Wellcome Library Transcribing Recipes Project: Final Report

Ben Brumfield and Mia Ridge

September 2015

Table of contents

Executive summary

The Wellcome Library seeks to run an innovative crowdsourced transcription project. In the first phase, the source material will be culinary and medicinal recipe books, in both print and manuscript formats. The project will also include c. 350 volumes of printed recipe books, and aims to include an equivalent number of books held by other institutions. The collection is mostly in English but includes European languages such as French or German. The overall goal is to create a platform and processes that can be used 'for any handwritten or early printed content'.

The 100,000 folios of manuscript recipes are only part of the corpus of material that is not suitable for Optical Character Recognition (OCR) held by the library. Future stages of the project will include other manuscript material. This creates a requirement for a robust system, able to integrate with other systems that may be swapped out or change over time, and with the ability to refresh the graphic design and/or crowdsourcing tasks as design requirements change over time.

Key findings

Challenges for the project include the separate end-to-end workflows required for print and manuscript sources, the integration of interfaces focused on specific tasks (such as full-text transcription, OCR correction or tagging) into a cohesive whole, and the integration of the resulting data into public-facing interfaces as appropriate.

In the current environment, successfully streamlining data flows within the Library's technical infrastructure to support end-to-end integration may be an innovative outcome, as an end-to-end transcription system is one that many current projects are yet to achieve. The project is an opportunity to optimise infrastructure for timely, automated data flow between systems, particularly in terms of the user experience for volunteers (e.g. they can quickly see the difference their contributions make reflected in the improved discoverability of material) and general users of the collections.

As such, we have recommended a staged approach. A staged process, broadly based on delivering one task at a time, should include the development of an end-to-end system as one of the first stages, and allow time for relationship with partners to develop at a realistic pace.

We have identified manuscript transcription as the essential data requirement for the Library, with OCR correction and indexed terms as important outputs, and translation and mark-up of textual insertions and deletions as ancillary outputs. Based on these priorities, we have surveyed the field of crowdsourced transcription tools and identified three very different tools which will satisfy essential outputs and provide a pathway for important and ancillary tasks.

- DIYHistory provides the quickest pathway to plain-text transcription and the best researcher support.
- FromThePage offers the broadest support for the Library's data requirements.
- Mirador matches the technical trajectory and organisational competencies of the Library's personnel best.

Each of these platforms has gaps in functionality that may need to be addressed through further software development or by passing their output to other systems to enable different tasks. In addition, they vary widely in readiness, pedigree, and availability of support.

About this report

The field of non-commercial crowdsourcing is relatively dynamic, with several crowdsourcing platforms now available, and with new projects launching during the writing of this report. A substantial body of previous practice from cultural heritage and citizen science, particularly public involvement in data transcription, categorisation and annotation, provides lessons for project structure, interface and task design, and support for volunteer communities. In addition to previous analysis and discussions with staff and researchers working on related projects, we have conducted peer interviews and drawn on formal and informal publications for this report.

We have used the term 'volunteers' to describe participants in non-profit crowdsourcing, in recognition of the voluntary nature of their contributions, and the subsequent importance of understanding their motivations and preferences. Following earlier Wellcome Library publications,[1] we have used the terms 'searching' for discovery interfaces such as catalogues, and 'viewing' for interfaces that display items. While terms such as annotation, mark-up, tagging and classification are sometimes used interchangeably, an important distinction is that classifications/tags can be applied at the page level (for example, as metadata about the digital image of a page, while mark-up refers to XML-style tags wrapped around transcribed text, and annotation refers to data linked to specific regions of an image. Indexed terms can be created through inline mark-up or through the selective transcription of text into database fields. Optical Character Recognition for printed books, and Handwriting Text Recognition for manuscripts, are referred to as OCR and HTR in this report. To match future usage, we have used the term 'Universal Viewer' to refer to the Wellcome Player. Other terms are outlined in our evaluation of the data requirements.

Following the project Brief, this report contains a critical review of the Library's high-level plan, stated outputs and outcomes; an evaluation of the data requirements; analysis of the software tools that most closely match the Library's requirements; a gap analysis; discussion of possible risks; and recommendations for contacts.

[1] Henshaw, Christy, and Robert Kiley. 'The Wellcome Library, Digital.' *Ariadne*, no. 71 (July 2013). http://www.ariadne.ac.uk/issue71/henshaw-kiley.

Critical review

This review includes both formal requirements outlined in the Consultancy Brief and additional requirements that emerged during the stakeholder interviews.

Sources used in this review of the project include five stakeholder interviews with eight Wellcome Library staff and one external agency representative. Five peer interviews with three organisations working on relevant crowdsourcing projects – the *Szathmary Culinary and Cookbooks* at the University of Iowa Library, the New York Public Library's *What's on the Menu* and *Ensemble,* and the Folger Shakespeare Library's *Shakespeare's World* – were conducted for this report. Two additional peer interviews were conducted with technical staff working with manuscript transcription and encoding technologies (the *Shelly-Godwin Archive* and Mirador*)*. Additional information was obtained from internal documents, email discussion, vendor websites and the Library website. The interview details are outlined in Appendix A: List of interviews, and referenced in this document as e.g. SI-2 for Stakeholder Interview 2.

Stated outputs and outcomes

The supplied Consultancy Brief lists the following outputs and outcomes:

- 100,000 manuscript pages digitised/identified
- 2 million printed book pages digitised/identified
- Transcription of 100,000 manuscript pages, and up to 500 early printed books
- User-friendly transcription interface
- Auto-transcription tools
- End-to-end crowdsourcing transcription system
- Infrastructure to provide discovery and access to content

The number of pages digitised and identified will depend in part on the relationships with external partners created through the project, and in part on the capabilities of the infrastructure around the crowdsourcing platforms to ingest material from other collections.

Transcription, mark-up and annotation of print and manuscript pages

The transcription outcomes are discussed in more detail in our evaluation of the data requirements below. Other outcomes listed in the Brief include the annotation of 'marginalia, corrections, strikethroughs, and illustrations'. Interviews with collections staff (SI-2) discussed the range of other material found within recipe books, including literature, poetry, hymns, family trees, annotations in different hands, 'plant matter, fabrics and other additions to the manuscript surface', suggesting additional potential items for annotation.

In addition to ingredients, quantities, timings, and methods/instructions, some recipes contain information on dosages, instructions, and the duration of medical treatments. Modern attempts to recreate culinary recipes might also lead to new annotations regarding methods, timings, ingredient substitutions, etc. Recipes may never be 'complete' in the sense that annotations or mark-up with additional information could be added over time, even after transcription is complete.

The interviews suggested a range of additional domains of interest within the documents, including diseases and names of people, places or species. Researchers with prior experience could provide additional classes of information including variant spellings and synonyms for diseases, methods or food substances. Researchers could potentially add links between individual recipes/books and records held in local record offices etc.

This rich set of possible annotations presents some challenges. Given the range of material within the corpus, ideally the tagging interface would be able to dynamically build term lists (project vocabularies), preferably using pre-existing cataloguing terms or Medical Subject Headings (MeSH)[2] for the initial list, and adding new terms as necessary. These dynamic tagging vocabularies may be a new requirement for most systems, and therefore presents an opportunity to fund innovative new features.

The Brief states that the results of any annotation, mark-up or classification processes should also be available through a single discovery system. As the stakeholder interviews revealed, this may require some adjustments to current digital systems.

User-friendly interfaces

A requirement from the Brief is that the interface should encourage a 'high volume of users to participate', creating 'useful and accurate data capture', and enhance the sense of community around the collections. Discussion with internal stakeholders further highlighted the importance of providing well-designed task interfaces to ensure a good user experience. Existing projects mentioned as good models include the Smithsonian's *Transcription Center*, the New York Public Library's *What's on the Menu* project, and the New York Times *Madison* project. Characteristics of these projects positively mentioned by Wellcome Library staff include: task simplicity; a lack of barriers (such as compulsory registration) before trying tasks; exposure to interesting content; a feeling that contributions are useful (best conveyed through both text and design elements); the provision of a range of types and/or topics of material; clear calls to action and direct access to tasks; personal/overall progress indicators; good project titles and descriptions; useful help (ideally short but sometimes necessarily longer); and sophisticated graphic design. These attributes echo other research and emerging best practice in UX for cultural heritage crowdsourcing, and established heuristics for web design.[3]

---

[2] http://www.nlm.nih.gov/mesh/

[3] For example, Lascarides and Vershbow discuss factors in the success of *What's on the Menu* and two PhD theses (Ridge, McKinley) have investigated potential heuristics for crowdsourcing in cultural heritage, supplementing existing website design heuristics. The growing body of literature on motivations for initial and sustained participation in voluntary crowdsourcing also provides insights into successful designs (for example, Rotman et al.). Much of the following section is based on consultant Ridge's previous research and a review of relevant projects for this report.

Lascarides, Michael, and Ben Vershbow. 'What's on the Menu?: Crowdsourcing at the New York Public Library.' In *Crowdsourcing Our Cultural Heritage*, edited by Mia Ridge. Digital Research in the Arts and Humanities. Farnham, Surrey, UK: Ashgate, 2014. http://www.ashgate.com/isbn/9781472410221.
McKinley, Donelle. 'Design Principles for Crowdsourcing Cultural Heritage: PhD Findings Report', July 2015. http://nonprofitcrowd.org/crowdsourcing-heuristics.
Nielsen, Jakob. '10 Usability Heuristics for User Interface Design', 1995. http://www.nngroup.com/articles/ten-usability-heuristics/.
Ridge, Mia. 'Making Digital History: The Impact of Digitality on Public Participation and Scholarly Practices in Historical Research.' Ph.D., Open University, 2015.

Other attributes of successful projects include 'mini projects', challenges and task 'ecosystems'. In 'mini projects', collections are divided into smaller sets (e.g. book by book or archive box by box). The natural history collections project *DigiVol*[4] groups material into 'expeditions', while the *Smithsonian Digital Volunteers: Transcription Center*[5] breaks material into 'projects' ranging in size from 1 to 500 pages. Dividing material into smaller groups has several advantages. Documentation can be tailored for the specific requirements of a group of documents. The descriptions for each project can include specific place, species, event or people's names, dates, and original purpose of the documents, providing specific hooks that may be more likely to grab a volunteer's attention. Smaller projects mean that any contribution is comparatively larger, and they are completed more quickly, providing many opportunities to celebrate the accomplishments of volunteers and thereby encourage others.[6]

Challenges (or 'missions') are activity drives based on targets set by crowdsourcing projects. Challenges usually have a target goal to be reached by a specific time - for example, completing a transcription task by a historic anniversary. Challenges are also a good way to focus on specific tasks that might not be part of the usual site activity. For example, *Ravelry* (a site for 'knitters and crocheters') held a highly successful week-long 'party' to enter metadata to support a structured search function.[7] It should be noted that both mini-projects and challenges require resources to set up and promote, and that mini-projects may place additional demands on the project administration interfaces.

In task ecosystems, related applications are combined into a single workflow to process different aspects of the same source materials. For example, the New York Public Library's *Building Inspector* offers five tasks based on historical maps, each embedded in an interface dedicated to that specific task, whether entering street numbers, classifying colours or transcribing place names.[8] Specialised interfaces seem to encourage greater participation by enabling volunteers to focus on one task at a time - for example, in specialist interfaces for transcribing text, annotating non-recipe items on a page, marking up words or phrases within text (such as ingredients or methods), validating previous transcription or correcting OCR/HTR text. The inclusion of simple tasks such as image classification also allows non-specialists to become familiar with the material before they try more complex tasks like transcription or mark-up. A task ecosystem may be particularly suited for the Recipes project, allowing specialist interfaces to be

---

Rotman, Dana, Jennifer Preece, Jennifer Hammock, Kezee Procita, Derek Hansen, Cynthia Parr, Darcy Lewis, and David Jacobs. 'Dynamic Changes in Motivation in Collaborative Citizen-Science Projects.' In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, 217–26. Seattle, 2012. doi:10.1145/2145204.2145238.

[4] http://volunteer.ala.org.au/ DigiVol also awards volunteers ranks on an expedition, the highest of which is 'Expedition Leader'

[5] https://transcription.si.edu/

[6] See for example Guralnick, Rob. 'Making Progress Clear on Notes from Nature.' *Notes from Nature*, 24 February 2014. http://blog.notesfromnature.org/2014/02/24/making-progress-clear-on-notes-from-nature/.

[7] rainydaygoods. 'Ravelers Rocked the Search Party!' *Unraveled*, 20 July 2010. http://blog.ravelry.com/2010/07/20/ravelers-rocked-the-search-party/.

[8] Task ecosystems can include computational processing such as OCR and HTR. In the case of *Building Inspector*, the maps have already been computationally processed to find probable building footprints and other attributes.

developed without creating an overly complicated task interface. As discussed previously, the source materials support many different crowdsourcing tasks including OCR or HTR correction and text mark-up with terms from various specialist domains. For example, documents may be pre-classified through a Zooniverse-style labelling task, allowing volunteers to highlight and annotate non-recipe items within documents, to rate them by relative difficulty (based on e.g. language, image legibility or the style of handwriting) or to geolocate place names mentioned. The staged development process we suggest would also support the incremental development of a task ecosystem.

A key factor in selecting crowdsourcing tools is the suitability of the user experience and task interfaces for both the source materials and the interests and motivations of typical volunteers. For example, volunteers may become attached to particular authors, locations, etc., and want to work more closely on that material rather than be assigned the next pages from a random queue of material. Some volunteers may prefer to collaborate on a single transcription, making successive passes until the best possible transcription has been recorded. Volunteers learning to transcribe particular hands, or working on subsequent pages of a book may strongly wish to 'go back' and edit their transcription or tags as they improve their skills or encounter contextual information on subsequent pages. The Recipes crowdsourcing platform must allow readers to browse back and forth through documents (for example, to read a recipe that spans one or more pages, or to follow a household over time). Ideally, volunteers should be able to manipulate document images as necessary, including rotating, zooming, and adjusting image contrast.

An emergent requirement was that public-facing sites must be fully responsive (SI-1). However, it is important to note that while viewing interfaces should be responsive, some crowdsourcing tasks are less suited to the limited screen space on mobile devices. Finally, ideally project administration interfaces (the systems used to set up a new project by importing a set of materials, creating documentation and other text and promoting it on the site) should also be as user-friendly as possible.

Auto-transcription tools

The integration of auto-transcription via Optical Character Recognition (OCR) for printed books and Handwriting Text Recognition (HTR) for manuscripts has been suggested as part of the project Brief. While OCR correction is a potential crowdsourcing activity, it is worth mentioning here that the OCR functions currently provided by the Library's digitisation partnership with the Internet Archive are optimised for modern printed materials, and do not adequately support blackletter fonts like Fraktur[9] or early Antiqua or Italic typefaces. Significant improvements in OCR quality over older materials might be attained by re-executing OCR via Tesseract, trained according to the eMOP methodology.[10]

For automated transcription of manuscripts via HTR, we expect the capabilities of these tools to improve considerably in future, as recent years have demonstrated progress

[9] The Internet Archive uses ABBYY FineReader through a white-label arrangement with LuraTech which limits the configuration options for the OCR algorithm. (http://archive.org/post/299315/ocr-on-fraktur)
[10] eMOP—the Early Modern OCR Project—has had real success performing OCR on 16th and 17th-century printed English books by building a specialised training regimen for the open-source Tesseract OCR software. (http://emop.tamu.edu/)

unforeseen even five years ago. At the time of writing, the HTR tools available are not suitable for public-facing projects,[11] but a continued dialogue with HTR researchers and developers could be facilitated by providing 'ground-truth' transcripts, which are essential to research in the field.

End-to-end crowdsourcing transcription system

The Brief defines end-to-end as 'the entire workflow from uploading images of text, to outputting the final data in the required data formats and standards'. Achieving this may require some changes to existing workflows (using existing systems such as Goobi where possible) and/or the development of new components. The end-to-end system will also need to integrate with preservation systems. The workflows between internal systems, or existing public searching or viewing interfaces should also generate outputs such as downloadable CSV files or a public-facing API.[12] The Library's systems will also change as the Digital Library Cloud Services (DLCS) platform is developed,[13] so the final project will need to adjust accordingly.

Stakeholder interviews discussed the challenges of moving data through Library systems. Challenges include the gradual, 'improvised' growth of the network of Library systems over time, and the relationship between the public-facing discovery system, Encore (which is maintained by a third-party vendor with many other clients) and other Library systems. The Library has more input into viewing interfaces once the user has been passed off to the Universal Viewer,[14] as changes can be commissioned from Digirati (with associated timeline and cost implications). Tags and full-text transcription could presumably be indexed by Encore for searching, and the Universal Viewer can currently 'search within'[15] full-text held in the Digital Delivery Service (DDS). Outputs such as transcriptions should be displayable in the Universal Viewer (not least because it would make earlier manuscript texts legible to non-specialists) but a specialised interface for browsing and searching recipes might be the best way to integrate structured data such as marked-up and tagged recipes while leveraging the ability to embed the Universal Viewer in other pages.

Our peer interviews revealed that – intentionally or otherwise – crowdsourcing interfaces tend to become a viewing interface for the collections they contain. Participatory interfaces tend to receive increased traffic, attention and inbound links driving Google PageRank. This is to be embraced, as chance discovery draws in volunteers who bring their own expertise in niche areas.[16] The Smithsonian

---

[11] See, for example, the *Transkribus* manual's opening statement: 'Transkribus is an **expert tool**. As with other feature-rich software, it is designed for users who "know what to do and how".' (Emphasis in original.) This suggests that the most promising HTR tool is not yet ready for novice users, although this may change in the next few years. (*How to use TRANSKRIBUS – a very first manual,* https://transkribus.eu/Transkribus/docs/How%20to%20use%20TRANSKRIBUS-0.1.7.pdf)
The *From Quill to Bytes* project at Uppsala University (http://www.it.uu.se/research/project/q2b) is still at an early stage of development.

[12] Comma-separated values (CSV) file store spreadsheet-style tabular data in plain text. Application programming interfaces (APIs) provide computational access to data sources.

[13] Kiley, Robert. 'Moving the Wellcome Library to the Cloud.' Wellcome Library, June 16, 2015. http://blog.wellcomelibrary.org/2015/06/moving-the-wellcome-library-to-the-cloud/.

[14] Following aspects of the data model documented at http://player.digirati.co.uk/data-model.html

[15] As documented at http://player.digirati.co.uk/search-within.html

[16] In one example, a super-volunteer first discovered a project by Googling his own name and finding transcripts mentioning a person – the diarist's postman, and the volunteer's great uncle – of the same

Transcription Center provides access to completed transcriptions and zoomable images of documents (as well as automatically-generated PDFs containing collections information, the original image and transcribed text) as well as to documents yet to be transcribed. However, while the site can be searched via external search engines, the site does not provide a built-in search. The challenge for a crowdsourcing tool then becomes providing a polished, focused task interface(s) while also supporting item discoverability (for example, through faceted browse and search) and strong links between the main Library catalogue search and display interfaces. We have taken this requirement into consideration when reviewing tools.

Infrastructure to provide discovery and access to content

Our stakeholder interviews suggest it should be possible to get data into the Library's Encore system for use in search discovery systems (for example, through the current ingestion process), but any weighting or structure in the data would be lost in the process, and that the discovery interface is unlikely to change in the short or medium term. If variant names (for diseases, ingredients, etc.) are incorporated into the search/discovery system, the viewing interfaces should display both the term used in the original document and the variant term, allowing the user to understand how variant names might have affected their search results.

The recipes, once transcribed and thus made more discoverable and accessible, could have a broad reach beyond existing users of the collections. However, some stakeholder interviewees expressed concern that the current discovery interfaces (such as the Encore catalogue search) were not easy to use and might not support the needs of casual and infrequent users. Ideally, the interface would also support the serendipitous discovery of related material (SI-2).

Creating mark-up in machine-readable formats, such as schema.org's Microdata, RDFa or JSON-LD[17] might help make material more discoverable in external search engines such as Google. This would require integration of the mark-up task into the crowdsourcing platform, and the incorporation of schema.org mark-up into the display interfaces. Ideally, a robots.txt file and sitemap would help search engines find the records.

The viewing interfaces (such as the Universal Viewer) will need to be updated to either link to specialist interfaces or to display data, such as tags or mark-up, created through crowdsourcing tasks. The search and viewing interfaces should ideally be able to detect when an item is available in the crowdsourcing system and present links to that interface. Similarly, the crowdsourcing system should link back to the main search/discovery systems to allow people to find records not available in the crowdsourcing system. This may require modifications to the DDS and crowdsourcing systems to incorporate links via the persistent, unique ID assigned by Sierra.[18] The ability to easily move between the viewing interfaces and the transcription system would enable discovery of additional interesting texts by transcribers, and draw

---

name. See 'Recruitment' in Brumfield, 'Best Practices for Engaging the Public at CCLA': http://manuscripttranscription.blogspot.com/2015/05/best-practices-at-engaging-public-at.html

[17] https://schema.org/Recipe, https://schema.org/Drug, https://schema.org/MedicalEntity

[18] As discussed in Henshaw, Christy, and Robert Kiley. 'The Wellcome Library, Digital.'

researchers to the transcription volunteer base. These issues are discussed further in Appendix B: System Integration.

It can be difficult to predict the 'long tail' of specialist uses of collections data, especially when it is exposed to a range of people through a crowdsourcing project. It may not be possible to develop interfaces or tasks that meet every specialist need, but the provision of downloadable data should allow researchers to use the data as they wish. Possible audiences beyond medical historians include people who might use the recipes to study domestic, science, trade, empire, local histories, linguistics, and those researching particular houses or families. Contemporary chefs or bakers researching and re-creating historical recipes might also use the resulting transcriptions. Audience research with scholars who already use the collections could be fruitful, as some have already created their own databases from the recipes (SI-2).

Other outcomes

The goal to create a platform that other institutions can use to transcribe and enhance their material entails further outputs and outcomes that should be listed in the project plan. Some external institutions may have limited technical or financial resources and hold only one or two recipe books; others might have substantial collections of their own and existing infrastructure requirements. The uptake of 'crowdsourcing as a service' will require specialist community management and documentation, including topics such as licensing, technical ingest requirements, data export options, expert input into community discussion, etc. The use of existing platforms, such as the Internet Archive, may alleviate some data management and documentation issues.

The use of existing standards should help other institutions that wish to use data generated through the project. The desire to include material from other institutions means the platform should not be too tightly coupled with the Library's own systems. This would also allow the Library to swap out parts of their infrastructure as necessary in the future, particularly as the project expands to include other types of material in future iterations.

In the peer interviews, the DIYHistory developers warned of the costs of supporting open-source software, whether developed from scratch or built on top of existing platforms. This may not be an issue for the Wellcome Library, which has a history of partnering with contract developers who may view support, extension, and customisation of open-source software as a business opportunity. For institutions whose development staff traditionally only support internal users, however, open-source software support is a serious distraction.

Finally, data sustainability and preservation is an outcome underlying the entire project. This will require decisions about the most appropriate system in which to store the canonical transcriptions and other data. Some possible solutions are presented for discussion in Appendix B: System Integration.

High-level plan

Our review included the Draft project plan presented in Annex 3 of the Consultancy Brief. An analysis of the stakeholder interviews and the project requirements highlights the importance of a staged development approach within the broad project plan.

This staged development process should include interim deliverables to make sure the internal data workflow is fit for purpose, so that validated input from volunteers quickly flows into search and viewing interfaces. The requirement to create appropriately effective and polished user experiences means that iterative development and testing (whether through paper prototypes, beta software or staged releases) is more likely to yield a successful project. A staged approach, focusing on one core goal or task in each phase, allows the Library to apply lessons from previous stages to the development of the next and to take advantage of improvements in related technologies over time. The Wellcome Library's Sandbox for experimental tools[19] suggests an open beta would be positively received.

The plan includes identifying key audiences, but our review suggests that identifying and communicating external stakeholders, including existing academic and other users of the collection, will also be important. These external stakeholders may be able to provide valuable insights into their uses of the collection in teaching or research,[20] may be able to help publicise the project or provide feedback on prototype interfaces. Some, such as the Early Modern Recipe Online Collective (EMROC)[21] may already have transcriptions and information they could share. These transcriptions, and the lessons these academics[22] learnt while transcribing with their students, might form the basis of the initial transcription standards and help documentation. Other external stakeholders and potential champions include people interested in making historical recipes, and institutions with related collections (including other recipe collections, the botanical collections at Kew Gardens, domestic collections and public programmes at the Geffrye Museum and academics at British universities such as Queen Mary University of London and the University of Westminster. This coordination will also have resourcing implications and may need to be considered when defining the 'community moderator' role. Planning the first timelines and outcomes for the staged process in consultation with various internal and external stakeholders, and taking into account the current state of related internal projects and external tools, should be one of the first tasks of the appointed project manager.

Some User Experience (UX) principles and audiences were discussed during our Stakeholder Interviews. Creating related use-cases would provide useful milestones for staged delivery process. We would also recommend that a Marketing and Outreach Plan and UX Guide be created early and reviewed as the project progresses. Coordinating Outreach and UX plans is important, as expectations about the tasks or typical materials created by marketing material will affect potential volunteers' experience of the crowdsourcing interfaces. Crowdsourcing projects have many stages, each of which should be represented in the UX Guide. Important stages include beta testing/early release, volunteer recruitment and onboarding, volunteer retention, provision for volunteer skills development, community management, and gracefully ending different stages of the project. The Outreach and UX plans should also include methods for rewarding and recognising volunteers, including super-users, those who take on non-

[19] http://wellcomelibrary.org/what-we-do/sandbox/
[20] While external scholars may already have their own bespoke databases based on the recipes, incorporating them is presumably out of scope for the project at this stage.
[21] http://emroc.hypotheses.org/ and the related http://recipes.hypotheses.org/
[22] SI-2 also suggested Claire Williams at QMUL who had worked on transcriptions with students.

core tasks such as researching or re-creating recipes, and scholarly users who may contribute variant terms and other references.

Budget

The proposed budget of £550,000 seem more than adequate for staff time (project manager, community moderator) associated with the project, as well as for crowdsourcing platform deployment, customisation, and system integration. However, we have no expertise in relevant digitisation costs, so cannot comment on the overall budget.

Evaluation of data requirements and crowdsourced outputs

Crowdsourced transcription is a broad category, which in the past has included volunteers in records offices entering data on spreadsheets and mailing them to genealogy organisations on CD-ROMS, scholars of ancient Greek palaeography using online XML editors to add transcribed papyrus fragments to a Git repository, and amateur military history enthusiasts drawing rectangles on page facsimiles to tag regions of an image with proper names. As such there is no standard definition of the product of a crowdsourced transcription project. Beginning with the Library's target sources – manuscript images and their metadata or OCR output files – we have analysed the possibilities in light of the stakeholder interviews, and classified the required data formats as *Essential*, *Important*, and *Ancillary*.

Essential Outputs

Full-text transcripts

Full-text transcripts of the manuscript recipes are the most important deliverable of the project, according to stakeholders. Such transcripts would enable the manuscripts to be findable via full-text search in the Library's discovery systems, would enhance the viewing experience in the Universal Viewer, and would allow text re-use by researchers.

Important Outputs

Indexed terms

Several stakeholders mentioned the desire for indexes of terms contained within the text: ingredients, maladies, preparation mechanisms, etc. Capturing these terms, their variants, and their association with individual pages of recipe documents would allow the Library to improve its authority files, support faceted discovery (i.e. "show me the recipes which mention 'potatoes' along with 'gout'"), and allow researchers to analyse the recipes as a structured dataset. Linking the terms to specific regions of a page image is less important to stakeholders than linking the terms to their context within a document transcript.

Clean OCR

Unlike manuscript material, the printed books the Library has digitised through the Internet Archive already have text associated with them. In many instances this is already adequate for full-text searches in the catalogue system or 'search within' functionality within the Universal Viewer. Clean transcripts produced by correcting the OCR provide an incremental improvement to both of those experiences, as well as allowing text re-use by researchers.

Ancillary Outputs

Translations

Many of the Library's recipes are not written in modern English. Asking users to translate recipes in foreign languages – particularly those widely studied in the UK – would produce modern English translations which could be used in discovery, delivery, and re-use. However, translations were not mentioned as a primary goal, and working with languages not widely studied in the UK presents challenges for volunteer outreach.

Genetic Mark-Up

The ways in which a document has changed over time – strike-throughs, insertions, marginalia, later additions in different handwriting – are represented by scholarly editors in 'genetic editions', which trace the genesis of a text.[23] While stakeholders and the project Brief have expressed interest in supporting genetic features within crowdsourced transcripts, and while excellent tools are available to encode such features, the Library does not have any system for integrating such features into search or viewing interfaces. If researchers cannot search for e.g. all texts which contain struck-through passages mentioning potatoes, then there may be little benefit to having volunteers encode strike-throughs. (See *Gap analysis* below for more detail.)

Handwriting Recognition 'Ground Truth'

The field of handwriting recognition (OCR for handwritten texts) is developing rapidly, with newly available tools like the tranScriptorium project's *Transkribus*. Stakeholders have indicated an interest in aiding such efforts, which may be best accomplished by making available 'ground truth' datasets – human-created transcripts associated with a particular image. Text-to-image linking on a word level may be particularly useful, but even plaintext transcripts associated with their page facsimile can contribute to HTR research.

Other Tool Requirements

Emergent requirements for the platform revealed during our review include:

- It must support Google Analytics.
- It should be archiveable by the Internet Archive and the British Library's UK Web Archive.

[23] 'Genetic mark-up' is certainly an oversimplification, as it conflates encoding genetic features of a text (strike-throughs, insertions, later marginal annotations) with other types of text encoding (abbreviations and their expansions, normalised and verbatim spelling) using the same kinds of tagging mechanisms described in the TEI Guidelines. However, as all of the specific encoding goals mentioned by stakeholders were genetic in nature we have chosen the term to clearly differentiate this data output requirement from other kinds of textual annotation.

Results of stakeholder and peer interviews

Stakeholder interviews were conducted in the Library's offices in London by Mia Ridge, and peer interviews were conducted over videoconference and telephone by Ben Brumfield. Discussion relevant to other headings has been integrated into the body of the report.

Stakeholder Interviews

Stakeholder goals for the project were in broad agreement, with variations in emphasis depending on their role within the organisation. Public-facing goals included encouraging a broader range of users, surfacing less well-known material, encouraging serendipitous discovery of items within the Library's interfaces, providing full-text search, and allowing people to access entire manuscripts. SI-1 emphasised that the transcribed text should be of a high quality. Given the richness of the material, tagging and/or annotating processes may continue once transcription is 'complete', as researchers bring different perspectives to the documents.

Information on the Library's digital infrastructure was provided in SI-1, SI-3, SI-4 and SI-5. Stakeholder goals for support public uses of the collections include streamlining internal systems to support a high-quality user experience. This includes the ability to make transcriptions and various forms of marked-up and annotated records available in the public discovery and access interfaces in a timely manner. The importance of the International Image Interoperability Framework (IIIF)[24] also emerged during these discussions.

The upcoming Zooniverse project, 'Diagnosis London', was discussed in more detail in two stakeholder interviews (SI-4, SI-5). The Zooniverse project shows how much thought can be required when designing workflows and managing data outputs. It also shows the importance of planning for staff engagement in project communities, and will provide valuable lessons for the Recipes project. This project, due to launch in October, may encourage the Library to consider the most efficient, user-friendly and sustainable processes for integrating crowdsourced data with existing internal and public-facing discovery and access systems. However, as the data created through 'Diagnosis London' will be in a different format and have different uses from transcription data, some infrastructure integration issues are likely to remain unresolved.

Peer Interviews

The most important finding of the peer interviews is that the Library must plan for the crowdsourcing platform to be used as a research and public access tool in addition to a data-entry tool. If the crowdsourcing project is successful, substantial new web traffic to the Library will arrive first at the Recipes crowdsourcing project. Much of that traffic will be from researchers new to the Library who are interested in the material, but may also constitute an entirely different population from the volunteers arriving to transcribe. The project must take care that those researchers have access to the additional context provided by the Library's discovery and presentation systems so that they do not reach a dead end doing research on the transcription platform alone.

[24] http://iiif.io/

Another recommendation from peers was that the Library carefully considers the integration of the user-generated content into its existing systems. It is possible to run a popular crowdsourcing project without making effective use of its products in finding aids or digital library systems, but this should be avoided. The recommendation should influence the selection of crowdsourced activities – if the library cannot make use of a particular data product, it is unwise to ask volunteers to waste their time producing it.

Finally, the peers at culinary projects expressed great enthusiasm about the outreach possibilities for the Wellcome Library's recipe project. The documents are intrinsically interesting, the potential audience is large and passionate, and the opportunities for press coverage are very good. In particular, the 'odd stuff' – ingredients and preparations unusual to modern eyes – make for attention-grabbing stories[25] with lots of appeal on social media.

[25] For example, Calf's Brain Soup as a headline in 'DIY History crowdsources the transcription of 17th century cookbook' (http://www.wired.co.uk/news/archive/2012-12/03/open-source-culinary-history)

SWOT analysis of existing tools

Research conducted for the Interim Report suggested that the Library's priorities for data requirements follow a hierarchy of importance. Accordingly, in this SWOT analysis, we have focused on platforms capable of performing the essential requirement of plain-text transcription, but which have capabilities to satisfy the important and ancillary data requirements. Tools that do not adequately support full-text transcription but which may address other data requirements are suggested in the *Gap analysis*. Systems developed in-house, which either are not available for other projects or which have been released only nominally but never deployed outside of the sponsoring institution have not been included, however well-designed those systems may be.

A notable omission from the transcription tool SWOT analysis is the Zooniverse platform. Until 1 September 2015, no Zooniverse projects had attempted full-text transcription. With the launch of AnnoTate[26] and the upcoming launch of Shakespeare's World,[27] the Zooniverse group will apply their considerable talents to the problems involved in transcribing unstructured textual sources. However, the Zooniverse approach may be quite limiting when it comes to the kinds of source material the Library works with. Generally speaking, full-text transcription has been best approached by allowing multiple users to collaborate on a single text. While variations exist among full-text platforms, the ability for users to see each other's transcripts and discuss the changes among themselves has proven itself effective as a quality-control mechanism for units of transcription longer than a name.[28] In addition, transcription tasks are strictly queued with Zooniverse platforms – each user is presented with an untranscribed[29] page, and after they have transcribed that page they are unable to revisit it to modify their transcript.[30] Finally, the stakeholder interviews have emphasised the importance of the crowdsourcing platform as a researcher tool, which is incompatible with the Zooniverse approach of focusing on data entry.[31] However, the addition of this project to the field, and their line-at-a-time, queue-oriented, multi-track transcription workflow, will be a rich source of data about the pros and cons of, and volunteer preferences for, different approaches to transcription.

---

[26] https://anno.tate.org.uk/

[27] Unreleased as of 11 September 2015. Source code and alpha site are available at https://github.com/zooniverse/shakespeares_world

[28] For an analysis of quality control methodologies for crowdsourced manuscript transcription, see Brumfield, 'Quality Control for Crowdsourced Transcription', http://manuscripttranscription.blogspot.com/2012/03/quality-control-for-crowdsourced.html

[29] Perhaps more accurately, 'insufficiently transcribed' page: in Zooniverse terminology, subjects for classification need *N* independent users to classify (in this case transcribe) them before they are removed from the queue of work to be done.

[30] Contrast this with the 'Story of Page 19' in Brumfield, 'Collaborative Digitization at ALA', http://manuscripttranscription.blogspot.com/2014/07/collaborative-digitization-at-ala-2014.html

[31] That said, one of the primary benefits of a crowdsourcing platform that provides researcher functionality is the ability to attract new volunteers who are researching related material. If the Zooniverse approach does not provide this benefit in volunteer outreach, a Zooniverse collaboration is still able to guarantee a substantial number of volunteers from among the extremely large Zooniverse user base. This benefit may obviate the need to do marketing for a crowdsourcing project – essentially that task is outsourced to the enormous, growing, Zooniverse volunteer pool.

Methodology

We tested the tools discussed in this section by reading documentation, trying other instances of the software, and by installing the software and checking claims/requirements. These methods were supplemented with interviews conducted with the authors (PI-4, PI-6).

Transcription Tools

Given the Library's primary goal of capturing the full text of the culinary sources to be transcribed, the Wellcome's infrastructure and integration needs, and the experience of peer institutions, there are three tools from which the Library should select for the transcription phase of the Recipes project. We have selected three very different types of tools from among the score or so open-source transcription tools, believing that they represent the best of three different approaches the Library may take.

DIYHistory

DIYHistory was developed by the University of Iowa Libraries after the success of a prototype crowdsourced transcription project based on unrelated technology.[32] Based on the Omeka digital exhibition tool created by the Roy Rosenzweig Center for History and New Media at George Mason University, DIYHistory first used using Scripto[33] for their transcription engine and a complex set of scripts to integrate the platform with the digital library system ContentDM.[34] Over the next three years, the DIYHistory platform evolved considerably, replacing Scripto and MediaWiki with their own freestanding transcription plug-in. In addition, other projects have deployed DIYHistory successfully; sometimes making substantial enhancements to the code, such as the Library of Virginia's export plugin for transcripts,[35] or the Letters of 1916 project's addition of the Bentham Transcription Desk's TEI Toolbar to the transcription editor.[36]

Strengths of DIYHistory

- **Track record:** DIYHistory has been supporting crowdsourced transcription since 2012. It recently passed 60,000 pages transcribed on the University of Iowa installation,[37] and has been installed at a couple of dozen institutions (PI-4) most

[32] 'UI Libraries launches new crowdsourcing site with manuscript cookbooks and more', Jen Wolfe, October 2, 2012 (http://blog.lib.uiowa.edu/drp/2012/10/09/diyhistory/)
[33] Scripto (http://scripto.org/) is a crowdsourcing transcription platform built on 'connector scripts' which provide a pipeline between a content management system like Omeka and a MediaWiki installation. This hides the MediaWiki interface behind the content management system, retaining the CMS branding and taking advantage of the version control and discussion features of MediaWiki pages, although losing features like wiki-links. Development appears to have stalled on the platform, with the last software update in August of 2013 ('Update link in ini': https://github.com/omeka/plugin-Scripto/commits/master) and loss of support for Drupal and Wordpress content management systems in the middle of 2014 ('Wordpress install – media viewer?' thread on the Scripto Dev Google Group: https://groups.google.com/forum/#!topic/scripto-dev/zxBohTWonNc)
[34] For details on the initial DIYHistory deployment and technical details on its integration challenges, see 'DIYHistory: Scaling Up with the Crowd at the University of Iowa' by Shawn Averkamp, November 7, 2012 (http://ir.uiowa.edu/cgi/viewcontent.cgi?article=1217&context=lib_pubs). For a more general background see Averkamp and Butler, 'The Care and Feeding of a Crowd' Code4Lib 2013 conference (http://code4lib.org/conference/2013/averkamp-butler)
[35] http://www.virginiamemory.com/transcribe/code
[36] http://dh.tcd.ie/letters1916/
[37] https://twitter.com/diy_history/status/628585808170094592

prominently the Library of Virginia's *Making History*[38] and Maynooth University/Trinity College Dublin's *Letters of 1916*.[39]

- **Active development:** DIYHistory has been maintained and extended continuously since its deployment in 2012. Recent changes have removed the dependence on the MediaWiki platform and added experimental support for structured data transcription for specimen labels from natural history collections. (PI-4)
- **Design for crowdsourcing:** The DIYHistory project emerged as a replacement for a crowdsourcing prototype, and has been designed explicitly for crowdsourced transcription as part of library digitisation.
- **Extensible underlying platform:** The discovery and presentation layer of DIYHistory is built on Omeka, an open-source digital exhibit platform. Omeka has a vibrant community of users and developers contributing their expertise through support, tutorials, and a wide array of plug-ins. Leveraging these Omeka plug-ins offers the possibility to turn DIYHistory into a robust system for presenting and analysing the collection for researchers and volunteers alike.
- **Integration-friendly API:** The Omeka 2.1 release includes an API which provides relatively easy and complete access to metadata about items in a collection, including custom fields used by DIYHistory for transcripts. Similarly, the CSV import tool reduces the labour involved in loading documents into the crowdsourcing platform.
- **Support for OCR correction:** Page transcripts may be initialised with existing text extracted from OCR, although loading and formatting such text may require pre-processing.
- **Possible support for genetic mark-up:** Customisation similar to that performed by Letters of 1916 would allow mark-up of 'genetic' features like strike-throughs or later annotations.

Weaknesses of DIYHistory

- **Plain-text only:** DIYHistory supports transcription in plain text. No facilities exist to export transcripts as ALTO or Open Annotation
- **No indexed terms:** Any effort to index terms within the text would currently need to be performed on a totally different platform.[40]
- **No translation:** Any effort to crowdsource the translation of texts would need to run on a different platform.
- **No IIIF integration:** Support for IIIF is under discussion,[41] but does not yet exist.
- **No Internet Archive integration:** Culinary books scanned by the Internet Archive must be ingested into DIYHistory for OCR correction, rather than using the Internet Archive API.[42]

---

[38] http://www.virginiamemory.com/transcribe/

[39] http://dh.tcd.ie/letters1916/

[40] In late June 2015, a job posting on the Code4Lib jobs list proposed hiring a developer to add support for indexed terms into Omeka (http://jobs.code4lib.org/job/21635/) based on detailed requirements for 'linked references' among other features desired by the Jane Addams Papers Project (https://docs.google.com/document/d/1juUmXxT_rjUZblV_ZLTZalu8bd9iFY_VDPogBi4-Ydo/edit)

[41] https://github.com/omeka/omeka-s/issues/182

[42] The nucleus of such support does exist in the BookReader Omeka plug-in (https://github.com/jsicot/BookReader), which, however, will require substantial development effort to make suitable for transcription purposes.

- **Limited support:** Supporting other institutions that are using the DIYHistory code-base has imposed a substantial and unexpected burden on the sole developer at the University of Iowa working on the project (PI-4). The publicly released code does not reflect the current state of the DIYHistory code base at the University of Iowa, indicating limited resources for support.[43]

***Conflict of interest warning**: The FromThePage tool discussed below was created by Ben Brumfield, one of the authors of this report. While the software is open-source, Brumfield sells services (custom development, hosting, support and installation) related to the software. Aware that Ben's insight into his own project might put others at a disadvantage, we sought to alleviate this by contacting project staff from other platforms to discuss the current state and future plans of their projects.*

FromThePage

FromThePage was originally developed by Ben Brumfield as a collaborative tool for transcribing, indexing, and presenting family diaries. Released as open-source software in 2009, it has since been deployed for herpetology field notes, literary drafts, historic diaries, and punk-rock fanzines. The software is a purpose-built, stand-alone platform depending only on Ruby on Rails and MySQL, with optional integration for transcribing documents hosted on the Internet Archive or within Omeka exhibits. In the last two years, new development efforts have updated the user interface and upgraded the underlying technical frameworks, while adding TEI-XML as a transcription export format and support for additional user tasks like translation and OCR correction.

Strengths of FromThePage

- **Track record:** Launched in beta in 2008, FromThePage was deployed by San Diego Natural History Museum in 2010[44] and has since been used by several libraries, archives and museums.[45] The SDNMH project alone has transcribed 14,101 pages, identifying 12,584 terms indexed to 49,217 locations within the text.
- **Active development:** Substantial enhancements and updates have been made in the last two years culminating in a major release in September 2015. New features have been committed for autumn 2015[46] and early 2016.[47]
- **Indexed term support:** FromThePage was designed to create indexes of subjects mentioned within texts through a simple wiki-link mark-up. Support for

[43] https://github.com/ui-libraries/DIYHistory-transcribe/commits/master shows only eight commits, none of which include the removal of Scripto/MediaWiki mentioned in PI-4.
[44] http://fromthepage.bpoc.org/
[45] Although it is impossible to track installations of open-source software, the tool has been used in private projects by Northwestern University Libraries (2012) and Rhodes College (2012) and in public projects at Pennsylvania State University (2013), University of Delaware (2013), Museum of Vertebrate Zoology (2013), Biodiversity Heritage Library/Missouri Botanical Gardens (2014), Fordham University (2015), and University of Texas (2015).
[46] A project with Adam Rabinowitz and the University of Texas Libraries will develop evaluation tools for classroom use, allowing transcription project moderators (teachers, in this case) to view all contributions made within a time range, grouped by contributor. This effort will also implement the ability to easily roll back unwanted edits for a particular page.
[47] As part of a US National Endowment for the Humanities-funded effort by the New York Philharmonic Leon Levy Digital Archives to encode manuscript collections, linkages between external authority files and transcripts will be implemented in FromThePage.

variation in spelling and terminology is extensive, including CSV export of indexed terms, automated mark-up suggestion for transcribers, and index-based browsing.[48]

- **Translation support:** Documents can be designated for translation, providing a second workflow after transcription and indexing to create parallel text editions.[49]
- **Internet Archive integration:** Books scanned by the Internet Archive can be ingested into the crowdsourcing platform via a built-in Internet Archive importer.
- **OCR correction:** Where the Internet Archive created OCR, the resulting text can be used to populate the initial transcripts of corresponding pages.
- **Design for amateurs:** FromThePage was designed for collaborative transcription and indexing by amateurs, and has a simple user interface.
- **Workflow/community management:** Community managers can track progress on transcription and indexing and can see user discussions.
- **Projected IIIF support:** FromThePage supports ingestion of IIIF manifests and display through the IIIF Image API in a working proof-of-concept. Further IIIF/Shared Canvas/Open Annotation integration is planned during the IIIFification Hackathon in September/October 2015.

Weaknesses of FromThePage

- **Limited volunteer discovery:** FromThePage has only been used by projects with small numbers of long documents.[50] As a result, volunteers are presented with a simple list of documents to transcribe – an interface which would require enhancement for the Recipes project.
- **Limited development community:** Although FromThePage is open source, to date the only developers contributing code have been Brumfield or staff employed by him.[51]

Mirador

Mirador emerged from the IIIF community (led by teams at Stanford, Yale, and Harvard) as a tool for viewing and comparing medieval manuscript images from multiple different institutional repositories simultaneously.[52] On-image annotation features were added quickly, based on the IIIF community's definition data models based on JSON-LD and Open Annotation. By the spring of 2015, some development efforts turned to adding transcription and translation features to the viewer/annotator Mirador core. Extensive use-cases were solicited from participating institutions, and prioritised and consolidated into a development road map. Development on the transcription interface to the tool began in August 2015.

---

[48] For more detail, see 'Wikilinks in FromThePage' at the iDigBio Original Sources Digitization Workshop: http://manuscripttranscription.blogspot.com/2014/03/wiki-links-in-fromthepage.html

[49] Translation was built for and funded by Fordham University's Center for Medieval Studies in order to support collaborative OCR correction, translation and annotation of the Assizes of Jerusalem (Old French) and transcription, translation, and annotation of the Codex Aubin (Nahuatl). (http://fromthepage.ace.fordham.edu/)

[50] A typical use case is Rhodes College's in-house project transcribing the diaries of historian Shelby Foote. In such diary series, tens of thousands of pages may reside in only a score of documents.

[51] While forks of the repository have been created by developers at Duke University and Northwestern University for maintenance and enhancement, those efforts were abandoned before producing substantial results.

[52] For the project background, status and vision ca. 2013, see Snydman, 'Interoperability in practice: a cross-repository image viewer' http://www.slideshare.net/StuartSnydman/interoperability-in-practice

Strengths of Mirador

- **IIIF support**: Mirador is designed to use IIIF natively, supporting both the IIIF Presentation API and Image API. In addition, much of the IIIF community have rallied behind Mirador to present manuscript material and to transcribe and annotate that material.
- **Open Annotation support:** Open Annotation is native to Mirador, so no transformation to Open Annotation will be needed.
- **Pedigree:** Mirador is being developed by leaders in the IIIF community, most prominently Stanford, Harvard, Princeton and Yale.
- **Active development:** Progress on Mirador's transcription capabilities has been substantial and shows no signs of slowing. In addition, development activities are spread across a number of institutions, collaborating via hack-a-thons, mailing lists, chat rooms, and distributed version control systems.
- **Source material:** All transcription tools are heavily influenced by the source material they were initially created to transcribe. Mirador was built around medieval and early modern material, rather than nineteenth- and twentieth-century documents, so may be a closer fit for the earlier material within the Recipes project.
- **Straightforward Player integration**: Because the underlying data models are compatible, displaying transcripts within the Wellcome Player should not be complicated.
- **Projected translation support:** Text translation is a core use-case for some of the Mirador partners, and is on the development roadmap.[53]
- **Annotation support:** Mirador supports on-image annotation, which might provide a basis for part of the ancillary data requirements to note changes in handwriting or marginalia.

Weaknesses of Mirador

- **Transcription is not ready:** Although Mirador has been deployed successfully as an image viewer and 'page turner', the transcription features have only been under development since Summer 2015. It has not been deployed beyond the development team.
- **No discovery/workflow/community features:** Because Mirador is designed to be a plug-in for multiple discovery systems, none of the development effort on the transcription tool has addressed mechanisms for letting volunteers discover material to work on. In the Princeton classroom use-case, a separate development effort was required to track transcription status of documents, assign documents to transcribers, and even list documents to be transcribed. (PI-6) Any end-to-end crowdsourcing system using Mirador for transcription will need to add a volunteer discovery layer, a workflow system for tracking status, and community coordination/communication features.
- **Not designed for crowdsourcing:** Mirador's origins as a page turner supporting annotation present serious usability problems for crowdsourcing. At present, the tool does not focus users on a single transcription task, which has been a source of confusion.[54]

---

[53] https://docs.google.com/document/d/1XAndHbaFzhu9LA-CwJIxw2AylNcjud-H_roLuovRzD8/edit

[54] The planned Autumn 2015 deployment of Mirador for classroom-based transcription at Princeton was cancelled largely due to time constraints. However, early user testing, as reported on the Mirador Tech Google Group, revealed that substantial simplification is needed to address the 'cluttered' interface. In addition, the conflation of *annotation* and *transcription* that is inherent to the Shared Canvas data model

- **Immature data model:** Any system like Mirador which relies on Open Annotation (OA) to represent transcribed text will need to define anew its own model for encoding a page transcript in an OA annotation list. The Shelly-Godwin Archive, which may have been the first adopter of OA and Shared Canvas for textual material, is in the process of revising their own data model entirely after initial experience. (PI-5) Substantial textual expertise exists within the IIIF/Shared Canvas community, but as current IIIF development focuses on the IIIF presentation and image API, work on text encoding remains to be addressed.

did not match user expectations: 'Even though there's not that much difference between Transcriptions and Annotations on a technical level, there is a big difference on a conceptual level, in the users' mental models, and how they expect the UI to behave for each. For example, a tab that says "Annotations" is meaningless to them and they don't want to see annotations mixed with transcriptions in this context. I'm still not sure how we are supposed to distinguish Transcriptions from other types of annotations.' (Quotes from Shaun Ellis, 'Princeton Transcription Update' post on Mirador Technical Working Group Google Group: https://groups.google.com/forum/#!topic/mirador-tech/7b4TmaiZ-xc)

Tools gap analysis

Given moderate resources to install, customise, and enhance existing tools and with the ability to combine different tools for different tasks into a crowdsourcing ecosystem, the Library faces no major gaps requiring development of substantial new systems from scratch. Open-source software packages already exist for collaborative manuscript transcription,[55] OCR correction, text indexing, translation, annotation, and tagging of genetic textual features. That said, each of the recommended tools will require technical effort to install and integrate with the Library's existing systems (see Appendix B), and each will require some amount of software development to be utilised most effectively.

As discussed earlier, while some crowdsourcing interfaces have become the de facto discovery interface for collections, there is a gap for platforms that provide direct access to participatory tasks and navigation that supports search/browse functions for large collections. The peer interviews also showed that other organisations have struggled to incorporate crowdsourced material back into their collections management systems.

There is room for valuable innovation in the design of crowdsourcing interfaces by developing tools that can provide personalised feedback to volunteers on the quality of their first contributions. 'Golden task' or machine learning software may help with this. Creating interfaces that encourage volunteers to gradually learn the palaeographic skills necessary to transcribe more difficult manuscripts would also be an innovative outcome.

OCR Correction

Both DIYHistory and FromThePage already support some form of OCR correction, and Mirador's transcription interface should be able to support it in future. Further enhancements may be needed to each tool, however. DIYHistory will require OCR text to be converted to page-specific plaintext from its existing format (likely ALTO, but possibly the DjVu files produced by the Internet Archive), and then loaded into the system via the Omeka CSV importer. If FromThePage is used for OCR correction on files loaded from the Internet Archive, no further development will be needed, but if the raw OCR text comes from a different source, custom development will be required to ingest that text. Mirador will likely need no internal modifications to load OCR text in its transcription tool, but converting ALTO or DjVu files into the Open Annotation annotation lists used for transcription will require effort.

Genetic mark-up

An apparent gap in the all of the abovementioned transcription tools is support for the genetic mark-up data requirement – identification of strike-throughs, insertions, and marginal additions within the text transcript. This is worth discussing at length, since there are straightforward technical ways to accomplish this, but careful thought must be put into task specification, tag selection, and making sure that the task matches the goals of the Library and the capabilities of the library's infrastructure.

[55] For an exhaustive list, see http://tinyurl.com/TranscriptionToolGDoc

Image tags or text tags?

The first question to be answered is whether mark-up should be applied to the transcript or to the image facsimile itself. Substantial work has been done by editors of digital scholarly editions to mark-up these kind features on the page facsimile itself. For example, the *Shelly-Godwin Archive* does this at the presentation layer, using TEI 'zones' which correspond to IIIF 'regions' to display changes of hand on the page image (PI-5). TextLab is a tool designed to create such on-image mark-up for the Melville Electronic Library.[56] However, such on-image annotations would need to be merged with the plaintext transcripts in order to be useful.

By contrast, embedding mark-up within the transcript is very common and well understood within textual editing projects. End-users with experience with HTML or the old WordPerfect 5.1 'reveal codes' mode find it easy to understand. Special-purpose scholarly tools support it (T-PEN,[57] the Papyrological editor,[58] Virtualles deutsches Urkundennetzwerk[59]) and crowdsourced versions have appeared with the Bentham Transcription Desk and the Folger's Zooniverse project *Shakespeare's World*. Adding the Project Bentham TEI Toolbar[60] to any plain-text transcription tool requires work, but is straightforward. The Letters of 1916 project did this for their Scripto/DIYHistory installation, technical designs for FromThePage have been created for (outstanding) grant applications, and TEI encoding is an important part of the roadmap for the Mirador transcription tool.

Tag selection and encoding

Although technically straightforward to implement, careful thought must be put into the choice of features users are asked to tag. Even highly trained researchers may flee from an over-broad choice of tags.[61] A model discussion of the compromises necessary in choice of features to tag is Paul Dingman's 'Tagging manuscripts: How much is too much?':

> [T]he EMMO team decided we would focus on tagging the text of the manuscripts for now and let the accompanying high-resolution images provide additional information about the page. Of course, even the adjustable image will not show everything about the actual manuscript, but we think the digital representation and the transcription text together will serve as a valuable resource.
>
> Accordingly, we set out to identify and test a tag set for primarily textual elements, but even this reduced scope contains many questions and possibilities. Would the lineation of the text be preserved? Would the original abbreviations, contractions, punctuation, and spelling (including apparent mistakes) be left as

[56] Melville Electronic Library, http://hofstradrc.org/projects/mel.html
TextLab was built by Performant Software Solutions, LLC, which includes a description of the tool in their portfolio: http://www.performantsoftware.com/portfolio/melville-electronic-library/
[57] http://t-pen.org/TPEN/
[58] http://papyri.info
[59] http://www.hki.uni-koeln.de/virtuelles-deutsches-urkundennetzwerk
[60] https://github.com/onothimagen/cbp-transcription-desk
[61] For more discussion on challenges of tagging within crowdsourced transcription projects see Brumfield, "What does it mean to 'support TEI' for manuscript transcription?", TEI Conference 2012: http://manuscripttranscription.blogspot.com/2012/11/what-does-it-mean-to-support-tei-for.html

written? Would cross-outs and corrections made by the scribe be reflected in the transcription?[62]

The EMMO (Early Modern Manuscripts Online) project settled on tags representing additions and deletions, abbreviations and their expansions, as well as a limited set of tags for non-textual elements like manicules, but avoided any attempt to record charts or illustrations in their encoding.

Purpose of mark-up

A vital consideration in considering mark-up of textual elements is the eventual use of that mark-up. If discovery and presentation systems cannot effectively utilise textual features encoded through mark-up, there may be little point in undertaking that effort. In systems in which transcripts are only used for full-text search, volunteer-created tags may even need to be stripped from the document.[63] If the Library is unable to make use of a particular kind of mark-up, asking volunteers to create that mark-up may violate an important crowdsourcing ethical principle, formulated by the Zooniverse as 'never waste people's time'.[64] Given the Library's existing discovery infrastructure, it appears unlikely that sophisticated analysis of genetic mark-up is possible, although presentation of marked-up transcripts within the Player may be possible. By contrast, the EMMO project at the Folger is building a new manuscript database explicitly to support searching and analysis of documents with extensive mark-up encoded using the tag set they have selected (PI-7).

Probably the best option for accomplishing genetic mark-up is to embrace the principle of an ecosystem of crowdsourcing tasks. If the Library launches an initial project collecting plain-text transcripts and the other important data output requirements, a subset of the resulting transcripts and images could be passed to a different system focused on genetic mark-up of those initial transcripts. We recommend exploring collaborations with EMROC and the Folger Shakespeare Library, using Dromio to add tags to the plain-text transcripts and EMMO to deliver and analyse the texts.

Indexed terms

Although FromThePage has full-featured support for indexing terms within transcripts, nothing similar exists for the other two recommended transcription tools. If transcripts are created using a tool which does not support this data output requirement, they may be ingested into a separate system for indexing. Unfortunately, options here are still quite limited. Possibly the most accessible option is to attempt an indexing project within the same Zooniverse system used for Diagnosis London. Asking users to identify diseases, ingredients, procedures, or names within the transcripts would be a good starting point for indexed terms, although consolidation of variants would need to be handled through post-processing.

[62] http://collation.folger.edu/2015/06/tagging-manuscripts-how-much-is-too-much/

[63] For example, one correspondent told us that one organisation they work with is only able to use plain-text transcripts for their full-text search engines, and therefore they remove the volunteer's mark-up from the copy of the text they receive from the transcription interface.

[64] cf. http://blogs.plos.org/citizensci/2015/01/12/coops-citizen-sci-scoop-try-might-like/

Translation

While FromThePage has support for translation, and translation is core to Mirador's development goals, DIYHistory has no translation functions. Translation is different from transcription in that it is only necessary for a subset of material, and only achievable by a subset of volunteers. If the source languages are common among English-speaking volunteers – as are Latin and Middle English in particular – there may be broad overlap within the volunteer pool. By contrast, attempting translation (and indeed transcription) of a large corpus in uncommon languages is likely to require a new plan for outreach and volunteer recruitment among speakers of the source language. Additional development effort may be necessary to support translation, depending on the transcription platform chosen, but the majority of the costs of a translation effort may be located in communication and coordination.

Mini-projects

As discussed earlier, mini-projects (or 'expeditions') have several advantages. However, none of the tools discussed currently provide good support for project administrators who want to set up specific goals for volunteers to complete targeted subsets of the recipe collection. In both DIYHistory and FromThePage, volunteer effort is focused on a document or a large, thematic collection of documents. Supporting expedition-style mini-projects will require either development investment or careful planning and communication to communicate project goals and focus through other channels.

Task ecosystems

While there are a number of examples of task ecosystems, supporting the diverse range of vocabularies from specialist domains that relate to the recipes could present a challenge for the user experience design of task interfaces. For example, the ability to tag text with terms from multiple vocabularies risks creating an overly complex interface that may reduce the number of volunteers. The Recipes project could make a useful contribution to the field by exploring options for creating and linking different interfaces for different types of tagging.

Tools 'state of readiness'

DIYHistory is clearly the tool which is most ready for rolling out a plain-text transcription project, and has the best support for researcher use. After the initial installation and ingestion of Recipe manuscripts into the system, it should be straightforward to operate for a large number of documents. However, it does not support the Library's additional data outputs,[65] and any attempt to use the platform for those tasks would require substantial effort to build new Omeka plug-ins for translation and indexing from scratch, or would require the plain-text transcripts to be transferred to a different platform which can support such tasks.

FromThePage has the best support for the Library's data output requirements, already supporting indexed terms, OCR correction, and translation in addition to plain-text transcription. The transition from plain-text transcription tasks to other outputs should not require any further customisation or development. However, the limited discovery interface is likely to require some enhancement in order to allow volunteers to navigate documents numbering in the hundreds or thousands.

Mirador is the closest technological match to the Library's current organisational/technical momentum. Its support for Open Annotation and IIIF is fundamental to the tool, and the support of Mirador among the IIIF community – of which both the Library and Digirati are active participants – is extensive. It is, however, the most immature of the tools we recommend, and would require substantial development effort not only to fill in the feature gaps of the transcription tool, but also to build infrastructure for workflow, community discussion, and incorporate appropriate interfaces for collections discovery and transcription tasks. Were the connection between the Library and the IIIF community not so strong, a tool at Mirador's level of readiness would not be a recommended option.

A strategy for ingesting documents into the chosen crowdsourcing platform will need to be developed. Similarly, each platform will need to export the data outputs in a format which can be processed by Goobi, and schedules and mechanisms for achieving that export will need to be defined. Both DIYHistory and FromThePage have existing mechanisms for import and export which can be utilised, but extensive development may be required for a Mirador-based platform.

[65] Apart from genetic mark-up using the TEI Toolbar, discussed above.

High-level risk analysis

Many potential issues encountered by other projects have already been mentioned in this report. For example, we discussed peer interviewees' observations that the crowdsourcing platform could become the most visible interface to collections, obviating existing catalogue interfaces. We suggest alleviating this by planning for strong integration between the task interfaces and discovery interfaces.

Technology

The stakeholder interviews showed that timely end-to-end integration could be difficult. Devoting one stage of the overall development process to implementing end-to-end integration would ensure that it is prioritised and that any emergent issues are addressed well before the project is officially launched. This would also provide some guidance on the workflows required to prevent the project from running out of material to transcribe (as happens for some highly successful projects).

One risk in building open source software for re-use by the wider community is that the Library spends disproportionate resources supporting the software at the expense of developing functionality to support the Library's goals.

Community

The most obvious risk is that the project could fail to attract volunteers. However, in addition to a marketing plan, the staged approach and a thoughtful approach to partnerships and outreach should help alleviate this, as existing users of the collections and early prototype testers can become the first 'word of mouth' promoters. Recruiting existing users of the collections could also help set a constructive tone for community discussion around the collections.

Other risks are more subtle. Organisations with a broad mission to reach the public must find the right balance between productivity and volunteer engagement. For example, volunteers often report researching the people, places, concepts (etc.) related to the items they are transcribing or marking-up. While this is a valuable activity, it also reduces the amount of time volunteers have available for the core project task. Similarly, community discussion can support volunteer learning and lead to new research questions, but it can also be time-consuming. Projects can influence the balance but ultimately participation in either activity is voluntary and volunteers may have strong preferences for some activities over others.

Selecting or building an appropriate community platform is not straightforward. Individual preferences depend on past experience, and they also change over time as new technologies introduce new modes of communication. Options include using off-site social media platforms such as Facebook or Twitter, forums or email discussion lists, and comments on individual item pages. Common issues include notifying participants of new posts and creating links between item pages and related discussion. The project must also find the right balance between reducing barriers such as compulsory pre-task registration, spam prevention, and collecting contact information for volunteers for newsletters and other centralised updates. Finally, the variety in discussion types may cause issues - while some volunteers might want to leave notes for

other transcribers, others may want to discuss the historical item itself, or link it to other resources.

Advantages of social media include updates being pushed to follower's streams or found via hashtags such as the Smithsonian's #volunpeers. However, updates are easily missed in a busy stream, and some volunteers may not want to blur the lines between their hobby and other identities on social media. Forums are flexible, provide stable URLs and are searchable over time;[66] but as forums don't notify people of new activity by default, people may forget to visit them. Email list discussion arrives directly in the users' inbox, but, like forums, some volunteers may expect a more 'modern' experience. Comments on individual pages are easy to discover from item pages but may be difficult to find without browsing individual pages. The use of any third-party system (liable to close or change their services without sufficient notice) is risky unless steps are taken to regularly harvest posts.

Consulting with volunteers about their preferences and platforms they currently use can help alleviate the risk that a chosen platform will not be used. However, preferences are likely to vary and attempting to use multiple platforms risks diffusing conversation.

User experience design

There are several models for crowdsourced transcription, from parallel transcriptions of one line-at-a-time to iterative, collaborative page transcription. There are also several models for marking up text, including annotations linked to the image or XML tags wrapped around transcribed text. Different volunteers will have different mental models about how interfaces and workflows operate on specific materials, depending on any previous experience with manuscripts or crowdsourcing projects. Crowdsourcing projects often develop jargon or shorthand terms which can be confusing for new participants; terms such as 'needs review' or 'indexing' should be explained in everyday but precise language.

There is a risk that the interfaces become overly complex or confusing as new tasks are added to the project. Mixing collections access interfaces with crowdsourcing tasks can introduce design issues. For example, DIYHistory-style interfaces in which volunteers browse through collections to find items to work on allow volunteers to follow their own interests or find material that suits their palaeographic skills, but it can be difficult to find items that need work. Information retrieval and discovery interfaces may be inadequate for researchers' needs if their design is an afterthought. While the Zooniverse's 'next item in the queue' method has some drawbacks (particularly for documents which people might want to follow sequentially), it makes it easy to find the next item for a given task. Ideally, the project's interface would address these issues and create a suitable hybrid browse/queue solution. Ongoing usability testing with target audiences at different stages of the design process should help address these potential risks.

Engaging the community with more difficult or less immediately appealing material or tasks can be hard. For example, some projects make good progress on transcription tasks but seem to struggle to move manuscripts from the 'needs review' to the

[66] However, not all forum software has the same capabilities.

'reviewed' stage. Mini-projects and challenges can help encourage volunteers to take on harder or dull tasks for a good cause. Strong, shared internal agreement on goals, and transparency in communications so that the value of more difficult tasks is communicated clearly to volunteers also helps.

Tools recommendation

Because of the differences in focus between the three recommended tools, and because of consultant Brumfield's conflict of interest, this report lays out three options for a crowdsourcing platform, for selection based on organisational priorities.

DIYHistory offers the quickest option for launching a project meeting the Library's essential data output requirement of plain-text transcription. Little work would be needed beyond integration and installation. In addition, the Omeka platform which underlies DIYHistory is extensible through plug-ins and has a vibrant development community. DIYHistory has by far the best support for the researcher uses of the crowdsourcing platform whose importance was revealed by peer interviews. However, it has by far the weakest transcription tool – any outputs beyond plain-text and OCR correction will need to be done in another system or developed from scratch. Arrangements for support should be explored in greater detail before DIYHistory is selected.

FromThePage provides the broadest support for the Library's data output requirements, with plain-text transcripts, indexed terms, OCR correction and translation already built-in. No other tool supports indexed terms which handle variants and automated tag suggestion. Researcher support is equivalent to that of a simple DIYHistory installation, through there is no plugin-based extensibility comparable to Omeka's. However, additional development of collections discovery features would be needed to support the number of documents the Recipes project is targeting, which would be necessary before launching anything beyond a pilot project. As with DIYHistory, arrangements for support should be explored in detail before FromThePage is selected.

Mirador's transcription tool is an excellent fit for the Library's existing technical momentum and the skills available through the Library's collaboration with Digirati. Its origin among the IIIF community pairs it nicely with the Library's existing involvement with that community. However, of the three transcription tools, it would require the most effort to get even a basic project started, and the annotation model may create UX challenges. Only if the Library is willing to commit to an extended development effort should Mirador's transcription tool-set be adopted. Substantial support may be available from other institutions within the IIIF community, but such support is likely to be limited to transcription/mark-up/translation functionality, leaving development of essential integration, community support and workflow features to the Library. However, if the Library is willing to invest in the platform, it would take the technical leadership role in a high-profile project benefiting its peer institutions.

Recommendations for contacts

Conversations with the following people would be a good starting point for potential collaborations and for sharing lessons learnt.

Recipe crowdsourcing and transcription

Colleen Theissen (University of Iowa)
Rebecca Federman (NYPL)
Rebecca Laroche, Elaine Leong, Hillary M. Nunn, Jennifer Munroe, Lisa Smith, Amy L. Tigner (various academic affiliations with the Early Modern Recipes Online Collective (EMROC))

Other early modern transcription projects

Heather Wolfe (Early Modern Manuscripts Online at the Folger Shakespeare Library)
Paul Dingman (Early Modern Manuscripts Online at the Folger Shakespeare Library)

Tools

DIYHistory

Matthew Butler (University of Iowa)

FromThePage

Ben Brumfield

Mirador

Shaun Ellis (Princeton)

Zooniverse

Victoria Van Hyning (Zooniverse)
Chris Lintott (Zooniverse)

Engagement, Outreach, and Workflows

Meghan Ferriter (Smithsonian Institution)
Paul Flemons (DigiVol)

Integration and Mark-up

Paul Dingman (Folger/EMMO)
Doug Reside (New York Public Library)

OCR and HTR Workflows

Paul Hagon (National Library of Australia, Trove)
Melissa Terras (Transcribe Bentham)

Cross-project organisation and Funding

Mary Flanagan (TILTFactor / Crowdsourcing Consortium for Libraries and Archives)
Neil Fraistat (Director of the Maryland Institute for Technology (MITH))

Appendix A: List of interviews

Peer interviews

PI-1: *DIYHistory*; Colleen Theissen, Special Collections Outreach & Instruction Librarian, University of Iowa; August 7 2015
PI-2: *Ensemble*; Doug Reside Digital Curator for the Performing Arts, New York Public Library; August 7 2015
PI-3: *What's on the Menu?*; Rebecca Federman, Electronic Resources Coordinator, New York Public Library; August 10 2015
PI-4: *DIYHistory*; Matthew Butler, Senior Developer, Media Production & Design, University of Iowa; August 11 2015
PI-5: *Shelly-Godwin Archive*; Raffaelle Vigilante, Research Programmer, Maryland Institute for Technology in the Humanities; August 21 2015
PI-6: *Mirador*; Drew Winget, Visualisation Engineer, Stanford University and Shaun Ellis, Digital Library Collections Interface Developer, Princeton University Library; August 24 2015
PI-7: *Shakespeare's World/Dromio*; Paul Dingman, EMMO Project Manager, Folger Shakespeare Library; September 9 2015

Stakeholder interviews with Wellcome Library staff

SI-1: Christy Henshaw, August 5 2015
SI-2: Jenny Haynes, Helen Wakely and Chris Hilton, August 5 2015
SI-3: Christy Henshaw and Dave Thompson, August 11 2015
SI-4: Jenn Phillips-Bacher and Alex Green, August 13 2015
SI-5: Robert Kiley and Tom Crane (Digirati), August 18 2015

Appendix B: System integration

Like other institutions, the Library has its internal systems for managing its collections and has separate, public-facing systems for delivering/displaying those collections to the public, making two integrated technical stacks. Any user-friendly crowdsourcing platform will be optimised for community experience, forming a third technical stack supporting transcription/OCR correction, volunteer discussion, and community coordination. The challenge will be to implement a crowdsourcing system that communicates effectively with the internal preservation, discovery, and delivery systems, so that a third data silo is not created.[67]

Integration Points

The project should consider the following points of system integration.

Basic Metadata & Images to Crowdsourcing System

For manuscript documents, the crowdsourcing platform will need to ingest enough metadata about the document and the page images sufficient to allow collections discovery and page image display to volunteers.

Crowdsourced Transcripts to Universal Viewer

Leveraging user-contributed transcripts in the digital library system requires exporting transcripts from the crowdsourcing platform in a format that can be consumed by the Universal Viewer for display and 'search within' functionality.

Crowdsourced Transcripts to Sierra

To support full-text search of manuscript and print material within the Library's discovery system (Encore), the project will need to export transcripts from the crowdsourcing platform, transform them into a format suitable for Sierra, then import them regularly.

Crowdsourced Transcripts to CALM

For preservation purposes as well as for enhancements to the Library's archival finding aid system, transcripts should be exported from the crowdsourcing system and ingested into the CALM archives system.

Crowdsourcing Item URL to Sierra/Encore

In order to invite researchers in the search interface to contribute to the crowdsourcing project, URLs for each document in the crowdsourcing system should be displayed along their corresponding entry in the discovery system.

---

[67] Integration of transcripts into existing Library systems need not be implemented perfectly at the first pass, however. The Folger Shakespeare Library's experience shows the benefits of a staged approach: transcripts produced collaboratively through Dromio (a web-based platform not currently used for crowdsourcing) are added to FolgerPedia, a MediaWiki installation only editable by Folger staff. These are displayed accessible for researcher purposes along with links to both the catalogue entry for the manuscript and the manuscript images on Luna, the library's image repository. The transcripts to be produced by the Zooniverse project Shakespeare's World will be added to the same system. As the new EMMO database software is developed, all these transcripts will be migrated to the new platform. Such a staged approach allows flexibility in choice of transcription tool while serving researchers at every stage. (PI-7)

Crowdsourcing Item URL to Universal Viewer

In addition to integration within the discovery system, researchers using the Universal Viewer to read documents should be able to add or correct the corresponding transcripts via a link to the corresponding page in the crowdsourcing platform.

Catalogue URL to Crowdsourcing System

If the crowdsourcing project is successful, many researchers will discover an item within the Library's collection via the crowdsourcing platform, and may be unaware of the additional context and resources provided by the Wellcome's digital library system. Providing them with a means to navigate from an item within the crowdsourcing system to the corresponding records within existing discovery systems allows them to access that context with a single click.

Indexed Terms to Discovery Systems

As with full-text transcripts, any structured terms identified by volunteers should be exported from the crowdsourcing platform that generated them for ingestion into CALM and Sierra. It is likely that this kind of data may require separate processing, as such terms may be registered as tags or subject headings in the discovery systems.

Transcripts to Rest of World

Allowing researchers to access user-contributed transcripts via APIs will provide them access to the Library's collection in formats immediately useful to them. For non-technical users, easily accessed formats should be provided, such as CSV downloads for structured data,[68] plain-text downloads for plain-text transcripts, and PDF,[69] rich-text format, or TEI-XML downloads for complex mark-up. The Medical Officer of Health project sets a precedent for the Library by providing some item-level downloads.

[68] The New York Public Library's *What's on the Menu?* API at http://menus.nypl.org/data provides an exemplar for structured data, as it includes both CSV downloads for analysis by non-programmers as well as an API for programmers to use. This API has resulted in the menus data being used and publicised by Digital Humanities projects neither affiliated with the NYPL nor culinary research, most notably *Curating Menus* (http://www.curatingmenus.org/) which uses the crowdsourced transcripts for a data curation course.

[69] The Smithsonian Institution's *Transcription Center* has offered the public the ability to download PDF versions of user-generated transcripts from the beginning, considering data reciprocity to be a moral imperative. (e.g. https://transcription.si.edu/pdf/7666/ECoR1)