Saving Working Notes for Future Use

Michael K. Buckland and Barry Pateman, University of California, Berkeley, USA.
Patrick Golden and Ryan B. Shaw, University of North Carolina, Chapel Hill, NC, USA.

Abstract

Scholarly annotated editions of historically important texts constitute an important foundation for understanding history and culture. Their preparation requires a sustained investment of highly specialized expertise and generates a rich corpus of research notes that are mostly not included in the eventual published volumes. Ordinarily these research resources are simply discarded. We examine the challenge of preserving working notes so that they could become the basis for future research and describe a shared website editornotes.org.

Motivation

Scholarly, annotated editions of historically important documents ("documentary editions") constitute an important resource for research and education in the humanities. The preparation of documentary editions requires expensive expert preparation over many years and funding is difficult. Extensive files of working notes are carefully compiled, including structured name authority files, itineraries, and chronologies. These editorial research resources are not shared with other scholars and are included in abbreviated form (if at all) in the eventual published volumes when the editorial staff are dispersed and the working notes are usually discarded. As with the data sets of scientific research, preserving these editors' research resources and making them openly available could significantly increase the return on the large investment in documentary editing projects. There is wide interest in developing digital editions and in techniques for attaching annotations to existing texts. Our concern is different and unusual. We are concerned with the working notes themselves.

Editorial Practices and the Web [3] is a collaborative effort with three major documentary editing projects with overlapping interests in late nineteenth-century and early twentieth century radical and feminist movements in the United States: The Emma Goldman Papers (University of California, Berkeley); The Margaret Sanger Papers (New York University); and the Elizabeth Cady Stanton & Susan B. Anthony Papers (Rutgers University). The initial problem was to make the editors' working notes accessible within projects, between projects, and to the public with minimal change in the work practices of the editors and their assistants. Other tasks focused on finding ways to integrate these research resources into the networked digital humanities environment are in progress.

Problem

The problem addressed here is how to sustain availability when projects have ended but might be resumed by other different editors. A focus on charismatic and heroic historical figures provides an attractive approach to history and appeals to funders, but what is also needed is an understanding of important movements and groups going beyond the roles of Great Men and Great Women. Because editing the papers of notable figures requires an understanding of the context and relationships of the individual being documented, the resources assembled could also support wider insights. Thinking tactically, we could examine what low-cost procedures could move these editorial resources into a preserved and accessible archive.

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. 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.

Scholarly communication could be greatly extended if scholars anywhere had *sustained access* to the working notes and if scholars anywhere could *add supplementary notes, corrections and additions to them* (with clearly separate attribution) in the future *as interest, ability, and resources allow*. This is a logical consequence of 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 which 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 seems little precedent except in local community archives and open note-book science.

The requirement that makes existing archival theory inadequate is that working notes might resume their role as working notes after a pause. The project might resume later with the same goals if future funding were found; or the corpus of notes could become the basis for a new project with a different but related goal. The corpus should remain usable by scholars anywhere during any pause and the curated notes could reflect the editors' knowledge of any particular topic at any point in time. Notes, like subject headings, are inscribed at point in time and obsolesce as scholarship and time flow forward.[2]

There is a steady move to digital notes. Retroactive digitization is unlikely and all such corpora will remain a combination of paper and digital records. Only very low cost procedures will be acceptable.

Approach

The work practices of the participating projects were evolved so that as working notes were written or revised they were stored in a shared website designed to support shared access within projects, between projects, and, now, open public access. The focus now is to make the highly isolated, unpublished research

resources of editorial projects available in the future using a three-part strategy: (1) The mostly paper-based resources of an editorial project now completed after thirty years are being processed for conventional archival deposit and the potential repurposing of specialized data sets examined; (2) The lessons learned from doing that are being retrofitted to the work practices of two on-going editorial projects; and (3) Low-cost, low-effort tools for routinely linking working notes with datasets being published by libraries, archives, and other digital humanities projects are being developed.

Organization

Users of the system are grouped into Projects. A Project might just have one User, and Users may participate in multiple Projects.

Data stored in the system is modeled as instances of three primary types: Document, Note, and Topic. See Figure 1 Data Model. Instances of these types are controlled by the project that created them. This allows projects to explicitly grant edit and access permissions for their original material as they see fit.

A Document is a bibliographic description belonging to a Project. It may have structured data (key-value pairs) compatible with the representation of Items in Zotero. This allows users to easily import their bibliographic descriptions from their Zotero library, and edit or reference them in Editors' Notes. A scan or a transcript of the cited Document may be added.

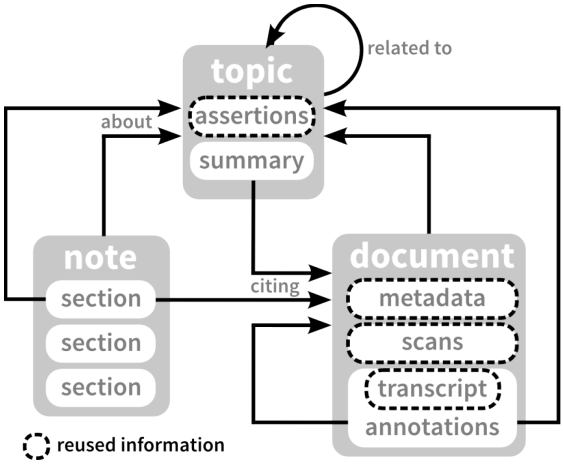


Figure 1: Data model. Notes, sections of Notes, and Topic summaries may cite Documents. Document annotations are linked to the Topics to which they relate. The Topic's "summary" is for free-form textual description of the Topic. Assertions are pieces of structured data created locally and/or imported from a trusted resource. These become a source for specialized search and visualization interfaces.

A Note is a record of a Project's attempts to articulate and possibly answer a research question. It also may be used to exhibit evidence of how a Project articulated and possibly answered a research question in the past, as when Editors' Notes are used to publish a Project's past notes (which might be stored in an archive or repository of scholars' work). A Note has a title, content stored as XHTML, and may have a number of Sections of different kinds. A Note Section may 1) cite a document and have as content notes associated with that document, 2) refer to another note and have as content notes associated with that other note, or 3) simply have some content unrelated to any specific document or other note. Notes are not necessarily annotations. Each Note has a status: "Open" when being actively worked on; "Closed" when considered sufficiently completed; and "hibernating" when inactive because of low priority or cannot be resolved. By default all Notes are copyrighted but Projects can choose to license individual notes more openly, e.g. under a Creative Commons license.

Topics are subject headings (including proper names) used to index, search, and sort Notes. Like other objects in the site, they are project-specific. However, Topics in different projects that refer to the same things (e.g. people, organizations, concepts) can be linked across projects. In this way, we are able to aggregate relationships between items site-wide while still allowing individual projects to maintain control over their own naming conventions and Topic descriptions. This architecture is similar to the "cluster" system used by the Virtual International Authority File <http://viaf.org/> to link the name authority files of different library systems.

Software

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. The goal of widespread adoption dictates the use of software that is already widely used, well-supported, economical, favored by local IT support services, and open source. As of March 2014 the Editorsnotes.org website uses

Django, the Python web framework
Postgres, using native support for XML fields
ElasticSearch, for full-text search
Zoom.it, for presenting 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/>.

Throughout our implementation of Editors' Notes, we have attempted to keep in mind the present and future use of the data that is produced by the system. Three areas in which this has come into play are text editing, bulk access, and version control.

The production of well-formed, predictable markup in authored content is important for ensuring the long-term preservation and reuse potential of data within Editors' Notes. To this end, we sought to include a web-based text editing interface that could produce quality markup while also being easy to use.

However, even standard word processing features such as links, headings, lists, quotations, and text emphasis are difficult to implement well when using the web as an editing platform. Prompted to complete the same actions (for example, "insert a heading, insert some text, insert a new line, insert a list, delete the list, insert more text"), different versions of different browsers will produce very different markup. While web developers have created many different solutions to this problem over the years, none are as sophisticated and mature as we would hope. Accordingly, we have been steady and deliberate in adding text editing features in order to normalize these inconsistencies.

Over the past year, we have developed an API (Application Programming Interface) which enables direct access to the data within Editors' Notes. This will aid in the preservation of working notes, since users and developers can access current dumps of data from the site without resorting to web crawlers or similar tools. This data is formatted as JSON (JavaScript Object Notation), which can easily be processed and converted to different formats based on archiving needs. The API also enables outside systems to both read from and write to Editors' Notes, allowing it to be integrated with existing tools and workflows.

Every time a user edits an item within Editors' Notes, a record of that change is saved along with a snapshot representation of the item at that time. While it has been difficult maintaining these snapshots in light of progressive schema changes, they have proven invaluable to us and to users for tracking and managing changes in data over time. We plan on making this version history more available in the future by exposing it in the API.

Use by Editors

A typical workflow might be:

1) Create a Document
2) Create a Note that draws on that Document
3) Create a Topic grouping Documents and Notes
4) Create datasets relating Topics

None of these things need happen in that order.

External Use

In August 2012 password control to the site was quietly removed making the site openly available to both humans and webcrawlers. 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.

Results

Now, the http://editorsnotes.org/ website achieves "open note-book science" in the Humanities for recent and revised notes. The growth of Notes is slow but cumulative, so the benefits are expected to increase gradually.

The notes in this form, however, constitute only a small fraction, the most recent portion, of the whole. When the editorial project has maintained well-organized series of files, a list of the files can be annotated by editorial staff into an acceptable archival finding aid.

A significant portion of digital notes created a decade or more ago using obsolete software and/or storage media are effectively lost. More recent digital datasets (name authority files, chronologies, itineraries, etc.) can be migrated, when necessary, to current version, of standard software. Shared access within a Project facilitates collaboration and supervision within the Project.

There were some 8,000 external visits to the website during February 2014. Emails expressing gratitude indicate that at least some visits were of practical benefit.

Future Plans

Initial deployment was significant because it moved routine day-to-day procedures from the isolated 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? Projects interested in using these technologies could enlist the help of specialists, but this sort of work would be more achievable, more affordable, and more sustainable if it could be done by the editors.

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. Software integration and interface design must lead 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 as Assertions. 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, a map display could show any location(s) mentioned in Notes, with options to display the locations in 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.[5]

Conclusions

Documentary editions are funded to achieve the eventual published editions, and the editorial staff and their working notes are merely means to that end. Funding does not (yet) support the preservation and access of the research resources generated but not incorporated into the published volumes. There are two disadvantages in this situation.

First, it reflects a short-term perspective that conflicts with the realities of scholarship in the humanities. Projects start and end, but scholarship continues, so the discarding of research resources is counterproductive. If it were known that future funding would become available and that the same editors could resume work on additional volumes for publication, the editorial debris might be left in situ. Such certainty is unlikely and the materials would be hard to use during the interval or later.

A conventional solution is to process these records for archival deposit and they could be consulted by occasional researchers. But if a new editor came along with significant funding, then the deposited archive, like Sleeping Beauty (Dornröschen) kissed by a handsome Prince, could have a new, exciting life. Orthodox archival theory considers it inappropriate to reactivate or reorganize an acquired archive, wanting to preserve the original condition, order, and authenticity of the materials. More recent theorizing, especially as it relates to born-digital, born-networked, and digitized materials, has been contemplating ways in which the organic "living" nature of the by-products of human activities can be preserved even after their transfer into archival control. A "hibernating" archive, such as that proposed here, might be one such approach.[1][4]

Second, it is incompatible with the data management plans increasingly mandated by research funders. The return on investment is greater if research materials can be repurposed by other subsequent researchers. When data management plans are required for this genre of material new work practices will be needed.

In the print environment prior research is incorporated by reference (citation), by quotation, and by summary. The technology does not allow otherwise. In a digital environment, one can do the same and also use "save as" as a basis for developing new, improved, derivative versions.

In the case of working notes as a genre the same techniques can be used (inspection, citation, quotation, and summarization), but if one thinks of a documentary editing project as a potentially enduring workshop, then a reversal of the relationship between working notes and published volumes is indicated: Instead of the working notes being a dispensable means to the eventual published volumes, the published volumes become valued, occasional by-products of an enduring "workshop" (research resources and expertise). Seen this way, the genre of working notes requires a new kind of archival practice, one that recognizes that this genre of material could and should be treated not as a static, frozen deposit but as a hibernating resource.

References

[1] Boles, Frank (1982). "Disrespecting Original Order". *American Archivist* 45 (1): 26–32.

[2] Buckland, Michael (2011). "Obsolescence in subject description." *Journal of Documentation* 68, no 2 (2011):154-161.

[3] *Editorial Practices and the Web.* [Website] http://ecai.org/mellon2010
Also: *Editors' Notes* [Website] http://editorsnotes.org/

[4] Fenyo, Mario (1966). "The Records Group Concept: A Critique". *American Archivist* 29 (2): 229–239.

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

Acknowledgments

The authors gratefully acknowledge the support of the A. W. Mellon Foundation, the Coleman Fung Foundation, and the School of Information and Library Science, University of North Carolina, Chapel Hill.

Author Biographies

Michael K. Buckland is Emeritus Professor , School of Information, University of California, Berkeley. He has worked as a librarian and academic administrator and is interested in the history of documentation and the intersection of digital libraries and digital humanities.

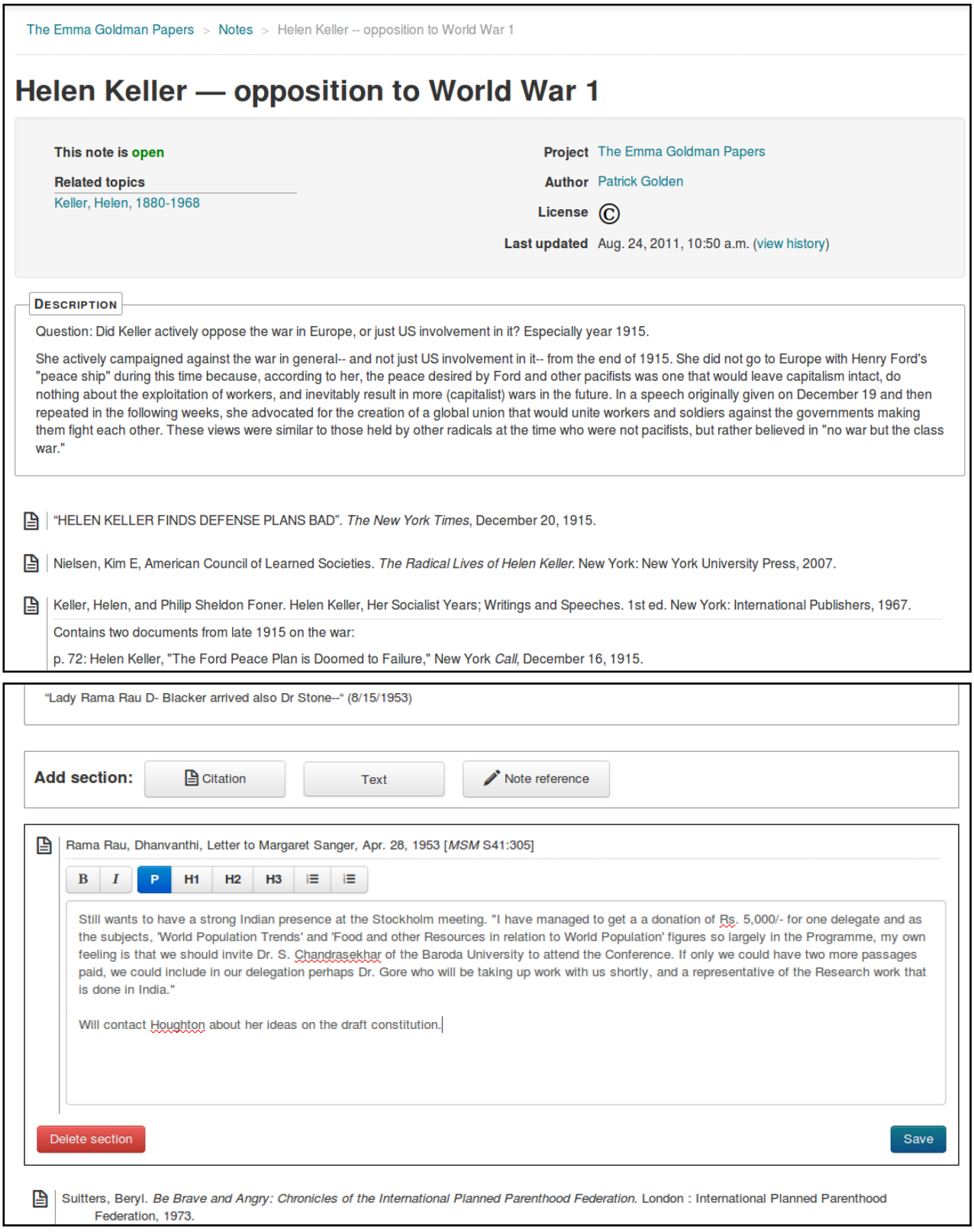
Barry Pateman, PhD, is Senior Editor, Emma Goldman Papers Project.

Patrick Golden, graduate student at the University of North Carolina, Chapel, formerly worked on the Goldman Papers.

Ryan B. Shaw, Assistant Professor, School of Information and Library Science, University of North Carolina, Chapel Hill, specializes in Web technology in the Humanities.

The four authors are collaborating in the *Editorial Practices and the Web* project. See http://ecai.org/mellon2010/

Appendix: Example of a Note and snapshot of a revision being made.



The Emma Goldman Papers > Notes > Helen Keller -- opposition to World War 1

Helen Keller — opposition to World War 1

This note is open

Related topics
Keller, Helen, 1880-1968

Project The Emma Goldman Papers
Author Patrick Golden
License ©
Last updated Aug. 24, 2011, 10:50 a.m. (view history)

DESCRIPTION

Question: Did Keller actively oppose the war in Europe, or just US involvement in it? Especially year 1915.

She actively campaigned against the war in general-- and not just US involvement in it-- from the end of 1915. She did not go to Europe with Henry Ford's "peace ship" during this time because, according to her, the peace desired by Ford and other pacifists was one that would leave capitalism intact, do nothing about the exploitation of workers, and inevitably result in more (capitalist) wars in the future. In a speech originally given on December 19 and then repeated in the following weeks, she advocated for the creation of a global union that would unite workers and soldiers against the governments making them fight each other. These views were similar to those held by other radicals at the time who were not pacifists, but rather believed in "no war but the class war."

"HELEN KELLER FINDS DEFENSE PLANS BAD". *The New York Times*, December 20, 1915.

Nielsen, Kim E, American Council of Learned Societies. *The Radical Lives of Helen Keller*. New York: New York University Press, 2007.

Keller, Helen, and Philip Sheldon Foner. Helen Keller, Her Socialist Years; Writings and Speeches. 1st ed. New York: International Publishers, 1967.

Contains two documents from late 1915 on the war:

p. 72: Helen Keller, "The Ford Peace Plan is Doomed to Failure," New York *Call*, December 16, 1915.

"Lady Rama Rau D- Blacker arrived also Dr Stone--" (8/15/1953)

Add section: Citation | Text | Note reference

Rama Rau, Dhanvanthi, Letter to Margaret Sanger, Apr. 28, 1953 [*MSM* S41:305]

B I P H1 H2 H3

Still wants to have a strong Indian presence at the Stockholm meeting. "I have managed to get a a donation of Rs. 5,000/- for one delegate and as the subjects, 'World Population Trends' and 'Food and other Resources in relation to World Population' figures so largely in the Programme, my own feeling is that we should invite Dr. S. Chandrasekhar of the Baroda University to attend the Conference. If only we could have two more passages paid, we could include in our delegation perhaps Dr. Gore who will be taking up work with us shortly, and a representative of the Research work that is done in India."

Will contact Houghton about her ideas on the draft constitution.

Delete section | Save

Suitters, Beryl. *Be Brave and Angry: Chronicles of the International Planned Parenthood Federation*. London : International Planned Parenthood Federation, 1973.