Evolution of biomedical communication as reflected by the National Library of Medicine*

Susan Y. Crawford, PhD, AHIP, FMLA

See end of article for author's affiliation.

DOI: http://dx.doi.org/10.3163/1536-5050.104.1.011

Keywords: *National Library of Medicine, National Center for Biotechnology Information, PubMed, Lister Hill National Center for Biomedical Communications, Human Genome Project*

OBJECTIVE

This commentary examines the evolution of the biomedical communications system in the Western world. The examination touches on many aspects, including the application of new technology, the interoperative relationship between publications and data, changes in the information infrastructure, the convergence of specialties, and consequences for research and health care [1–3].

METHODS

As an overview of communication in the biomedical sciences, this commentary draws upon studies of how science is practiced and how information is produced. Thomas Kuhn introduced the notion of paradigms, scientific models that provide solutions to problems [4]. The adoption of paradigm changes in methods that control the flow of information in the digital age has taken place in many data-rich disciplines [5]. For this examination, I selected as a focus the biomedical information programs of the National Library of Medicine (NLM). This public-service organization within the US National Institutes of Health (NIH) is representative of Western biomedical information management and has produced widely used communication tools.

To address the hypothesis of paradigm change, data were collected through site visits over a three-month period with NLM staff. Socioeconomic issues were probed for insights into the support of science and the role of public and private sectors

* This study was supported in part by the Eugene Garfield Research Fellowship awarded by the Medical Library Association.

A BRIEF HISTORY

In 1879, the Library of the US Army Surgeon General's Office started publishing a monthly index to medical literature under the title *Index Medicus*. Dr. John Shaw Billings, the first editor, counted approximately 85 medical periodicals published each year, with an average of 20,000 substantive articles [6]. This appears to be the first attempt to develop a system for managing the world output of medical information.

By the middle of the 20th century, it was apparent that the formal communication system of the biomedical sciences was in deep trouble. Then published by the American Medical Association and produced by manual manipulation of over a million cards, the index had to cope with an annual output of more than 1,300 medical periodicals. The system had reached the limits of its capability, and the index was some 3 years behind.

In the late 1950s, NLM began experimenting with a mechanized system for composing a similar index. The fledging technology of IBM punched cards eased the filing, and an Eastman Kodak listomatic camera photographed the entries. The result proved successful, and in 1960, the new *Index Medicus,* second series, replaced its predecessors, *Current List of Medical Literature* and *Quarterly Cumulative Index Medicus*. Thus, began an era of research and development toward an automated information system, a paradigm change in managing the universe of biomedical information [7].

NLM's objective now is "connecting and making the results of research from scientific data to published literature to patient and consumer health information [more] readily available" [8]. To meet this challenge, NLM has five major operational

divisions: Library Operations, the Lister Hill National Center for Biomedical Communications (LHNCBC), the National Center for Biotechnology Information (NCBI), Specialized Information Services, and the Office of Computer and Communications System. The divisional functions interact: for example, building databases relies on library operations and on the application of computer technology. I chose four program areas to illustrate NLM's approach to managing systems for communicating biomedical information:

1. Access to health sciences journal publications
2. Interoperative relationship of data and publications
3. Access to molecular and genetic processes
4. Improvement of access to health care information

1. Access to health sciences journal publications

The patient care algorithm requires identifying a set of signs and symptoms, making a diagnosis, and developing a plan of action. Information is needed for each step of decision making. Information from the biomedical literature is now collected, organized, and shared under the NLM PubMed program, which incorporates MEDLINE, successor to *Index Medicus*.

Some 5,600 medical journals are processed for MEDLINE to extract bibliographic data: title, authors, abstracts, and affiliations. It is difficult to recall when practicing physicians relied on scanning columns of references in printed indexes and on the public mail system to receive journal articles from their medical societies or hospital libraries. Often it took 1–2 weeks to receive an answer. Computerization of the *Index Medicus* in the 1960s, and later advent of the Internet, revolutionized the process of information transfer. Online information retrieval systems started in the 1970s and became ubiquitous by 1990.

2. Interoperative relationship of data and publications

Traditional science is based on formulating hypotheses and developing experiments to test them. With the arrival of new information technology, the process of scientific discovery has expanded. Data can now be captured from different experiments and many sources, national and international. Computer scientist Jim Gray said, "Basically, we get data from a bunch of instruments into a pipeline, which calibrates and 'cleans' the data, filling in gaps as necessary, then re-grid the information and essentially put it into a database, which you would like to 'publish' on the Internet for access" [9]. Automated tools are being developed to support the research cycle from data capture and curation to data analysis and visualization.

As scientific papers are produced in digital format, both data and publications are integral parts of the scientific record. The challenge of linking all relevant biomedical information sources into an interoperating system becomes possible. An example is the collection of more than forty databases that NCBI has created and maintains. A list of databases at NCBI is available at http://www.ncbi.nlm.nih.gov/guide/all/#databases.

3. Access to molecular and genetic processes

Understanding nature's mute but elegant language of living cells is the quest of modern molecular biology. From an alphabet of only four letters, representing the chemical subunits of DNA emerges a syntax of life processes whose most complex expression is humans. The unravelling and use of this "alphabet" to form new "words and phrases" is a central focus of the field of molecular biology. The staggering volume of molecular data and its cryptic and subtle patterns have led to an absolute requirement of computerized tools. The challenge is in finding new approaches [10].

James Watson and Francis Crick's discovery of the DNA structure in 1953 ushered in a new era in the evolution of biology and medicine. This means probing the biology of the cell and how genetic information is communicated. In 1990, the plan for a joint Human Genome Project was started by the Department of Energy and NIH and completed 13 years later. The ultimate goal was to generate a high-quality reference sequence for the entire human genome and to identify all 20,500 genes in human DNA [11]. As massive quantities of data would be generated from this initiative and to cope with the volume and complexity, Dr. Donald A. B. Lindberg, with leading scientists, developed and sought support for creating NCBI. Approved by Congress in1988, NLM was chosen to establish and direct the center, which was charged [12]:

- to create automated systems for sorting and analyzing knowledge about molecular biology, biochemistry, and genetics
- to facilitate the use of such databases and software by the research and medical community

- to coordinate efforts to gather biotechnology information both nationally and internationally
- to perform research into advanced methods of computer-based information processing for analyzing the structure and function of biologically important molecules

NCBI subsequently created multidisciplinary research groups composed of computer scientists, molecular biologists, mathematicians, biochemists, research physicians, and structural biologists to focus on basic and applied research in computational molecular biology.

Basic and applied research: program in molecular biology. Initially, the focus was on creating and maintaining databases, developing software for analyzing data, and conducting research on computational biology. The program has branched into research methods for analyzing the function of macromolecules and providing analysis and computing tools for researchers and for the public.

Building the GenBank database.

- Data submitted to NCBI, such as a genome sequence from an organism, is reviewed. If accepted, the data are curated, that is, identified, cross-indexed, and codified to transform disparate sets of research into a cohesive standardized database
- Analysis and annotation add value to the data, find relationships to other sequences, cut across species, synthesize into the larger context, and create hypotheses for further research
- The data are accessible through Entrez, which links NCBI databases to searching algorithms. The challenge is to analyze and connect data from the research community with published records, add value to the data, and link all sources of information into an integrated service.
- The Basic Linear Alignment Search Tool (BLAST), a data-analytic software tool for searching for sequence similarity and for identifying genes and genetic features, can execute searches across the entire DNA database in less than fifteen seconds.

4. Improvement of access to health care information

While programs such as Entrez and BLAST make information available, it is not always assured that the information is readily usable by the lay and science public. Recognizing that access to research information is important for public health, Congress created LHNCBC in 1968 to develop and obtain quality biomedical information, to improve its access, and to optimize its dissemination.

Medical language processing. The Unified Medical Language System (UMLS), developed by the center, identifies and brings together more than 3 million health and medical concepts and 11.9 million terms. The system enables integration of all biomedical information services and bioinformatics research from PubMed to genomic data to patient records.

Visual presentation of information. The focus is how to represent, display, and present biomedical information and to build advanced tools for research, training, and clinical assessment. Visualization and immersive display, high-resolution microscopy at nanometer scales, three-dimensional (3D) printing, quality biomedical imagery on the molecular level, and imaging tools for cancer are among the research projects. A notable achievement supported by NLM is the Visible Human Project, developed by scientists like those at the University of Colorado, Denver. The images are complete anatomically detailed 3D representations of the normal male and female bodies. The project, produced by slicing cadavers at millimeter and below sections and digitally photographing the sections, is used worldwide.

Cognitive science. How technology can simulate and improve the processing and understanding of information is the objective of the Cognitive Science Branch of LHNCBC. The complex aspects of human information processing—perception, concept formation, pattern recognition, and language—are approached by multidisciplinary teams.

From research to the public. NLM has launched a consumer health site on health topics, Medlineplus. There are clinical data standards, electronic medical records, and plans for personalized medicine related to genetic factors. The Collaboration Technologies program shares a library of leading-edged software (InsightToolkit) with research organizations and with industry. International partnerships with the United States Agency for International Development (USAID) assist in health problems outside the United States.

TOWARD A NEW PARADIGM

Thomas Kuhn, in *The Structure of Scientific Revolutions*, defines paradigms as universally

recognized scientific achievements that, for a time, provide model problems and solutions for a community of practitioners. As anomalies arise or if a methodology is no longer capable of solving problems of a new era, the model is replaced by what he called a paradigm shift [13].

Change in how science is practiced

The mid-twentieth century (von Neumann) model of how science is practiced is based on hypothesis testing. Experiments are designed by individuals or laboratories, and data are collected to test assumptions that produce theoretical explanations. This research model is now supplemented by technology: data from numerous sources are captured by instruments or generated by simulation, then processed by software and stored in computers to be mined.

The information infrastructure and interoperability

With data-intensive science, a new research infrastructure emerged that scientists at the Massachusetts Institute of Technology (MIT) have called "convergence." Convergence embraces two procedures: integration of contributions from different disciplines and integration of technology to achieve interoperability [14]. Phillip Sharp has emphasized the need for an informatics infrastructure to incorporate new types of data and to navigate across tiers and domains of knowledge [15]. Both procedures were implemented by NCBI, which organized interdisciplinary teams in the 1980s and 1990s to address molecular and genetic research. Entrez is an example of an interoperative system.

Future of the scientific paper: the global digital archive

As scientific papers are produced in digital format, the traditional print-based scientific record is transformed into a medium for computation. The electronic scientific journal, which applies digital storage and delivery technologies to articles that are essentially printed pages, is being replaced by hybrid collections of text, data, and algorithms to operate the data. Tony Hey has predicted that the "cloud" of magnetic polarizations that encode data and documents in the digital library will become the modern equivalent of miles of library holdings [16].

A deluge of data has resulted from invention of advanced technologies like next-generation sequencing machines, sophisticated data collection techniques, the contribution of many specialties, convergence of research from discipline-centric and independent laboratories, and international participation.

CONCLUSION

Over the past fifty years, there have been profound changes in the way that science is practiced and how information is produced, captured, organized, and used. The concept of information transfer has expanded from managing published papers to understanding molecular and genetic communication on the cellular level.

ACKNOWLEDGMENTS

The author acknowledges contributions and support from the NLM division staffs who provided extensive interviews: Dr. Michael Ackerman, Joyce Backus, Dr. Dennis Benson, Dr. Oliver Bodenreider, Gale A. Dutcher, AHIP, Betsy L. Humphreys, FMLA, Sheldon Kotzin, FMLA, Dr. Barbara Rapp, and Dr. George Thoma. Special assistance with the manuscript was given by Dennis Benson, Betsy L. Humphreys, and Barbara Rapp. Dr. Alfred Soffer, former editor of the American Medical Association's *Archives of Internal Medicine,* and Ann C. Weller, emeritus professor, University of Illinois at Chicago, supported with expertise in their fields of specialization.

REFERENCES

1. Merton R. The sociology of science: theoretical and empirical investigating. Chicago, IL: University of Chicago Press; 1973.
2. Price D. Little science, big science. New York, NY: Columbia University Press; 1963.
3. Crane D. Invisible colleges. Chicago, IL: University of Chicago Press; 1973.
4. Kuhn T. The structure of scientific revolutions. Chicago, IL: University of Chicago Press; 1970.
5. Benson D. Personal communication. Bethesda, MD: National Library of Medicine; Nov 2014.
6. Introduction in Index Medicus. 2nd series. Chicago, IL: American Medical Association; 1960. p. 1–3.
7. National Center for Biotechnology Information. Twenty-five years of growth: NCBI data and user services. Washington, DC: US National Library of Medicine; 2014.

8. National Library of Medicine. Charting a course for the 21st century—NLM's long range plan 2006–2016. Washington, DC: National Library of Medicine; 2007.
9. Hey T, Stewart T, Tolle K, eds. Jim Gray on eScience: a transformed scientific method. In: Hey T. The fourth paradigm: data intensive scientific discovery. Redmond, WA: Microsoft Research; 2009. p. xvii–xxxi.
10. National Center for Biotechnology Information. Our mission [Internet]. The Center [cited 29 Sep 2015]. <http://www.ncbi.nlm.nih.gov/About/glance/ourmission.html>.
11. US Department of Energy Human Genome Program. Genomics and its impact on society: a primer. Washington, DC: US Department of Energy; 2001.
12. National Center for Biotechnology Information. A brief history of NCBI's formation and growth. Bethesda, MD. Government Bulletin (no date).
13. Kuhn T. The structure of scientific revolutions. Chicago, IL: Chicago University Press; 1973.
14. Massachusetts Institute of Technology. The power of "convergence" [Internet]. White paper. Cambridge, MA: The Institute [cited 29 Sep 2015]. <http://newsoffice.mit.edu/2011/convergence-0104>.
15. Sharp PA. Meeting global challenges: discovery and innovation through convergence. Science. 2014 Dec 19; 346(6216):1468–71.
16. Bell G. Foreword. In: Hey T. The fourth paradigm: data intensive scientific discovery. Redmond, WA: Microsoft Research; 2009. p. xi–xvi.

AUTHOR'S AFFILIATION

Susan Y. Crawford, PhD, AHIP, FMLA, sjcrawf@aol.com, Emeritus Professor and Director, Library and Biomedical Communications Center, Washington University School of Medicine, 2418 Lincoln Street, Evanston, IL 60201