

ONLINE

Exploring Technology & Resources for Information Professionals

From Concept to Content:

The Genesis of USGENE

by Suzanne Sabroski



Martin Goffman

Author's note: This article is an unofficial follow-up to an interview I did with Martin Goffman in 2000 for my book, Super Searchers Make It on Their Own (Information Today, Inc., 2002). As the 10th book in the Super Searcher series, it featured extensive profiles of 11 information entrepreneurs, with Goffman included as the patent information expert. It was a pleasure to revisit his business life and successes 8 years later.

WITH internet content exploding, Web 2.0 and its requisite search tools emerging, and journalists no longer defining the news, information professionals are adjusting to a world where anyone with access to a keyboard is not only an author but can also be an "expert." Searchers stay ahead of the game by constantly viewing content with a skeptical filter: What is it, who wrote it, when was it written, why was it said, and where was it published? If information does not set off any whacko alarm bells, it passes this unofficial test and may just be worthy to pass along to a patron or client.

Enter the content aggregators: Long before everyone's desktop was connected to the internet, industry-specific information was gathered, coded, and distributed via Dialog, LexisNexis, Orbit, STN, and Dow Jones Factiva. Although these vendors must now compete harder for their customers' attention, information professionals continue to rely upon the credibility that comes with hosted content—especially when it comes to intellectual property and patents.

THE PATENT PROCESS

Patent literature commands the respect of any searcher who has ever gone near it. Intellectual property experts say that about 80% of the information published in a patent document is not

available anywhere else. There are no lexicographers who monitor the publication process, so inventors can use whatever language they choose. If you are scanning the patent literature for a type of table, you could miss important documents simply because the inventor describes her device as "a planar surface with three or four perpendicular members." Equally important to understand is the issue of timing. A patent application is usually published 18 months after filing. Once published, the application is available for public inspection, while under examination by the patent authorities in its country of origin. When the examination is complete, and if the applicant is successful, the patent is granted and issued. This legal process may take several years to reach a conclusion.

The stakes are particularly high when it comes to patent research in the pharmaceutical, biotechnology, and agricultural industries. The importance of DNA and protein research is critical to the discovery of new drugs and vaccines, genetic therapies, and sustainable agriculture. Missing details in genetics research is not an option. A key part of this process is known as genetic sequencing, or sequence data.

Last year a groundbreaking database known as USGENE was released on STN International, a service of FIZ Karlsruhe and Chemical Abstracts Service (CAS). USGENE allows searchers to

perform freedom-to-operate, prior art, validity, and infringement patent sequence searches in U.S. gene patent publications, and provides the most up-to-date sequence data from the United States Patent and Trademark Office (USPTO). While one would expect the source of this database to be a Fortune 100 company or an oversized government agency, USGENE in fact originates from a one-person firm known as SequenceBase Corp., located in the home office of New Jersey resident Martin Goffman.

BACKGROUND

Goffman began his career as a Ph.D. chemist; he worked in the corporate sector for more than 20 years as both a laboratory researcher and inventor. In 1985 he started his own patent research and consulting business, which he still operates today as Martin Goffman Associates. His client base runs the gamut from small inventors to the Fortune 500, and he has enjoyed tremendous success over the years. He is also the co-founder of StockPricePredictor, LLC, which offers automated patent valuations.

In order to stay connected as a sole practitioner, Goffman is active in the Patent Information Users Group (PIUG) and the Association of Independent Information Professionals (AIIP). An expert searcher first and foremost, he has been online since the early days of Dialog and Chemical Abstracts Service (CAS). As a pioneer in the work-from-home movement of the 1990s, he has truly found what works for him. "I love having my office at home", says Goffman. "I tried having an office outside the house at one point and found that I really resented the time it took to get there!"

IDENTIFYING THE NEED, SEEING AN OPPORTUNITY

Goffman specializes in scientific and technical patent searching for legal and competitive intelligence applications. In the course of his work, he would routinely need to search several separate sources in order to construct the information needed for his clients. While attending a training session on sequence searching led by Robert Austin of FIZ Karlsruhe, the two began a discussion of what was missing in available sources. The major sequence databases in existence at the time were lacking a significant portion of U.S. granted patents and applications. Goffman noted that a proficient sequence searcher would need to go to five different sources, at a minimum, in order to construct the proper results.

As a trainer specializing in scientific and patent searching in the STN databases, Austin was also keenly aware of the need for a single source for U.S. patent sequence searching. Although the USPTO held the data, many of the scientists in Austin's courses did not know how to obtain it or thought it was all freely available on the web. Out of pure necessity he had developed a cumbersome methodology for sequence searching involving numerous government and commercial sources. In 2002 he co-authored a paper titled "Information Resources for Biotechnologists Part 1: Sequences," in which the whole research area was described as a "minefield for the unwary" (*Chemistry in Australia*, August 2002; www.stn-international.de/training_center/bioseq/information_for_biotechnologists_p1.pdf).

According to Austin, "Others had been expressing the need for a single source of sequence data for some time. The amazing thing is that the solution was created by one individual and not some giant corporation."

"I never set out to start another company," says Goffman. "I wasn't looking to grow my business or anything like that. I simply saw a need that was unfulfilled in the marketplace and went to work on

creating it! Rob Austin was an invaluable technical advisor throughout the development of USGENE up until its launch on STN and continues to provide excellent support and workshops for our customers in the U.S."

Since all of the sources that provide U.S. patent data are publicly available, Goffman says it was a matter of extracting useful portions of the information and combining it into one database. This included the main database put out by the USPTO, known as the Publication Site for Issued and Published Sequences (PSIPS). This database contains all sequences from patents that are more than 300 pages in length. In fact, many patents run more than 10,000 pages long. The USPTO full text was also included. Numerous proprietary computer algorithms were created to locate, parse, and extract the sequence information. Goffman then worked his way through the National Center for Bioinformatics (NCBI) Genbank patent division database, as well as the European Molecular Biology Laboratory of the European Bioinformatics Institute (EMBL-EBI) patent division database. After combining the sources, he added a careful deduplication process, as there is considerable overlap between the various USPTO databases. An enormous amount of quality-control measures followed to ensure clean and consistent data.

THE LAUNCH ON STN

In July 2007, 18 months after the creative process began, USGENE launched as an official database on STN International. Throughout the development Goffman maintained contact with Austin and others at FIZ Karlsruhe and found it was a natural fit to work together. Sample data exchanges and testing took place behind the scenes. When it came time to launch the product, a joint press release was issued by FIZ Karlsruhe and SequenceBase (www.stn-international.com/archive/pressroom/pressreleases/2007/usgene-new-db_en.html).

"The reason to go to STN was pretty clear—if you are a sequence searcher, that is where you go, as they hosted all the relevant content already assembled. I was able to offer them a database that would complement—not compete with—their current offerings," Goffman commented. "From a personal perspective they have been absolutely wonderful to work with."

STN describes the addition of USGENE as "a major enhancement to patent sequence searching on STN, where it features the same powerful command line based sequence searching options available in DGENE and PCTGEN." STN training courses have now been updated to include USGENE, and have come full circle, being taught by Austin (*USGENE on STN Workshop Manual*, updated May 2008; www.stn-international.com/archive/presentations/USGENE_workshop_manual.pdf).

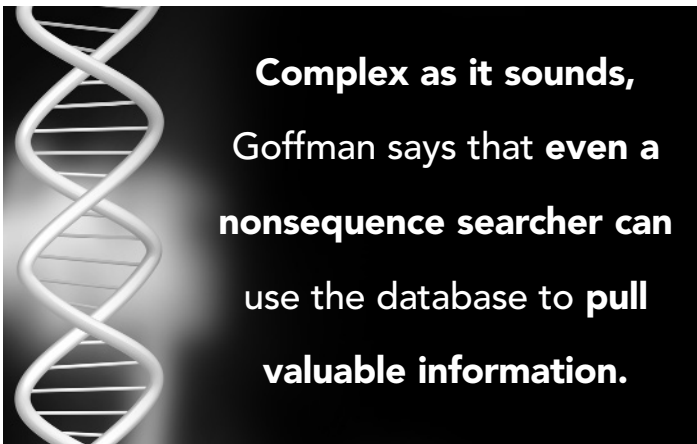
THE PRODUCT SPECS

USGENE covers all available peptide and nucleic acid sequences from the published applications and issued patents of the USPTO dating back to 1982. Each database record contains a sequence and related data including organism name, sequence length and tables for modifications, and other features. Bibliographic and text search options, including publication title, abstract, patent assignees at issue, full inventor names plus the complete set of publication, application, and parent case WIPO/PCT numbers and dates are also provided.

Complex as it sounds, Goffman says that even a nonsequence searcher can use the database to pull valuable information. If one is researching breast cancer, for example, they can go in and find

company names and treatments that are discussed in the context of a patent application—information that they are not going to find published anywhere else.

Another highlight is the currency of the data—USGENE is updated weekly on STN and provides USPTO patent sequence data within 3 days of publication. Numerous algorithms are in place to automate the process and to make it run smoothly.



“We are thrilled that we are the fastest in the industry with updating our data” says Goffman. “Patents that are granted on Tuesday are online at STN by Friday morning U.S. time. Applications that are published on Thursday are available the next morning, so that’s less than 24 hours.”

IMPORTANCE IN THE MARKETPLACE

Sequence data searching is a critical piece of the research that drives several major industries. It is an essential element to any type of genetics research and development.

“The need to conduct a thorough patent sequence search is of great significance in three areas,” according to Austin. “Biotechnology intellectual property in general, the science of genetics and biology because of the large volume of unique scientific information found in patents, and finally, the software of bioinformatics. However, patent sequence data represents a unique IT challenge. Users typically need to be able to search in a scientific way, but for legal reasons. The tools and databases they use must meet both requirements.”

In the pharmaceutical industry, sequence data is critical to developing new drugs and therapies. In the field of biomedicine, it is very common for patent applications to disclose nucleic acid and amino acid sequences. This is important in the hardware and software of genetics research. One example is the development of biochips to detect certain genetic traits. In agriculture, large chemical companies are constantly developing genetically modified plants and seed varieties in the quest for sustainability and environmentally sound practices. Disease resistance, pesticide resistance, and harsh environment resistance are areas of concern and form the basis for large investments in agricultural research and development.

CUSTOMER PERSPECTIVE

USGENE user James Coburn is president and CEO of Harbor Consulting IP Services, Inc. (<http://seqidno.com>), an outsourcing provider serving law firms and patent attorneys. Coburn and his staff review patent applications to identify all relevant DNA and amino acid sequences, and prepare a valid sequence listing

that meets U.S. and World Intellectual Property Organization (WIPO) requirements. Launched in 1995 as a general patent search firm, specializing in sequence searching was a logical extension of their services as a response to market demand. Prior to the launch of USGENE, Coburn was also working around the gap in sequence searching of U.S. patents.

“One of the key things missing for our business was a straight route to the U.S. data,” said Coburn. “Nobody was providing access to PSIPS and the huge listings that resided there. Some of our clients only wanted U.S. sequences and there was just no way to hit the data they needed.”

As principal of Harbor Consulting IP Services, Coburn attended training sessions offered by STN, and became yet another voice in Austin’s courses expressing the need for a product such as USGENE. Once development was underway, Coburn actually became a beta tester for USGENE and today is a satisfied customer. He and Goffman were introduced recently at the SequenceBase-sponsored PIUG 2008 Boston Biotechnology Meeting (www.piug.org/Biotech/2008/biotech08meet.php). “It was great to meet the one guy who created this product,” Coburn commented. “USGENE has really helped our business grow.”

DABBLING IN CURRENT DEBATES

Taking a large step back to view the complex current debates that are affected by genetics research, the “Aha!” factor is huge. Consider the stakes with public policy issues such as biomedical research into Alzheimer’s disease, cancer, Parkinson’s disease, and others. Consider the debate on the patenting of the human genome and where that may end up. What about the cloning of sheep? Contentions in bioethics abound, and with good reason, as so much is at stake in the quest for new discoveries.

The patenting of genetic organisms began only 25 years ago, with the intention of advancing new drug therapies. In the past year, the U.S. Supreme Court has begun to review more patent cases and has indicated that, in a variety of fields, the USPTO and Federal Circuit have incorrectly interpreted patent law by granting the right to patent genes. A common argument is that without patents for genetic sequences there would be no motivation or incentive to conduct research. This raises the question: Is the motivation pure scientific discovery or is it profit? Do gene patents actually prevent access to patented tests or therapies because of higher costs, or do they protect the rights of patent owners and drive the market toward more discovery?

In 2005 the ethics of gene patenting was raised in *Science* magazine (Jensen, Kyle and Fiona Murray “Intellectual Property Landscape of the Human Genome,” *Science*, 14 Oct. 2005: Vol. 310 No. 5746, pp. 239–240; www.sciencemag.org/cgi/reprint/310/5746/239.pdf) and was later picked up in the mainstream media by *The Guardian*, (Ravilious, Kate, “Private companies own human gene patents,” *The Guardian*, Oct. 14, 2005; www.guardian.co.uk/science/2005/oct/14/genetics.research), bringing the scientists’ debate into public view.

Last year *The New York Times* (“Patenting Life,” *The New York Times*, Feb. 13, 2007; www.nytimes.com/2007/02/13/opinion/13crichton.html?_r=1&oref=slogin) ran an op-ed piece by Michael Crichton (of *Jurassic Park* fame), who raised the issue of a specific medical test for breast cancer that is three times the actual cost in the U.S. because of a gene patent held by a large corporation. Recently the University of Sussex in the U.K. released a report on The PATGEN Project (“The Patenting of

