

**NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION  
NATIONAL LIBRARY OF MEDICINE, NIH**

**BOARD OF SCIENTIFIC COUNSELORS  
MEETING MINUTES**

**April 24, 2018**

**9:00 a.m. – 2:00 p.m.**

The Board of Scientific Counselors of the National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), convened on April 24, 2018, in the NLM Board Room, Bethesda, Maryland. The meeting was open to the public. Dr. Valerie De Crecy-Lagard presided as Chair.

**BSC Members Present**

Valerie De Crecy-Lagard, Ph.D., University of Florida (*BSC Chair*)  
Michael Boehnke, Ph.D., University of Michigan  
Kiplin Guy, Ph.D., University of Kentucky  
David Relman, M.D., Stanford University  
Steven Salzberg, Ph.D., Johns Hopkins University  
James Ostell, Ph.D., NCBI, NLM (*BSC Executive Secretary*)

**NLM Staff Present**

Jeff Beck, NCBI, NLM  
Dennis Benson, Ph.D., NCBI, NLM  
Patricia Brennan, Ph.D., NLM (*participated by phone*)  
Janet Coleman, NCBI, NLM  
Kathi Canese, NCBI, NLM  
Larry Fitzpatrick, NCBI, NLM  
Al Graeff, NCBI, NLM  
Elizabeth Kittrie, NLM  
Bill Klimke, Ph.D., NCBI, NLM  
David Landsman, Ph.D., NCBI, NLM  
Zhiyong Lu, Ph.D., NCBI, NLM  
Kim Pruitt, Ph.D., NCBI, NLM  
Valerie Schneider, Ph.D., NCBI, NLM  
Jerry Sheehan, NLM  
Steve Sherry, Ph.D., NCBI, NLM  
Bart Trawick, Ph.D., NCBI, NLM  
Eugene Yashenko, NCBI, NLM

**I. Welcome and Introductions**

Dr. De Crecy-Lagard called the meeting to order at approximately 9:00 a.m. Members introduced themselves. Dr. Ostell thanked the Board for their work and offered special thanks to

the members who would be cycling off the BSC after this meeting: Drs. De Crecy-Lagard, Guy, and Green (who was not in attendance). Dr. Brennan, who joined the meeting by phone, added her thanks to the BSC, emphasizing the importance of their input.

Dr. Landsman reported that the next two BSC meetings are scheduled for November 13, 2018, and April 9, 2019. He noted that four new members will be joining the BSC: Kateryna Makova, Ph.D., Penn State University; Katie Pollard, Ph.D., University of California-San Francisco; Donna Slonim, Ph.D., Tufts University; Pamela Soltis, Ph.D., University of Florida.

Dr. Boehnke – who together with Dr. De Crecy-Lagard represents the NCBI BSC on a recently formed Blue Ribbon Panel that is reviewing NLM’s intramural research program – raised the possibility of adding another BSC member to the panel, or replacing him, because he is going to miss the only in-person meeting, on May 24-25. Dr. Brennan said she would support adding an additional person and suggested that Drs. Ostell and Landsman confer with Jerry Sheehan, who could propose the addition to the NIH Deputy Director for Intramural Research. Dr. Relman was suggested as a potential member; he said he would check his calendar before committing.

A BSC member asked about a recent announcement from China regarding required sequence depositions to a database in that country. Dr. Ostell said that he had no details and thought it best not to speculate at this point about whether there might be any impact on deposition to databases in the International Nucleotide Sequence Database Collaboration.

## **II. NLM Strategic Plan – Dr. Patricia Brennan**

Dr. Brennan described the NLM 2017-2027 Strategic Plan, which resulted from an effort undertaken over the last 18 months and represents the NLM Board of Regent’s advice to NLM. The Strategic Plan has three key pillars, each of which has several parts. The pillars, and their components, are as follows:

### **Accelerate discovery and advance health through data-driven research**

- Connect the resources of a digital research enterprise
- Advance research and development in biomedical informatics and data science (Dr. Brennan noted that NLM received approximately a 4% increase in its FY18 appropriation and would be allocating \$5M each to intramural and extramural research. Within intramural, NLM has received approval for four new investigators, each of whom will be supported by 2 staff, for a total of 12 new researchers.)
- Foster open science policies and practices
- Create a sustainable institutional, physical, and computational infrastructure

### **Reach more people in more ways through enhanced dissemination and engagement**

- Know NLM users and engage with persistence
- Foster distinctiveness of NLM as a reliable, trustable source of health information and biomedical data
- Support research in biomedical and health information access methods and information dissemination strategies

- Enhance information delivery

### **Build a workforce for data-driven research and health**

- Expand and enhance research training for biomedical informatics and data science
- Assure data science and open science proficiency
- Increase workforce diversity
- Engage the next generation and promote data literacy

### Q&A/Discussion

Discussion following Dr. Brennan’s presentation focused on the role of NLM in personal genomics. BSC members commented that in the near future, whole genome sequencing and/or exome sequencing likely would become a standard of care, and that it is important to think about how best to take advantage of that shift in the context of research.

Dr. Brennan said that while NLM does not anticipate storing individually identifiable genomes in the near future, its growth in this area may be through partnerships, especially the ability to locate and connect datasets. She noted, for example, that there have been some requests from individuals who have data in dbGaP to have their addresses attached to their sequences so they may be contacted by researchers interested in knowing how they are managing their diseases. Dr. Brennan added that one of the areas NLM is investing in is understanding its role with research data from electronic health records.

Dr. Brennan noted that NLM, through its network of libraries of medicine, has taken an important role in NIH’s All of Us program, providing informational materials as well as specialized training to librarians who may get questions from patrons. She added that two goals of the precision medicine initiative – providing health information back to study participants and stimulating citizen scientists – are going to require that NLM builds informative and useful tools.

Dr. Brennan also cited a program NLM launched last year in its extramural program called Personal Health Libraries. The program funded 8 projects in FY17 and probably will fund 10 in FY18 that are either developing tools, or visualizations, or interesting displays that address a range of personal health information.

Dr. Brennan indicated that a number of the issues involved in the discussion about personal genomes and personal health data are included in the NIH Strategic Plan for Data Science; an “almost final version” of the document is now circulating through HHS and should be available in mid-May.

### **III. NCBI Today: A Finite Resource in an Expanding DataVerse – Dr. James Ostell**

Dr. Ostell presented information about the growth of data and its usage at NCBI and some of the changes NCBI is making to handle that growth with limited resources. Metrics and data he presented include:

- NCBI resources are used by 5.5 million people/day and receive 23 million page views/day
- Users download 80 terabytes of data each day
- Peak web hits number 7,000/second
- PubMed Central page views have roughly doubled over the last 3 years, from about 600 million in 2014 to 1.3 million in 2017
- ClinVar page views have gone from 1.5 million in 2014 to 10.6 million in 2017, despite the primary use of this resource being data downloads

Dr. Ostell noted that NCBI has seen exponential growth in services such as GenBank and the Short Read Archive (SRA), and that the growth curve for SRA is exceeding NCBI's resources such that it cannot pay for disk storage anymore, let alone computing on the data. Despite the growth in data and usage, NCBI's staff and budget has not been increasing in recent years, creating a challenging environment. NCBI had been facing a budget shortfall, but NLM provided NCBI with additional funds following NLM's budget increase, and NCBI can now cover its obligations, though it is still facing the situation of having to handle exponential growth in data with the same number of staff.

Dr. Ostell described how NCBI has changed its process for managing its programs and making deliberate decisions about where to dedicate resources. As part of that effort NCBI recently reorganized. Dr. Ostell outlined the current NCBI organizational structure, which involves three new divisions under the Information Engineering Branch, which is headed by Dr. Kim Pruitt, who is acting branch chief. The three new divisions under IEB are the Data Services Division (also headed by Dr. Pruitt), the Customer Services Division (headed by Dr. Bart Trawick), and the Software Division (headed by Lawrence Fitzpatrick, on an acting basis).

Dr. Ostell went on to describe some of the elements of the reorganization and how they are helping NCBI manage its workload. One big shift relates to how software is developed. Previously software developers were often on a project, such as GenBank, for the long term. While this allowed for easy communication between the software developers and content staff on the same project, it meant that other project teams might not learn of a new tool or technology that might be useful to them, and it led to software developers creating more and more features for a resource ("feature creep") when they had extra time. In the new model, NCBI is making decisions across the organization about the projects and features to focus on, how much investment to make, and then assigning software developers and other staff to those projects with a set timeframe and milestone goals. NCBI also has added project and product managers to keep on track with the milestones.

Dr. Ostell gave a couple of examples of the accomplishments NCBI has made using its new approach, the first being dbSNP. The database had been using a 15-year-old system that was originally designed for 10,000s of variants, and it was now faced with over one billion submissions grouped into 400 million reference variants. Over the years additional views of the data were added, and users started to complain that views of the data were inconsistent, which was due to staggered and manual releases. In addition, because of limitations of the software used, the database was not able to properly group two submissions that were the same, and production costs were exploding.

NCBI addressed the dbSNP problems by adopting a formal technology, with a scheduled release and using limited resources. The new technology is based on a MapReduce framework that allows for a different logical structure. Dr. Ostell explained that in creating the revised database one of the first tasks was to map variants to a genome and then to find variants at the same location and reduce them to a common form. The results have been very successful and the cost is dramatically decreased, he said. The new version is currently entering production.

Dr. Ostell also described NCBI's automation of rRNA sequence submissions, which has resulted in processing that used to take months now only taking 10 minutes. Other efforts include automation of processing for whole genome sequence submissions and using ANI for bacterial genomes to reliably cluster species.

#### **IV. NCBI Tomorrow: A Finite Resource in an Expanding DataVerse – Dr. James Ostell**

Dr. Ostell began his presentation with the metaphor that NCBI historically has been a walled garden, where data suppliers throw their data over the wall and NCBI does its magic to organize it and make it accessible on an ongoing basis and at no cost to the submitters. That model has become problematic with the large growth of sequencing data. As the new director of NCBI, Dr. Ostell said his approach is to move out towards the data, instead of bringing it all within NCBI. This effort includes engaging more with NIH, such as participating in the Scientific Data Council, the NIH Cloud Commons Pilots, Identity and Access Management (with CIT), and NIH cloud planning. It also involves engaging more with NLM, for example with efforts to consolidate IT, incorporating NLM's ToxNet resource into NCBI's PubChem and literature resources, and working with Library Operations. In addition, NCBI is engaging with commercial cloud platforms and using standard frameworks.

Dr. Ostell briefly described the NIH Strategic Data Plan, a first draft of which was released in March for comments. The draft cited five overarching goals:

- Support a Highly Efficient and Effective Biomedical Research Data Infrastructure
- Promote Modernization of the Data-Resources Ecosystem
- Support the Development and Dissemination of Advanced Data Management, Analytics, and Visualization Tools
- Enhance Workforce Development for Biomedical Data Science
- Enact Appropriate Policies to Promote Stewardship and Sustainability

Two strategic objectives of the first goal are to 1) optimize data storage, access and security, relying on the private sector, where possible, and 2) to connect NIH data systems. Dr. Ostell noted that NCBI has been making use of the private sector (e.g., for software). Regarding the connection of NIH data systems, he noted that the draft plan specifically mentions that NLM and NCBI can be used as hubs.

Dr. Ostell described two use cases for a cloud platform. In one case, NCBI uses the cloud to deliver its services to the public; for example, NCBI expects to be delivering PubMed from the cloud within 12 months. In the second use case, the cloud is used to provide access to NCBI or

non-NCBI data. For data housed at NCBI, users would be able to access the data in the cloud and compute on it there without having to download it. For non-NCBI data, NCBI can provide the indexing and access permissions without holding the data.

NCBI has been working on putting BLAST on the cloud in a way that would make it very convenient and accessible to users but allow NCBI to avoid bearing the cost of users accessing the data and running services. He noted that this is a model for how NCBI could move other data and services to the cloud.

### Q&A/Discussion

BSC members asked about the mix of researchers/labs using large versus small amounts of data. Dr. Ostell replied that the large data usage is by a small minority of users but that their usage is very expensive. He commented that over time he expects that the cloud will become very convenient to use and that even smaller labs likely will switch to using it instead of local computers.

BSC members also asked whether NCBI views its role as a repository of record or as an organization that develops ways to bridge between the research world and the usage world. Dr. Ostell said that NCBI's role is somewhere between those two spaces; it is the repository of record for some things, but it cannot have that role for everything produced by science.

In response to a comment that NCBI seems to be in the position of having to provide data forever without receiving the necessary resources, Dr. Ostell said that the cloud model NCBI is working on would enable it to be in a much better position for dealing with the ever-increasing quantity of data. In the cloud model NCBI is proposing, which was described by Dr. Sherry in the following presentation, the funding NIH institute would put its data in the cloud and buy the storage, while NCBI would set up the permissions that enable access. The funding IC would then be the one to determine how long to continue to pay for storage for the full or partial data sets. One BSC member commented that the value of data goes beyond the funding IC and submitters, and that there may be future uses such as meta analyses.

### **V. Cloud Strategy and Distributed Data – Dr. Steve Sherry**

Dr. Sherry described the pilot program NCBI is engaged in with MITRE Corporation – a federally funded research and development organization – for delivery of SRA sequence data in the cloud.

The infrastructure is being organized around two types of cases. One case is from the perspective of the research investigator or team who wants rapid access to data that might be distributed over multiple administrative programs, funded by different NIH ICs, and with different permission requirements. The aim would be to quickly provide access to the data, for example the subset of people with a certain expression profile in multiple studies, without months of moving data and the costs that typically would be involved.

The second case is from the NIH perspective. NIH provides millions of dollars of funding for large cohort sets, such as NHLBI's Trans-Omics for Precision Medicine (TOPMed) study and NIA's Alzheimer's Disease Sequencing Project (ADSP). Currently storage of the data from many large studies is siloed. If samples across programs could be combined there would be greater statistical power, particularly with the effect of rare variants, and NIH could better leverage its investments.

Dr. Sherry outlined four complex problem areas related to providing cloud access to such data:

- 1) Identity – Who is approved to get the data and how do they present their credentials in the cloud environment
- 2) Authorization – Is the user authorized to access a particular object on the cloud
- 3) Discovery – What data is available, where is it, and in what formats
- 4) Delivery – How to access the data and track who used it and when

In order to tackle these four areas NCBI needs to build a couple of services, he said. First, NCBI needs to find a way to take its knowledge about where artifacts or objects are stored on the cloud (whether it be Amazon, Google or another provider) and create a relationship that can get to the data in a trustworthy and user-specific way. Second, this needs to scale computationally so that it will work for thousands of users. He noted that NCBI is looking at how to take its SRA run selector, which is how SRA presents metadata about information, and enhance that in the cloud environment. In addition, NCBI is working with NIH's CIT on a centralized service for identity and authorization management that uses NIH's eRA IT infrastructure.

Dr. Sherry described some of the backend engineering that is being done to enable functions such as access authorizations for data in the cloud. He noted that NCBI worked with MITRE to write code, called Fusera, that creates a virtual directory to user-approved data that has been authorized to a dbGaP user. The software mounts the data the user wants to see out of the hundreds or thousands of genomes that they might have available to them. The system uses signed URL technology, through encrypted URLs on the backend, to create virtual access so that the software can natively deliver the subset of objects wanted based on the user's data access approvals. That information all gets logged so that NCBI knows who is obtaining what data.

Dr. Sherry characterized the pilot project as an extension of SRA into cloud-based storage. Authorized users can access data in the cloud just like they could if they were downloading the data from SRA. The searching is fully integrated with a unified catalog. He commented that this model is aligned with Dr. Brennan's aspiration that NLM is a consolidated hub; NCBI will be providing a directory of data and carrying information about the quality, source, expiration date, and terms of access without storing the raw data.

Submitters of data held in the cloud would provide metadata to SRA that would include attributes about the data such as locations where it resides and the file type. NCBI is not constrained about the data types since it is not extracting or normalizing the data, and, for example, can include VCFs in addition to BAM files.

Dr. Sherry showed a slide of how users would come in through dbGaP and their project would be linked to a collection of datasets they requested. With one approval they could have an identity

key that says they are authorized to access a broad array of data across studies. The identity function is currently part of dbGaP, but NCBI wants to replace this with a virtual identity service from CIT.

### Q&A/Discussion

BSC members generally commented that the cloud model is sensible and scalable. Issues raised following Dr. Sherry's presentation primarily focused on use of the eRA system, privacy management, and combining datasets.

One BSC member commented that with the ability to aggregate data, multiple data sets will be connected that were not previously connected, which raises potential concerns about privacy management. Dr. Sherry responded that the umbrella protection is that users of the datasets must promise they will not try to re-identify anyone, and that users are not supposed to apply biometrics to connect people. IRBs that approve studies also have the role of protecting the privacy of study subjects.

The board briefly discussed the possibility of concerns with data being created for a particular purpose and then being combined with other data for an entirely new purpose. Dr. Ostell commented that that is a "really big discussion" and that the issue is at the NIH level and outside the bounds of NLM's purview.

A BSC member pointed out that the cloud model entails increasing dependency on data residing elsewhere and asked about the government's responsibility for understanding where data resides, particularly if outside the U.S. Dr. Sherry replied that the pilot program currently is limited to domestic cloud providers, but that NIH has never restricted investigators from putting copies of data in other locations; those discussions generally occur at the time of a funding award, he noted.

One BSC member raised concern that using eRA for identity verification would make it useful only for NIH grantees. Dr. Ostell responded that this issue was also raised when eRA was used for dbGaP access, and it was discovered that many foreign investigators who are partners on NIH grants are in eRA. eRA also agreed that it would add people if necessary to support use for identity management. While it is not perfect, he said, it is a solid management system that serves a large portion of potential users, and it is a good place to start and then improve upon with incremental changes.

### **VI. Tier 1 Project Updates – Dr. Kim Pruitt**

Dr. Pruitt described NCBI's current Tier 1 projects, which are high-priority projects that NCBI has chosen to focus on for a year, with specific deliverables and metrics of success. NCBI made an informed decision to put more energy and effort into these projects, and reallocated staff from other projects to these efforts, Dr. Pruitt explained.



Five projects were selected for Tier 1 status in August 2017. A sixth project, putting sequence reads in the cloud, was added later following discussions with NLM and NIH. The five initial Tier 1 projects are as follows:

#### Pathogen Detection project

The strategic goal of the Pathogen Detection project is to integrate bacterial pathogen genomes originating in food, environmental sources and patients, and then cluster and identify related sequences to uncover potential food contamination sources. Partners in this project – CDC, FDA, USDA, state health labs – submit the sequences of bacterial samples to NCBI, which integrates the data to provide high-resolution rapid clustering of related isolates to aid in traceback and outbreak investigations.

NCBI's goals for the project include improving the pipeline processing time to provide Rapid Reports within an hour of sequence submission and to provide SNP trees within 24 hours, Dr. Pruitt said. The Rapid Reports system already has been launched, with some improvements being made as the scale of data grows. A new clustering tool will soon be implemented that will further improve turnaround time. NCBI aims to accommodate more than 90,000 submissions/year.

Another goal is to improve the usability of the Pathogen Detection browser to enable rapid identification of isolates that need further investigation. Towards this end, enhancements have been made to the browser. New cluster match tables show the number of matched isolates per cluster, the total number of isolates, and the minimal SNP distance of any matched isolate to any other in the tree. A new tree viewer was implemented in February 2018 that allows epidemiologists to use the tool without having to go through intermediate steps. Improvements to the tree viewer are ongoing.

#### Virus annotation & access

The goal of this project is to improve quality and accessibility of viral sequence data, Dr. Pruitt said. The initial effort is focused on supporting public health users, but the project will also benefit the wider research community. The scope includes improved ease and quality of sequence submissions, normalizing data to enhance usability, and improved search and retrieval. The deliverables include a viral sequence search and retrieval interface with a customer-centric design, new submission and annotation tools, and usage analytics such as dashboards and surveys.

Reporting on the progress of the goals, Dr. Pruitt said a prototype virus sequence search interface with integrated BLAST searching is undergoing testing in NCBI Labs, where feedback is being gathered. Two new flu submission tools have been deployed: a web wizard and a programmatic interface. More than 17,000 influenza sequences have been submitted using the tools. NCBI has been engaging with FDA and CDC, and these interactions are driving the development efforts. Norovirus and Ebolavirus annotation tools are in development.

## PubMed 2.0

The scope of the PubMed 2.0 project is to update PubMed with a new relevance-based search and retrieval system and a redesigned user interface that is informed by user needs and feedback. Benefits include delivering the specific results that customers need, making the features intuitive and easy to learn, and optimizing the system for use on mobile devices.

The new interface and default use of the new relevance-based search are being tested in PubMed Labs, a site introduced in October 2017 to test potential new PubMed features and designs. Dr. Pruitt noted that feedback has been overwhelmingly positive. PubMed Labs has been introduced quietly, and the plan is to gradually increase exposure. Currently there are 3,000-4,000 users a day, and 60% are repeat users.

NCBI is partnering with GSA's 18F design team on a project to uncover the needs of PubMed users, test what is and isn't working, and create a roadmap for turning PubMed Labs into the new PubMed. The Phase 2 goal of the 18F engagement is to provide the user research and design requirements that will allow the PubMed development team to complete PubMed 2.0 to satisfy 90% of user needs – based on feedback and research – by the end of 2018.

## Known Item Search (KIS)

The goal of this project is to deliver expected, high-value results for text searches where users have known sequence items in mind, regardless of the queried database. KIS is intended for users who are searching for a “known item” (e.g., BRCA1 or E. coli genome) but who may not understand which NCBI database to search to find their preferred result. Dr. Pruitt noted that some users may not understand the structure or content of NCBI databases, may be overwhelmed by the number of results, and may not know which result is best for their search. Databases in scope for this project are Nucleotide, Protein, Gene, Assembly and Genome.

NCBI used user interviews, surveys, and query reviews to guide development. Based on the user data and the timeline, it was decided to focus the initial release on reducing the incidence of zero-hits and promoting RefSeqs. Also, the decision was made to implement a new search service that responded to more natural language-like queries and to display results in the form of a sensor. In order to assess whether KIS is achieving its stated goal, logging and dashboards are being used to monitor defined success metrics, including a decreased count of zero-results queries, increased clicks on the KIS sensor, and increased clicks on RefSeq Select.

## BLAST in the cloud

NCBI's move to the cloud has several dimensions that will provide a roadmap to our long-term IT strategy. Three notable cloud programs are PubMed, which involves many transactions but is a typical web application with a relatively small amount of data; human reads in the cloud, which involves big data storage and retrieval; and BLAST, which is a computationally intensive application. The rationale for having BLAST in the cloud includes the growth in BLAST searches and the difficulty of maintaining performance as searches are conducted against an ever-expanding pool of data. NCBI has managed to maintain performance through code refinement and by outsourcing compute power to the Amazon cloud, but this will not be sustainable going forward.

Objectives of the project are to deploy Web BLAST on the Google cloud platform without any runtime dependencies on other data centers, to scale with demand, to replace custom code with open-source commodity code, and to enable users to run BLAST under their own cloud account. This will allow NCBI to meet current and future performance needs of its users, and users will be able to scale their own BLAST system to meet their performance, security and database needs.

Developers from NCBI are working together with Google developers and have daily online meetings. Code is shared between NCBI and Google via an online repository. Functioning standalone components for web, orchestration, and map/reduce execution have been built. Dr. Pruitt presented a slide showing that a prototype is expected in the next few months. It is expected to take another nine months to rollout to production, after which BLAST would be migrated to other cloud vendor platforms for redundancy and sustainability.

### Q&A/Discussion

Much of the discussion following Dr. Pruitt's presentation centered on how consumers will get information about things like their genome as sequencing becomes more common. Dr. Ostell noted that consumer health information has unique needs, and that NCBI is doing less in this area but tries to provide support to NLM consumer health activities.

## **VII. NLM Strategic Activities and NCBI – Dr. James Ostell**

Dr. Ostell listed six trans-NLM strategic activities that include NCBI, two of which he only briefly mentioned: establishment of an NLM User Interface/User Experience team, and consolidation of NLM's toxicology resources into NCBI resources. The other four activities are:

### Blue Ribbon Panel Review of Intramural Research Program

The Blue Ribbon Panel will be reviewing the strengths and weaknesses of the current NLM intramural research and training initiatives for the period of 2008-2018. Their charge is to:

- Consider the optimal balance of and interaction among research (basic and applied), development, and provision of services and tools used by the biomedical community;
- Identify priority areas of biomedical informatics and biomedical data science research that NLM's intramural research program should pursue to advance biomedical research and health, considering relevant research activities in academia, industry, and elsewhere at NIH;
- Recommend ways in which NLM intramural research program can best support training to advance the fields of biomedical informatics and data science;
- Recommend (as warranted) changes to NLM's organizational structure, budget, staffing, internal and external partnerships, or other factors that could enhance the ability of NLM's intramural research program to advance NLM's mission in the next 5-10 years;
- Suggest approaches NLM can use to assess the outcomes and impact of its intramural research and training investments;
- Consider how to align NLM's intramural research activities with the goals articulated in NLM's new Strategic Plan, the NIH Data Science Strategic Plan and the recommendations of the 2015 NIH Advisory Committee report to the Direct

### Portfolio Analysis and Establishment of Performance Metrics

NLM is working with MITRE corporation on this effort, which includes establishing metrics and attributes for evaluating NLM resources, and creating an inventory of NLM resources.

Associating items in the inventory with the metrics will enable development of an NLM portfolio dashboard with information such as how many people work on a resource, how many people use the resource, system availability, etc. The MITRE analysis also will provide recommendations for moving forward. Dr. Ostell said the analysis is expected in the coming year.

### IT Assessment

Among other things, this effort is aimed at adopting common tools and a common Help Desk across NLM. Currently questions about PubMed are triaged in Library Operations, and they use a help system that is different from NCBI's. Once there is a common Help Desk, inquiries would come in, be answered, and be recorded through the same system. The IT assessment will also include performance and reliability statistics. In addition, common practices will be developed for how NLM manages servers and uses cloud services.

### Streamline and Integrate Deposit, Curation, and Discovery

Dr. Ostell showed a slide listing a number of items in this category:

- PubMed Data Management (PMDM) System – this has already been implemented and turned separate NCBI and LO systems into a single system that streamlines the process for adding and modifying citations
- Investigate the use of authoritative vocabularies in MEDLINE indexing in addition to, or as a partial replacement for MeSH, for some topics or types of metadata, for example, chemical names
- Implement a range of indexing methods to ensure the timely assignment of MeSH or terms from other approved vocabularies to MEDLINE citations
- Support the discoverability of ClinicalTrials.gov content
- Support the pharmacology and toxicology research communities by sustaining and improving the discoverability of chemical information in PubMed/MEDLINE
- Support NIH and other funding organizations by ensuring the discoverability of funding information in PubMed/MEDLINE

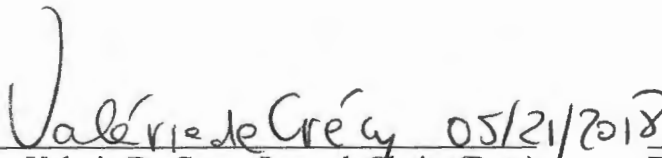
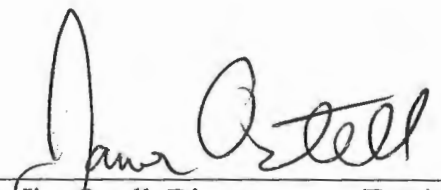
### Q&A/Discussion

Following Dr. Ostell's presentation the BSC discussed the issue of storage of supplementary data. Dr. Ostell noted that PubMed Central can store up to 2 gigabytes of data associated with a study, and that when authors submit through the manuscript submission system they are asked for supplementary data. However, publishers who submit on behalf of authors are not consistent in providing supplementary data. While PubMed Central provides a venue for data up to 2GB, the answer for bigger files likely will be the cloud, Dr. Ostell said. He added that even if the data is provided there is the question of its quality and how to index it so that it is useful. He noted that PMC has basic metadata, such as whether the data is an image, but that there is a need for a more flexible metadata description.

One BSC member commented that certain types of supplementary data, such as that related to structures and sequences, uses standardized formats that aren't platform dependent, making the data more useful, and suggested that NCBI consider such factors in determining the types of data for which it will be the repository of record.

### VIII. Adjournment

The BSC adjourned at approximately 2:00 p.m.

	
Dr. Valerie De Crecy-Lagard, Chair (Date)	Dr. Jim Ostell, Director (Date)
Board of Scientific Counselors	National Center for Biotechnology Information