

This is a preprint of a paper presented at the international workshop and conference entitled "Modern Information Technologies and Written Heritage: From Ancient Manuscripts to Electronic Texts" [«Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам»], held on July 13-17, 2006, in Izhevsk, Udmurt Republic, Russia.

Creating a Digital Scholarly Edition of the British Library's *Cotton MS.*
Vitellius F. v.
Kevin S. Hawkins

Summary in Russian: Отдел научных публикаций Библиотеки Мичиганского университета (г. Анн-Арбор, США) планирует издание «Хроники» Мэйчина—рукописи, являющейся ценным источником сведений об истории Лондонского быта 16-го века. Электронное издание будет состоять из сканированных изображений рукописи, восстановленного текста рукописи в оригинальной и современной орфографии, а также из изображений списка «Хроники», датируемого 19-м веком. Доклад представит общую деятельность Отдела научных публикаций и историю данного проекта.

Abstract in English: The Scholarly Publishing Office of the University of Michigan University Library will soon publish an online scholarly edition entitled *A London Provisioner's Chronicle, 1550-1563, by Henry Machyn*. This edition will include a detailed introduction, images of the manuscript, a transcription including supplied text, a modernization of the text into contemporary English, and images of a 19th-century handwritten transcription of the original manuscript. After explaining the role of the Scholarly Publishing Office in promoting new models for scholarly communication, we will discuss the construction of this edition: the acquiring of images, creation and proofreading of the encoded text, customization of the interface, and digitization and cataloging of the 19th-century manuscript. Particular attention will be paid to the difficulties created by intellectual property rights.

Introduction

Henry Machyn's *Chronicle* of daily life in London from 1550-1563 is a unique resource for historians and linguists studying the 16th century. The *Chronicle* is especially notable because it covers the entire reign of Mary I, a tumultuous period in English history. No modern edition of the text has ever been published, and the last edition, published in 1848, was immediately recognized as deficient for scholarly analysis.

A scholarly edition entitled *A London Provisioner's Chronicle, 1550-1563, by Henry Machyn* will soon be published by the Scholarly Publishing Office (SPO),¹ a division of the University Library of the University of Michigan, Ann Arbor (U-M). This project is unlike other SPO publications in that it presents primary rather than secondary or tertiary (bibliographic) sources and utilizes markup and the delivery system architecture in innovative ways. This edition will include a detailed introduction, images of the manuscript, a transcription including supplied text from other sources, a modernization of the text, and images of a 19th-century handwritten transcription of the original manuscript. The project has been marked by complex relationships among the various stakeholders, which have evolved over the years and contributed to the project's long "gestation period". It is hoped that SPO's experience with this project will help all embarking on electronic publishing projects avoid similar problems in the future.

¹ See <<http://spo.umdl.umich.edu/>>.

The Scholarly Publishing Office

SPO's mission is to foster scholarly communication by using technology to provide a more effective and cost-efficient means of disseminating scholarly information than through commercial publishers and even university presses. SPO works with faculty both within and outside its host institution (as a university press would), with a special focus on advising and enabling publishing by the University Library itself and by university faculty members. In addition, SPO also acts as a hosting service for two subscription-based digital collections, the ACLS History E-Book Project² and LLMC Digital,³ a massive collection of digitized legal materials.

SPO supports the tenets of the open access movement. SPO primarily publishes online, making its journals, monographs, and other digital projects freely available whenever possible. Some content providers wish to restrict access in order to sell subscriptions to the resource; in this case, SPO can enable access by password, IP range, or both. SPO never requires the transfer of copyright from the author; instead, the content provider signs an agreement with SPO, stating that he or she has the right to publish the work and giving SPO an irrevocable license to publish the work online.

Nearly all SPO publications are hosted online using the DLXS suite,⁴ which includes XPAT, an XML-aware search engine, and middleware to provide a browser-accessible interface. Digital objects in DLXS can have page images with lightly encoded text or more fully encoded text,⁵ plus the software can also contain collections of fielded data and images not associated with text.

Creation of the content

Many people have contributed to the content and design of the Machyn project. Work began in 1994, when a group of graduate students attempted to transcribe microfilm of the Machyn manuscript. Unfortunately, the film was of poor quality, making deciphering Henry Machyn's handwriting even more difficult than it already was.

Richard W. Bailey, professor of English language and literature at U-M, visited the British Library in London to inspect the original manuscript. He organized the digitization of the manuscript by the British Library, and a student spent a summer correcting mistakes and completing the transcription begun from the microfilm. Chris Powell, of the Digital Library Production Service at U-M, helped in the early conception of the project before SPO was formed.

Marilyn Miller, a former editor of the Middle English Dictionary and an expert in medieval handwriting, spent many hours revising the transcription using her own tagging system, and William Ingram, a retired professor at U-M and specialist on 16th-century London, made valuable suggestions.

Colette Moore, then a graduate student and now assistant professor at the University of Washington (in Seattle), edited the text further and oversaw the work of an undergraduate student who enhanced the transcription with supplied text from other

² See <<http://www.historyebook.org/>>.

³ See <<http://www.llmc-digital.org/>>.

⁴ See <<http://www.dlxs.org/>>.

⁵ The "lightly encoded text" and "more fully encoded text" each conform to Level 1 and Level 4, respectively, of the TEI Text Encoding in Libraries Guidelines for Best Encoding Practices, available at <<http://www.diglib.org/standards/tei.htm>>.

sources. She and Richard Bailey also wrote an introductory essay to accompany the scholarly edition.

In short, the many contributors transcribed, folio by folio, Machyn's original spelling and noted text struck through, underlined, or written in superscript or subscript by Machyn.

Involvement of the University of Michigan Press and ACLS

Richard Bailey began discussions in 2000 with the University of Michigan Press (UMP) about publishing the material prepared so far. UMP felt that a modernization of the text (a version in contemporary orthography) would be more widely read than a transcription of the original, so Richard Bailey and Colette Moore prepared this version. In January 2003 Richard Bailey signed an agreement with UMP to publish "the work," which was taken to mean just the modernization. By signing this agreement, the editors transferred their copyright to UMP, and UMP agreed to edit and publish the work.

UMP later entered into an agreement with the ACLS History E-Book Project (which SPO coincidentally hosts), by which ACLS would distribute this book online to subscribers of the project while UMP retains the copyright. While the cost of paper was no longer a concern for UMP after this agreement, it still did not want to invest in editing the transcription and supplied text and therefore was not interested in publishing this portion unedited.

The role of SPO

UMP's decision not to publish the transcription with supplied text led Richard Bailey to approach the Scholarly Publishing Office about publishing this other material separate from the UMP/ACLS version. In addition, he wanted to deliver the images of the original manuscript online, plus images of a 19th-century handwritten transcription he had recently purchased. SPO accepted this project, acting in its role as advisor to faculty members with publishing projects and because the DLXS suite is suited to delivering both page images and more heavily encoded text. The source files were converted into an XML markup language devised by Brian Rosenblum, formerly of SPO. Kevin Hawkins took over this project, converting the material into Text Class⁶ XML, the markup needed for use with the DLXS suite, and to UTF-8 encoding, allowing rendering of the long "s" character (ſ) by contemporary web browsers.

ACLS, being aware of SPO's role in publishing a parallel edition, was excited at the opportunity to experiment with linking between the editions, matching the transcription and modernization of each entry and providing links to the page images in both editions. While the two editions would be stored separately, users could view them together for scholarly analysis, assuming that they have access to the ACLS History E-Book Project, which is available only to institutional and individual subscribers. Richard Bailey hoped to overcome this barrier to access by having SPO publish the entire text on its site, making it freely available. However, in order to allow ACLS/UMP to keep value in their content, the modernization would be interspersed with the transcription, thereby making the modernization difficult to read alone in the SPO version. (For each entry in the *Chronicle*, there would be a transcription followed by a modernization.) ACLS

⁶ See <<http://www.dlxs.org/docs/12a/class/text/index.html>>.

rejected this plan, and the agreement for parallel, non-overlapping editions was reaffirmed.

SPO finished its version before UMP/ACLS did theirs, and, at the urging of Richard Bailey, was prepared to release it before the UMP/ACLS version, including a note that the latter was forthcoming. At this point, ACLS decided that its involvement in the project was not benefiting anyone, so ACLS broke its contract with UMP, agreeing to compensate UMP for its expenses in editing the text. Pending payment from ACLS, UMP has agreed to relinquish its copyright to the introduction and modernization and give the text, with all of UMP's revisions, to SPO for publishing on its site along with the rest of the edition. SPO has decided not to release its edition until this text is received from UMP in order to release a complete publication all at one time.

Structure of the online scholarly edition

The online scholarly edition is not yet available to the public, as explained above. It currently contains all components except the detailed introduction and modernization, which are expected from UMP. The planned structure is discussed below.

The main entry point of the scholarly edition was designed by Eric White, a freelance graphic designer hired by SPO. It uses an image copied from Early English Books Online (EEBO)⁷ and also contains brief introductory text about the *collection*, a term referring to the way DLXS organizes digital objects into collections.

From here the user can enter either (a) the table of contents for the scholarly edition itself, containing the introduction, images of the manuscript, the transcription including supplied text, and the modernization of the text into contemporary English, or (b) access the images of a 19th-century handwritten transcription of the original manuscript.

The scholarly edition itself can be browsed or searched. Browsing shows a table of contents dynamically generated from the hierarchy of text divisions. The front matter includes the introduction, with illustrations,⁸ and information on how to use the online edition. Following this is the enhanced transcription, with supplied text in red and other presentational effects preserved as best possible: text is struck through, superscripted, and subscripted to imitate the appearance of the manuscript. The dates were regularized and supplied by the editors. Access to the manuscript images is through the enhanced transcription, by clicking on a link to a folio image at a "page break" in the transcription. The use of red for supplied text is a change we introduced into the DLXS middleware (it used to have no distinctive appearance), and SPO further customized the interface of this DLXS collection to have links to the folio images in the encoded text.

Searches can be restricted to the transcription only, the enhanced transcription (the transcription plus supplied text), or the modernization only. Due to a limitation of the XPAT search engine used in DLXS, searching matches only exact phrases in the XML,

⁷ Permission was obtained both from ProQuest, which scanned the original, and the British Library, which holds the original. According to US law, ProQuest does not hold a copyright to the work, but they do to the image that they made from it. The British Library holds no copyright; however, ProQuest required SPO to obtain the British Library's permission as well.

⁸ Permission was obtained from the holding institution for all of these images as well, though it's questionable whether this is necessary since copyright claims to reproductions of public domain works are tenuous. Often the holding institution instead charges for the right to reproduce the image, or charges a fee for them to reproduce it for you.

so intervening XML tags limit the recall in phrase searches.⁹ However, an asterisk can be used at the end of a stem to find variants with different endings (inflections), a distinct advantage when working with languages with inflectional affixes and non-standard spelling.

The 19th-century manuscript, as mentioned previously, was purchased by Richard Bailey. He donated it to U-M, and the University Library created bitonal images of it using the existing procedures for scanning print material. Since these images could not be OCRd and no one has transcribed this manuscript, the text is not searchable, and the images are simply provided for additional reference. A short explanatory text about this manuscript is found linked from the main entry page described above.

Thoughts for future projects

The experiences of SPO and the project editors yield some valuable lessons in intellectual property for any digital scholarly edition. We are reminded that signing an academic book contract with a publisher usually means the authors or editors must turn over their copyright, putting significant restrictions on their freedom to reuse their scholarship elsewhere. As for the failure of the plan of cooperation among SPO, ACLS/UMP, and the editors, it is difficult to determine a single cause of the breakdown in communication. SPO and the editors attempted to make their plans clear to ACLS/UMP on many occasions, but with all parties lacking a common language for discussing the structure of the edition and lacking a common business model, more thorough discussions were needed at an earlier stage to prevent so much lost time. All stakeholders in such a complicated project need to understand which components will be available to whom and agree on the degree of autonomy allowed by each party.

Nevertheless, the customizations made to DLXS will allow SPO and other organizations using the DLXS suite to publish other scholarly editions with similar components. This project shows how an electronic resource can be made available through a mixed model of free and restricted access or entirely free, using software specifically designed for hosting digital collections. SPO hopes to publish more open-access primary source material to make our written heritage available to scholars and to the public.

⁹ This is a known shortfall in XPAT, but building a work-around is difficult and has not yet been accomplished.