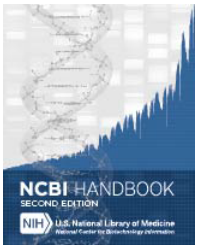
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: Smith K. A Brief History of NCBI's Formation and Growth. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013-.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



A Brief History of NCBI's Formation and Growth

Kent Smith[1]

The establishment of the National Center for Biotechnology Information (NCBI) in November of 1988 occurred primarily through the convergence of three independent but related actions. They were:

- 1984-86—Advocacy groups convened meetings on Capitol Hill to educate legislators and their staffs on the value of supporting genomic research.
- 1986—NLM's Long Range Plan was completed; it contained a recommendation that a new NLM Division be created to manage and process molecular biology information.
- 1987—The House Select Committee on Aging, Chaired by Senator Claude Pepper, introduced a Bill to establish the NCBI.

In 1984, the Delegation for Basic Biomedical Research began briefing sessions on the Hill, using Nobel winners like Dr. James Watson and Dr. David Baltimore to inform legislators about the importance of genomic research as a new and integral part of the advancement of scientific research. These briefing sessions were thought to be critical in creating an atmosphere in Congress that was receptive to the creation of a biotechnology information center like that of the NCBI.

Also in 1984, Dr. Donald A. B. Lindberg became the director of the NLM, and soon thereafter led the National Library in a major long-range planning effort. Over 100 leaders in the biomedical community participated in this rigorous process, forming 5 panels covering the principal domains of NLM. Of particular note was panel 3—titled "Obtaining Factual Information from Data Bases"—which would prove to be the source from which the idea for NCBI was initially conceived.

The germination of the idea for a center emanated in great part from Dr. Allan Maxam, a professor of biological chemistry at Harvard who was a pioneer of molecular genetics and served as a key member of the 1986 NLM Long Range Planning Panel. He instructed his fellow panel 3 members on the importance they should assign to the field of biotechnology and informed them of the country's need to harness the large volume of data that would be generated by the oncoming genetic revolution in science. The Planning Panel, and the NLM Board of Regents, embraced the idea of the need for an organization that could serve as both a repository and distribution center for the growing body of genomic and genetic knowledge and also serve as a unique resource for developing new computer analysis and communication tools for managing molecular biology information.

The newly formed Friends of the NLM saw this as an opportune time to approach the Congress on the need for a biotechnology information center and sought out Senator Claude Pepper, a major champion for medical research. Realizing his congressional colleagues would need to be educated about the benefits of biotechnology research, Senator Pepper asked that NLM develop a document that could be used to explain the need for a

Author Affiliation: 1 NCBI; Email: kents@ncbi.nlm.nih.gov.

Corresponding author.

center. The resulting document, known as "Talking One Genetic Language: The Need for a National Biotechnology Information Center," formed the background for the introduction of the initial bill (H.R.5271) to create the NCBI as part of the NLM. The bill was introduced late in the congressional session, and no action was taken on it, but it was reintroduced in the following session by a determined Senator Pepper.

On March 6, 1987, Senator Pepper, as Chairman of the House Committee on Aging, introduced his new bill (H.R.393) to establish NCBI, stating that the center would deal "with nothing less than the mystery of human life and the unfolding scroll of knowledge, seeking to penetrate that mystery, which is life itself." The hearing had a compelling slate of 15 witnesses, including senior federal and academic health officials as well as five patients who had benefitted from biotechnology.

Although the bill encountered a number of legislative obstacles, Senator Pepper kept the effort alive by securing an appropriation of $3.85 million to begin the biotechnology information program at NLM. During this timeframe, Dr. Daniel Masys, director of the Lister Hill Center for Biomedical Communications, and his branch chief, Dr. Dennis Benson, initiated NLM's early biotechnology information activities.

The following year Senator Pepper, with the help of Congressman Henry Waxman and Senators Edward Kennedy and Lawton Chiles, incorporated H.R.393 into the NIH reauthorization bill known as the Health Omnibus Extension Act P.L.100-607. It passed the Congress and was signed into law by President Reagan on November 4, 1988.

Following enactment, Senator Pepper, in ceremonies conducted in the Capitol's Mike Mansfield Room, said about biotechnology and the new center: "I hope and pray it's going to realize the dreams that many of us have cherished for a long, long time, by being able to prolong the lives and promote the health and happiness of human beings."

The act stipulated the following functions for the new National Center for Biotechnology Information:

1 *design, develop, implement, and manage automated systems for the collection, storage, retrieval, analysis, and dissemination of knowledge concerning human molecular biology, biochemistry, and genetics;*

2 *perform research into advanced methods of computer-based information processing capable of representing and analyzing the vast number of biologically important molecules and compounds;*

3 *enable persons engaged in biotechnology research and medical care to use systems developed under paragraph (1) and methods described in paragraph (2); and*

4 *coordinate, as much as is practicable, efforts to gather biotechnology information on an international basis.*

Beginning with a modest budget of $8 million and a dozen staff members, NCBI began its journey to become a national resource for molecular biology information. Dr. David Lipman, a key developer of the FASTA algorithm, was recruited from NIDDK and was appointed as NCBI Director. Along with the support of his three key appointees—Dr. Dennis Benson, Chief, Information Resources Branch; Dr. David Landsman, Chief, Computational Biology Branch; and Dr. James Ostell, Chief, Information Engineering Branch—Dr. Lipman rapidly grew the center into a major information hub in the molecular biology revolution.

NCBI is now a leading source for public biomedical databases, software tools for analyzing molecular and genomic data, and research in computational biology. Today NCBI creates and maintains over 40 integrated databases for the medical and scientific communities as well as the general public. There are over 3 million visitors daily to its website, approximately 27 terabytes of data downloaded per day, and the number of users as well as downloads increases dramatically each year.

Listed below are some of the major milestones from the many that have occurred over the past quarter of a century:

- **1990—BLAST**—the Basic Local Alignment Search Tool (BLAST) is introduced; optimized for speed, the sequence comparison algorithm quickly finds similar sequences to one's query.
- **1991—Entrez**—The search and retrieval system for NCBI's linked databases is introduced in CD form, allowing users to easily find related information from different databases.
- **1992—GenBank at NCBI**—NCBI assumes responsibility for GenBank, a database of nucleotide sequences, and collaborates in its development with international partners at the European Molecular Biology Laboratory (EMBL) and the DNA Data Bank of Japan (DDBJ).
- **1993—Network Entrez**—Network Entrez, a client-server version of the CD-ROM, is introduced, bringing Entrez to the Internet.
- **1994—NCBI Website**—NCBI establishes its own website, mounting initially BLAST, Entrez, **dbEST** (Expressed Sequence Tags), and **dbSTS** (Sequence Tagged Sites).
- **1995—Genomes**—This new resource organizes information on genomes, including sequences, maps, chromosomes, assemblies, and annotations.
- **1995—Bankit**—The online tool is introduced to facilitate submissions to GenBank.
- **1996—OMIM**—NCBI mounts the Online Mendelian Inheritance in Man (OMIM), a directory of human genes and genetic disorders.
- **1997—PubMed**—NCBI introduces PubMed, a freely accessible bibliographic retrieval system to the entire MEDLINE database. The new service is launched at a Capitol Hill event by Vice President Al Gore and the ranking Labor/HHS Appropriation Subcommittee members, Senator Tom Harkin (D-IA) and Senator Arlen Specter (R-PA), highlighting its significance.
- **1998—New NIH Disease-Based Services**—Collaborations with NIH Institutes for Disease-Based Services are established such as **CGAP**, the Cancer Genome Anatomy Project, to identify the human genes expressed in different cancerous states.
- **1999—Human Genome**—Human Genome Project researchers completely sequence the first human chromosome (#22) and deposit the sequence data at NCBI. A working draft of the entire human genome is completed the following year and made freely accessible from NCBI.
- **1999—Suite of Genomic Resources**—NCBI releases a number of resources to support comprehensive analysis of the human genome, including: **LocusLink**—key descriptors of genetic loci; **RefSeq**—a non-redundant set of human reference sequences; and **dbSNP**—a collection of data on human genetic variation.
- **2000—PubMed Central**—NCBI debuts its free full-text digital archive of biomedical and life sciences journal literature. PubMed Central (PMC) serves as an online counterpart to NLM's extensive print journal collection and is in keeping with the National Library's legislative mandate to collect and preserve the world's biomedical literature.
- **2000—GEO**—The Gene Expression Omnibus database is launched in response to community interest in a public repository for data generated from high-throughput microarray experiments.
- **2001—Bookshelf**—The new Entrez database is introduced to provide free access to books and documents in the life sciences and healthcare fields.
- **2002—WGS**—GenBank begins including Whole Genome Shotgun sequences, which are generated by a semi-automatic technique.
- **2003—DTDs**—NCBI Introduces Document Type Definitions (DTDs) for archiving and exchanging journal content.
- **2003—Entrez Gene**—The Entrez Gene database (formerly known as LocusLink) is developed to supply key connections between maps, sequences, expression profiles, structure, function, homology data, and the scientific literature.
- **2004—PubChem**—The PubChem database is released, providing information on the chemical structure and biological activities of small molecules.
- **2005—NIH Public Access**—NIH develops a Public Access Policy to provide scientists, researchers, and the general public with access to the published results of NIH-funded research through NCBI's PubMed

Central. NCBI develops a NIH Manuscript Submission System that allows researchers to submit their published papers to PubMed Central.

- **2005—My NCBI**—NCBI introduces the My NCBI tool, which retains user information and database preferences to provide customized services for many NCBI databases.
- **2006—dbGaP**—NCBI launches the database of Genotypes and Phenotypes (dbGaP) to archive and distribute the results of studies that investigate the interaction of genotypes and phenotypes. Studies include Genome-Wide Association Studies **(GWAS)**, medical sequencing, and molecular diagnostic assays, among others.
- **2007—Genome Reference Consortium**—The Consortium of NCBI, EBI, Sanger Institute, and the Genome Institute is created to improve the sequence quality and accuracy of the human reference genome. It takes on the task of improving the reference sequences for other model organisms, including the mouse and zebra fish, often used as models for human disease.
- **2008—Discovery Initiative**—NCBI embarks on a program to help users better explore the myriad of data contained in NCBI's resources. Automated methods are employed to surface related data that may not be apparent to the user in their original search query but which could lead to serendipitous discoveries.
- **2008—Public Access Becomes Mandatory**—Congress enacts legislation mandating that scientists submit final peer-reviewed journal manuscripts that arise from NIH funding to PubMed Central. The policy requires that the papers be made public on PubMed Central no later than 12 months after publication.
- **2008—1000 Genomes Project**—NCBI archives and distributes data from this international public-private consortium, which aims to build the most detailed map available of human genetic variations. In **2012**, NCBI improved the accessibility of these data by collaborating on an effort to make them available on the cloud through Amazon Web Services.
- **2010—dbVar**—NCBI establishes the dbVar archive of large scale genomic variation data and associated defined variants with phenotypic information.
- **2010—My Bibliography**—NCBI introduces the My Bibliography tool to simplify the process for gathering one's published articles and other materials. The tool, which is connected to NIH's grants management system, also assists researchers in complying with the NIH Public Access Policy.
- **2011—PubMed Health**—The service was introduced to provide information for consumers and clinicians on prevention and treatment of diseases and conditions, with an emphasis on reviews of clinical effectiveness research. It was discontinued on October 31, 2018, so NLM could consolidate its consumer health information and comparative effectiveness resources to make them easier to find in PubMed, MedlinePlus, and Bookshelf.
- **2012—Genetic Testing Registry (GTR)**—NCBI, in collaboration with NIH, launches the GTR to address the need for information about genetic tests for healthcare providers, researchers, patients, and others. The database provides information about the availability, validity, and usefulness of genetic tests.
- **2013—ClinVar**—NCBI creates the ClinVar database to aggregate information about sequence variation and its relationship to human health. The database includes submissions from outside research and testing groups as well as internal data drawn from such sources as dbSNP, dbVar, dbGap, and Gene Reviews.
- **2013—PubReader**—NCBI develops a new presentation style that optimizes reading of PMC articles through a browser on a desktop, laptop, or tablet computer.

An in depth account of the history of NCBI and NLM can be found in a published version of the Joseph Lieter NLM/MLA lecture presented at the annual meeting of the Medical Library Association in 2007 (1).

References

1. Smith KA. Laws, leaders, and legends of the modern National Library of Medicine. J Med Libr Assoc. 2008 Apr;96(2):121–33. PubMed PMID: 18379667.