Digital Research in the Arts and Humanities

Series Editors
Marilyn Deegan, Lorna Hughes, Andrew Prescott and Harold Short

Digital technologies are becoming increasingly important to arts and humanities research, expanding the horizons of research methods in all aspects of data capture, investigation, analysis, modelling, presentation and dissemination. This important series will cover a wide range of disciplines with each volume focusing on a particular area, identifying the ways in which technology impacts on specific subjects. The aim is to provide an authoritative reflection of the 'state of the art' in the application of computing and technology. The series will be critical reading for experts in digital humanities and technology issues, and it will also be of wide interest to all scholars working in humanities and arts research.

Other titles in the series

Digital Archetypes
Adaptations of Early Temple Architecture in South and Southeast Asia
Sambit Datta and David Beynon
ISBN 978 1 4094 7064 9

Paradata and Transparency in Virtual Heritage
Edited by Anna Bentkowska-Kafel, Hugh Denard and Drew Baker
ISBN 978 0 7546 7583 9

Art Practice in a Digital Culture
Edited by Hazel Gardiner and Charlie Gere
ISBN 978 0 7546 7623 2

Digital Research in the Study of Classical Antiquity
Edited by Gabriel Bodard and Simon Mahony
ISBN 978 0 7546 7773 4

Crowdsourcing our Cultural Heritage

Edited by

MIA RIDGE
Open University, UK

ASHGATE

Quinn, Michael, ed. *Writings on the Poor Laws: Volume 2* (*Collected Works of Jeremy Bentham*). Oxford: OUP, 2010.

Robinson, Peter. 'Why Digital Humanists Should Get Out of Textual Scholarship'. Social, Digital, Scholarly Editing conference, University of Sasketchewan, July 11–13, 2013. http://www.academia.edu/4124828/SDSE_2013_why_digital_humanists_should_get_out_of_textual_scholarship (last accessed July 30, 2013).

Schofield, Philip. *Bentham: A Guide for the Perplexed*. London: Continuum, 2009.

Schofield, Philip. *Jeremy Bentham: Prophet of Secularism*. London, 2012. http://discovery.ucl.ac.uk/1370228/ (last accessed March 22, 2013).

Schofield, Philip, Catherine Pease-Watkin and Cyprian Blamires, eds. *Rights, Representation and Reform: Nonsense upon Stilts and Other Writings on the French Revolution* (*Collected Works of Jeremy Bentham*). Oxford: OUP, 2002.

Schofield, Philip, Catherine Pease-Watkin and Michael Quinn. *Of Sexual Irregularities, and Other Writings on Sexual Morality* (*Collected Works of Jeremy Bentham*). Oxford: OUP, 2014.

Semple, Janet. *Bentham's Prison: A Study of the Panopticon Penitentiary*. Oxford: OUP, 1993.

Stephen, Leslie. *The English Utilitarians: Volume I – Jeremy Bentham*. London: Duckworth and Co., 1900.

Weinberger, David. 'Crowdsourcing Transcription'. Too Big to Know blog. http://www.toobigtoknow.com/2012/09/04/2b2k-crowdsourcing-transcription-2/ (last accessed February 20, 2013).

Zou, Jie Jenny. 'Civil War Project Shows Pros and Cons of Crowdsourcing'. *Chronicle of Higher Education*, June 14, 2011 (updated June 21, 2011). http://chronicle.com/blogs/wiredcampus/civil-war-project-shows-pros-and-cons-of-crowdsourcing/31749 (last accessed February 7, 2013).

Chapter 4

Build, Analyse and Generalise: Community Transcription of the *Papers of the War Department* and the Development of *Scripto*

Sharon M. Leon

On the night of 8 November 1800, fire devastated the United States War Office, consuming the papers, records and books stored there. Two weeks later, Secretary of War Samuel Dexter lamented in a letter that 'All the papers in my office [have] been destroyed'. From the perspective of historians, the loss was monumental since in many respects the documents lost in the fire constituted the first 'national archive' of the United States. The *Papers of the War Department, 1784–1800* (PWD, wardepartmentpapers.org) is an ongoing digital editorial project at the Roy Rosenzweig Center for History and New Media (RRCHNM) (chnm.gmu.edu) that encourages historical scholarship of this lost period in the Early American Republic by restoring the archive online and making the entire collection accessible and fully searchable to a wide audience of scholars, students, teachers and the general public (see Figure 4.1). This website officially debuted in June 2008 with nearly 45,000 document images and basic metadata. Unfortunately, no realistic prospects existed to fund the transcription of this unique digital archive.

These circumstances presented an opportunity for innovation in digital humanities work and software. With the support of the National Endowment for the Humanities Office of Digital Humanities (NEH-ODH) and the National Archives and Records Administration's National Historical Publications and Records Commission (NHPRC), the team at RRCHNM devised a system to allow the existing user community to begin transcribing the materials from PWD. At the time that we launched the transcription project in March 2011, we knew very little about our user base. Over the course of the project, we have learned a great deal about those initial users and the many, many individuals attracted to the collection by the opportunity to transcribe. Together those users came to transcription with six areas of interest and motivation for their work: (1) a general interest in early American history; (2) a sense of civic duty; (3) a specific point of scholarly research; (4) genealogical and family history questions; (5) various educational assignments; and (6) a curiosity about how the transcription tool and process worked. Based on the lessons learned through developing the community

transcription tool and working with our PWD volunteers, we generalised the software into an open-source tool. *Scripto* (scripto.org) is available for use by other projects through a customisable version or through extensions for a range of popular content management systems.

Figure 4.1 ***Papers of the War Department, 1784–1800*** **website**

RRCHNM's foray into community transcription with PWD and the development of *Scripto* offers some significant lessons for cultural heritage institutions and professionals who want to engage with their constituents in meaningful ways. Primarily, we gained a dedicated and engaged audience for PWD, and a tremendous insight into their motivations. Equally important, the development process for the generalised tool, and its role in the larger ecosystem of open-source software that enables widespread user participation in cultural heritage projects, points to viable directions for the development of subsequent tools. Together the case study of PWD and the story of the creation of *Scripto* suggest that a wide range of cultural heritage organisations can launch and sustain lightweight transcription projects that encourage increased engagement with core audiences.

Papers of the War Department, 1784–1800

The *Papers of the War Department, 1784–1800* has never been a traditional documentary edition project. Decades in the making, the work on the collection began before many scholars were aware of the world wide web and certainly before the majority had even begun to consider the ways that the internet would change our relationship to research methods, content access and engagement with the larger public. The conditions under which the archive was reconstituted – assembling a collection of photocopies and high-resolution scanned document copies – allowed for an experimental approach. While the original documents resided in archives and special collections around the world, PWD itself has no original holdings. With a completely virtual collection, RRCHNM had the opportunity to err on the side of immediate and open access, making the digital copies of the documents directly available to the public via the web.

At the dawn of the project, documentary collections and editions were the purview of academic researchers with access to well-funded research libraries. In 1989 when Ted Crackel, the first Editor-in-Chief for the project, proposed the idea of reconstituting PWD, he initially thought in terms of a traditional print edition.[1] In 1993, Crackel began the initial planning to undertake the project himself with the support of East Stroudsburg University in Pennsylvania. With funding from the NHPRC and eventually the Department of Defense Legacy Project, beginning in 1994 Crackel and his staff visited over 200 repositories and consulted over 3,000 collections in the United States, Canada, England, France and Scotland, copying, scanning and processing nearly 50,000 documents. At this tremendous volume, Crackel soon realised that a print publication would be unreasonable, and turned his sights towards producing a CD-Rom that would include the document images. In 2004, when Crackel accepted the position as Editor-in-Chief of the George Washington Papers, PWD faced a crisis point: without Crackel at the helm, the project lacked the professional support and leadership to continue at East Stroudsburg. Responding to a call from the NHPRC and at Crackel's urging, RRCHNM applied to adopt the project in the summer of 2005, and began work on producing the digital documentary edition in spring 2006 under the direction of Editor-in-Chief Christopher Hamner.

RRCHNM built a website with nearly 45,000 documents that launched in June 2007. The archive contains materials ranging from several years before to several years after the heart of the materials, thus spanning 1781 to 1803. This includes 42,887 documents with scanned images and 2,482 additional citations for documents that do not have the accompanying image due to some rights or permissions issue. Moreover, the database includes listings for nearly 4,180 people or groups who were listed as sender or recipient of a document or were explicitly mentioned therein. Given the scope of materials and seeking to be realistic about the resources available to process the collection, the editorial team

1 Crackel, 'The Common Defence'.

had to balance their approach to describing the materials. On the advice of Max Evans, then the Director of the NHPRC, the editors opted initially to index the documents with basic metadata about author, recipient and date. With this as their charge, Editor-in-Chief Hamner, Assistant Editor Ron Martin and nearly a dozen graduate research assistants from George Mason University's History Department systematically reviewed and described the papers. By 2010, these basic metadata made the documents searchable to the extent that if a researcher knew what she was looking for, she could extract it from the corpus. For researchers with less concrete demands, the index proved less usable. Routes to access materials on a range of subjects in the history of the Early American Republic, such as the handling of Indian affairs, pensions, procurement, the relationship of the first American citizens with the new federal government, and conflicts including the Whiskey Rebellion and the Quasi-War with France, were extremely difficult to navigate. As a result, in the subsequent years, the editors and assistant editors on the project have added an additional layer of metadata – names, places, things, ideas mentioned and a brief abstract – for the most significant elements of the collection, roughly one-third of the documents. When the funding for the project came to a close in June 2013, the team of editors at RRCHNM had produced this two-tiered level of description for the entire collection.

Looking for Precedents for Community Transcription

While the staff at PWD would never have the capacity to transcribe a significant number of the documents in the archive, steady site traffic suggested that we had an untapped resource of scholars and researchers who could help with this task. Every day researchers examined documents in the archive, regularly making rough transcriptions to use in their own work. In the years before we launched the transcription project, the editors routinely got emails from researchers suggesting improvements to the archive's metadata based on their work with the materials. Thus, we proposed to build an open-source transcription tool to allow users easily to submit those transcriptions and their knowledge back to the archive. The resultant tool would allow PWD, and eventually other digital archival projects, to draw upon the wisdom of the thousands of interested researchers, scholars and students who work with these materials. Gradually, users' combined work would enhance the discoverability and usefulness of the archive without significantly adding to the costs of the project.

We envisioned this tool as a response to Max Evan's 2007 call for commons-based peer production as a way to create 'Archives of the People, by the People, for the People', where he points to 'the concept of commons-based peer-production as a means of turning collections inside out. It encourages archival institutions to reinvent themselves, and, in collaboration with other archives and with other types of organizations, to organize archival work in concert with a curious and

interested public'.[2] Evans' approach to openness and user engagement matched nicely with the philosophy of public history that undergirds all of RRCHNM's work. Moreover, Evans was not alone in his vision for participatory archives. J. Gordon Daines and Cory L. Nimer's 'The Interactive Archivist: Case Studies in Utilizing Web 2.0 to Improve the Archival Experience' published by the Society of American Archivists in 2009, provided a useful summary of the interests of the archivists in social networking and the usefulness of tagging, commenting, reviewing and rating services, but did not mention strategies for incorporating users into the archival process at the transcription or description level that remains the domain of archivists. These pieces suggested that the team at RRCHNM was at the leading edge of an emerging push to encourage more significant user engagement in archival and documentary editing projects.

While key individuals in the archival profession and the documentary editing world spoke directly to the needs of our papers project, in proposing to open the transcription process up to crowdsourcing, RRCHNM also drew upon the example and success of a host of successful ventures in the wider realm of digital culture. Though the goals and needs of digital archive and documentary editing projects are distinct from the goals of these community-driven ventures, they did offer promising glimpses of the efficiencies and outcomes for our purposes.[3] Each tapped into the interests and passions of a segment of the public who contributed to the accumulated value and content of a project – whether software development or knowledge aggregation. The open-source software movement – with successes like the Linux operating system and Mozilla's popular Firefox web browser – contains lessons for those of us interested in harnessing the expertise of a particular community to build a successful project. The developers who participate in open-source software projects do so for their own reasons, but they contribute to the common good by applying their expertise to the demands and problems raised by software innovation. There were also several content-focused examples of crowdsourcing that were particularly revealing.

Wikipedia, the free online encyclopaedia that is written and maintained by users around the world, is by far the most well-known instance of successful crowdsourcing. Launched in 2001, using a simple authoring and versioning software to manage articles, Wikipedia thrives on the contribution of thousands of anonymous users. As of July 2013, the site had more than 77,000 active contributors who created and edited over 22 million articles in 285 languages – more than four million of which are in English.[4] From its founding, teachers, parents and scholars have worried about the accuracy and content of Wikipedia, especially since articles from the free encyclopaedia have long been the first result to appear in most search engines. Yet, the source of the occasional errors and incoherence of Wikipedia is also its power – the tremendously fast-moving contributions and

2 Evans, 'Archives of the People', 387.
3 The term 'crowdsourcing' was coined by Jeff Howe in 'The Rise of Crowds'.
4 'About', Wikipedia http://en.wikipedia.org/wiki/About_Wikipedia.

collaboration of users means that errors are unlikely to remain for long before they are corrected. Similarly, the growth of articles and topics covered in the free encyclopaedia reflects the energy and interests of the open-source community of users. An assessment of the creation and revision history of articles on historical topics could provide a tremendous wealth of information about the concerns of the public with the past.[5]

In a similar, albeit much less extensive way, Flickr Commons provided another example of tapping the collective knowledge of interested members of the public.[6] Beginning with a seed contribution by the Library of Congress in January 2008, cultural institutions from around the world, including the New York Public Library, the George Eastman museum, the Powerhouse Museum in Australia and the Smithsonian Institution, contributed images and associated metadata to the Flickr Commons collection. The collections are freely available to the public, who can tag and comment on each image. In October 2008, the Library of Congress assessed their participation in the pilot project, reporting that over 2,500 Flickr users had left more than 7,000 comments on almost 3,000 images. Library staff selected important corrections and additions to captions, titles and the identification of individuals that were then incorporated back into the metadata of more than 500 items by August 2008. Based on these positive interactions with the public, the staff advocated that the library continue to draw upon public knowledge through the Flickr Commons project and other Web 2.0 interactive projects, declaring: 'The benefits appear to far outweigh the costs and risks.' The Library of Congress' success helped to encourage the Smithsonian Institution to join Flickr Commons, and as a result, in the period between June and December 2008, their photographic contributions received over 625,000 views with many user comments and tags. Thus, trusted cultural heritage organisations were beginning to recognise the powerful ways that interested members of the public, scholars and educators can contribute to the knowledge base related to collections.[7]

Additionally, Zooniverse has supported a large number of 'citizen science' projects in the last several years. Most focus on the crowdsourcing of big data, such as the *Galaxy Zoo* identification projects, but others have a more historical focus. The *Old Weather* project (see also Chapter 2 in this volume) is helping scientists recover weather observations included in the logbooks of Royal Navy vessels during the First World War era. The results of this work help climate scientists build better models and provide historians with access to new data about the ships and their sailors. Similarly, the *Ancient Lives* project has made the fragmentary Greek texts, the *Ocyrhynchus Papyri*, available for transcription. This work will facilitate the identification of known texts and the isolation of new ones, in turn

5 Rosenzweig, 'Can History Be Open Source?'.

6 Flickr Commons, http://flickr.com/commons/.

7 Springer et al., 'For the Common Good', 36; Kalfatovic et al., 'Smithsonian Team Flickr'.

contributing to a greater understanding of Greco-Roman Egypt. Zooniverse has released the code for their generalised transcription tool, *Scribe*.[8]

At the same time, a number of universities and national libraries began running several very successful crowdsourcing projects. First, in August 2008 the National Library of Australia started a project where members of the public corrected the results of Optical Character Recognition (OCR) software for their digitised newspapers. By then-project manager Rose Holley's estimates, there were roughly 6,000 participants who had corrected over seven million lines of text by November 2009.[9] That project then became part of the *Trove* project, which allows users not only to correct OCR, but also to contribute historic images, tag materials and link a host of other cultural materials.[10] As of June 2013, *Trove* provided users with access to nearly 350 million digital items. The results of these National Library of Australia projects show the remarkable range of contributions that public users can make to historical material, especially when they have the capacity to work across large collections.

Second, university and public libraries have also had good success with projects that ask users to participate in direct transcription of document images.[11] Of these, the University College London's *Transcribe Bentham* project is probably the most well known (Chapter 3 in this volume).[12] In the course of its public work, the project has allowed for the transcription of close to 2,000 manuscripts from Jeremy Bentham's published and unpublished works. The transcriptions will form the foundation for future work on the *Collected Works of Jeremy Bentham*. MediaWiki is the system that underlies the *Transcribe Bentham* transcription work, and the project team has released the code for its MediaWiki plugins.[13] This code does not allow MediaWiki to interact with existing content management systems. The project team at University College London, who began work on *Transcribe Bentham* in April 2010 and launched the site in September 2010, has offered ample insights from their work about the ways that it enhanced their relationship to various interest communities, even if it did not necessarily speed the process of transcription or reduce the cost.[14] Their results point to a primary good of opening

8 Zooniverse, http://www.zooniverse.org/; *Old Weather* project, http://www.oldweather.org/; *Ancient Lives* project, http://ancientlives.org/; *Scribe*, https://github.com/zooniverse/Scribe.

9 Holley, 'Crowdsourcing'.

10 *Trove*, http://trove.nla.gov.au/.

11 See also Ben Brumfield's 'FromThePage', http://beta.fromthepage.com/?ol=1_hd_logo; New York Public Library's 'What's on the menu?', http://menus.nypl.org/; and the University of Iowa Libraries' 'Civil War Diaries and Letters Transcription Project', http://digital.lib.uiowa.edu/cwd/transcripts.html.

12 *Transcribe Bentham*, http://www.ucl.ac.uk/transcribe-bentham/.

13 *Transcribe Bentham* MediaWiki Transcription Desk toolbar, http://code.google.com/p/tb-transcription-desk/.

14 Causer et al., 'Transcription Maximized'.

community transcription work: building dedicated user communities for digital cultural heritage projects.

All of these projects suggest that community-sourcing transcription for digital collections can provide significant benefits to cultural heritage institutions. First, and foremost, public contributions provide transcriptions where there once were none, and where there likely would be none in the future. Second, in the case of documentary editing projects, staff can draw on the publicly contributed transcriptions to form a base for their editorial work. Editors may start with a rough transcription provided by interested users, and then apply the techniques and expertise of their training to produce a corrected transcription quickly. Moreover, allowing the public to contribute document transcriptions to digital collections has real-time benefits for the accessibility of the materials. Each transcription contributed to the archive can then be made available to the collection management system's search engine. The result is ever-improving discoverability. The full text of the documents allows the search engines to surface the most relevant documents by better weighting their results. Furthermore, as more technically astute scholars increase their reliance on computational text analysis, legions of documents with digital images that lack transcriptions will be off limits to their examination and processing. Growing the field of transcribed archival materials can only benefit humanities scholars.

Community-contributed transcriptions also allow editors to understand better the ways in which some users interact with their archive. In a collection where the volunteers can select any document for transcription, often the documents that are of the most interest to users will be transcribed most quickly and fully. Thus, the public contributions can serve as a barometer of the most interesting materials within a particular collection. This convergence of volunteer interest and the collection coverage points to perhaps the most important reason for launching collaborative work with the public: community building. As Trevor Owens has noted, the concept of crowdsourcing and the related turn towards gamification, can seem like an effort by projects to take advantage of public contributions simply as a free labour force. But, transcription volunteers make the contributions that they do because they find the work meaningful.[15] They are contributing to the usefulness of the collection and learning about history at the same time. This type of community building around collections increases investment in cultural heritage and points to long-term gains for the humanities.

Implementing Scripto *with the* Papers of the War Department, 1784–1800

Given the positive outcomes from these early community-sourcing and open-source projects, RRCHNM decided in 2009 to apply for support from both the NEH-ODH and the NHPRC to design and build an open-source tool that would enable

15 Owens, 'Meanification and Crowdscaffolding'.

the community sourcing of transcription.[16] The result of those applications was a Digital Humanities Start-Up Grant from NEH-ODH to build the basic tool and implement it with the *Papers of the War Department*, and a grant from the NHPRC to generalise the tool, do user testing and develop support and documentation so that other projects could launch community transcription projects. Together, these two grants enabled RRCHNM to build and refine *Scripto*, offering it as a customisable software library connecting a repository to an editing interface, and as extensions for three popular web-based content management systems (*Omeka*, omeka.org; *Drupal*, drupal.org; and *WordPress*, wordpress.org).

The first step in designing the transcription tool for the *Papers of the War Department* was to map PWD's idiosyncratic data model to *Scripto*. To do this, Jim Safley (RRCHNM Digital Archivist and Web Developer) wrote a small software function implementing *Scripto*'s adapter interface that responded to requests by the transcription service. He then linked the existing PWD website to a custom web application containing a document image viewer and input forms for transcription and discussion. Safley chose to use the software library OpenLayers as the image viewer because of its ability to render high-resolution image files directly in the web browser without the need for those files to be converted into another format for viewing.[17]

Safley purposefully developed a very limited feature set for *Scripto* so that it could easily be generalised for work beyond the PWD case. In addition to the image viewer and the transcription form, he included a discussion form, where transcribers could ask questions and clarify their work, and where administrators could answer questions and ask for clarification. Since PWD's transcription application requires users to register and log in, transcribers can view a list of document pages to which they have contributed. This makes it easy for users to return to their previous work. Administrators have the authority to protect pages from further edits and to export the document transcription from the MediaWiki database to the PWD database.

Beyond the core operational features of the system, the RRCHNM team did some work to assure that the tool meshed with the needs of the PWD archive. Thus, Ken Albers (RRCHNM Web Designer) created a number of mock-ups for the transcription interface. Using both paper prototypes and unstyled builds of the interface, Albers and Safley did user testing with PWD editors, RRCHNM graduate research assistants who had worked with the PWD archive and with several individuals who had no familiarity with the system. In August 2010, once Safley had sufficiently mapped out the functional requirements for the tool, the *Scripto* team met with the editorial team from the *Papers of the War Department* to review wireframe drawings and possible layouts for the tool's functional web interface. Albers proposed a layout that was quite similar to the design mock-

16 NEH-ODH, http://www.neh.gov/divisions/odh and NHPRC http://www.archives.gov/nhprc/.

17 OpenLayers, http://openlayers.org/.

ups submitted with the initial grant proposal, which included a vertical split screen with the digital image to be transcribed in a left window and the editing/transcription window on the right. After interacting with this basic layout the PWD editors expressed their concern about the narrow image viewer and the ways that it would constrain a user's ability to view a whole line of script at once. This critical feedback resulted in the first revision to the tool's user interface.

As a result of this user testing, we fundamentally reoriented the transcription interface, rejecting the common side-by-side document and transcription window positioning in favour of a top and bottom orientation (see Figure 4.2). In September, Albers and Safley returned to the editors with a functioning mock-up of the tool, which included a horizontal split screen with the image viewer on the top and the editing/transcription window on the bottom. By positioning the OpenLayers image viewer on top with the transcription window below, we maximised the width of the viewer window. This decision dramatically increased the efficiency of transcribers by allowing them to view a complete line of text while zooming in on an image. Next, with this orientation, we narrowed the width of the transcription window to make it comfortable for typed text. Finally, we positioned the list of document page images to the left of the window, allowing a volunteer to proceed easily through multi-page documents. During this round of testing, the users offered suggestions about the location of the page navigation and the links to 'help' materials such as the style guide. Albers and Safley integrated this feedback into the subsequent build of the tools, which they then integrated with the administrative interface of PWD. This implementation represented the third iteration of the initial user testing.

Developing a functioning tool with a logical workflow was only part of the task of launching community transcription with PWD. We also had to create a support apparatus for contributors that meshed with the content and the character of the repository and its users. In preparation for launching *Scripto* with PWD, we created a registration workflow for new volunteers. Although MediaWiki can be configured to allow users to create their own accounts or to edit documents without being logged into the system, we felt strongly that it was important to maintain editorial control over the transcription system through user accounts and logins. This required login would give us the peace of mind that we would not have to deal with significant spam users and vandalism. Thus, we configured MediaWiki to prohibit document edits by anonymous users, and placed the process of account creation in the hands of the PWD editorial team. To manage that process, we created a Google form to gather registration information from volunteer transcribers. That form requires the minimal data for account creation (username and email address), but requests a full name, affiliation, country, zip code and the reason the user is interested in working with the PWD archive as optional fields. The results of these form submissions are gathered in a Google spreadsheet, which a PWD editor uses to hand-create MediaWiki accounts for each user. Once the user has verified the account by setting her password, she is set to begin transcription work.

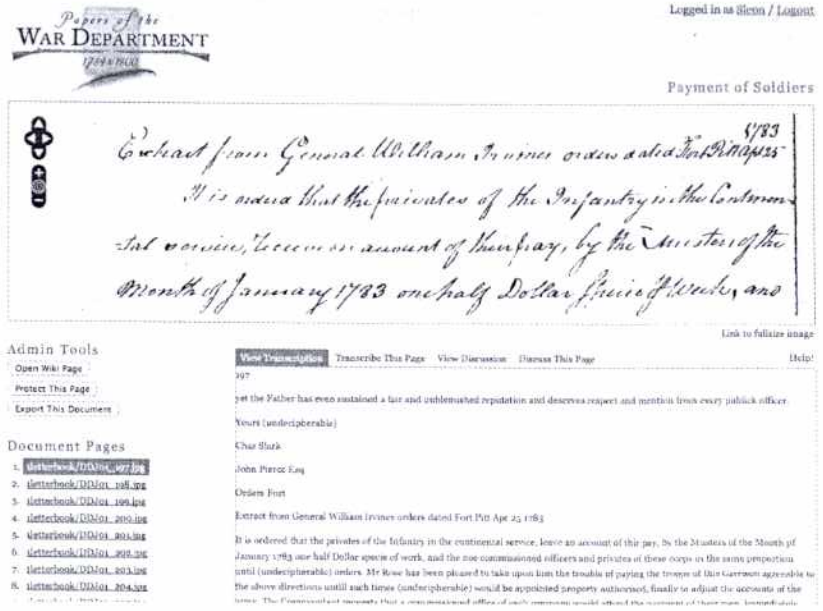

Figure 4.2 PWD transcription interface

In addition to the basic transcription interface of the document viewer and the transcription window, the PWD implementation of *Scripto* required us to create a number of support structures. We developed a set of static text pages that offered potential contributors a clear and concise introduction to the transcription project and its role in the larger PWD project. These included both the invitation to participate as transcribers and the short guidelines for creating good transcriptions within the conventions of the *Scripto* system. Also, we provided contributors with a large list of potential documents that were good candidates for transcription. While many of our participants were drawn to the project by their own research interests and had clear ideas about the documents they wanted to work with, we realised that we were likely to attract many volunteers who simply wanted to aid the progress of the project and who would welcome our suggestions to direct their work. Thus, PWD assistant editors continually nominate documents for transcription from their ongoing work with the collection. At any given time, there are roughly 200 nominated documents for users to choose from if they do not have their own specific research interests.

Each day a PWD editor spends some time working with the volunteers and the transcription submissions. First, he monitors the registration list and creates new accounts. Next he surveys all of the newly created page transcriptions, making some corrections and edits. Then, he reads the discussion pages associated with the transcriptions. Finally, he spends some time blogging and tweeting about the

nominated documents, the completed transcriptions and other project progress. As a result, we have a very good sense of the volume of interest and activity amongst transcribers for the first two years of the project.

Building the Community of Transcribers

The efforts to recruit community members to participate in transcribing the *Papers of the War Department* was jump-started by early national press recognition that preceded the launch of the transcription facilities. In December 2010, *Scripto* received press coverage in the *New York Times*.[18] The article generally dealt with efforts of documentary projects to experiment with crowdsourcing, and did not refer to *Scripto* by name, but this initial mention of the *Papers of the War Department* work generated a significant amount of interest from potential transcribers and members of the documentary editing community. Since the article was published before the release of the transcription functionality, the project did not reap the significant bump in participation that the *Transcribe Bentham* project gained from the exposure, but it certainly laid the groundwork for a successful launch.

In March 2011, when we officially launched the transcription facilities with PWD, the website received visits from roughly 3,800 unique users, which was a fairly typical number for that point in the life of the project. From this base, we set out to attract a new set of users to the work by coordinating a publicity campaign that included blog posts, twitter coverage and direct messages to email discussion lists with high traffic from early Americanists, those teaching the US history survey and a full range of genealogical organisations. These efforts produced notice in some unlikely places, such as a post by Curt Hopkins entitled 'Crowdsourcing the Preservation of the U.S. War Papers' on *Read, Write, Web*, which placed the effort in front of an audience who primarily identified as being interested in technology rather than history.

Due to this outreach and press recognition, the project got off to a swift start. Within the first week of launch, we had 120 transcribers request accounts and transcribe roughly a dozen documents. As the months progressed, the momentum continued. By May, the site had 170 transcribers who had completed 80 documents. Within six months, those numbers had increased to 308 users, roughly 70 of whom had been active transcribers, and who had finished 450 documents. By the close of the second year, the project included 1,345 registered transcribers, 227 of whom had been active within the last 90 days, and who had completed just over 2,000 documents. This range shows a relatively slow and steady increase in the number of active transcribers over time, but the amount of completed documents and registered users exhibits a more swiftly increasing rate (see Figure 4.3). Additionally, in March 2013, the website received roughly 11,450 unique visitors, an increase of over 7,600 users when compared to two years earlier. Generalising from these numbers is difficult, in part because of what

18 Cohen, 'Scholars Recruit Public for Project'.

constitutes a document within the PWD collection. Some documents consist of a single page with a few sentences and others are letter books with hundreds of pages. The degrees of difficulty in transcribing the material also varies greatly due to the fact that the documents come from many hands, and frequently include difficult to transcribe tabular data.

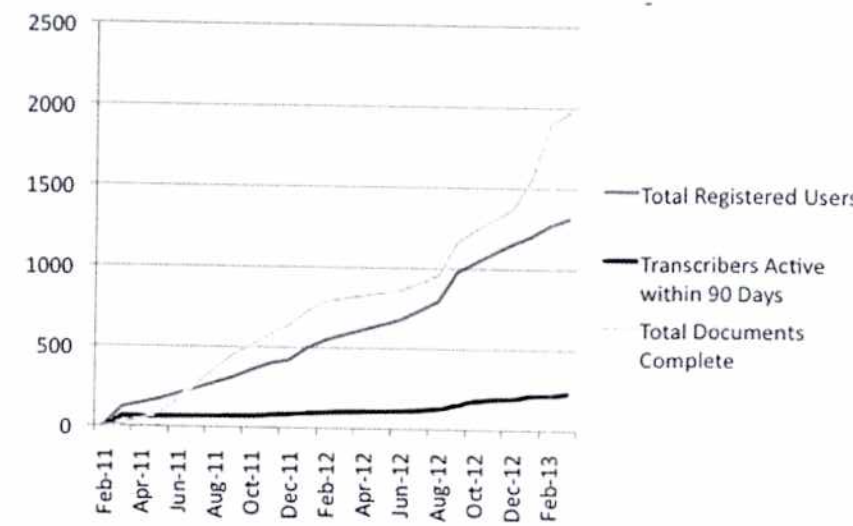

Figure 4.3 Total registered users and documents complete in comparison to active transcribers over 90 days

Nonetheless, the important data generated over the course of the first two years of the project seem, unsurprisingly, to reveal a great deal of productivity on the part of a small number of dedicated volunteers. In the three months leading up to the two year mark, of the 226 active transcribers, only 17 had made more than 100 edits, with the most active contributor, Paulmd199, making more than 2,600, followed by transcriber HollyPBrickhouse with just over 1,500 edits, and then a drop to around 675 from Prosenbloom, and eventually a levelling off where approximately 200 somewhat active users had similar numbers of edits (see Figure 4.4). This curve is familiar to those who work with volunteer editors. As Ben Brumfield has noted, transcription projects and other crowdsourcing ventures tend to follow the power-law distribution, suggesting that 90 per cent of the edits are done by 10 per cent of the users.[19] This projection is generally borne out with the participants in the PWD transcription project.

19 Brumfield, 'Crowdsourcing IMLS WebWise 2012'.

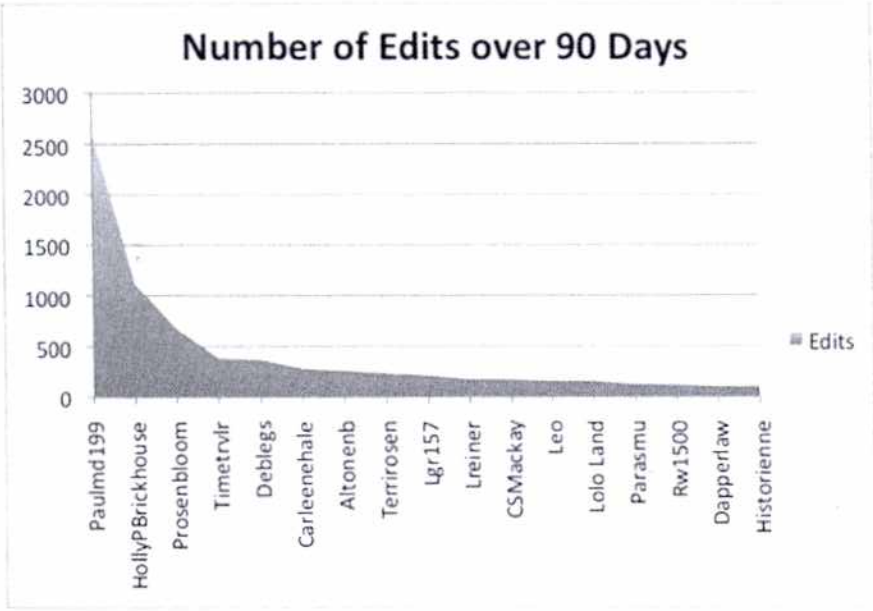

Figure 4.4 Number of edits from the most active users

More significant, however, than the bulk numbers of transcription contributions from our volunteers, is the range of important information they have offered us about themselves. RRCHNM's previous extensive experience with digital collecting projects, such as the *September 11 Digital Archive* (911digitalarchive.org) and the *Hurricane Digital Memory Bank* (hurricanearchive.org), has taught us that requiring too much information from contributors is a sure-fire way to encourage them *not* to participate.[20] As a result, we generally try to keep our sign-up forms to a bare minimum length, and only make the absolutely necessary fields required. In the case of the transcriber account sign-up, we only required volunteers to provide us with a username and an email address, but we requested a full name, zip code, an affiliation and the reason they wished to participate in the project. These optional fields on the form have provided us with a wealth of data about our contributors.

Out of the 1,328 transcribers who had requested accounts by the close of the project's second year, 74 per cent had offered some information about why they wanted to transcribe documents from PWD (see Figure 4.5). This is quite a remarkable response rate for a non-required field. Analysis of the content of those responses reveals six general types of volunteers. The largest group of contributors

20 For more on these projects, see Cohen, 'The Future of Preserving the Past'; Brennan and Kelly, 'Why Collecting History Online is Web 1.5'.

(34 per cent) came to PWD with very specific historical research agendas, searching for material on a particular person, place or event. Those who were explicitly engaged in genealogical research (29 per cent) were a close second. Some 14 per cent of volunteers expressed a more general interest in the American Revolution and the early national period while a full 10 per cent of respondents noted that they felt they were making a civic contribution by working to help make the papers accessible. Many of the librarians, archivists and information professionals (8 per cent) requested accounts because they were interested in *Scripto* as a transcription tool. Finally, 6 per cent of accounts were requested by teachers designing activities for students or by students fulfilling an assignment.

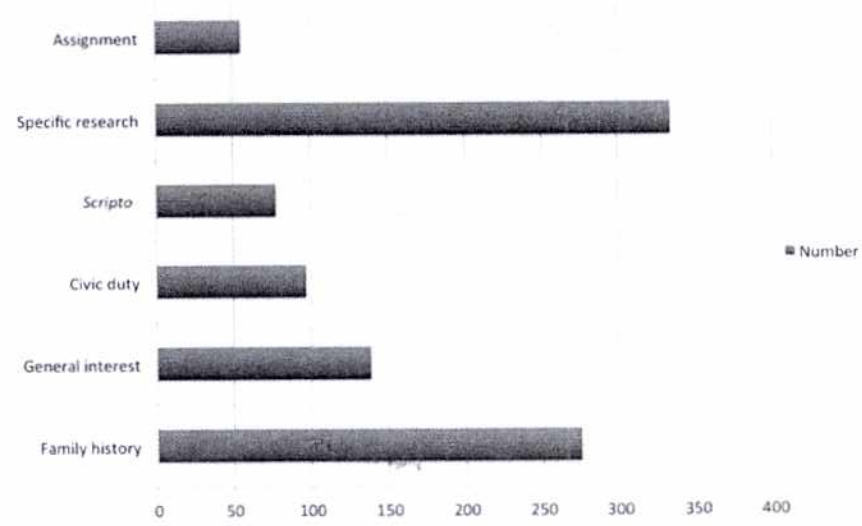

Figure 4.5 Reasons for requesting a transcription account

With transcription volunteers hailing from at least 19 non-Native American nations and 12 Native American communities, contributors came to the documents with an array of research interests. Even focusing just on the 25 most frequently used words in the explanations for participation, several key areas of interest emerge (see Figure 4.6) – some more predictable than others. Those enthusiastic about family history list themselves as genealogists, independent researchers, descendents, relatives and members of the Daughters of the American Revolution, or the Society of the Cincinnati, organisations that focus on members having family ties to the American Revolutionary Era. Another cohort is focused on military history and the American Revolution, including the development of the Navy, and key events like the Whiskey Rebellion. This focus matches nicely the large number of volunteers who claim affiliation with the United States armed

services, either as someone on active duty or a retiree. Those not conversant with the *Papers of the War Department* might be surprised by the fact that a significant concentration of transcription volunteers are focused on Native American history and Indian affairs, with a healthy interest in Cherokee and Creek tribal affairs. Of those mentioning an interest in Native American history, volunteers reported having affiliations with the Brothertown, Cherokee, Chicksaw, Choctow, Creek, Dakota, Miami, Mohawk, Notaweega, Seneca, Wyandot and Yuchi bands and nations. With key negotiations between the War Department and Native Americans taking place during this period, the repository is filled with important and revealing materials for research on these topics and of interest to these communities.

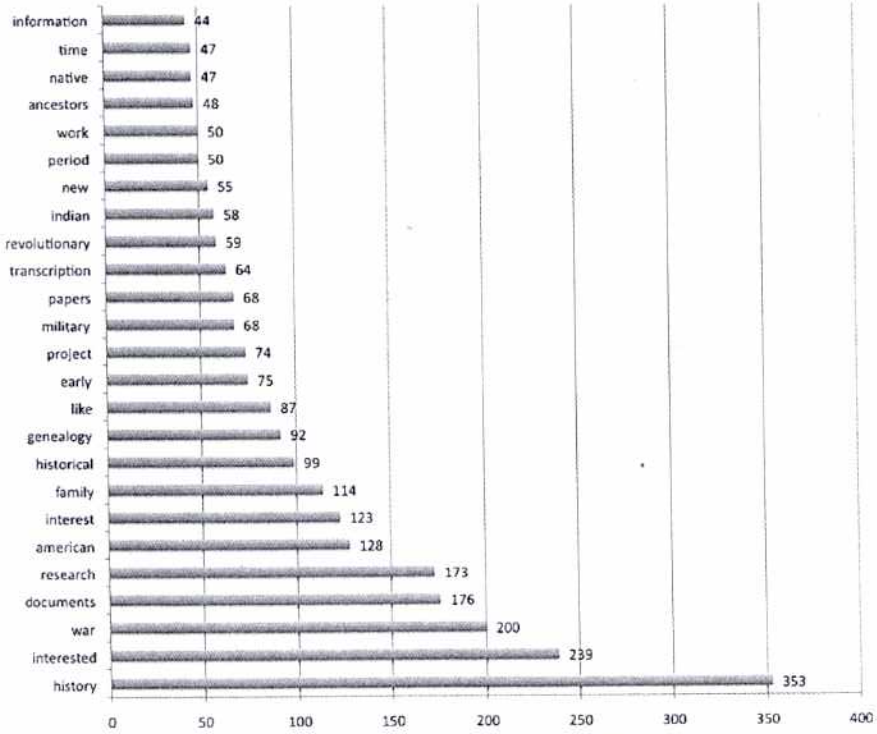

Figure 4.6 Word frequency within stated transcriber interests

In sum, the two years of experience with community transcription for PWD has yielded a number of important gains for the project. First and foremost, the project has brought many, many more people in contact with the papers, not all of whom have ended up as transcribers. Simply publicising the transcription work has yielded a tremendous increase in website traffic. In 2010, the site saw just over 36,500 unique users. In 2011, the first year of the transcription project, that

number rose to just over 51,000, and in 2012 to nearly 90,000 unique visitors. In the first six months of 2013, the site has roughly 62,000 visitors, putting it on track to reach over 120,000 for the year. Second, the editors have a much better sense of the kinds of research interests that are driving the active users who contribute transcriptions. This information will be essential in the planning for the kinds of narrative historical interpretation RRCHNM plans to add to the papers in the coming years. Finally, every completed document transcription increases the ability of users to find the documents they are looking for because it gives the site search and presumably search engines like Google more data to work with, and those texts are also accessible to screen readers that support vision-impaired users. So, in many respects, these contributions offer every PWD user a slowly and steadily improved experience with their research.

Generalising Scripto *for Widespread Use*

While the data that RRCHNM has gained from opening the *Papers of the War Department* to community transcription has provided a range of insights about our users and their interest in the content, we remain committed to building tools that other scholars and cultural heritage organisations can use to advance their own digital work. As a result of that commitment, after the implementation of the transcription facility, the team began the process of enabling connections between the tool and several common content management systems. Unlike many of the other recently developed transcription tools, *Scripto* was designed specifically to work with existing content management systems, rather than to replace them with a second source repository. The rationale behind this choice was two-fold. First, we firmly believed that cultural heritage institutions need to employ standardised metadata systems when they provide web content. The structured data provided by a standardised metadata system dramatically increase the possibility that the data can be exported to a new system, thus making that content as interoperable as possible. We did not want to create a tool that would impede interoperability by forcing users to separate their source material from their metadata schema. Second, we wanted to offer the lowest possible barrier to use. Hence, the idea that users would have to duplicate their sources in a second system seemed like an unnecessary and unwise step.

The software's architecture resulted from a process of considering the needs of the *Papers of the War Department* project, and generalising from that case. At its most basic, *Scripto* is a software library that mediates the communication between a content management system, a custom *Scripto* application and MediaWiki. The content provider serves the content, usually a corpus of digital material (images, video, sound) that can be transcribed; the *Scripto* application juxtaposes that content (via a media viewer/player) with a transcription form; and MediaWiki serves as the transcription database, revision engine and user account administrator (see Figure 4.7). We designed the *Scripto* library to be compatible with potentially any content provider. We accomplished this in two ways. First, the library defines

an adapter interface that is used to establish two-way data mapping between the content provider and *Scripto*. Second, the library normalises the content provider's identification scheme (no matter how informal and inconsistent) to enable fail-safe data transport between the content provider and MediaWiki.

What are the *Scripto* Components?

Scripto provides the engine for crowdsourcing transcription through creating a relationship between the organisational framework for the website (CMS) and a collaborative editorial platform (MediaWiki).

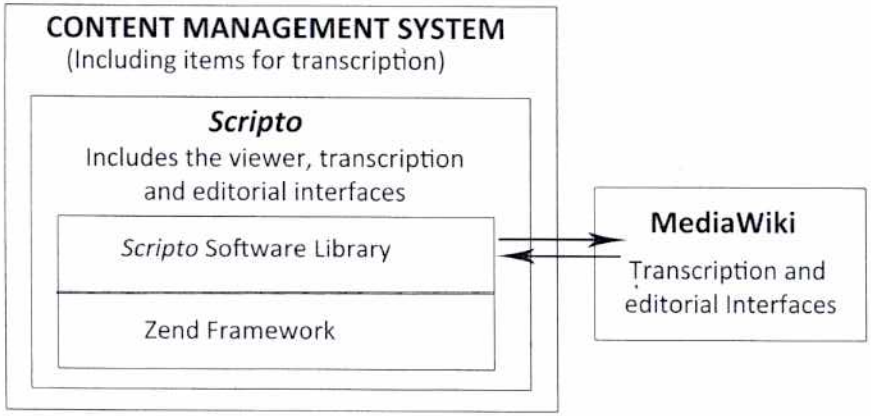

Figure 4.7 ***Scripto*** **architecture schema**

We chose MediaWiki for the transcription database for several reasons: it is the most popular Wiki application and has a sizable and active developer community; Wiki mark-up is relatively easy to learn and there are useful editors available; it offers helpful features, such as discussion pages and user administration; and it comes with a powerful, fully featured application programming interface (API) that offers technical hooks which *Scripto* uses to interface with the content provider and transcription application. Transcriptions are stored in the MediaWiki database until an administrator exports them to the content provider's database. This is done to ensure that transcriptions are complete and vetted before they are added

to the content provider's database. Each project can then evaluate the contributed transcriptions according to its own criteria.

Scripto is interface-agnostic, meaning that content providers are not tied to a single user interface or a predetermined feature set. They have full control over how their transcription application looks and functions. This makes it possible to embed a fully customisable user interface into an existing context, such as a website or standalone application. It does require some technical proficiency to build a transcription application, but *Scripto*'s API is straightforward and well documented.

Since the majority of projects do not have the time or technical expertise to create their own transcription application, under the *Scripto* flagship project, web developer Jim Safley created a set of connector scripts that enable users of common content management systems to implement *Scripto* with their work. The resultant scripts for *Drupal*, *Omeka* and *WordPress* are each available on Github as open-source code, and in zipped versions ready for installation from the *Scripto* site.[21] The development community that coalesced around the software through a Google Groups developers' discussion list offered crucial feedback on features and functionality during the connector development and testing process. The development group included 23 active members who participated in 18 support and feature conversations on the email list during the initial release of the testing versions of each connector. As a result, each connector had several releases and updates in response to developer community interaction with Safley.

In support of the release of the stable connector scripts, the *Scripto* team fully redesigned the software's website in June 2012 (see Figure 4.8). During the life of the project, the site had received just over 13,000 unique visitors, and the new 'look and feel' made the site a much more welcoming place for those visitors. The new design highlighted the software's functionality and offered project administrators easy access to a 'User's Guide' for working with *Scripto* that details the installation process, the editor's role and the transcriber's role.[22] The guide provides non-technical users a step-by-step introduction to working with *Scripto* in each of the content management system environments. It also contains tips on project organisation, volunteer management, transcription oversight and outreach.

While the 'User's Guide' offers potential users a great deal of information about the software and its implementation, our experience in software development tells us that users want to experiment with a system before they install it on their own sites. As a result, we also set up 'sandbox' sites for users to work with both an *Omeka* implementation and a *WordPress* implementation.[23] We did not offer a *Drupal* sandbox because the *Drupal* system is so flexible and customisable that

21 *WordPress+Scripto* plugin, https://github.com/chnm/scripto-wordpress-plugin; *Drupal+Scripto* extension, https://github.com/chnm/scripto-drupal-module; *Omeka+Scripto* plugin, http://omeka.org/add-ons/plugins/scripto/.

22 *Scripto* 'User's Guide', http://scripto.org/documentation/.

23 *Omeka+Scripto* sandbox, http://scripto.org/omeka/; *WordPress+Scripto* sandbox, http://scripto.org/wordpress/.

Figure 4.8 ***Scripto*** **website**

it is almost impossible to create an environment that would realistically mirror an actual individual user's experience with that content management system.

Since the public release of the connector scripts, many testing sites and *Scripto* implementations have sprung up across the web. The most extensive work has come out of university libraries. For example, while the University of Iowa Libraries might be best known for their initial crowdsourcing venture 'The Civil War Diaries and Letters Transcription Project', which launched in spring 2011 using a rudimentary transcription submission form, the Libraries have since launched a more extensive transcription project called 'DIY History', which uses *Omeka* and *Scripto* as its software platform (see Figure 4.9).[24] The project offers users a chance to transcribe materials from the Libraries' culinary manuscript collections and the Iowa Women's Archives. To date, the site includes over 35,000 transcribed pages. Similarly, the University of Alabama Libraries has offered over 500 items from their collections for transcription using *Omeka* and *Scripto*. The materials are drawn from the Manly Family Papers collection, which relates to the early history of the university, and the Meriwether Family Papers collection, which centres on the correspondence between an Alabama Infantryman and his wife during the

24 University of Iowa Library's 'DIY History', http://diyhistory.lib.uiowa.edu/.

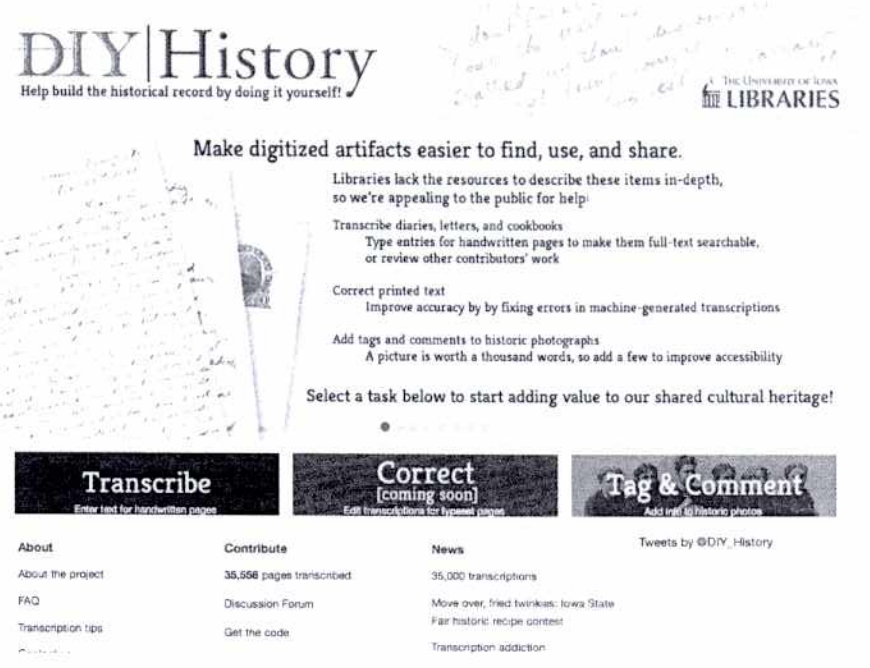

Figure 4.9 **DIY History website**

Civil War.[25] On a somewhat smaller scale, the College of William and Mary's Swen Library Digital Projects site uses *Omeka* and *Scripto* to add crowdsourced transcription to its materials. Their Special Collections have digitised materials from the Civil War period for the 'From Fights to Rights: The Long Road to a More Perfect Union' project and from a number of collections related to the College itself to build a site that enables the transcription of nearly 3,800 items.[26] The site 'Texas Manuscript Cultures' is using *Omeka* and *Scripto* to offer crowdsourced transcription of a variety of handwritten manuscripts related to Texas social and cultural history before 1950. Currently there are over 75 manuscripts available for transcription, and the site is supported by staff from a wide range of Texas libraries and historical societies.[27] Finally, Dr William B. Hafford, an archaeologist at University of Pennsylvania, has launched 'Crowdsourcing Ur', a site to solicit public assistance in transcribing the documents related to the excavations of Ur in Mesopotamia. The joint expedition of the British Museum and the University of

25 University of Alabama Libraries' 'Transcribe', http://transcribe.lib.ua.edu/.

26 College of William and Mary's 'Swen Library Digital Projects', http://scrcdigital.swem.wm.edu/.

27 'Texas Manuscript Cultures', http://writingstore.com/txmsc/.

Pennsylvania Museum occurred between 1922 and 1934, and resulted in a cache of documents about the excavations.[28]

Together, these examples show the early work that the *Scripto* extensions have enabled for organisations with document collections. But community transcription need not be limited to documentary collections. Since *Scripto* was designed to be as flexible as possible, projects could implement it to assist with the transcription of any range of file types, including audio and video files. This flexibility points to a bright future for the tool and for the range of projects that might choose to engage their users and constituents in contributing to making cultural heritage more accessible. In turn, this engagement can lead to a strengthening of the bond between core audiences and cultural heritage institutions. And, if those institutions learn as much about their volunteers as the PWD editors learned about their transcribers, they will be in good stead to continue to develop programs and applications that address their users' needs and interests.

References

Brennan, Sheila A. and T. Mills Kelly. 'Why Collecting History Online is Web 1.5', March 2009. http://chnm.gmu.edu/essays-on-history-new-media/essays/?essay id=47.

Brumfield, Ben. 'Crowdsourcing IMLS WebWise 2012'. *Collaborative Manuscript Transcription*, March 17, 2012. http://manuscripttranscription.blogspot.com/2012/03/crowdsourcing-at-imls-webwise-2012.html.

Causer, Tim, Justin Tonra and Valerie Wallace. 'Transcription Maximized; Expense Minimized? Crowdsourcing and Editing *The Collected Works of Jeremy Bentham*'. *Literary and Linguistic Computing* 27, no. 2 (2012): 119–37.

Cohen, Daniel J. 'The Future of Preserving the Past'. *CRM: The Journal of Heritage Stewardship* 2, no. 2 (2005): 6–19. http://chnm.gmu.edu/essays-on-history-new-media/essays/?essayid=39.

Cohen, Patricia. 'Scholars Recruit Public for Project'. *New York Times*, December 28, 2010. http://www.nytimes.com/2010/12/28/books/28transcribe.html.

Crackel, Theodore J. 'The Common Defence: The Department of War, 1789–1794'. *Prologue* (Winter 1989): 331–43.

Daines, J. Gordon, III and Cory L. Nimer. 'The Interactive Archivist: Case Studies in Utilizing Web 2.0 to Improve the Archival Experience'. Society of American Archivists, May 18, 2009. http://lib.byu.edu/sites/interactivearchivist/.

Evans, Max J. 'Archives of the People, by the People, for the People'. *American Archivist* 70, no. 2 (2007): 387–400.

Holley, Rose. 'Crowdsourcing: How and Why Should Libraries Do It?'. *D-Lib Magazine* 16, no. 3/4 (2010). Available at http://www.dlib.org/dlib/march10/holley/03holley.html.

28 'Crowdsourcing Ur', http://urcrowdsource.org/omeka/.

Hopkins, Curt. 'Crowdsourcing the Preservation of the U.S. War Papers'. *Read, Write, Web*, March 18, 2011. http://www.readwriteweb.com/archives/crowdsourcing_us_war_papers.php.

Howe, Jeff. 'The Rise of Crowds'. *Wired* 14, no. 6 (2006). http://www.wired.com/wired/archive/14.06/crowds.html.

Kalfatovic, Martin et al. 'Smithsonian Team Flickr: A Library, Archives, and Museums Collaboration in Web 2.0 Space'. *Archival Science*, October 2009. http://dx.doi.org/10.1007/s10502–009–9089-y.

Owens, Trevor. 'Meanification and Crowdscaffolding: Forget Badges'. *Playing the Past*, March 17, 2011. http://www.playthepast.org/?p=1027.

Rosenzweig, Roy. 'Can History Be Open Source? Wikipedia and the Future of the Past'. *Journal of American History* 93, no. 1 (2006): 117–46. http://chnm.gmu.edu/essays-on-history-new-media/essays/?essayid=42.

Springer, Michelle et al. 'For the Common Good: The Library of Congress Flickr Pilot Project, Final Report', October 30, 2008. http://www.loc.gov/rr/print/flickr_report_final.pdf.