The *Profiles in Science* Digital Library: Behind the Scenes

Marie E. Gallagher
U.S. National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894 USA
+1 301 435 3252
mgallagher@mail.nih.gov

Christie Moffatt
U.S. National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894 USA
+1 301 496 9136
moffattc@mail.nih.gov

ABSTRACT
This demonstration shows the *Profiles in Science*® digital library. *Profiles in Science* contains digitized selections from the personal manuscript collections of prominent biomedical researchers, medical practitioners, and those fostering science and health. The *Profiles in Science* Web site[1] is the delivery mechanism for content derived from the digital library system. The system is designed according to our basic principles for digital library development [1]. The digital library includes the rules and software used for digitizing items, creating and editing database records and performing quality control as well as serving the digital content to the public. Among the types of data managed by the digital library are detailed item-level, collection-level and cross-collection metadata, digitized photographs, papers, audio clips, movies, born-digital electronic files, optical character recognized (OCR) text, and annotations (see Figure 1). The digital library also tracks the status of each item, including digitization quality, sensitivity of content, and copyright. Only items satisfying all required criteria are released to the public through the World Wide Web. External factors have influenced all aspects of the digital library's infrastructure.

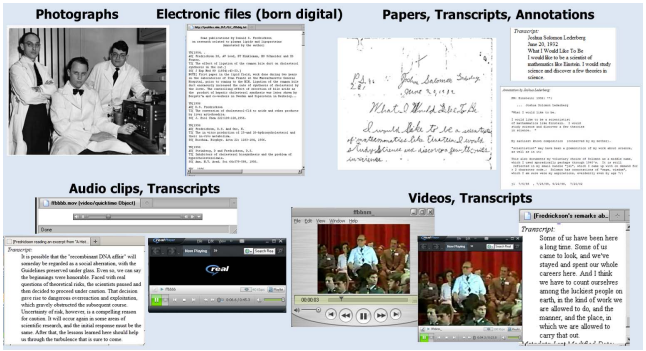

Figure 1. Samples of items in *Profiles in Science*

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *Collection, Dissemination, Systems issues*

General Terms
Design, Legal Aspects

Keywords
Digital Library, Information System, Digitized Manuscripts, Adaptability

1. INTRODUCTION
The U.S. National Library of Medicine (NLM) is the world's largest biomedical library. In addition to published books and journals, NLM collects unpublished items belonging to groups and individuals who have made significant contributions and discoveries. NLM's History of Medicine Division (HMD) is the repository for these special collections. The Web provides an ideal route for making NLM's unique and extraordinary collections available worldwide to researchers, educators and students. Making available digital reproductions of historical items also decreases some of the need for handling the fragile originals. In 1997, the HMD and NLM's Lister Hill National Center for Biomedical Communications (LHNCBC) partnered to build a prototype digital collection that demonstrated the promise of delivering historical manuscript collections over the Web. Having anytime/anywhere access to high quality digital copies of their selected manuscripts, laboratory notebooks, drawings, diaries, photographs, audiovisuals and other materials encouraged the Nobel Prize-winning National Institutes of Health (NIH) scientists to donate their papers to NLM for digitization.

The prototype digital collection grew into a digital library system designed to manage digitized items and their associated metadata for the long-term. The digital library system became known as *Profiles in Science* and is delivered to the public through its Web site. NLM continues to organize, describe, digitize and add collections to *Profiles in Science*. The additions expanded to include other manuscript collections donated to NLM as well as collections held by other institutions through collaboration. Some of the items in *Profiles in Science* are: Joshua Lederberg's earliest known autograph, a photograph of Julius Axelrod at age 3, Francis Crick's drawing of the double helix, Albert Szent-Gyorgyi discussing cancer research, and Marshall Nirenberg's laboratory notebooks showing his work deciphering the genetic code. Currently thirty-six collections are available to the public. These include 142,180 image pages that constitute 27,042 digital items. Users can browse by navigating through the Web site, or search the metadata and OCR text to find items of interest. The project is staffed with computer scientists, archivists, a historian, and a library technician.

[1] http://profiles.nlm.nih.gov/

JCDL'12, June 10–14, 2012, Washington, DC, USA.
ACM 978-1-4503-1154-0/12/06.

2. BEHIND THE SCENES

2.1 Data Creation

At the heart of the *Profiles in Science* digital library system is the metadata database [2]. The database contains descriptive, administrative and technical metadata about each item in the digital library. Software detects technical information about each file and updates the database automatically. Trained digital archivists enter most descriptive information and some administrative information using the *Profiles in Science* data entry program. The program guides the archivist by providing options and pull-down lists of standard vocabularies rather than free text input wherever possible, alerting the user when required information is missing, tracking workflow such as quality control status and copyright research status, and providing customized access so each user can edit only what s/he is authorized to edit. The program automatically generates unique identifiers that forever bind the metadata to the digital item. Item-level metadata includes most of the fifteen Dublin Core metadata elements[2] as well as extensive metadata regarding sensitivity, copyright permission status, quality control, physical condition, and final disposition. Collection-level metadata describes the subject and overall contents of the collection, exhibit sections, a brief chronology, and the collection's physical (boxes and folders) and logical (series and folders) structure. Some metadata spans all collections such as lists of organization names, personal names, journal names, languages, Medical Subject Headings (MeSH)[3] and document types[4].

2.2 Quality Control

Profiles in Science has a quality control and reporting tool known as the Diagnostic Server. The Diagnostic Server provides a read-only view of the metadata from various perspectives. The archivist uses the tool to compare and spot inconsistencies within the detailed information associated with an item, within collections, and across collections. For example, within any collection or across all collections, the archivist can see which metadata records and which digitized items are pending review. The archivist may also see which items have a particular copyright research status. S/he may choose to include all items or filter for items that are marked for public release. The archivist may also use the Diagnostic Server to review items to ensure that their descriptions are standard, consistent, and meet internally developed guidelines.

2.3 Data Dissemination

When we add items to the *Profiles in Science* digital library, we intend for them to become publicly available. Copyright law places U.S. Government works, including the notebooks and correspondence authored by the NIH Nobel laureates, in the public domain per U.S. Code Title 17 Section 105. When manuscript collections are donated to the NLM, the copyrights of the donor are usually transferred to the public domain through the Deed of Gift. Copyrights held by others, such as authors of correspondence sent to the NIH Nobel laureates, remain unchanged. Permissions to release these items to the public must be sought. Occasionally, although infrequently, the request is denied. In a few cases the copyright status of an item has changed when a previously unidentified copyright holder came forward. Over time we discovered that some items could not be released until sometime in the future. Such works include personnel recommendations, voting records, grant applications, research proposals, inquiries from patients, and items whose copyright status will change. Among the most complex changes to our digital library system are the additional rules that prevent accidental release of protected items and implement the permissions status tracking workflow.

2.4 External Factors

In response to external factors, the *Profiles in Science* digital library system's software components undergo constant modifications. One external factor we have experienced is software obsolescence. For example, our original search engine was a commercial product whose support ended. It was replaced by in-house software and will eventually be replaced by open source software. Another factor is security policies that dictate or prohibit the use of certain software, hardware, and configurations. Because security policies necessarily evolve, we must change or replace our software and configurations to comply; sometimes this happens on short notice due to imminent unforeseen security threats. We expect that policies and regulations requiring more space-saving and energy efficient hardware will become increasingly important for future hardware choices. Another external factor is the rapidly increasing number and variety of mobile devices; rather than providing custom interfaces for different devices, we wish to develop a universally-usable, standards-compliant, device-independent interface. We must also ensure that we comply with standards for implementing Section 508 of the Rehabilitation Act which are also undergoing change[5].

3. ACKNOWLEDGMENTS

We thank our dedicated and creative colleagues at the National Library of Medicine for their work on this project, as well as the collaborators who made it possible to include their collections in *Profiles in Science*. This work is supported by the Intramural Research Program of the NIH, National Library of Medicine. The Library, the world's largest library of the health sciences, is a component of the National Institutes of Health, U.S. Department of Health and Human Services.

4. REFERENCES

[1] McCray, A. T. and Gallagher, M. E. 2001. Principles for digital library development. *Commun. ACM* 44, 5 (May 2001), 48-54. DOI= http://doi.acm.org/10.1145/374308.374339.

[2] McCray, A. T., Gallagher, M. E. and Flannick, M. A. 1999. Extending the Role of Metadata in a Digital Library System. In *Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries* (Baltimore, MD, March 19 - 21, 1999). ADL '99. IEEE Computer Society, Washington, DC, USA, 190-199. DOI= http://doi.ieeecomputersociety.org/10.1109/ADL.1999.777714.

[2] http://dublincore.org/documents/dces/

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

[4] http://www.getty.edu/research/tools/vocabularies/aat/

[5] http://www.access-board.gov/508.htm