CASE STUDY

DOI: dx.doi.org/10.5195/jmla.2017.47

Our journey to digital curation of the Jeghers Medical Index

Lori Gawdyda, MLS; Kimbroe Carter, MD, MT; Mark Willson; Denise Bedford, PhD

See end of article for authors' affiliations.

Background: Harold Jeghers, a well-known medical educator of the twentieth century, maintained a print collection of about one million medical articles from the late 1800s to the 1990s. This case study discusses how a print collection of these articles was transformed to a digital database.

Case Presentation: Staff in the Jeghers Medical Index, St. Elizabeth Youngstown Hospital, converted paper articles to Adobe portable document format (PDF)/A-1a files. Optical character recognition was used to obtain searchable text. The data were then incorporated into a specialized database. Lastly, articles were matched to PubMed bibliographic metadata through automation and human review. An online database of the collection was ultimately created. The collection was made part of a discovery search service, and semantic technologies have been explored as a method of creating access points.

Conclusions: This case study shows how a small medical library made medical writings of the nineteenth and twentieth centuries available in electronic format for historic or semantic research, highlighting the efficiencies of contemporary information technology.

BACKGROUND

Harold Jeghers was a well-known medical educator of the twentieth century who contributed to defining Peutz-Jeghers syndrome [1, 2]. During his career, he collected approximately one million medical articles dating from the late 1800s to the 1990s. The collection followed him to several academic centers and teaching hospitals, and it eventually came to Youngstown, Ohio, in 1982. In his later years, Jeghers joined the faculty of the Internal Medicine Department at Northeast Ohio Medical University (NEOMED) and medical staff at St. Elizabeth Youngstown Hospital (SEYH), a community teaching hospital affiliated with NEOMED. His collection continued to be a teaching resource at SEYH and became known as the Jeghers Medical Index (JMI). The JMI has been self-supporting through a trust shortly after its purchase by the Department of Education at SEYH. An advisory board of physicians, hospital administrators, and community professionals has directed the activities of the JMI, aligning goals to the trust's directive, and has safeguarded the collection for three decades.

Throughout the 1980s and early 1990s, Jeghers oversaw the continued accumulation of the collection, with articles from about seventy academic clinical journals. Internal medicine and surgery residents as well as medical students rotated through the JMI for one month. They were assigned a clinical problem to research, created an annotated bibliography, and presented their findings at the end of their rotations. These residents and students, assisted by JMI staff, learned the skills of finding medical literature before the widespread use of personal computers. In the early 1990s, the JMI was used less frequently in residence training; however, the medical staff continued to use the collection as a source of background information for lectures, poster presentations, and manuscript preparation.

By the late 1990s, the collection had grown to fill more than 165 fireproof cabinets holding 45,000 folders containing approximately 1 million medical articles consisting of 5 million pages. Initially, Jeghers used an anatomic scheme to categorize articles, but by the 1990s, the filing system had features of the Medical Subject Headings (MeSH) scheme used by the US

DOI: dx.doi.org/10.5195/jmla.2017.47

National Library of Medicine. As the collection grew, Jeghers' indexing limited the efficiency of article retrieval. Additionally, notes written on the pages by patrons and indexers, along with paper deterioration, started to limit legibility of the articles.

STUDY PURPOSE

A solution for dealing with the paper deterioration, legibility, retrieval efficiency, and continued relevance of the JMI in medical teaching required a modern approach. The authors planned to store the text of the articles in a logically organized, searchable digital database system, which would enable standards-based bibliographic metadata searches in both specific fields as well as the text. Herein, we describe the two-decade journey of converting a large paper collection to a specialized digital database.

CASE PRESENTATION

Electronic archiving of the collection involved the Preservation Project, followed by the eJeghers Content Project. The Preservation Project included about 120,000 articles that predated 1965, which were chosen due to their deterioration. The collection contained copyrighted work subject to allowances permitting reproductions for preservation purposes. Although JMI staff consisted of only two full-time workers—a librarian and an assistant—digitization of the paper collection began in the late 1990s with little forethought that these efforts could last for decades.

Articles were scanned to tagged image file format (TIFF). A Microsoft SQL database was created, and bibliographic data were manually entered over two years with the assistance of five part-time students from the regional university. Concern about preserving the remaining print collection mounted as experience was gained in working with the pre-1965 articles. The brittleness of the paper sometimes slowed progress. After completing the Preservation Project, the advisory board decided to digitize the remaining collection due to paper deterioration and safety inspection concerns related to an emergency exit that was impeded by the filing cabinets. This part of the collection also contained copyrighted work, but the deteriorating condition of the material was considered to be a copyright exception that permitted reproduction for preservation. A detailed project management document was created to guide tasks, timelines, and milestones.

Subsequently, the eJeghers Content Project included digitizing the remaining 880,000 print articles, merging them with the digitized data of the Preservation Project, and integrating the collection into a flexible data management system specialized for storage and searching. Since inception of the first project, noticeable improvements evolved in scanning, optical character recognition (OCR) processing, and archival document formats. Experience gained in the first project showed the importance of using OCR software that met the project specifications. OCR software ABBYY Recognition Server, ABBYY Finereader Corporate, Cvision Maestro, Expervision Typereader, and AnyDoc were compared for speed, accuracy, and pricing. The price of purchasing software and hardware was compared to that of using a third-party vendor. ABBYY Recognition Server best met the project's criteria. Similarly, data management vendors Google, GoPubmed, Exalead, Thunderstone, Greenstone, and dSpace were compared for cost, relevance, user interface, methodology, location of corporate headquarters, and willingness to customize software. Thunderstone and its TEXIS product most suitably met the project's data management needs.

In developing the eJeghers Content Project, PubMed XML citations and a table of anatomic terms that Jeghers created to use in retrieving articles from the original collection were added to a specialized relational database, TEXIS. TEXIS is an SQL relational database with powerful text searching capabilities with its own scripting language, Vortex [3, 4]. PubMed records were loaded into TEXIS, while maintaining PubMed's field structure.

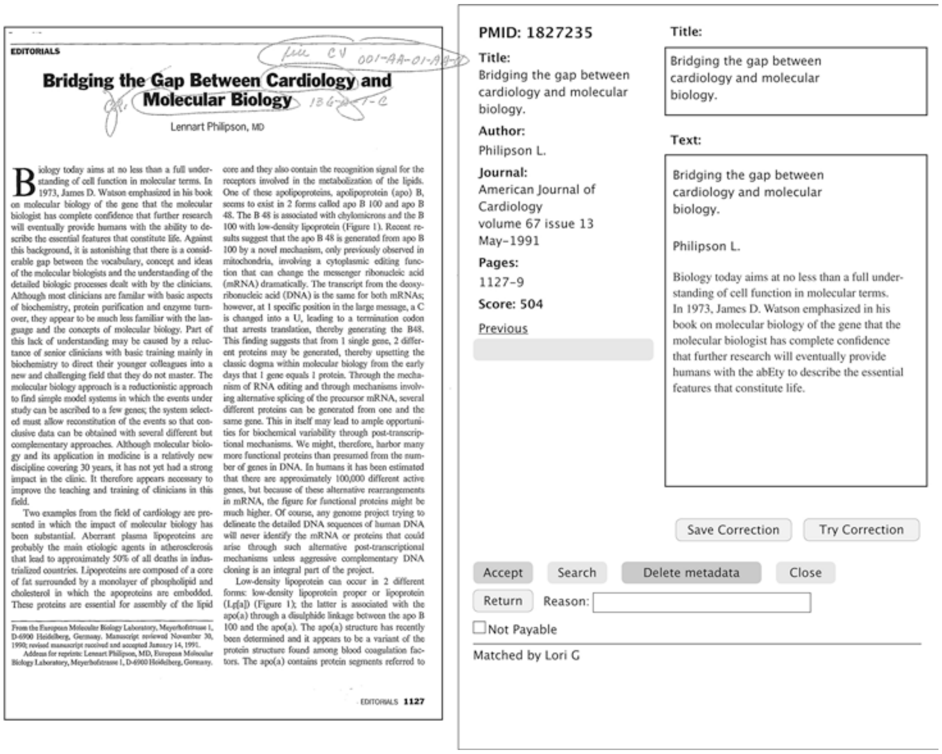
Several computational techniques were used in extracting the needed bibliographic data from each article. These included rules using the position of alphanumeric characters and publication symbols. For example, a number following "Volume" was interpreted as the publication volume number and a number following "#" was interpreted as an issue number. Page numbers were considered more likely when at the top or bottom of a page than in the body of the text. Another important rule that was applied was clustering author names on the first page of an article.

DOI: dx.doi.org/10.5195/jmla.2017.47

At the time of the project, PubMed citations did not exist for many articles published before 1960. An algorithm assigned a score that predicted if a PubMed citation matched a JMI article. Some articles required human review, based on the value of the match score. A user-friendly interface was developed to facilitate this review. Features of the interface are shown in Figure 1. The staff compared an article on the left-hand side of the screen with the algorithm-assigned PubMed record on the right-hand side. The reviewer had the option of accepting the record, searching for the correct PubMed record, or adding bibliographic data. Six people spent more than 2 years reviewing approximately 300,000 articles. At the project's end, more than 900,000 articles had bibliographic metadata.

The completed TEXIS system manages about one million articles, comprising five million pages and containing over two billion words. The database, which is hosted online, offers three search options to assist in article retrieval. One search approach uses Boolean logic, with search and retrieval behaviors similar to PubMed. Users can query author, title, journal, abstract, and MeSH. A second approach—proximity search—uses the occurrences of terms and location of words in sentences, paragraphs, pages, or documents to retrieve articles. This approach takes advantage of the relationships between words and the clustering of words with concepts. Proximity search can extract concepts contained in the body of the articles. A third approach uses the indexing method that Jeghers developed that relies on anatomic and disease concepts and retrieves folders containing the article. Jeghers' method is influenced by disease, organ system, physiology, and anatomy [5]. The public view of the database displays results initially by relevance. The display of results from a query is shown in Figure 2, with several options available for organizing the bibliographic data. Users can choose to list results by year of publication, with or without abstracts, or by study type. A fair use copy may be requested from the JMI librarian in compliance with copyright law.

Figure 1 Reviewer interface for determining PubMed match accuracy



EDITORIALS

CV 001-AA-01-AA

Bridging the Gap Between Cardiology and Molecular Biology

136-A-1-C

Lennart Philipson, MD

Biology today aims at no less than a full understanding of cell function in molecular terms. In 1973, James D. Watson emphasized in his book on molecular biology of the gene that the molecular biologist has complete confidence that further research will eventually provide humans with the ability to describe the essential features that constitute life. Against this background, it is astonishing that there is a considerable gap between the vocabulary, concept and ideas of the molecular biologists and the understanding of the detailed biologic processes dealt with by the clinicians. Although most clinicians are familar with basic aspects of biochemistry, protein purification and enzyme turnover, they appear to be much less familiar with the language and the concepts of molecular biology. Part of this lack of understanding may be caused by a reluctance of senior clinicians with basic training mainly in biochemistry to direct their younger colleagues into a new and challenging field that they do not master. The molecular biology approach is a reductionistic approach to find simple model systems in which the events under study can be ascribed to a few genes; the system selected must allow reconstitution of the events so that conclusive data can be obtained with several different but complementary approaches. Although molecular biology and its application in medicine is a relatively new discipline covering 30 years, it has not yet had a strong impact in the clinic. It therefore appears necessary to improve the teaching and training of clinicians in this field.

Two examples from the field of cardiology are presented in which the impact of molecular biology has been substantial. Aberrant plasma lipoproteins are probably the main etiologic agents in atherosclerosis that lead to approximately 50% of all deaths in industrialized countries. Lipoproteins are composed of a core of fat surrounded by a monolayer of phospholipid and cholesterol in which the apoproteins are embedded. These proteins are essential for assembly of the lipid core and they also contain the recognition signal for the receptors involved in the metabolization of the lipids. One of these apolipoproteins, apolipoprotein (apo) B, seems to exist in 2 forms called apo B 100 and apo B 48. The B 48 is associated with chylomicrons and the B 100 with low-density lipoprotein (Figure 1). Recent results suggest that the apo B 48 is generated from apo B 100 by a novel mechanism, only previously observed in mitochondria, involving a cytoplasmic editing function that can change the messenger ribonucleic acid (mRNA) dramatically. The transcript from the deoxyribonucleic acid (DNA) is the same for both mRNAs; however, at 1 specific position in the large message, a C is changed into a U, leading to a termination codon that arrests translation, thereby generating the B48. This finding suggests that from 1 single gene, 2 different proteins may be generated, thereby upsetting the classic dogma within molecular biology from the early days that 1 gene equals 1 protein. Through the mechanism of RNA editing and through mechanisms involving alternative splicing of the precursor mRNA, several different proteins can be generated from one and the same gene. This in itself may lead to ample opportunities for biochemical variability through post-transcriptional mechanisms. We might, therefore, harbor many more functional proteins than presumed from the number of genes in DNA. In humans it has been estimated that there are approximately 100,000 different active genes, but because of these alternative rearrangements in mRNA, the figure for functional proteins might be much higher. Of course, any genome project trying to delineate the detailed DNA sequences of human DNA will never identify the mRNA or proteins that could arise through such alternative post-transcriptional mechanisms unless aggressive complementary DNA cloning is an integral part of the project.

Low-density lipoprotein can occur in 2 different forms: low-density lipoprotein proper or lipoprotein (Lp[a]) (Figure 1); the latter is associated with the apo(a) through a disulphide linkage between the apo B 100 and the apo(a). The apo(a) structure has recently been determined and it appears to be a variant of the protein structure found among blood coagulation factors. The apo(a) contains protein segments referred to

From the European Molecular Biology Laboratory, Meyerhofstrasse 1, D-6900 Heidelberg, Germany. Manuscript reviewed November 30, 1990; revised manuscript received and accepted January 14, 1991.

Address for reprints: Lennart Philipson, MD, European Molecular Biology Laboratory, Meyerhofstrasse 1, D-6900 Heidelberg, Germany.

EDITORIALS 1127

PMID: 1827235

Title:
Bridging the gap between cardiology and molecular biology.

Author:
Philipson L.

Journal:
American Journal of Cardiology
volume 67 issue 13
May-1991

Pages:
1127-9

Score: 504

Previous

Title:
Bridging the gap between cardiology and molecular biology.

Text:
Bridging the gap between cardiology and molecular biology.

Philipson L.

Biology today aims at no less than a full understanding of cell function in molecular terms. In 1973, James D. Watson emphasized in his book on molecular biology of the gene that the molecular biologist has complete confidence that further research will eventually provide humans with the abEty to describe the essential features that constitute life.

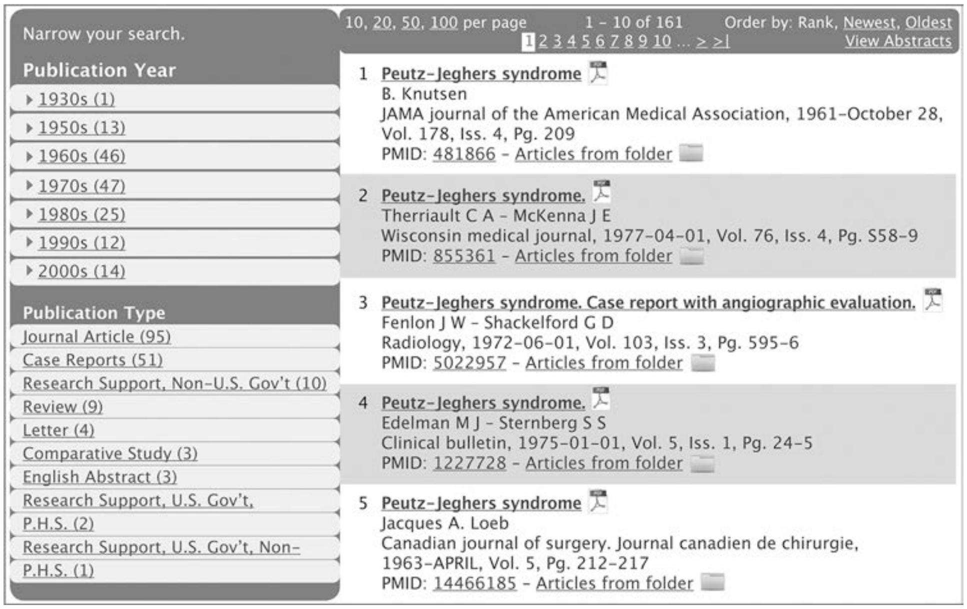
Reason:

□ Not Payable

Matched by Lori G

DOI: dx.doi.org/10.5195/jmla.2017.47

Figure 2 Display of query results



Narrow your search.

Publication Year

- 1930s (1)
- 1950s (13)
- 1960s (46)
- 1970s (47)
- 1980s (25)
- 1990s (12)
- 2000s (14)

Publication Type

Journal Article (95)
Case Reports (51)
Research Support, Non-U.S. Gov't (10)
Review (9)
Letter (4)
Comparative Study (3)
English Abstract (3)
Research Support, U.S. Gov't, P.H.S. (2)
Research Support, U.S. Gov't, Non-P.H.S. (1)

10, 20, 50, 100 per page 1 - 10 of 161 Order by: Rank, Newest, Oldest
1 2 3 4 5 6 7 8 9 10 ... > >| View Abstracts

1 Peutz-Jeghers syndrome
B. Knutsen
JAMA journal of the American Medical Association, 1961-October 28, Vol. 178, Iss. 4, Pg. 209
PMID: 481866 - Articles from folder

2 Peutz-Jeghers syndrome.
Therriault C A - McKenna J E
Wisconsin medical journal, 1977-04-01, Vol. 76, Iss. 4, Pg. S58-9
PMID: 855361 - Articles from folder

3 Peutz-Jeghers syndrome. Case report with angiographic evaluation.
Fenlon J W - Shackelford G D
Radiology, 1972-06-01, Vol. 103, Iss. 3, Pg. 595-6
PMID: 5022957 - Articles from folder

4 Peutz-Jeghers syndrome.
Edelman M J - Sternberg S S
Clinical bulletin, 1975-01-01, Vol. 5, Iss. 1, Pg. 24-5
PMID: 1227728 - Articles from folder

5 Peutz-Jeghers syndrome
Jacques A. Loeb
Canadian journal of surgery. Journal canadien de chirurgie, 1963-APRIL, Vol. 5, Pg. 212-217
PMID: 14466185 - Articles from folder

DISCUSSION

History plays an important role in anchoring the perspectives of researchers and practitioners alike in approaching medicine. In the context of medical advancements, knowledge of past questions and their bygone solutions are important in understanding the medical choices of today. Medicine and the concepts of disease are not fixed, as they change in response to social and environmental conditions and are understood at any time in strikingly different ways. The JMI now is a window into the past.

Although it is no longer directly involved in a teaching program, the JMI continues to be a source of articles and information for scholarly purposes, providing background to current questions. In a review of recent publications by SEYH staff, about 16% of publications referenced medical literature that was more than 20 years old. Many of these references came from the JMI collection. Use of the JMI is not restricted to SEYH staff. About 5% of interlibrary loan requests to the SEYH Medical Library can be found in the collection. Increased usage of the collection has been the result of inclusion in a discovery search service of the SEYH Medical Library. Over the past year, more than 2,000 hits or searches have been recorded.

Techniques to improve recall and accuracy of queries in the collection are now under consideration. Shifts in health care terms, meanings, and concepts over the years sometimes limit keyword queries that use current natural language expressions. Retrieval accuracy could be improved using semantic methods such as entity extraction and classification to create additional metadata from text in the collection. The metadata must be tagged to make it available for search. With such a large collection, meeting these requirements manually, with indexers and historians, is impractical [6].

Therefore, a pilot study was undertaken to investigate the automated gathering of supplemental metadata through entity extraction and classification using Reuter's Open Calais [7] and SAS Content Categorization [8]. These methods rely on authority lists and rule sets used at the World Bank [9]. Terms and concepts of disease, body parts, symptoms, and treatments are drawn from Jeghers' indexing scheme, parts of the MeSH tree, and the International Classification of Diseases, Tenth Edition. The results of the pilot study for this collection demonstrated that automation can create useful semantic metadata, but refinements are needed.

DOI: dx.doi.org/10.5195/jmla.2017.47

The hundreds of person-hours of work in creating an online data source of searchable bibliographic metadata from this paper collection have resulted in a low-maintenance, compact, and searchable collection of select medical literature. Today, the collection is curated by one full-time librarian and has worldwide viewings. The JMI database permits viewing of past medical terminologies, bygone concepts, faded ontologies, and narratives of watershed events as expressed in their original nineteenth and twentieth century writings.

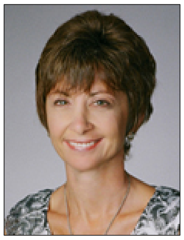
ACKNOWLEDGMENTS

We acknowledge the enduring support, wisdom, and encouragement of past and current Jeghers Advisory Board members, J.P. Morgan trustees, the Ralph H. Ewe Trust, and the Mercy Health Foundation Mahoning Valley for funding support; Terry Lisko for years of project work; Frank Castro of the Medical Decision Making Society of Youngstown, Ohio, for technical consultation; and Sammy Davidson and Robert Kindry, graduate students in the Information Architecture and Knowledge Management master's program at Kent State University's School of Library & Information Science, for contributing to a proof-of-concept use of Reuter's Open Calais and SAS Content Categorization.

REFERENCES

1. Peutz J. Very remarkable case of familial polyposis of mucous membrane of intestinal tract and nasopharynx accompanied by peculiar pigmentations of skin and mucous membrane. (Dutch). Nederl Maandschr Geneesk. 1921;10:134–46.
2. Jeghers H, McKusick VA. Generalized intestinal polyposis and melanin spots of the oral mucosa, lips and digits: a syndrome of diagnostic significance. N Engl J Med. 1949 Dec 29;241(26):1031–6.
3. Thunderstone Software. TEXIS overview [Internet]. [cited 11 Oct 2016]. <https://www.thunderstone.com/texis/site/whitepaper/texisoverview.html>.
4. Kuhns RJ. A survey of information retrieval vendors. Technical report SMLI TR-96-56 [Internet]. 1996 Oct. <http://dl.acm.org/citation.cfm?id=974960>.
5. Jeghers H. Medical care, education and research philosophy and technics of self-education of the medical student and physician. N Engl J Med. 1964 Dec 17;271:1297–301.
6. Thompson P, Batista-Navarro RP, Kontonatsios G, Carter J, Toon E, McNaught J, Timmermann C, Worboys M, Ananiadou S. Text mining the history of medicine. PLOS One. 2016 Jan 6;11(1):1–33.
7. Butuc M. Semantically enriching content using Open Calais. Editia. 2009:77–80.
8. SAS. SAS content categorization [Internet]. [cited 11 Oct 2016]. <http://support.sas.com/software/products/ccs/>.
9. Bedford DAD, Gracy KF. Leveraging semantic analysis technologies to increase effectiveness and efficiency of access to information. Qualitative Quantitative Methods Libr. 2012:1(1):13–26.

AUTHORS' AFFILIATIONS



Lori Gawdyda, MLS, lori_gawdyda@mercy.com, Medical Librarian, Jeghers Medical Index, St. Elizabeth Hospital, Youngstown, OH

Kimbroe Carter, MD, MT, kjcarter@mercy.com, Director, Health Sciences Libraries, Mercy Health, Youngstown, OH; Adjunct Instructor, Computer Technology, Regional College, Kent State University at Trumbull, Warren, OH; and Assistant Professor of Clinical Pathology, Northeast Ohio Medical Universities, College of Medicine, Rootstown, OH

Mark Willson, support@thunderstone.com, Programmer, Thunderstone Software, Cleveland, OH

Denise Bedford, PhD, db233@georgetown.edu, Adjunct Professor, Communication, Culture and Technology, Georgetown University, Washington, DC

Received October 2016; accepted January 2017

ISSN 1558-9439 (Online)