THE MEDICAL HERITAGE LIBRARY

AN INFORMATION SOURCE AND A COMMUNITY SPACE

WWW.MEDICALHERITAGE.ORG

A White Paper Prepared in Partial Satisfaction of a NEH Digital Humanities Start Up Level 1 (Grant # HD-51239-11)

Lori M. Jahnke
The College of Physicians of Philadelphia
University of Pennsylvania

June 5, 2012

CONTENTS

EXECUTIVE SUMMARY

This Level 1 Start Up grant was to support planning for the development of the Medical Heritage Library (MHL) as a multi-institution digital collaboration. We used this grant to engage our stakeholders and to incorporate the information from these discussions into our planning process. Since the challenges faced by users of digital content and those who build collections resonate across professions, this whitepaper is largely a description of the themes that emerged from our interviews with various stakeholders[1]. My colleagues and I interviewed more than 50 researchers, faculty, students, administrators, and technologists on the use of digital content in research and teaching. Our discussions touched on a variety of topics, such as current research projects, methods of content discovery and use, anticipated future research, the influence of technology on their research areas, desired types of content that are currently difficult to find or access, and the use of digital content in teaching.

USER NEEDS AND PRESENTING DIGITAL CONTENT

The researchers interviewed during the MHL site visits fit into two broad categories with differing needs for content access. Interface design must allow flexibility in the way that content is presented and accessed.

- *Data Oriented Scholars* are primarily interested in the content of the object regardless of its form.
  Key access features include: a variety of downloadable formats, the ability to create lists with download links, a viewing interface that facilitates rapid scanning of content (no page turning interfaces), options for metadata download and management, and the ability to search across formats and collections.
- *Context Oriented Scholars* are interested in the materiality of the object and its context.
  Key access features: The ability to zoom in, page turning interfaces or other functions that mimic the physicality of the book, organization of content into coherent collections that are thematically based (single or multiple format), and image or video displays that mimic the convenience of using a light box, for example.

BARRIERS TO THE USE OF DIGITAL RESOURCES

Researchers generally agreed that digitization offers numerous advantages for both research and teaching, but much could be done to improve access to digital content, as well as its general utility. Additional effort should be expended to:

- Aggregate high quality digital content, digitize additional content to create more intellectually robust content pools, and link relevant digital resources (e.g. health or geographic datasets, and images)
  - For example, disaggregated or incomplete content is not only inefficient, but it can promote facile interpretations of historical information, particularly among less experienced researchers.
- Improve or enrich metadata in ways that will help users identify:
  - Whether or not they are viewing is a trusted resource
  - Why and how a collection was created

[1] Many of the project guidelines that resulted from this grant are available on our website (www.medicalheritage.org), or contact us at medicalheritage@gmail.com and we would be happy to share examples.

    - Items that include original covers and advertisements
    - Images or data tables within texts (e.g. explore the potential for applying document layout analysis tools such as OCRopus (https://code.google.com/p/ocropus/), or OCRFeeder (http://live.gnome.org/OCRFeeder) to index images within the text)
- Apply Semantic Web or other technology to improve search performance over linguistically heterogeneous content.

Despite the many advantages offered by digitized content, several scholars pointed out the limitations of digitized content and emphasize that it is an adjunct to the physical object rather than a replacement. The physical book is particularly important for the classroom. Although students may have high expectations for the inclusion of digital media in the classroom, several faculty members reported that physical books and other examples from archives were most effective for engaging students in the historical material.

GROWTH OF THE MEDICAL HERITAGE LIBRARY

The Medical Heritage Library (MHL, www.medicalheritage.org) was initiated in 2009 with a grant from the Alfred P. Sloan Foundation. This grant provided five institutions with the funds to digitize approximately 30,000 volumes in the history of medicine. From this small start the MHL has already grown into a collaboration of more than 10 partner institutions with several additional content contributors. The Sloan grant provided a much needed opportunity to foster institutional collaboration in the history of medicine, and in some ways, it also provided an experiment that would gauge whether such diverse organizations could work together to create a resource that is invaluable for the public good. The answer is, undeniably, yes. The goal of the MHL is to create a content centered digital community supporting research, education, and dialog that enables the history of medicine to contribute to a deeper understanding of human health and society. This NEH Start Up grant has provided the MHL partners with a key opportunity to engage our stakeholders and to incorporate this information into our planning process. While the diversity in organizational structures and cultures creates a steep learning curve for project planning, we are well on our way to combining the best from each institution in terms of expertise, flexibility, and stability.

Since we submitted the NEH Start Up proposal in 2010, the landscape of collaborative digital projects has changed dramatically with the development of the Digital Public Library of America (DPLA, http://dp.la/). Although the creation of the DPLA did not change our activities significantly, it has changed the perspective of many of our discussions. The development of the DPLA may provide new opportunities for collaboration in terms of technological infrastructure that the MHL would like to support. Such collaboration could allow continued flexibility in MHL structure and activity that, thus far, has been critical to our success. With these new developments in mind, we turned our attention to the unique needs of history of medicine content, its users, and the potential for new audiences.

The MHL partners have devoted significant time to project planning that has resulted in a series of guidelines for our future activities. Since many of these guidelines are specific to the MHL project, I do not discuss them in this report[2]. Rather, this whitepaper is dedicated to the themes that emerged from our interviews with researchers and other stakeholders. The challenges faced by users of digital content, as well as those who digitize and provide access, resonate across subject areas and professions.

In order to gain a better understanding of user practices and how the MHL might serve unmet needs, my colleagues and I interviewed more than 50 researchers, faculty, students, administrators, and technologists on the use of digital content in research and teaching (Table 1). The institutions visited include Harvard University, Yale University, Columbia University, Johns Hopkins University, the National Library of Medicine (NLM), and the New York Academy of Medicine (NYAM). I attempted to visit the New York Public Library (NYPL), but circumstances conspired to prevent this. Although NYPL was not part of the NEH planning grant, they were an original participant in the MHL digitization project funded by the Alfred P. Sloan Foundation and we would like to encourage their continued involvement.

[2] Many of the guidelines are available on our website (www.medicalheritage.org), or contact us at medicalheritage@gmail.com and we would be happy to share examples.

Table 1. Summary of participation in the MHL site visit interviews.

Site	Visit Date	Dept/Divisions Included
Columbia University	10/11 - 10/12/11	School of Public Health CU College of Physicians and Surgeons Pediatrics, Columbia University Medical Center Religion Journalism Library Sociomedical Sciences Art History and Archaeology
Harvard University	10/25 – 10/26/11, 11/16/11	History of Science Global Health and Social Medicine Anaesthesia, Medical School Internal Medicine HMS Center for Bioinformatics Graduate School of Arts and Sciences
Johns Hopkins University	12/1 – 12/2/11	School of Medicine University Libraries and Museums Health Sciences Informatics, SOM Alan Mason Chesney Medical Archives Bloomberg School of Public Health Institute of the History of Medicine Art as Applied to Medicine Bioethics and Health Policy
Yale University	11/14/2011	History of Science and Medicine History of Art Psychiatry, School of Medicine Dept of History Medical Library World Oral Literature Project
New York Academy of Medicine	5/4/2012	History of Medicine Library Finance and Administration
National Library of Medicine	10/20/2011	Profiles in Science Library Operations History of Medicine Division Exhibition Program Digital Repository Implementation Group Public Services

USERS AND UNMET NEEDS

Users interviewed during these site visits primarily included university faculty, graduate students, and affiliated researchers[3]. Our discussions touched on a variety of topics, such as current research projects, methods of content discovery and use, anticipated future research, the influence of technology on their research areas, desired types of content that are currently difficult to find or access, and the use of digital content in teaching.

Broadly speaking, the users described in this report fall into one of two categories: those who are primarily interested in the content of the object regardless of its form, and those who are interested in the materiality of the object and its context. As shorthand, I refer to these two groups as data oriented and context oriented, respectively. Data oriented scholars are primarily interested in the information contained in the object rather than the object itself. Thus, they prefer to access large amounts of content expediently in order to assess its relevance and download the qualifying content. A data oriented scholar may manipulate and transform content for use in a scholarly product or integrate the content in a local information management system. For example, a data oriented scholar may develop datasets using multiple sources and formats, extract data for maps, or create digital projects of their own.

- Key access features include: a variety of downloadable formats, the ability to create lists with download links, a viewing interface that facilitates rapid scanning of content (no page turning interfaces), options for metadata download and management, and the ability to search across formats and collections.

Context oriented scholars tend to use digitized content as a surrogate for the analog object. They are also more likely to be interested in digital resources as collections and seem to view them as analogs to traditional library collections.

- For these researchers key features include: the ability to zoom in, page turning interfaces or other functions that mimic the physicality of the book, organization of content into coherent collections that are thematically based (single or multiple format), and image or video displays that mimic the convenience of using a light box, for example.

Context oriented scholars reported that they are also interested in downloading digitized content, but this is typically for convenience when working offline or when traveling. For this purpose, a searchable PDF or high-resolution image is usually sufficient.

Not surprisingly, the two types of users diverge in their descriptions of a high quality interface and it would be a mistake to provide an interface that requires users to interact with digital content in uniform ways. For researchers who need to view hundreds of items (i.e. data oriented scholars) page turning interfaces or features like Zoomify are frustrating and unnecessarily time consuming. As one researcher commented:

> *For big images, they use Zoomify. There's a big image file on the server and you can zoom in and out directly in the web browser. But this drives me bonkers because I want to download the images and zoom in and out myself and use them in my lecture slides rather than always going to the web. If they're trying to keep control of it [the image], I understand, but if they're doing it to try to be user-friendly it would really be more convenient just to let me download it. (Assistant Professor, History of Science)*

[3] The NLM visit also included discussions with exhibits staff.

Conversely, for researchers who are conducting a close study of a few items, a page turning interface is an asset that helps provide some sense of the context for an image or passage of interest.

> *The page turning interface of [one site on illuminated manuscripts] is invaluable for letting you see how the images relate to one another within the context of the book. Since you don't see everything all at once when you look at a real manuscript, you can't make the side-by-side comparisons art historians love to make between one page at the front and another at the back; when you look at things in sequence, flipping from one page to the next and back again, they take on new meaning... so I really love that function in dealing with the digitized manuscripts. (Assistant Professor, Medieval Art and Architecture)*

However, researcher attitudes converge when it comes to navigating the proliferation of resources and information available to them. Many interview participants expressed frustration over the lack of information regarding the context of the digital resources, which would help them evaluate the quality of the content (e.g. who created the collection, why, and how). Aggregation of high quality content by trusted parties and providing additional input for vetting sources would add significant value for scholars. These features are equally important for supporting non-scholarly audiences (e.g. K-12 students, casual researchers, undergraduates, and the general public) in exploring history of medicine content.

DISCOVERY

The discovery process was relatively uniform among the researchers. Most of the researchers interviewed use simple keyword searches in a range of environments with a preference for systems that offer full-text searching.

> *It's very rare that I do what I think a lot of these interfaces are asking for, which is to string together a bunch of keywords and then browse through results sorted by relevance. And often I think the relevance algorithm doesn't actually work very well. If I have 1,000 results arranged from most to least relevant, that's not as helpful to me as a finite list, for example 60 results of the word 'economist' that I can go through completely. I don't know how it happens, but no matter what I'm searching for I seem to get the same things popping up as the "most relevant." These algorithms are good at finding two or three things you want to look at right away, but not so good for getting deep into a collection, and certainly not good for making me feel like I've exhausted my possibilities. (Assistant Professor, History of Science)*

Nearly all of the researchers interviewed reported using Google Books and Google Images as their primary resources for discovery. Scholars invest the majority of their labor in reviewing and vetting sources. When exploring a brand new topic most of the researchers interviewed attempt to identify key figures in the field and then review the associated publications and cited works. Some of the researchers also identify key figures by searching Wikipedia topics, asking a colleague, or consulting resources like the NLM Syllabus Archive (http://www.nlm.nih.gov/hmd/collections/digital/syllabi/). Other commonly used resources include JStor, the Library of Congress, sites such as Gallica (http://gallica.bnf.fr/?lang=EN), and library staff (especially for archives). One researcher emphasized the culture of the archive and knowledgeable staff as being particularly important to the research endeavor:

> *Why I'm even here in the United States in a university like that is really because of the librarians. Honestly, it's any researcher's paradise. I go to the archives and you encounter people with extreme competence and it's just really fun to do research under such circumstances. Where I grew up everything is being withheld from you and you are always a burden to an institution. You go to the National Library in Vienna*

> *and they say, 'Well, there are all these reasons why you can't see this.' It's a completely different culture... (Research Professor, German and Romance Languages and Literatures)*

The extensive use of Google Books and Google Images bears further discussion, since it may illustrate a shortcoming of local systems and the difficulty in navigating fragmented digital resources. While researchers acknowledge problems with metadata accuracy, relevance algorithms, a lack of search transparency, and the spotty or limited catalog of books, the benefit of simultaneously searching across collections and within items typically outweighs these concerns. In fact, the researchers interviewed for this report only mentioned their local systems when expressing frustration with its limitations (e.g. finicky treatment of search terms, irrelevance of LC subject headings) and the regrettable need to consult multiple library catalogs. One researcher went on to say he would be happy to move away from Google and use a search interface that was more suited to his needs, but he would not be willing to use multiple search interfaces. As a short-term solution, he suggested making local library holdings searchable in Google, even when they are not digitized. This would save considerable time in the discovery process.

Other researchers echoed his sentiments and only preferred searching within a collection when they are certain it is high quality and relevant to their research (e.g. Life Magazine, http://www.life.com/, David Rumsey Map Collection, http://www.davidrumsey.com/). However, focused and carefully constructed digital collections can be of limited use for the non-specialist if adequate attention isn't given to the presentation of content and metadata.

> *Right now I'm teaching a course on the body and medieval art. There are 15 grad students, and most of them are not medievalists. If someone says, 'Alright, I want to write my paper on depictions of dancing! Where do I find them?' I can point her to a couple of things to read, and if she looks at those books there she'll find footnotes to that'll lead her in the right direction. There's the Index of Christian Art [http://ica.princeton.edu/], which is a fantastic database of iconography, but it's still basically a card catalog transposed into digital form, so if you type in 'images of dancing 13th century all media' you get this list of things, but no images and not even a clear description. Unless you're really a specialist, it becomes hard to use. If there were more sites like the various ones that deal with different library collections or topics, where you could do a search for 'images of tooth drills' or something like that and be able to come up with good images and descriptions, that would be so nice! (Assistant Professor, Medieval Art and Architecture)*

Although several researchers reported that they like the HathiTrust discovery interface and its visualizations of search returns, few use it regularly due to the limited content. While Internet Archive content is searchable through Google, the lack of precision and opaque item rankings hinders scholarly work and ultimately the utility of the MHL content. The scholars discussed above seem to be making do with Google Search until something better comes along that preserves the value of curatorial expertise and offers the efficiencies of the digital environment.

> *I've actually been really impressed by JStor. It manages to be monopolistic in a good way. It really has most of the big stuff. And the search function is great – it has lots of options, by field, by journal, etc. And the OCR is good and the metadata is good. I'm very happy to go to JStor instead of the web or Google books. It would be great if there was a way to have a similar sensibility for collections of things like images, maps, datasets, and books, whatever it might be. With JStor it's very clear that pieces have been added one by one and not just dumped in* en masse *with the hope that it will work. (Assistant Professor, History of Science)*

CONTENT

Researchers point to a number of content areas that could be improved, but the most common requests were related to popular culture sources (e.g. magazines, television, and radio), advertisements, and complete runs of series. Advertisements and popular culture sources are particularly problematic since they have rarely been collected by libraries and they now provide important cultural context for research problems. Scholars have resorted to creative strategies such as using E-bay, thrift stores, and farmers' markets to find old magazines of various types, as well as books with the covers intact. Researchers lament the fact that advertisements and covers were routinely excised during binding and they will make extra efforts to identify collections that might have journals with the advertisements and cover illustrations intact. The MHL can provide some support in this area by identifying and aggregating content already in Internet Archive and facilitating links with additional resources. Adding a metadata field to indicate when advertisements are present would also be of great benefit to researchers.

> *Sometimes I find myself wanting to see the cover of the book, but usually it doesn't exist. For example, for the lecture I'm giving tomorrow I'm talking about a book published in the 1970s and I want to show the cover image from the book. But if you go to Amazon or Google all they have is covers from later editions; if you go to the library, the cover's been cut off. (Assistant Professor, History of Science)*

Researchers also expressed frustration with a lack of access to long runs of series of various types (e.g. trade journals, government reports, and academic journals). If a series is available digitally, it is often incomplete or it is presented in a way that does not allow the researcher to locate and view the entire series efficiently. This is a known issue to the MHL that we hope to begin addressing it in our recently funded journal digitization project ("Expanding the Medical Heritage Library: Preserving and Providing Online Access to Historical Medical Journals", NEH Division of Preservation and Access, Grant #PW-51014-12).

Additional content areas of interest include oral histories (both collecting and analyzing), images (very high demand), archives, and indexes of digital resources including exhibits. Researchers also reported using a variety of geographic and health data sources, including:

- *Health, Demographic, or Other*
    - US Census, http://2010.census.gov/2010census/data/
    - CDC, http://www.cdc.gov/datastatistics/
    - World Bank, http://data.worldbank.org/
    - JStor DFR, http://dfr.jstor.org/
- *GIS Data*
    - Harvard Geo-spatial library, http://dixon.hul.harvard.edu:8080/HGL/hgl.jsp
    - National Historical GIS, https://www.nhgis.org/
    - Gebco, http://www.gebco.net/
    - GeoData@Tufts, http://geodata.tufts.edu/ - part of the OpenGeoportal Consortium, http://opengeoportal.org/
- *Tools and Software*
    - National Atlas, http://www.nationalatlas.gov/
    - Global Mapper, http://www.globalmapper.com/
    - USGS Seemless, http://seamless.usgs.gov/

SCHOLARLY PRODUCTS AND TEACHING

I have grouped scholarly products with teaching to reflect the trend and general interest in producing multimedia digital projects through the classroom. The researchers reported that their primary products are books and articles since these items count toward tenure and professional advancement[4]. Other outlets for research include blogs oriented to casual readers, producing materials for adult education, and writing for public health policy professionals.

Several researchers expressed interest in producing multimedia publications and cited the addition of oral histories, film, or other items as significant enhancements to more traditional scholarly communications. Aside from receiving little or no professional credit for this work, researchers cited a lack of time and expertise to create these products on their own. Several of the researchers also draw an explicit connection between research, multimedia scholarly products, and teaching. They discussed the importance of using various kinds of media to engage students and the potential for adopting new scholarly formats that could decrease the distance between active research and the classroom.

> *I love sending students to go and explore this or that website, or watch this video that shows you how, say, the process of bronze-casting works. It's a million times better than reading some encyclopedia article about it, which is what we used to do. The fact that people are continuing to go forward with the question of 'How do we present this information in the most rich and detailed way?' is interesting. That's the sort of thing that I'm going to latch onto if it's done really well… It's so exciting—the possibilities that are out there that are allowing people to write new narratives about their materials. It's just really, really thrilling. Even as my precious books are becoming obsolete. (Assistant Professor, Medieval Art and Architecture)*

Several faculty members also expressed interest in facilitating student-produced multimedia digital projects, but they are hindered by a lack of resources or technical support.

Nearly all of the researchers interviewed use Google Images as their primary discovery tool when preparing for lectures. They also utilize sites such as the NLM (http://www.nlm.nih.gov/hmd/ihm/) and the Wellcome (http://images.wellcome.ac.uk/) image repositories, and Adam Matthew Digital (http://www.amdigital.co.uk/). If a digital copy of suitable size cannot be located, then some individuals will scan the image from a book themselves. Faculty members recognize that expectations for good visual content in the classroom have risen significantly over the last several years with a concomitant increase in prep time. Many faculty members would welcome improved access to images including larger databases, improved metadata, indexing of images within books, aggregation of existing image resources, and alternatives to YouTube for video in the classroom.

> *…it [indexed images] could be very useful for preparing a class lecture, where I use images I can take from online databases. That would be extremely valuable, actually, because that's what I do—I Google. When I prepare my PowerPoint now, I have to be content with whatever I find. Whereas if there was a database from the university's libraries with good quality pictures, that was searchable, it would be a very interesting tool for preparing class lectures. (Assistant Professor, History of Science and Medicine)*

[4] Managing permissions was a recurring frustration and many researchers were at a loss for how to even begin the process.

Faculty members seem to be using Google Images for reasons similar to those discussed above regarding Google Search. While they may realize that there are better resources available, they do not have time to search multiple disconnected portals.

> *For teaching, I've been impressed by how image driven it is. To write the lecture I'll work with a couple of books, big synthetic books to organize the topic, but when I'm actually preparing the PowerPoint what I need are images… What I've been doing this semester is Google Image search for the most part, which isn't fantastic because it's mostly little teeny web graphics. My sense is that there are lots of really good digital image initiatives out there in the scholarly realm but I've found them to be rather disconnected. (Assistant Professor, History of Science)*

INTERFACE

Although several interface features are discussed above, a few points are worth emphasizing. Researchers are requesting:

- Flexibility in the modes of presenting search results and accessing content
  - Both the ability to search across formats and collections, and to search within a collection, format, or other feature
  - The ability to explore content in non-linear ways
- Full-text search of in copyright material even if the digital copy is not available
- Additional search tools that improve simple term searches (e.g. Google Ngrams, or the Amazon affinity search)
- Better aggregation of content and features for vetting sources

ADVANTAGES AND LIMITATIONS OF DIGITIZED CONTENT

Researchers generally agreed that digitization benefits both research and teaching. The efficiencies created by digitization are numerous and scholars acknowledged several of these:

- *Viewing content*
  - Images can be manipulated to improve detail and one can zoom in and see details of an image or handwriting more clearly than with the eyes alone.

> *This is one of the great things digital tech has to offer: you can see things better, and closer, and more sharply than when you're looking at the manuscript in person. This is quite aside from the conservation issues involved in putting your face up close to this old parchment! Being able to get into details is great. At the same time being able to see where images fit within the larger book as a whole is important. (Assistant Professor, Medieval Art and Architecture)*

- *Engaging students in the course material*
  - Thematic indexes and exhibits help to engage students more deeply (e.g. Historical Anatomies on the Web, http://www.nlm.nih.gov/exhibition/historicalanatomies/home.html)
  - Using digitized content in the classroom helps prepare students for the library and encourages use of non-digitized collections

> *We could also think of ways in which having these treasures, rare books, available online would get students interested in this kind of material. That would be great actually because they don't even imagine what's in here [the library]. You have to introduce them. (Assistant Professor, History of Science and Medicine)*

- *Searching:* Fact checking is much easier and one can easily and accurately source quotes.
- *Enables new research that was impossible with paper only*
- *Promotes conservation of the physical object*
- *Reduces travel time, which lowers the cost of research*

Despite the many advantages offered by digitized content, several scholars were careful to point out the limitations of digitized content and emphasize that it is an adjunct to the physical object rather than a replacement. Limitations cited by researchers include:

- *Poor digitization quality* (mostly referring to early digitization efforts and scanning performed by Google)
  - Plates are often not in color.
  - Foldout maps and illustrations are not always scanned.
  - Advertisements are often missing and when present they are hard to identify in the metadata.
- *Lengthy items are unwieldy in digital format*
  - It is hard to skim and find information quickly, but this could be alleviated through interface improvements.
- *Losing the material culture of the book*
  - Cannot identify watermarks, chain marks, or other artifacts of construction
  - Multiple versions of a text are rarely digitized and therefore cannot be compared
  - Provenience of the object (e.g. who owned which books)
  - Cannot use digitized content to introduce students to the material culture of the book

> *Because to me it doesn't make too much of a difference if the book is physically available or if I can read it online. For teaching it's different, though. When the book is there, the students realize that size matters, that papers are different, the covers, the binding – the material culture of the book all of a sudden acquires incredible relevance to them. Also they realize that there can be old books that are so valuable even more expensive than their father's car, or even an apartment! This is big surprise and is one of the catchy tricks that we use to get them interested in anything [historical]. (Assistant Professor, History of Science and Medicine)*

- *Teaching reading and interpretation skills*
  - Digitized content changes the kind of reading one does and digital tools can encourage students to use materials out of context. A situation that could be improved with more sophisticated discovery and interface enhancements.
  - Digital only research can lead to facile interpretations since the majority of content has not been digitized.

AREAS FOR PROJECT GROWTH

Numerous topics emerged during our discussions on the site visits and I have attempted to summarize three areas: collection development, metadata management and technology, and collaboration.

Collection Development

Although the MHL already consists of more than 36,000 items, we have barely scratched the surface in terms of the volume of content held by history of medicine libraries and archives alone. In 2009 the MHL partners estimated that there were approximately 300,000 historically important books that should be prioritized for digitization, a figure that did not include journals, archives, images, film (moving images), and ephemera. Our site visits have also confirmed what we already suspected regarding the physical disaggregation of history of medicine materials—relevant, and often important, collections reside in main university libraries, public libraries, museums, and small subject libraries (e.g. the Burke Theological Library at Columbia, http://library.columbia.edu/indiv/burke.html). Thus, there seems to be no end in sight regarding digitization projects for the MHL. Our greatest challenges will be to set coherent priorities for content selection and to maintain a stream of funding that will help us accomplish this work.

Some topical areas of interest that emerged during our site visits include:

- Hospital reports before 1850; or the earliest hospitals, such as Pennsylvania General Hospital
- Botanic and alternative medicine, including journals (Good collections can be found at The College of Physicians of Philadelphia, Lloyd Library and Museum http://www.lloydlibrary.org/, and University of Michigan)
- Public health nursing, or nurses in the community (e.g. The Bates Nursing Archive at University of Pennsylvania)
- Origins of women in medicine, focusing on the earliest period
- Origins of African-Americans in medicine, focusing on the earliest period
- All editions of the Osler textbooks
- Health policy texts: hospital management, health care delivery, etc.
- Serials that include advertising
- 19th century and prior pamphlets (the most important items are cataloged, but many have poor bibliographic control)
- Clinical photos or illustrations in a discrete area (e.g. a specialty, a period of time, or a key artist)
- Origin of film in medicine
- Building an international collection of public health reports
- Pharmaceuticals

Additional suggestions can be sent to medicalheritage@gmail.com.

Metadata and Technology

Thus far the MHL has discovered only minor issues regarding metadata consistency. However, as the MHL continues to aggregate existing digital content and the pool of content becomes more complex, these issues are likely to multiply. Enlisting a part-time data analyst who could identify issues as content is added and who can manage metadata changes with Internet Archive would benefit the project. In addition to the fundamentals of maintaining consistent metadata, the MHL could explore more experimental methods of metadata enrichment that would address some of the concerns expressed by researchers above. For example:

- Indexing images, advertisements, and data tables within books and journals
    - Explore the potential for applying document layout analysis tools such as OCRopus (https://code.google.com/p/ocropus/), or OCRFeeder (http://live.gnome.org/OCRFeeder) to identify and index images within the texts. Several proprietary software packages are also available.
- Connect historical materials with health datasets[5]. Develop a meta-thesaurus of archaic medical terminology that can be used to associate historical content with contemporary medical concepts. This could be used to improve search performance over heterogeneous content.

Collaboration

The MHL has made a significant step toward long-term growth with the adoption of our tiered participation model (http://www.medicalheritage.org/2011/08/new-contributors-sought/) and the addition of several new Content Contributors from public and private institutions within the US and abroad. We hope that contributing content to the MHL through Internet Archive is only the beginning of these relationships and that these new contributors will be inspired to take an active role in project development. In addition to developing new digitization projects, we will continue our Internet Archive tagging project as a means to aggregate existing content (i.e. project members search Internet Archive for high quality content that is within scope and then work with the institution to add it to the MHL collection).

The MHL will also continue to prioritize its program of outreach and engagement with history of medicine stakeholders (e.g. librarians, archivists, researchers, curators, and the public). Thus far, input from the community has been invaluable in helping us identify priorities for project development and we plan to continue these discussions. In addition to making history of medicine content freely accessible, we hope to facilitate discourse around these materials that engages not only the scholarly community, but also students, the public, and policy-makers.

CONCLUSION

> *Most people in policy, or science, or business want to change the world in which they operate. They want to impact their immediate environment and they have a sense that their immediate environment is naturally occurring. That it [their environment] is a sort of pre-discursive fact of life. If you can't explain that the environment you're trying to change is the product of historical processes, then you won't be able to avoid repeating history. If you can help people understand that the world is a product of historical processes—we don't have health care because of the following reasons, or people haven't bought toasters*

[5] A senior administrator at Johns Hopkins Medical School emphasized the importance of understanding and managing historical medical terminology for contemporary medical education. For example, the sequencing of the human genome has radically changed the way disease is conceptualized and classified prompting major revisions of the hierarchically organized International Classification of Diseases (ICD). Disease classifications will need to be translated to a matrix or an open network and the successful translation of historical terminology is of critical importance.

> *because of these following reasons—now you know why you're here. You can change it. (Research Scientist, Bioethics and Health Policy)*

With the application of informatics approaches to mine and reanalyze scientific findings, history of medicine content is more broadly relevant to contemporary discussions of health and society than ever. Meta-analysis of research findings can help us identify unstated assumptions, theories, or disciplinary paradigms that influence scientific conclusions and the adoption of procedures into the mainstream[6]. For medicine this could mean identifying areas where premature certainty has halted promising investigation, or identifying overly narrow perspectives on the literature that have resulted in citation entrenchment (i.e. the more often a work is cited the more likely it is to be cited again generating a feedback loop) that inadvertently excluded valuable lines of inquiry.

Much of this report has focused on the day to day activities of researchers and their use of digital content, but I would like to step back for moment and address the importance of building a community around the digital content. This is not only because 'community' is a nice word. Rather, it is because we are moving into new territory where the consequences of the digital transition for our intellectual heritage are unknown. This report only alludes to the larger issues confronting libraries and higher education. We are faced with many questions regarding the way in which the digital transition will affect the scholarly process and public access to our intellectual heritage. Can we resolve the legal quagmire of copyright before we are left with a nearly century long digital gap in the scholarly record? In the meantime, how do we create digital resources that are sufficiently comprehensive, yet are coherent enough to support meaningful research and education?

This report outlines several opportunities for growth of the MHL project and challenges us to think of the ways we might further develop our technical and organizational infrastructure to encourage broad participation of stakeholders. Situating the MHL within a community of users provides the platform on which to build collaborations among curators, researchers, and exhibition staff that would support multiple audiences for history of medicine content (e.g. casual user, high school students, scholars, and policy makers).

[6] Evans, James A, and Jacob G Foster (2011). Metaknowledge. *Science* 331(6018): 721-5.