnobiticam in Italia vitam inſtituit, Thomas Aquinas in philoſophia ac theologia ſummus, & pleriquealij, maximopere ſunt delectati, prodeunt ex officina noſtra. Atque ea licet doctorum hominum labore & induſtria ab innumeris pene mendis collatione vetuſtiſſimorum codicum expurgata & ſcholijs illuſtrata, priores omnes editiones multis partibus ſuperent, ve-

† 2 recum-

Cassian, 1588

Cepari, Virgilio SJ. ***Vita B. Aloysii Gonzagae…hac secunda editione accuratius in capita & paragraphos distincta.*** **Trans. by Johannes Horrion. Valenciennes: J. Vervliet, 1609. 8vo. xxvi + 452 pp.**

Inside the front cover, Shapiro writes, "This is one of the books frequently cited by Donne in *Pseudomartyr*. His ref[erence]s show that he read it very carefully to the end."

Cepari, Virgilio SJ. ***Vita del beato Luigi Gonzaga della Compagnia di Giesu.*** **Rome: L. Zannetii, 1606. 4to. xxiv + 344 pp.**

Donne owned a number of biographies of early Jesuits, including this popular text by Virgilio Cepari, which documents the life of the early Jesuit and saint, Luigi Gonzaga. He also refers to this text in *Pseudomartyr*: "And were there not some degrees of spiritual pride in Gonzaga, who is praised because *he had a paire of patched hose in Deliciis?*"

Clavius, Christopher SJ. ***In sphaeram Ioannis de Sacro Bosco commentarius… Accessit geoemetrica…de crepusculis tractatio.*** **S. Gervais [Geneva]: S. Crespin, 1608. 4to. viii + 598 pp.**

Widely circulated in Europe and beyond, this text, by the famous Jesuit mathematician and astronomer, is described by Shapiro as "the most important astronom[ical] treatise of its time." Donne refers to Clavius and this edition in particular in *Pseudomartyr*, "so that they

Clavius, 1608

[astronomers] must tell us, how much the Pope exceedes a Prince: which were a fit work for their *Jesuite Clavius,* who hath expressed in one summe, how many granes of Sand would fill all the place within the concave of the firmament. . . .”

Cutsem, Peter. ***Vivum speculum, in quo vera et apostolica Christi ecclesia cuivis introspicienti ad oculum clare apparet. Pontificiorum, Lutheranorum, et Calvinistarum trino calculo approbatum.*** **Cologne (S. Hemmerden for) B. Walther, 1610. 8vo. xvi + 254 pp.**

This collection of texts by Cutsem, a German who converted to Calvinism in 1608, covers a diverse range of religious topics. Donne refers to Cutsem’s *De desperata Calvini causa* (Mainz, 1609) in *Ignatius His Conclave.*

Daneau, Lambert. ***Politices christianae libri septem… Additae sunt peculiars aphorismi de optimo princippe, & eius officio ex C. Plinii Panaegyrica ad Traianum per eundem Lambertum Danaeum … Editio secunda.*** **[Geneva:] J. Vignon, 1606. 8vo. xvi + 573 pp.**

This collection of political pieces, assembled by Lambert Daneau, contains selections by various ancient writers with the significant exception of Machiavelli. In *Pseudomartyr,* John Donne refers to several of the pieces in this collection as well as other works by Daneau.

I. DRVSII
RESPONSIO
AD
SERARIVM
De Tribus Sectis Iudaeorum.

Accessit
IOSEPHI SCALIGERI
Elenchus Trihaeresii
NICOLAI SERARII
etc.

FRANEKERAE,
Ex Officina AEGIDII RADAEI,
Ordinum Frisiae Typographi.
cIɔ. Iɔ. cv.

Drusius, 1605

Drusius (Van den Driesche), Johannes. *Responsio ad Serarium de tribus sectis Judaeorum. Accessit Iosephi Scaligeri elenchus Tribaeresii Nicolaii Serarii, etc.* Franeker: Gilles Rade, 1605. 8vo. xxxii + 726 pp.

This religious tract by great early modern Hebrew scholar, Drusius, contains an important section on the divine name, Yahweh. Discussing the name of God in *Essays in Divinity,* Donne refers specifically to Drusius.

Dulcis, Catharinus. *Schola italica in qua praecepta bene loquendi… proponuntur… Editio altera.* Frankfurt: widow of M. Becker f. P. Musculus, 1614. 8vo. viii + 698 pp.

Espence, Claude d'. *De coelorum animatione… collectanea, cum resolutione catholica.* Paris: M. Sonnius, 1571. 8vo. xvi + 136 pp.

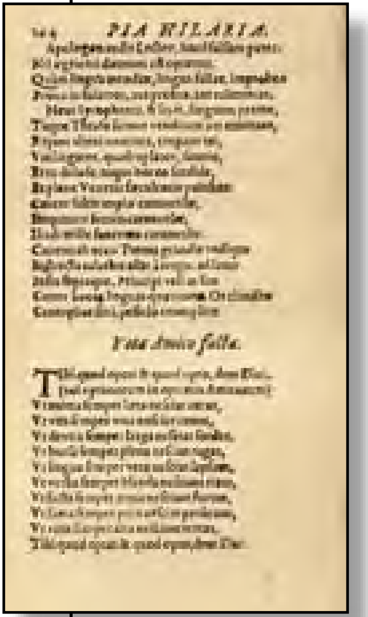
Espence's biblical commentaries were widely read during the 16th and 17th centuries. Donne refers to Espence's works in *Pseudomartyr.*

Espence, Claude d'. *De eucharistia, eiusque adoratione, libri quinque, etc.* Edited by G. Génébrard. Paris: P. L'Huillier, 1573. 8vo. xvi + 233pp.

Donne studied and referred to this work on the Eurachrist in a variety of places, including *Biathanatos*. Published posthumously, it was edited by the noted French Benedictine Hebraist and scholar Gilbert Génébrard.

Filesac, Jean. *De sacra episcoporum auctoritate. Seu ad tit. De off. Iud. Ord. lib. I Decretal. Commentarius.* Paris: B. Macé, 1606. 8vo. xxviii + 352 pp.

Donne cites this commentary on a small section of the Decretals (papal letters which formulate ecclesiastical law) three times in *Biathanatos.* Paris theologian Filesac's discussion of the authority of bishops was particularly controversial in the 16th and 17th centuries.



PIA HILARIA.

[illegible]

Vota Amica facta.

[illegible]

Gazaeus, 1625

Gazaeus, Angelinus. *Pia hilaria.* Pont-a Mousson: S. Cramoisy, 1625. 8vo. xii + 180 pp.

This collection of religious verse is based on stories from chronicles and contemporary sources. One comical poem recounts how a Belgian woman cured her husband of drunkenness by nearly treating him to a "premature burial." Donne translated the poem "Vota amica facta" into English.

Gazaeus, Angelinus. *Pia hilaria…Nova edition…correctior.* Antwerp: B. Moretus, 1629. 12mo. xxiv + 311 pp.

Gentili, Alberico. *De iure belli libri III.* Hanau: W. Antonius, 1598. 8vo. viii + 715 pp.

Considered one of the founders of modern international jurisprudence, Gentili moved from Italy to England in 1580 after being forced into exile by his father. This work was printed in Hanau but intended for the English market. In chapter 10 of *Pseudomartyr,* Donne cites Gentili, "as one notes him to say, *Fateor plane te mentitum, Gratiane*: and sometimes hee doth positively teach the just contrarie to Gratian, in matter of faith…" This copy contains a manuscript index at the end of the book and frequent underlining in ink throughout the text.

Gothofredus, Dionysius, ed. *Auctores linguae latinae in unum redacti corpus… Notae Dionysii Gothofredi j.c. ad Varronem, Festum, & Nonium. Variae lections in Fulgentium & Isidorum. Index generalis, etc.* [Geneva]: G. Lemaire, 1595. 4to. viiii + 1924 pp.

Gregoire, Pierre. *Syntaxis artis mirabilis in libros septem digestae, [etc].* Venice: G.D. de Imberti, 1588. 8vo. xvi + 582 pp.

Gretser, Jacobus, SJ. ***Basilikon doron sive commentarius exegeticus in Regis Jacobi praefationem monitoriam; in apologiam pro iuramento fidelitatis.*** **Ingolstadt: A. Sartorius, 1610. 4to. viii + 258 pp.**

BOUND WITH:

Gretser, Jacobus, SJ. ***Catalogus librorum quos Iacobus Gretserus… evulgavit, usque ad octobrem anni 1610.*** **Ibid, 1610. 4to. 16 pp.**

Gretser, Jacobus, SJ. ***Summula casuum conscientiae de sacramentis pro sectariis praedicandis…. Ex Luthero, Calvino, et Beza…collecta, [etc].*** **Ibid, 1611. 4to. 293 pp.**

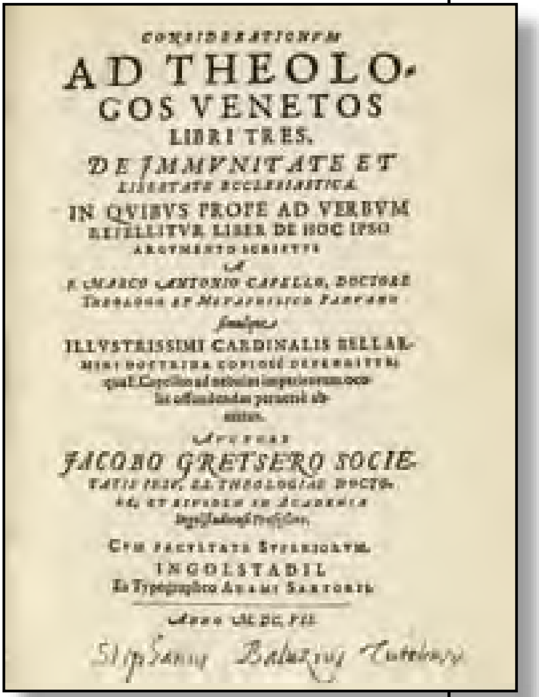
Gretser, Jacobus, SJ. ***Lixilivium pro abluenda male sano capite anonymi cuiusdam fabulatoris… qui caedem…Henrici IV in Jesuitas partim aperte, partim tacite confert. (Christliche Antwort Henrici des vierden Königs, [usw.].*** **Ibid, 1610. 4to. 35 pp.**

Gretser, Jacobus, SJ. ***Vindiciae Bellam inianae et muricum praedicanticorum a crimitationibus et inscitia Lutherani cuiusdam Magisletti Zephyrii, [etc].*** **Ibid, 1611. 4to. viii + 99 pp.**

Gretser, Jacobus, SJ. ***Commentariolus. De Imperatorum, Regum ac Principum Christianorum in Sedem Apostolicam Munificentia.*** **Ibid, 1610. 4to. 134 pp.**

These texts, bound together, represent six religious and historical works by Gretser, one of the great Jesuit scholars of the late 16th and early 17th centuries. Shapiro notes that Donne refers to the fourth work in this collection, *Lixilivium*, in *Ignatius his Conclave*.

Gretser, Jacobus, SJ. ***Considerationum ad theologos venetos libri tres, de immunitate et libertate ecclesiastica.*** **Ingolstadt: A. Sartorius, 1607. 4to. 384 pp.**



CONSIDERATIONVM
AD THEOLO-
GOS VENETOS
LIBRI TRES.
DE IMMVNITATE ET
LIBERTATE ECCLESIASTICA.
IN QVIBVS PROPE AD VERBVM
REFELLITVR LIBER DE HOC IPSO
ARGVMENTO SCRIPTVS
A
F. MARCO ANTONIO CAPELLO, DOCTORE
[illegible]
simulque
ILLVSTRISSIMI CARDINALIS BELLAR-
MINI DOCTRINA CONFOSSI DEFENDITVR,
quam F. Capellus ad nebulas imperitorum ocu-
lis offundendas peruerse ab-
utitur.
Auctore
IACOBO GRETSERO SOCIE-
TATIS IESV, SS. THEOLOGIAE DOCTO-
RE, ET EIVSDEM IN ACADEMIA
Ingolstadiensi Professore.
Cum facultate Superiorum.
INGOLSTADII
Ex Typographeo ADAMI SARTORII.
Anno M. DC. VII.
Stephani Baluzii Tutelensis

Gretser, 1607

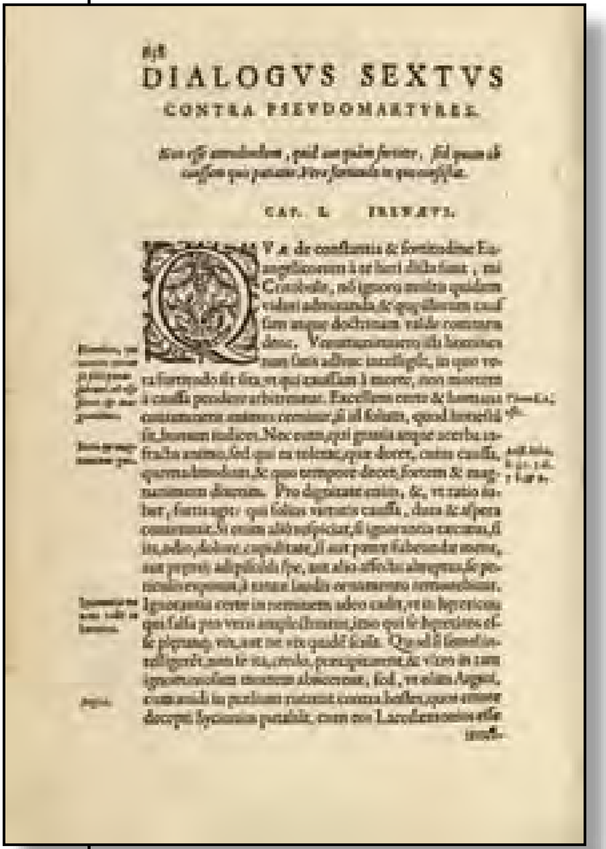
According to a note by Shapiro, *Considerationum* is "one of the books studied by Donne before he wrote *Pseudomartyr*, [and] is an important contribution to [the] controversy between Venetians & [the] Pope [in] 1605."

Gretser, Jacobus, SJ. *Libri quinque apologetici pro vita Ignatii Loiolae. Edita* a P. Ribadeneira. Ingolstadt: A. Sartoius, 1599. 8vo. xviii + 542 pp.

Gubert, A. *De sponsalibus matrimoniis et dotibus commentarius.* Marburg: P. Egenolff, 1597. 8vo. viii + 358 pp.

BOUND WITH:

Paleotti, Gabriele, Cardinal. *De nothis … tractatus singularis.* Frankfurt: N. Bassaeus, 1587. 8vo. xvi + 432 pp.



DIALOGVS SEXTVS
CONTRA PSEVDOMARTYRES.

[illegible]

CAP. I. [illegible]

[illegible]

Harpsfield, 1566

Harpsfield, Nicholas. *Dialogi sex contra summi pontificatus, monasticae vitae sanctorum sacrarum imagines oppugnatores, et pseudomartyres… ab Alano Copo Anglo editi.* Antwerp: C. Plantin, 1566. 4to. xxxii + 1050 pp.

This work in the form of a dialog between English and German orators defends aspects of Catholicism and serves as a rebuttal against Protestant histories of the 16th century. The sixth chapter is devoted to "False Martyrs" (i.e. "Pseudomartyrs"). As one might expect, Donne refers to this work in *Pseudomartyr.*

Heurn, Otto. ***Barbaricae philsophiae antiquitatum libri duo: I Chaldaicus II Indicus.*** **Leiden: ox off. Plantiniana, C. Raphelengius, 1600. 12mo. xxii + 362 pp.**

BOUND WITH:

Scaliger, Julius Caesar. ***Oratio pro M. Tullio Cicerone contra Ciceronianum Erasmi… Hymni sacri, et poemata sacra [not included].*** **Cologne: B. Walther, 1600. 12mo. viii + 141 pp.**

Heurn's work focuses on the history, customs, legends and beliefs of the ancient world, including the Babylonians, the Egyptians, and Zoroaster. Indirectly quoting it in *Biathanatos,* Donne also refers to this work in *Essays in Divinity* when he considers the oracles of Zoroaster.

Ignatius Loyola, St. ***Exercitia spiritualia.*** **Rome: Jesuit Press, 1615. 8vo. iv + 150 pp.**

This celebrated spiritual text, first printed in 1548, forms the basis of all Jesuit training. Donne frequently refers to both this text and the Jesuit *Constitutiones.* Though the focus of Donne's *Ignatius his Conclave* focuses on the followers of St. Ignatius, Ignatius is also mentioned in various places throughout the work. Donne's library also included lives of several Jesuit saints, notably Aloysius Gonzaga, Peter Canisius, Francis Xavier, and Ignatius himself.

Joannes ab Indagine. ***Introductiones apotelesmaticae in physiognomoniam, astrologiam naturalem complexiones hominum, naturam planetarum… Gulielmi Grataroli…opuscula… Pomponii Gaurici… tractatus de symmetriis… & physiognomia.*** **Strassburg: heirs of L. Zetzner, 1622. 8vo. 384 pp.**

Menghi, Girolamo, OFM. ***Flagellum daemonum, exorcismos… remediosque…ad malignos spiritus expellendos … complectens…Accessit par secunda, quae Fustis daemonum inscribitur.*** **Lyon: P. Landry, 1604. 8vo. xvi + 354 pp.**

pro Capuccinis [illegible]

FLAGELLVM

DAEMONVM.

EXORCISMOS TERRIBILES,

[illegible]

[illegible]

AVCTORE R. P. F. HIERONYMO MENGO

[illegible]

LVGDVNI

SVMPTIBVS PETRI LANDRY

M DCIIII

Menghi, 1604

This popular religious text by Girolamo Menghi, a 16th-century Franciscan friar, describes the formula of using the beginning of St. John's Gospel to banish the devil. Donne refers to this text in various works, including *Ignatius His Conclave* and *Essays in Divinity*, where he writes "But Saint John's *In the Beginning*, hath ever had strength against the Author of all error, the Divell himself, if we may believe the relations of the exorcists."

Menghi, Girolamo. *Flagellum daemonum. Exorcismos…ac doctrinam singularem in malignos spiritus expellendos…complectens… Acccessit…Fustis daemonum.* Venice: D. Maldura, 1608. 8vo. xvii + 471 pp.

Mercurius Gallobelgicus. *Mercurii Gallobelgici sive rerum a Gallia e & Belgiopotissimum: Hispoiania quoque,Italia, Anglia, Germania, Polonbia, vicinisque locis… gestarum…nuncii tomus primus.* Cologne: G. Kempen, 1596. 8vo. xxxii + 735 pp.

This is volume one of an eighteen volume set published between 1587 and 1630. This volume deals with events of 1587 to 1593, and, according to a note by Shapiro, is "frequently mentioned by Donne in his *Satires* & other early writings." "Mercurius Gallo-Belgicus" is also the title of one of Donne's epigrams.

Paleotti, Gabriele, Cardinal. ***De nothis spuriisque filiis liber.*** **Venice: G. Leoncino, 1572. 8vo. xxiv + 343 pp.**

Studied and cited by Donne, this book was an influential text for lawyers and was used in the 17th century by Thomas Salusbury in connection with Galileo's alleged illegitimacy.

Paleotti, Gabriele, Cardinal. ***De nothis… cum indice… Accessit… tractatus… de libera hominis nativitate seu de liberis naturalibus auctore Ponto Heutero Delfio.*** **The Hague: J. Verhoere, 1655. 8vo. xvi + 325 pp.**

Paleotti, Alfonso, Archbishop of Bologna. ***Iesu Chriusti crucifixi stigmata sacrae sindoni impressa… explicate. Mellifluis elucidationibus… accommodata… Auctore F. Daniele Mallonio… Adiectus est index quintuplex, [etc].*** **Douai: B. Bellère, 1607. 4to. xxxii + 479 pp.**

First published in Bologna in 1598, this volume contains both the Latin version of Paleotti's work by Laurentis de Arrighis and

De Cruce & Titulo.
Cap. VI.
SERPENS AENEVS CRVCEM PERFICIT.

Paleotti, 1655

the commentary of Daniel Mallonius. Donne refers to this work in both *Biathanatos* and *Pseudomartyr*. In the latter, he refers to Mallonius's commentary when he writes, "So also must the Scriptures afford prophesies for every ragge and inch of the Sindon, which wrapped our Saviour in the Sepulchre. . . ." This edition also contains 14 full-page engravings.

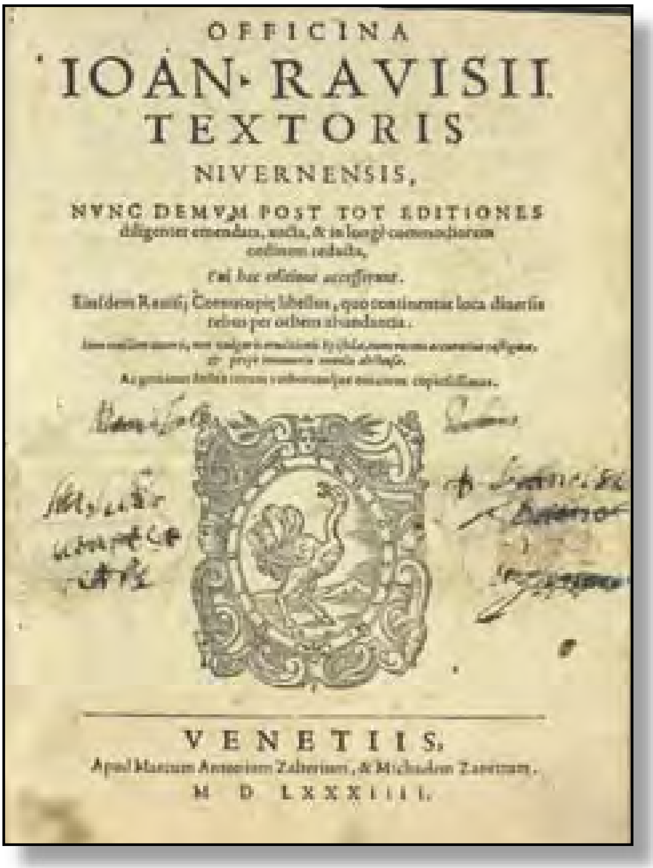
Pithou, Pierre, editor. *Epigrammata et poemata veteran.* Paris: N. Gilly, 1590. 12mo. xii + 683 pp.

Plinius Caecilius Secundus, Caius, known as Pliny the younger. *Epist. Lib. IX. Eiusdem & Traiani imp. epist. Amoebaeae. Eiusdem Pl. et Pacatii Mamertini, Panegyrici. Item, Claudiani Panegyrici. Adiunctae sunt I. Casauboni notae [etc].* [Geneva: Estienne], 1599. 12mo. 448 pp.

Small editions of classical writers like this edition of Pliny's *Letters* were very popular in the 16th century, primarily for their portability. In *Biathanatos,* Donne cites the Panegyricus of the emperor Trajan twice. This copy contains a late 19th-century folded manuscript index pasted inside the rear cover.

Prateolus (Du Préau), Gabriel. *De vitis, sectis, et dogmatibus omnium haereticorum. . . elenchus alphabeticus.* Cologne: C. Calenius (Typis G. Kempensis), 1583 (1581). 4to. xcii + 519 pp.

Donne refers to this religious compilation by Gabriel Prateolus in both *Pseudomartyr* and *Essays in Divinity*.



OFFICINA
IOAN. RAVISII
TEXTORIS
NIVERNENSIS,
NVNC DEMVM POST TOT EDITIONES
diligenter emendata, aucta, & in longè commodiorem
ordinem redacta,
cui hac editione accesserunt.
Eiusdem Rauisij Cornucopiæ libellus, quo continentur loca diuersis
rebus per orbem abundantia.
[illegible]
[illegible]
[illegible]

VENETIIS,
Apud Marcum Antonium Zalterium, & Michaelem Zanettum.
M D LXXXIIII.

Ravisius, 1584

Ravisius Textor, Joannes. ***Officina.*** **Venice: M.A. Zaltieri & M. Zanetti, 1584. 4to. xvi + 338 pp.**

Existing in a number of editions printed throughout Europe, this popular collection contains short stories categorized under a number of headings, including "On the Gods," "On the World," "On Time," "On Man," "On the Magistracy," and "On Virtues and Vices." It also includes a group of stories about death, specifically a section devoted to suicides. Though he may have used the 1590 Lyons edition, Donne clearly drew on these stories for many of his examples in *Pseudomartyr* and *Biathanatos.*

Reboul, Guillaume, sieur de. *Les Salmonees… Le premier contre les ministres de Nismes. Le second contre les ministres du Languedoc.* Arras: G. de la Riviere, 1600. 12mo. 503 pp.

Donne refers to this religious text in the beginning of *Ignatius His Conclave* when he writes, "I remembered him also how familiar a fashion this was amongst the papists themselves; and how much Rebullus that Run-away, had done in this kind, as well in those books, which he calls Salmonees"

Ribadeneira, Pedro, SJ. *Illustrium scriptorium religionis societatis Jesu catalogus. Hac seconda editione auctior.* Lyons: J. Pillehotte, 1609. 8vo. 312 pp.

This catalogue of Jesuit writers also includes lists of martyred Jesuits and Jesuit provinces and colleges as well as an index of surnames, nationalities, and subjects. Donne refers to this work several times in *Pseudomartyr.*

Rinaldi, Gratiano. *Lia, e Rachele, overo Marta, e Madalena, cioè La vita humana divisa in attiva, e contemplative.* Rome: A. Bernabo, 1673. 12mo. xxiv + 600 pp.

Sa, Manoel de, SJ. *Notationes in totam scripturam.* Lyons: J. Cardon, 1609. 4vo. viii + 534 pp.

Donne uses this brief commentary on individual words and expressions in the Bible in *Pseudomartyr, Essays in Divinity,* and *Biathanatos.* In the latter, Donne writes "and Emanuel Sa, who in his notes is more curious and superstitious in restoring all the Hebruismes, and often tymes their interpretations, then perchance that Churche would desire at his hands, offers at no other sense, then the Words present."

Sa. Manoel de, 1599. 1600, 1607, 1609

Sa, Manoel de, SJ. *Aphorismi confessariorum e doctorum sententiis collecti.* Antwerp: I. Trognaesius, 1599. 8vo. 312 pp.

Sa, Manoel de, SJ. *Aphorismi … Opusculum theologis … utile ac necessarium.* Cologne: P. Amerfort, 1600. 8vo. 384 pp.

Sa, Manoel de SJ. *Aphorismi confessariorum.* Rome: B. Zaneti for V. Pelagalli, 1607. 8vo. 502 pp.

Sa, Manoel de SJ. *Aphorismorum… Editio nova, secundum correctius Romanum exemplar edita.* Cologne: J. Crithius, 1609. 8vo. iv + 395 pp.

Originally published in Italy, these collections of quotations and definitions of religious and legal matters are cited several times in both *Pseudomartyr* and *Biathanatos.*

Serarius, Nicolaus, SJ. ***Rabbini, et Herodes, seu de tota Rabinnorum gente… maxime de Herodis tyranni natalibus…& regno libri tres: adversus Jos. Scaligeri Eusebianas annotations, & Jo. Drusii responsionem.*** **Mainz: B. Lippius, 1607. 8vo. 292 pp.**

BOUND WITH:

Bosendorff, Hermann. ***Credo Calvinisequarum… in gratiam Calvinistarum tertio disputatum. In quo Calvini de primariis fidei christianae capitibus… dogmata execranda proponuntur.*** **Mainz: L. Rassfeldt, 1607. 8vo. 19 leaves.**

Considered to be one of the best commentators of his period, Serarius was a proficient linguist and the author of various works of Hebrew scholarship. Donne frequently cites Serarius's work, specifically when referring to Herod's daughter in *Biathanatos*.

Sigonio, Carlo. ***De republica Hebraeorum libri VII.*** **Frankfurt: heirs of A. Wechel, 1583. 8vo. 393 pp.**

Addressed to Pope Gregory XIII, this work focuses on Jews and Jewish religion, calendar, rites, rabbis, and governance, but cites Christian rather than Jewish sources. Donne studied and cited this text frequently, and he refers to it specifically in *Biathanatos* in a section on the burial rites of Jews. An early reader has covered the front endpaper in writing.

Simancas, Jacobus (Diego), [successively Bishop of Ciudad Rodrigo, Badajoz and Zamora.] ***Theoricae et praxis haereseos sive enchiridion iudicum violatae religionis.*** **Venice: G. Ziletti, 1573. 8vo. xlviii + 310 pp.**

This text provides a useful summary of the theological and practical position of heretics: how they were tried, tortured, and buried. Donne refers to this text and other works by Simancas multiple times and, specifically in *Biathanatos*, when he defines heresy as "any thing which is against Catholique faith, that is Scriptures rightly understood; or the traditions and definitions of the Church."

Soto, Domingo de O.P. ***De iustitia et iure libri decem… accessit liber octavus, de iuramentis, & adiuratione, etc.*** **Venice: P. M. Bertano, 1608. 4to. lxii + 1006 pp.**

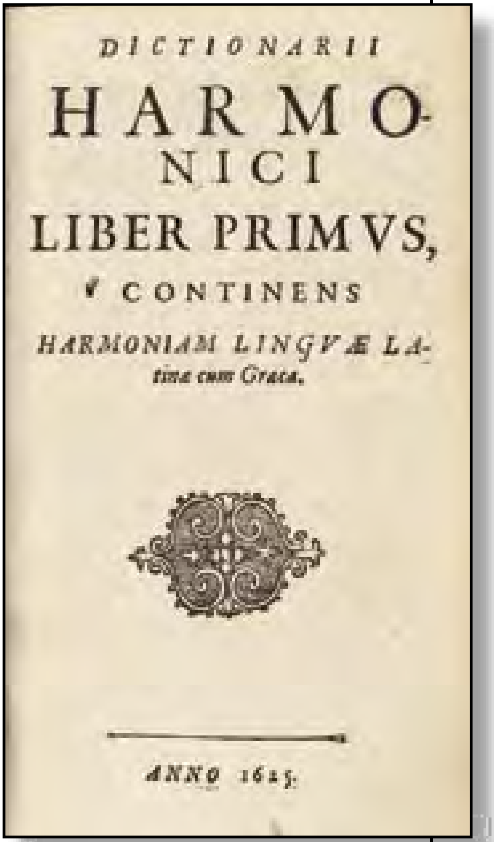
Distinguished for his important legal, theological, and philosophical commentaries, Soto is referred to by Donne in both *Pseudomartyr* and *Biathanatos*. Specifically, in the latter he cites Vitoria, Soto, and Valentia (all of whom follow Aquinas) when asserting that suicide is acceptable in situations where it is the only available means of avoiding idolatry.

Surius, Laurentius. ***Commentarius brevis rerum in orbe gestarum 1500-1574.*** **Cologne: A. Quentell, 1602. 8vo. xvi + 838 pp.**

This historical text chronicles international events by year. Donne satirizes Surius's "history" in *Satire IV* and, in other texts, refers to Surius' other works – most notably his work on saints' lives, *De probates sanctorum historiis.*

Turnemann, Matthaeus.
Dictionarium harmonicum.
Frankfurt: G. Bezerus, 1625. 8vo. viii + 647 pp.

This extremely uncommon work was used to help the reader refine his or her language skills in Greek, Latin, French, and Italian. According to Shapiro's notes in the book, "Matthew Turnemain was a Protestant theologian who had met Donne and corresponded with him."



DICTIONARII
HARMONICI
LIBER PRIMVS,
CONTINENS
HARMONIAM LINGVÆ LAtinæ cum Græca.

ANNO 1625.

Turnemann, 1625

Vazquez Menchaca, Fernando. *Controversiarum illustrium aliorumque usu frequentium libri tres.* Frankfurt: G. Corvinus for S. Feyerabend, 1572. folio. ix + 582 pp.

Vasquez incorporates elements from history, the Latin poets, and even part of *Orlando furioso* into this primarily legal work, which Donne cited frequently. Rebound in 1968, this copy preserves the vellum manuscript pieces which were used as protection in the original binding.

Vitoria, Francisco de OP. *Relectiones theologicae… Opus novissime iuxta ingolstadiensem editionem… repurgatum.* Lyons: P. Landry, 1586. 8vo. xvi + 531 pp.

This legal text was particularly influential for its advocacy of the idea of *ius gentium,* the commonly held belief that "infidelity" (i.e. lack of the Christian faith) and sin deprived a civil system of legitimacy and destroyed any true rights of ownership. Donne refers frequently to this work in his *Pseudomartyr,* and in one passage writes "even in the Roman Church a great Doctor of eminent reputation there, agrees (as he says) *Cum omnibus sapientibus,* that *this Regall Jurisdiction and Monarchie* (which word is so odious and detestable to *Baronius) proceedes from God, and by Divine and natural Law, and not from the State or altogether from man.*"

Venetian Interdict. *Pieces du memorable process esmeu l'an 1606 entre le Pape Paul V et les seigneurs de Venise. Touchant l'excommunication du Pape publiee contre iceux Venitiens.* S. Vincent: P. Marceau, 1607. 8vo. viii + 690 pp.

Venetian Interdict. *Raccolta degli scritti usciti fuori in istampa, e scritta a mano, nella causa del Paolo V. Co' Signori Venetiani.* Coira: P. Marcello, [Venice?], [1607]. 8vo. iv + 405 pp.

Misdated "1507" on the title page, this collection of texts by numerous Italian writers was actually published in 1607. Donne refers to this collection frequently in *Pseudomartyr.*

NOTES: