PROPOSAL NARRATIVE [EXCERTS]

The Problem Addressed

Scholarly annotated editions of historically significant texts constitute an important foundation for learning and research in the Humanities. Scholarly editing requires a sustained investment of highly specialized expertise, but long-term funding is difficult. Existing editorial procedures are still rooted in the pre-digital work practices and space constraints of the printed codex.

The expert editors of scholarly editions and their carefully trained assistants spend a good deal of time researching people connected with their papers in one way or another. Notes on these time-consuming investigations are typically kept in folders in the project offices or hard drives and may result in a few lines of footnote in the eventual published volume. The reality is that most of what is learned about these individuals (and all of what is learned in some cases) is not included the published volumes, is not shared with other researchers, and is discarded when grants for publication expire.

This situation is the more poignant for two reasons: First, editorial work tends to be duplicated in the parallel editorial efforts of different projects with overlapping scope. For example, Emma Goldman and Margaret Sanger knew each other and were active in some of the same circles, so the editors of the Goldman papers and the editors of the Sanger papers, located over 2,000 miles apart, often research the same individuals, unwittingly – not to mention overlap with editorial work at the Samuel Gompers papers, the Eleanor Roosevelt papers, the Bertrand Russell papers, the Wellcome Library in England, and on and on, let alone scholars working in other capacities: historians, archivists, and curators of special collections.

Second, projects expire, but scholarship continues. The ideal would be if the editorial "workshop" could remain ready to support resumed scholarship as and when labor and funding allow.

Current editing work practice is rooted in the lingering impact of pre-digital work practices and the space constraints of the printed codex. In the project *Editorial Practices and the Web* three major documentary editing projects undertook a fundamental change in how scholarly editors' document and share their work. The three collaborating editing projects have adapted their work practices to take advantage of Web technology as a way to share their working notes with each other and with the world.

Achievements in Phase 1

The present situation can be described by considering procedures into three stages:

1. *Initial Research Notes*: Ideas and notes are remembered and recorded in (often handwritten) notes. With the reducing costs of laptops, notebooks, scanners, and OCR software, the trend is away from paper pads towards scanned documents and keyed notes.

2. *Editors' Working Notes*: Notes, collected data, lists, references, clippings, photocopies, etc., are mostly stored in topical folders in filing cabinets, but there are also specialized locally-developed tools such as itineraries, chronologies, and legislative histories. Editors can and do ask editors elsewhere for help on specific topics, but answering may be time-consuming and they do not want to burden over-worked colleagues with repeated requests for help. Importantly, the *Editors' Working Notes* could include notes on the many of unresolved problems that researchers accumulate: reasons to question published accounts; why a claim might be suspect; known false leads; promising clues and lines of inquiry that might be followed up later; notes that someone

else knows about some point; references to documents not yet located; citations known to be garbled; unresolved queries; and so on. One might hesitate to publish such working notes openly, but some editors and researchers in other projects could find them very useful and others might already know the answer or, at least, be able to suggest where to look. This revives in a small way the nineteenth-century "Notes and Queries" genre.

There are also locally-made tools. Editors of personal papers usually need to create a detailed itinerary of that person's movements, which were complex in the case of Emma Goldman's lecture tours. Similarly, editors might create uniquely detailed legislative and legal histories of specific topics as the Stanton and Anthony project editors have.

3. *Editors Notes*, appear, if at all, in very concise footnotes, endnotes, and appendixes in the eventual published volumes. In contrast, if explanatory notes were written at whatever length an editor considered justified and helpful, with sources clearly stated, and promptly posted on the project's website, they would be more informative, would be immediately available to everyone, would soon be indexed by Google and other search engines, and would easily be found by inquirers. Dated and signed, these notes can provide a steadily growing population of trustworthy research reports that all kinds of scholars and students can benefit from. In particular such notes could facilitate a higher level of scholarship in popular compilations such as the Wikipedia.

What is needed is a sustained move from the upper row below to the lower row.

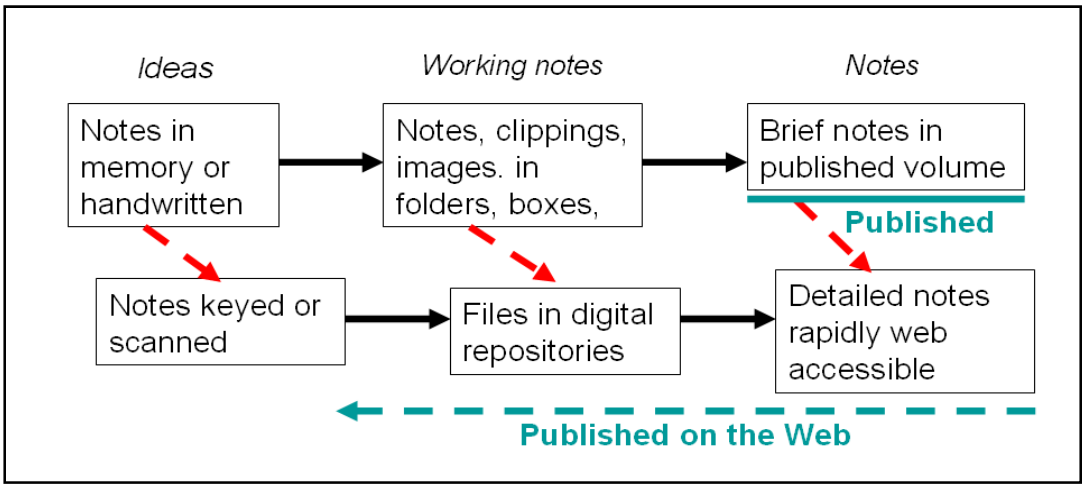

Figure 1: The transition to digital notes

A more systematic, more digital approach to the first and middle stages is mainly a matter of accelerating the existing shift from handwritten to keyed notes and adopting a more structured arrangement of material. Each stage feeds the next.

A website named *editorsnotes.org* was created for shared use by the three editing projects. The architecture of the site is shown in Figure 2:

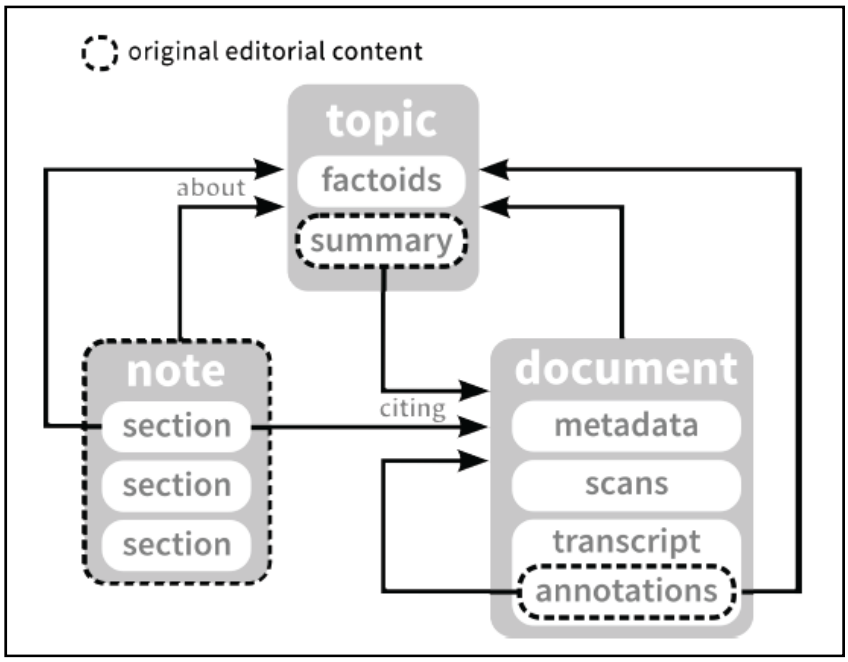

Figure 2: Part of the Editors' Notes data model. Notes, sections of Notes, and Topic summaries may cite Documents. Document annotations are linked to the Topics to which they relate. The Topic's "summary" is for free-form textual description of the Topic. "Factoids" are the pieces of structured data created locally and/or imported from a trusted resource. These would be the source for specialized search and visualization interfaces.

In August 2012 password control to the site was quietly removed making the site openly available to both humans and webcrawlers. Because the server is at Berkeley the site gets priority attention. By September, after Web search engines, including Google, Bing, and Baidu (China), had indexed the contents, the resources on the Editors' Notes site were being viewed by scholars from around the world.

PROJECT DESCRIPTION

Phase 1 constituted a pivotal shift in editorial work practice in the three editing projects. Phase 2 builds directly on this achievement with four components: Adding advanced Digital Humanities technology; preservation and access; extending this approach to archivists' notes; and documentation and dissemination.

1. Incorporating Digital Humanities Technology

Phase 1 was significant because it moved routine day-to-day procedures from the desktop to a Web environment. Nevertheless, editors and their staff continue to work primarily with simple flat text files and scanned images. Here, as elsewhere, there is a chasm between the daily routines of ordinary scholars and the impressive technical achievements of experts in the large-scale, complex projects reported at Digital Humanities conferences with dazzling visualizations created from complex databases by experts using sophisticated software. How might the latter be harnessed for use by the former, who have so little capacity for absorbing additional workload or complexity? The usual response is that projects interested in using these technologies could enlist the help of specialists, either for ongoing consultation or for individual applications.

However, this sort of work would be more achievable, more affordable, and more sustainable if it could be done by non-specialist researchers within projects.

The problem is not a lack of tools for using name authorities or generating map displays, timelines, prosopographies, and the like, but, rather, how to incorporate such tools into the work routines of hard-pressed editors and their assistants with an acceptably low threshold of learning and effort. We address that severe challenge in Phase 2 as essentially a matter of software integration and interface design to very low thresholds of user effort. We see this task as having three components: Making links; enriching data; and invoking visualizations.

1. *Making links*: The Editorsnote.org site will be enhanced such that when editors write or note a place name, person, organization, or selected other entities, the interface will offer elective autocompletion from a ranked list of matching candidates from the existing list of Topics and/or an external authority list and store that selection as linked data mark-up. (We will start with Geonames, VIAF, and Wikidata, and add other resources as deemed desirable after consultation with the editors.) Entities previously unused within the site would be established as new Topics, thereby building a larger and more authoritative vocabulary of place names, people, organizations, etc.

2. *Efficient enriching of local data*: We will build tools to allow users to add and maintain geospatial or prosopographical information, events including dates, and other structured data to Topics. Through a combination of importing data from external links and entering it locally, Topics would be gradually and incrementally enriched from a mere list of generic "things" to a structured group of semantically distinct and descriptive entities suitable for advanced querying and manipulation. Importantly, researchers would have full editorial control over this data, ensuring its high quality and compatibility with their painstaking scholarship.

3. Visualizations: A simple interface would allow users to invoke three kinds of visualizations based on targeted Topics: Maps, timelines, and network graphs. These correspond most naturally to places, events, and personal relationships, but any Topic which has coordinates can be mapped, any Topic with time points or ranges can be put on a timeline, and any relationships among Topics can be visualized as a network. These three together are, therefore, broadly applicable to any kind of structured data about Topics that might be gathered. Documents have equivalent data (when and where published; authorship) allowing the same types of visualizations for them too. So, for example, a map display would show any location(s) mentioned in a Note, with options to display the locations of other related Topics and Documents in any number of ways as determined by the interests of the editors.

These tools would have an added benefit of removing some tedious, duplicative work from everyday research. Editors would be able to import contextual details of, for example, persons (e.g. birth and death dates, place of birth, other names) or of places (alternative names, containing jurisdiction, latitude and longitude) without researching or transcribing these details at every mention. Using a link can bring the benefit of automatic updating as additions and corrections are made to the resources to which they are linked.[1]

Benefit 1: Improved filtering and sorting of lists of notes and documents

[1] Ryan Shaw & Michael Buckland. 2011. *Editorial control over linked data.* American Society for Information Science and Technology Annual Meeting, New Orleans, Oct 10, 2011. Preprint http://metadata.sims.berkeley.edu/posterasist2011.pdf

A major challenge for Editors' Notes has been how to enable fine-grained addressing and indexing of notes without making users feel as if they are working with "a million little pieces"? One the one hand, we want researchers to be able to search for and link to "atomic" notes taken on a single source document as it relates to one narrow topic. Yet we also need to allow researchers to work with aggregations of these small "atoms" in ways that feel natural to them. For example, notes taken while researching "the status of birth control in India in the 1930s" might reference dozens of documents that encompass several more specific Topics such as the Indian birth control activist Dhanvanti Handoo Rama Rau, birth control clinics, and the Bombay Municipal Corporation. Researchers can work with these notes in the context of the broader research task, or they can pull together all the notes about Rama Rau, whether or not these were taken in the course of researching "the status of birth control in India in the 1930s."

Thus many interactions with Editors' Notes involve working with lists. Upon browsing to the Topic "India, birth control movement in", one sees lists of notes and documents related to that Topic. The note on "Dhanvanthi Rama Rau & the Fourth International Conference on Planned Parenthood" is a list of "atomic" notes on various letters and reports that reference that event. A search for "Rama Rau" returns a list of notes and documents with that name in their text. Working fluidly with lists like these requires the ability to filter and sort them. Currently we allow filtering and sorting of lists of documents (and the notes that cite them) using the bibliographic metadata associated with the documents. This was a major step forward for the usability of Editors' Notes, but we envision going much further.

When not only documents, but also places, people, organizations, and events have structured data associated with them, the facilities we can provide for filtering and sorting notes and documents will be far more powerful. Given a list of notes on "India, birth control movement in," users will not only be able to filter and sort them based on the dates of the cited documents (as they currently can), but will also be able to use, for example, the locations and birth and death dates of the people referenced in the notes, the locations and dates of existence of the organizations referenced, or the locations and dates of the events referenced. Furthermore, by taking advantage of structured spatial and temporal metadata, we will no longer be restricted to presenting lists of notes and documents textually. The note on "Dhanvanthi Rama Rau & the Fourth International Conference on Planned Parenthood" will be viewable not only as a text document, but also as a map of specific locations in Stockholm and Bombay, or as a timeline of dates associated with the conference.

Benefit 2: Improved naming control via linked authority data

Names are always a challenge for historical research, and this is especially true for the projects we have worked with. For example, many of the women researched by the Margaret Sanger Papers had multiple different married names and alternated between using their maiden name and their married name at a given time. Dorothy Hamilton Brush, a close friend of Sanger's, appears in documents variously as Dorothy Adams Hamilton, Dorothy Brush, Dorothy Dick, and Dorothy Wamsley. The Sanger Papers also research individuals with Chinese and Indian names that have been transliterated in a wide variety of ways in their documents. Nicknames and children who share names with their parents or grandparents are also sources of confusion.

The most technologically sophisticated projects, such as the Sanger Papers, deal with these issues by maintaining name authority databases.[2] A name authority database records

[2] Hajo, Cathy Moran. 1991. "Computerizing Control over Authority Names at the Margaret Sanger Papers." *Documentary Editing* 13(2), 35-39.

preferred names, variant names, and identifying codes for every individual researched by the project. Each primary and secondary document controlled by the project is examined for personal names, and each name found is searched for in the name authority database. If a matching individual is found, the document is tagged with an identifying code for that individual, and if the form of the name found in the document is a new variant, it is added to the name authority database. If no match is found, a new authority is created. To verify spelling and usage, periodically the name authority database will be manually checked against external authority files such as the Library of Congress Name Authority file.

This is an effective system, but it could be improved upon by linking the local name authority database into a network of name authority files. Rather than each project maintaining an isolated authority file with its own identifying codes, each project would link its name records to those of other projects and external authorities such as VIAF. This would enable the process of checking spelling and usage to be further automated, and every project would benefit whenever any linked authority file was updated. Transforming the stand-alone name authority database into an integrated web service will have other benefits as well. Currently name authority databases like Sanger's are used primarily when ingesting documents so that they can be indexed properly. A web service, on the other hand, could be used to power an auto-completion function that would allow editors to quickly and easily link names to identifiers in any context, such as authoring a research note. And rather than using the separate name authority database to produce reports of variant names to be manually consulted when searching, a web service could be used to automatically expand searches to include variant names. Finally, all of these amenities could be extended beyond personal names to other kinds of names (events, organizations, places, subjects).

Benefit 3: Structured documents / chronologies

Projects often want to create visualizations displaying temporal, geospatial, or prosopographical information to concisely express large datasets, reinforce explanations, or offer aesthetically pleasing displays for public exhibition. However, to do so, they typically have to enroll the help of experts who have mastered the use of specialized tools for text processing and data visualization, thereby decoupling the visualization of information from the process of research that created it. This becomes a problem if, for example, a researcher who is a technological layperson wanted to correct or amend a timeline created by a since-departed specialist. Giving researchers without specialized knowledge the ability to author, edit, and control structured documents and derived visualizations would change visualizations from static, one-off exhibitions to works in progress, able to be controlled with the same editorial touch as the textual record.

Documentary editing projects already create ordered chronologies in their day-to-day work. For example, the Margaret Sanger Papers and Stanton-Anthony papers keep records of their subjects' travels and actions in Microsoft Access and dBase III databases, respectively. Researchers at the Emma Goldman Papers have made chronologies inside Microsoft Word and Word Perfect for events such as the assassination of President William McKinley, the Preparedness Day Bombing of 1916, and the rise of the Industrial Workers of the World. Additionally, there are hundreds of informal chronologies spread across the research notes of each project. Especially in the case of the word processor files, these documents require non-trivial amounts of special post-processing in order to turn them into semantically rich expressions of events which can then be fed to visualization libraries. We will enable researchers to directly and routinely enter structured data to take full advantage of the work that they are

already doing in creating these sorts of chronologies. With this rich data to draw from, we will integrate existing visualization libraries to create map, timeline, and network displays as described in the *Architecture and* Software section below.

Along with allowing visualizations to be changed and updated, this integration of easy visualization of work would provide new ways of seeing research. For example, a researcher could aggregate the chronologies of several different events together into a larger timeline to look for connections or patterns that might not have been evident when viewing each separately. Reducing the expertise required would enable and encourage experimentation with research data in addition to the creation of exhibits for predetermined purposes.

Data accumulation will be incremental and come as a *by-product* of the routine editing effort. If procedures at this level of convenience can be achieved, it would be an amenity likely to be desired by individual humanities scholars everywhere and, ultimately, functionality that could and should become standard in academic word-processing software.

2. Preservation and Access.

Our focus on editors' working notes has drawn attention to the loss of resources when editorial projects end. The funding for documentary editions is narrowly limited to support for the eventual published edition. When the manuscript of the final volume is ready for publication, the editors and staff retire or move on and their working notes become effectively inaccessible if not discarded.

Grants do not (yet) fund the preservation of the editors' working papers. Elite projects associated with founding fathers and presidential papers generally have strong continuing institutional support but they are not typical[3] and nobody seems to know about the rest. [. . .]

The Stanton-Anthony Papers Project, for example, expired on September 30. These editorial working notes include unique resources, including the painstaking reconstruction of the detailed legislative history of women's franchise in each and every state, yet there was no funding (or plans) to do anything with them.

Projects end, but scholarship continues! Could the legacy of working notes of completed projects cost-effectively support future scholarship instead of being discarded? What if, after the Stanton-Anthony project had ended, a new editor with new funding sought to revive the project and prepare additional volumes? How feasible would that be? To the extent that editors' working notes have been handled digitally along the lines of our Phase 1, they could remain available and accessible with minimal overhead, but complete retroactive digitization for a project of that size would be unaffordable.

Thinking tactically, we could examine what low-cost procedures could keep these editorial resources accessible as a more-or-less arranged and preservable archival resource.

Thinking strategically suggests that the relationship between the editorial working notes and the published editions should be reconsidered. Currently, the published editions are the one and only product. The editorial expertise and project working resources are treated as expendable means to that sole objective. But changed technology makes it imaginable to reverse that relationship. In this view the editorial "workshop" (expertise and working notes) could be enduring assets and the published editions would become intermittent valued by-products.

---

[3] The annual survey in *Documentary Editing* (2011) lists 66 published volumes from 65 different editing projects. Nine volumes were from founding father and/or president papers projects, with an 18th century emphasis. The remaining 56 volumes were from projects covering a wide range of individuals, groups (e.g. Cherokees, German Immigrants), and themes (e.g. Ballads, Mexican-American War, Vaudeville) with, mostly, a 19th and/or 20th century emphasis.

Scholarly communication could be greatly extended if it were feasible not only for scholars anywhere to have *sustained access* to the working notes, but also for scholars anywhere to *add supplementary notes, corrections and additions to them* (with clearly separate attribution) in the future *as and when interest, ability, and resources allow*. Useful research programs should not disappear irrevocably! This is a logical consequence of the rationale for our existing project and follows from the move to digital technology and a networked environment.

The ambition would be to move beyond a short term tactical solution (graceful retirement into a passive archival collection) toward a working collection in which at least the finding aids and research guides could be updated and enriched as scholarship continues, a new genre somewhere in between a conventional (static) archive, a library special collection, and an ongoing research program. There appears to be little precedent for this, except in local community archives and open note-book science.

Note that there has a been a strong trend in public sector archives to reduce expenses and to speed up access through more limited processing under the slogan "More Product, Less Process."[4] This means that existing arrangements of documents, boxing and labeling are retained as far as possible; description is at the folder or container level, not each sheet of paper; and not every last paperclip is removed.

One corollary of the “More Product, Less Process” approach is the archival principle that the quality and clarity of arrangement of working notes *during* a project – prior to deposit – through good records management greatly facilitates good and economical archival processing when papers are transferred.[5]

Another consequence of the move to more meager More Product, Less Process finding aids is that topical research guides not limited to a single records group become more important because they can, in part, compensate for limitations in individual finding aids. [. . .] In library terminology there is a comparable relationship between the detail in catalog records and the provision of literature reviews: They are different in kind but can complement each other. Finding aids usually remain unchanged, but literature reviews, pathfinders, and research guides could and should be regularly updated as new relevant publications and archival groups become known. (One might add that in a digital environment updating and enriching finding aids becomes more viable.)

At the California State Archives *both* the economic imperative to adopt More Product, Less Process finding aids *and* the two corollaries were accepted. There was already a good informal collaboration with State agency records managers leading to archival deposits being well-arranged and well-documented prior to receipt. The conclusion was that the Archives could not achieve their public mission without a complementary investment in research guides. Visitors to the Archive’s research room are usually interested in a topic (commonly biography or railroad history) that cuts across multiple archival groups or for which individual series are too cumbersome for unassisted use or where other institutions elsewhere are known to hold significant relevant resources. Learning of the *Editorial Practices and the Web* project they requested our help because, as a practical matter, in a large archive (as with a documentary editing project) multiple staff would need to make, share, and accumulate informal working notes if research guides are to be prepared cost-effectively.

The conjuncture of our three collaborating editing projects and the situation at the California State Archives provides an exceptional basis for a practical examination of these issues.

---

[4] Mark A. Greene & Dennis Meissner. More product, less process. *American Archivist* (Fall/Winter 2005): 208-263.

[5] Weideman, Christine. Accessioning as processing. *American Archivist* (Fall/Winter 2006): 274-283.

The Stanton-Anthony project ended after 20 years on September 30, 2012 without plans or funding for the disposition of the working notes, much of which could not be incorporated into the published volumes. Editorial projects with time remaining can plan and prepare for an anticipated closure (e.g. the Margaret Sanger Papers Project) and those at risk of premature closure for lack of sustained funding (e.g., regrettably, the Emma Goldman Papers Project) would be helped by knowledge of best practices in archiving the resources of documentary editing projects. Involving the California State Archives would not only provide them with assistance they greatly need, but would broaden the testing of our ideas, and test-drive the Editors Notes approach for wider adoption among archives.

We propose to proceed as follows:

1. The Stanton-Anthony project papers would receive conventional archival processing: accessioning, selection/weeding, arrangement (if and as needed), a finding aid, and precautionary preservation measures. This work would be documented as a case study in the archival treatment of this genre: the residual resources of documentary editing projects. Lessons learned would be offered as suggested guidelines for handling the resources of other completed documentary projects.

2. Accordingly, the Goldman and Sanger papers projects would examine the implications for their project procedures with a view to ascertaining what steps, if any, would prepare their records for easy archiving when the projects eventually end.

3. The retired Stanton-Anthony editor [. . .] will draw on her experience to create research guides that include but are not limited to the archived Stanton-Anthony working papers.

4. During the second year, some additional documentary editing work would be undertaken at Rutgers to see how usable the archived resources are in practice.

5. The California State Archives will create and publish research guides in whatever form and format they find best, privileging topics of interest to the collaborating editing projects. They will use the shared Editors' Notes website for their working notes. [. . .]

The intention is not simply to rescue the working papers of the three editing projects from oblivion, but that "best practices" recommendations based on experience can be provided for the entire field of documentary editing. None currently exist.

Comment: The rationale for including the Labadie library collection in Phase 1 was not simply the usefulness of making detailed, expert curators' notes openly available but that this course of action opens up a possible renaissance in the active curation of library special collections. Working with archivists at the Rutgers University Archives and at the California State follows the same goal.

3. Documentation and Dissemination

Dissemination will be through multiple channels: an informative project website; technical reports; papers presented at conferences and published in conference proceedings; and articles in leading professional and technical journals, dealing with inquiries, and offering advice for others interested in doing the same or similar. (More below).

A checklist of tasks and a timetable are provided below.

ARCHITECTURE AND SOFTWARE

The Central Team established an Editors' Notes database and website hosted at Berkeley and designed to generate the varied forms of notes: *editorsnotes.org*.

The underlying data structure has three kinds of records:

1. *Notes*. Notes consist of text written by an editor. They are stored as html so that they may have hyperlinks and all the other features that html enables. Each Note is categorized based on its completeness: Notes are "Open" when they require more work; "Hibernating" when a resolution remains desired but appears impractical or of low priority; and "Closed" when deemed completed. Any Note can be revised at any time and all prior versions are retained and could be restored. The intended separate category of *Queries* is adequately handled by the "Open" and "Hibernating" categories.

2. *Documents*. Documents are records of source material that may be cited by an editor. We have used Zotero to manage Document metadata (e.g. item type, author, title, archive), enabling the input and output of Documents as structured bibliographic records. Documents may have attached Scans, Transcripts (with optional annotations), and hyperlinks to external websites.

3. *Topics*. All Notes and Documents are indexed using terms drawn from a controlled vocabulary of Topics which the interface uses to aggregate the Notes and Documents relevant to a specific person, organization, place, event, publication, or theme.

Topics may be person names, organization names, place names, event names, publication names, or names of topics or themes. We can think of these as subject authority records, with support for variant spellings, aliases, etc., but they can go beyond that, with support for various kinds of relations among Topics, e.g. personal relations between persons, involvement of persons and organizations in events, and so on. Just as Topics are used to index Notes and Documents, relationships between Topics might also be used for indexing and discovery. For instance if John Doe and Fred Smith were both schoolmates and business partners, their business partnership relation might be used to index a Document that is relevant to the latter relationship but not the former. Refinement of Topic records is central to Phase 2.

Permissions to view or edit different items are handled by Projects, which are made up of one or more users. Each Note, Document, and Topic is associated with all users who have edited it, as well as the Projects those users belong to. Projects may choose to restrict public access to certain items that they "own" as they see fit. Full records of changes to each item are stored, making it possible to view or revert to earlier versions.

This not a software development project, but an experiment on changing work practices in keeping with the general move from a print on paper to a digital environment. Numerous more or less suitable software tools are available and more will doubtless emerge during the next two years. The goal of widespread adoption dictates the use of software that is already widely used, well-supported, economical, and favored by local IT support services. The Editorsnotes.org website uses

- Django, the Python web framework
- Postgres, using native support for XML fields
- Xapian, for full-text search
- South, for database migrations
- Disqus, for discussion threads
- Zoom.it, for high resolution scans

- Zotero, for input and editing of bibliographical data
- Open Refine (formerly Google Refine), for duplicate detection

Django is an open source web application framework originally developed for the rapid production of news reports. Its primary goal is to ease the creation of complex, database-driven websites. Django follows the model-view-controller architectural pattern, emphasizes reusability and pluggability of components, rapid development, and the principle of DRY (Don't Repeat Yourself). Python is used throughout, even for settings, files, and data models. Django is opensource software now administered by the non-profit Django Software Foundation <http://www.djangoproject.com/>.

The Xapian search engine is fast, flexible, well-documented, and fully open-source. It is also well-integrated with the Django web framework that we plan to use as a platform <http://xapian.org >.

Design and architecture of new software

To support the scenarios envisioned above, we plan on adding the following components to the Editors' Notes system architecture:

1. Note-making tools
2. Reconciliation & linked data import tools
3. Backend storage of arbitrary linked data
4. Linked data editing and authoring tools
5. Sorting, filtering, and visualization

In the following sections we further explain each of these components with a typical scenario.

1. Note-making tools

An editor is creating a note about Bill Haywood's reaction to the Bisbee Deportations of 1917. After she types the letters "H-a-y-w-o-o-d" she presses the Tab key, triggering a search of the Editors' Notes Topic index for Topics with labels containing those letters. The preferred Topic label "Haywood, William 'Big Bill' (1869-1928)" appears in a small menu below the cursor, and the editor presses enter to confirm that this is the "Haywood" referred to in the note. The text of the note is left unchanged, but a link is created between the "Haywood" in the text and the Topic labelled "Haywood, William 'Big Bill' (1869-1928)".

Subsequently the editor types the letters "B-i-s-b-e-e" and presses Tab, again triggering a Topic search. This time no results are returned, so instead of candidate Topic names, a small "Create Topic?" menu appears with the options "Person," "Place," "Organization", "Event," and (selected by default) "Topic". She selects "Place" and is prompted to edit the preferred Topic label "Bisbee." She adds ", Arizona" to complete the preferred Topic label, and a link is created between the mention of "Bisbee" in the text of the Note and the newly-created Topic labeled "Bisbee, Arizona."

We currently plan on modifying the open-source wysihtml5 editor to provide this functionality.

For the most part the notes created are short self-standing textual documents always assigned Topics and often referring to other Notes and/or Documents. Our architecture also supports Annotations of Documents in the usual sense of a comment on a specific fragment of text. (See Fig 2). Both Notes and Annotations are compliant with the Open Annotation Collaboration's Open Annotation Principles http://www.openannotation.org/documents/GuidingPrinciplesRevised.pdf . Our software is currently and is intended to remain interoperable with the evolving draft Open Annotation Core Data Model. Community Draft, 09 May 2012 at http://www.openannotation.org/spec/core/. Notes and annotations will be able to be exported as OAC-compatible RDF if desired.

2. Reconciliation and linked data import tools

Note that the annotation tools described above simply provide a streamlined version of the indexing and annotation functionality already available in Editors' Notes. The truly new functionality comes into play with the reconciliation and linked data import tools.

After the editor creates the new Topic labeled "Bisbee, Arizona" she continues with her research and note-taking. In the background, however, Editors' Notes has begun querying a configurable set of linked data endpoints, trying to reconcile the label "Bisbee, Arizona" with identifiers managed by those endpoints. It discovers the following candidate identifiers: <http://viaf.org/viaf/137912931> (VIAF geographic name "Bisbee, Ariz."), <http://sws.geonames.org/5284905> (Geonames administrative division "Bisbee, Cochise County, Arizona, United States"), and <http://dbpedia.org/resource/Bisbee,_Arizona> (DBpedia city entity "Bisbee, Arizona"). Once this search for candidate identifiers has finished, an unobtrusive notification appears, letting the editor know that she can finish the reconciliation process when she wishes. Not wanting to interrupt her work, the editor ignores the notification.

The next day, the editor logs into Editors' Notes again. On her dashboard she sees a reminder that the new Topic "Bisbee, Arizona" has not yet been reconciled. She clicks on the reminder and is taken to page displaying data from VIAF, Geonames, and DBpedia. Scanning the data, she confirms that the entities identified within each of these services are equivalent to the Editors' Notes Topic, "Bisbee, Arizona." She clicks a button to finish the reconciliation process, thereby associating the three external identifiers with the new Topic.

In most cases this reconciliation process can be implemented by directly using the APIs provided by external linked data repositories, or by using a service such as sameas.org. However it may ease implementation to use middleware that provides a uniform interface for reconciling against any number of repositories, such as OpenRefine (formerly Google Refine) or Apache Stanbol Enhancer.

3. Backend storage of arbitrary linked data

Once a Topic has been reconciled against external identifiers, Editors' Notes will download and store any linked data associated with those identifiers. The external data sources will be also monitored for changes to the data associated with the external identifiers, so that new linked data may be added periodically. This data will not be stored in the relational database currently used to store Editors' Notes data, but in a graph database (or triple store) suitable for storing and query arbitrary linked data.[6] Each Editors' Notes Topic will have a (possibly empty) set of linked data assertions associated with it.

We will also investigate moving some of the data currently stored in the relational database to the graph database. For example, the bibliographic metadata associated with Editors' Notes documents, currently stored as JSON in a relational database field, might be better handled similarly to the linked data associated with a Topic and stored in the graph database.

There are a number of implementation options here to be investigated. We plan to explore two. The first option is directly accessing from Django a triple store (such as Virtuoso, 4Store, Sesame, or Jena) or graph database (such as Neo4J, OrientDB, or Titan). The second option is to use a higher-level framework such as Apache Stanbol (specifically the Entityhub and Factstore).

4. Linked data editing and authoring tools

[6] A triplestore assumes that all data conforms to the RDF data model. A graph database is capable of storing graph-structured data whether or not it strictly conforms to RDF.

Once linked data has been stored locally, editors should be able to make changes or additions to it. Furthermore, editors should be able to create structured data "from scratch" for Topics that are not reconcilable against external data sources.

Browsing the Topic page for "Haywood, William 'Big Bill' (1869-1928)" and selecting the Facts (i.e. linked data) tab, the editor sees all the structured data associated with the Topic, including Haywood's birthdate, birthplace, date of death, and place of death. She notices that the place of death is given as "Russia" and decides that more specifics are needed here. She presses the "Edit" button and is presented with an editable tabular view of the facts. She selects the "place of death" row and adds the values "Moscow" and "Soviet Union." She also adds a row asserting that Haywood was general secretary-treasurer of the Industrial Workers of the World (a separate Topic). Although this information was added to the "Haywood, William 'Big Bill' (1869-1928)" Topic, it will now also appear among the facts related to the "Industrial Workers of the World" Topic.

In addition to editing linked data associated with Topics, editors will also be able to create notes consisting of structured data rather than free text. (See the "Structured documents / chronologies" scenario above.)

This is one area where there is a distinct lack of suitable tools. We expect that we will need to build our own custom editing interface in Javascript, which would communicate with a linked data store following the recommendations made in the evolving W3C the Linked Data Platform Working Group's Linked Data Platform draft at http://www.w3.org/TR/ldp/ .

5. Sorting, filtering, and visualization

The sorting, filtering and visualization tools envisioned are described more fully in the scenarios presented above. In addition to the tools already identified, we plan on using jQuery to implement faceted sorting and filtering (building upon the code we have already written to sort and filter documents based on bibliographic metadata), and D3.js for visualizations of lists of items with associated temporal, spatial, and relational data. The intent here is not to compete with dedicated exhibition and visualization tools such as Neatline or Exhibit, but to integrate lightweight but useful visualization and organization functionality into the note-finding interface.

These choices could change if more satisfactory options emerge. Where collaborating institutions prefer other tools, we will work with them to migrate their data from Django to their local software of choice. We seek to avoid local software development except in cases where it is necessary. Any new software would be made freely available as open source at no cost.

[. . .]

SUMMARY AND OUTCOMES

Scholarly annotated editions of historically significant texts constitute an important foundation for learning and research in the Humanities. Scholarly editing, however, requires a sustained investment of highly specialized expertise and long-term funding is difficult.

Given the right software, minor changes in work practices can make the painstaking editorial research much more organized, convenient and rapid, widely accessible, and permanent, thereby increasing utilization, efficiency, and the return on investment. The move to web-accessible in Phase 1 was a pivotal change. We now seek to bring the functionality of advanced Digital Humanities projects to the editor's workbench and to ensure that resources so carefully assembled by editors remain capable of supporting future projects.

The direct outcomes of his project are (1) the incorporation of digital humanities technology into everyday editing practice through making links, efficient enriching of local data,

and visualizations; (2) the development and demonstration of efficient procedures for archiving editors working notes for future use; and (3) extension of these technique to the work practices of professional archivists.

The broader outcomes include technology transfer of Digital Humanities techniques from demonstration projects to ordinary scholars, the sharing of research materials beyond the academy.