NIH Workshop on Trustworthy Data Repositories for Biomedical Sciences

Apr. 8th 9:00 am - 5:15 pm, Apr. 9th 9:00 am - 12:15 pm

5601 Fishers Ln, 1D06AB, Rockville, MD 20852

The NIH Data Science Strategic Plan aims to modernize the ecosystem of biomedical data repositories. The overarching goal of the workshop is to explore principles of 'trust' as they relate to management of data repositories. Through this workshop we aim to better understand the principles that are most important for biomedical data repositories, and to explore interest in available options for acquiring independent certification of repository management practices. NIH is interested in promoting good data management practices by encouraging use of core principles for biomedical NIH-funded repositories.

The workshop will introduce the CoreTrustSeal, which is an emerging, international, community-based standard, and use it as an example to evaluate the management of data repository operations in biomedical sciences. Workshop speakers include advocates for using Trustworthiness Standards to support data sharing and data transparency (including CoreTrustSeal board members) and representatives from biomedical data repositories that already have trustworthiness certification. The targeted audiences are key data repository management personnel and NIH Program, Scientific Review and Policy Staff who work with data repository related programs.

The workshop will focus on high-level management of data resources, not technical details of operations. The first morning will be webcast and will introduce (1) the role of repository trustworthiness in supporting an effective data sharing ecosystem and (2) the proposed Core Requirements. The remainder of the workshop will not be webcast but will include hands-on exercises to enhance understanding. The workshop includes time to consult on specific topics to prepare a certification application. CoreTrustSeal reviewers will be available to answer repository-specific questions by appointment on the second day.

The in-person workshop is an invitation-only event due to limited space, and part of the workshop will be provided by webinar to those are interested in the workshop.

Please register for the webinar for Trustworthy Data Repository (TDR) Workshop on Apr 8, 2019 8:30 AM ET at:

https://attendee.gotowebinar.com/register/973713601040153857

After registering, you will receive a confirmation email containing information about joining the webinar. The webinar will be available on Apr. 8th 9:00am – 12:00pm, 2:00pm – 3:00pm.

We welcome live tweet of the workshop for the two webcast sessions. The Twitter hashtag is #NIHTDR

Please contact Dr. Fenglou Mao at fenglou.mao@nih.gov if you have any questions or comments.

Agenda

The workshop sessions between 9:00am – 12:00pm, 2:00pm – 3:00pm on April 8, 2019 will be Webcast at https://attendee.gotowebinar.com/register/973713601040153857

April 8, 2019

Session 1 - TRUST Concepts and Standards (webcast)

Chair: Dawei Lin (NIH/NIAID)

09:00-09:30 Opening Remarks, *Dawei Lin (NIAID) & Susan Gregurick (NIH/OD)*
09:30-10:00 Perspectives on Data Preservation, *Jonathan Crabtree (UNC Chapel Hill)*
10:00-10:30 International Standards for Trustworthy Data Repositories, *Robert Downs (Columbia Univ)*
10:30-11:00 **Break**
11:00-12:00 Introduction of CoreTrustSeal (CTS), *Ingrid Dillo (DANS)*
12:00-13:00 Breakout groups on reviews Trustworthiness requirements (no webcast)
13:00-14:00 **Lunch NIAID Cafeteria**

Session 2 - TRUST Examples (webcast)

Chair: Kim Pruitt (NIH/NLM)

14:00-14:30 ICPSR - A certified Social Science Repository by other standards, *Jared Lyle*
14:30-15:00 PDB - A certified Biomedical Data Repository by CTS standards, *John Westbrook*
15:00-15:20 **Break**

Session 3 - TRUST Challenges and Opportunities (no webcast)

Facilitators: Jonathan Crabtree, Robert Downs, Ingrid Dillo, Dawei Lin

15:20-16:30 Break out groups on applying requirements to Biomedical Repositories
16:30-17:10 Break out groups reports
17:10-17:15 Evaluation and close day 1

April 9, 2019

Session 4 - TRUST Community (no webcast)

Facilitators: John Westbrook (PDB) and Ingrid Dillo (DANS)

09:00-10:30 Hands-on session: applying requirements to Biomedical Repositories. What are the challenges and how can you help yourselves by helping each other?
10:30-11:00 Examples of collaborations from different disciplines around the world. *Ingrid Dillo on RDA, Robert Downs on Geoscience, Jonathan Crabtree and Jared Lyle on the Data Preservation Alliance for the Social Sciences (Data-PASS)*
11:00-12:00 Networking session - Forming a trust community
12:00-12:15 Evaluation and workshop close

Speaker Biographies

Dr. Susan K. Gregurick

Dr. Susan K. Gregurick is the Division Director for Biophysics, Biomedical Technology, and Computational Biosciences (BBCB) in NIH's National Institute of General Medical Sciences (NIGMS). Her mission in BBCB is to advance research in computational biology, biophysics and data sciences, mathematical and biostatistical methods, and biomedical technologies in support of the NIGMS mission to increase understanding of life processes.

Dr. Gregurick also serves as the Senior Advisor to the Office of Data Science Strategy, a newly formed office within the Office of the Director at NIH.

Prior to joining the NIH, Susan was a program manager for the Department of Energy where she oversaw the development and implementation of the DOE Systems Biology Knowledgebase, which is a framework to integrate data, models, and simulations together for a better understanding of energy and environmental processes. During Susan's academic career she was a Professor of Computational Biology at the University of Maryland, Baltimore County and her research interests include dynamics of large biological macromolecules. Susan holds a Ph.D. in Computational Chemistry and her areas of expertise are computational biology, high performance computing, neutron scattering and bioinformatics.

Dr. Dawei Lin

Dr. Lin joined the Division of Allergy, Immunology and Transplantation at the National Institute of Allergy and Infectious Diseases (NIAID), NIH, as a Senior Advisor to the Director and Associate Director for Bioinformatics in February 2013. He is a member of NIH Big Data to Knowledge Initiative (bd2k.nih.gov) and Program Officer for the BD2K Data Discovery Index (DDI) program (bioCADDIE.org). Prior to joining NIH, Lin was the founding Director of the Bioinformatics Core at the University of California Davis Genome Center; and before that, he led a Bioinformatics group at the Southeast Collaboratory for Structural Genomics (SECSG) at University of Georgia. Lin also spent time at the Brookhaven National Laboratory in New York, where he played a key role in the modernization and operation of the Protein Data Bank (PDB). Lin received his Ph.D. in Physical Chemistry with an emphasis on Computational Biology at Peking University, Beijing, China in 1996. Lin is widely recognized for his contributions to various "Big Data" initiatives and for his expertise in complex data analysis, bioinformatics, and high performance computational infrastructure. He is the elected Board Member of CoreTrustSeal. In addition to his work at NIH, he teaches at conferences on Next Generation Sequencing Technology and maintains a twitter handle Twitter (@igenomics).

Dr. Ingrid Dillo

Dr. Ingrid Dillo is Deputy Director at DANS (Data Archiving and Networked Services) in the Netherlands. She holds a PhD in history and has worked in the field of policy development for the last 30 years, including as senior policy advisor at the Dutch Ministry of Education, Culture and Science and the National Library of the Netherlands (KB). Among her areas of expertise are research data management and the certification of digital repositories. Ingrid is Co Chair of the

Research Data Alliance (RDA) Council. She is also Treasurer of the Board of CoreTrustSeal (CTS) and Vice Chair of the Scientific Committee of the ISC/World Data System (WDS).

Mr. Jonathan Crabtree

Dr. Jonathan Crabtree is the Director for Cyberinfrastructure at the Odum Institute for Research in Social Science at UNC Chapel Hill and helps lead the Global Dataverse Community Consortium (GDCC). The institute's social science data archive is one of the oldest and most extensive in the United States. As director, Crabtree completely revamped the institute's technology infrastructure and has positioned the institute to assume a leading national role in information archiving. He is president of the International Federation of Data Organizations (IFDO) and leads a development group supporting the use of Dataverse for data publication and verification workflows for journals.

Crabtree's experience in information science, information technology and networking as well as his engineering background bring a different perspective to his current role. Crabtree joined the institute over twenty-five years ago and is responsible for designing and maintaining the technology infrastructure that supports the institute's wide array of services. Before moving to the social science side of campus he was an information systems technologist for the University of North Carolina at Chapel Hill School of Medicine. His grounding in medical information technology adds to his education and training in electrical engineering, library and information science, digital preservation, computer science, economics, geographic information systems, hydrology and geomorphology. He is currently enrolled in the UNC School of Information and Library Science doctoral program with his research focuses on the auditing of trusted repositories.

Dr. Robert R. Downs

Dr. Robert R. Downs serves as the senior digital archivist and acting head of cyberinfrastructure and informatics research and development at CIESIN, the Center for International Earth Science Information Network, a research and data center of the Earth Institute of Columbia University. He is the co-chair of the Columbia University Morningside Campus Institutional Review Board, an elected member of the CoreTrustSeal Standards and Certification Board, co-leader of the Group on Earth Observations System of Systems (GEOSS) Evolve Data Management Principles team, co-chair of the Research Data Alliance (RDA) Interest Group on Repository Platforms for Research Data, and co-chair of the RDA Data Versioning Working Group. He serves on the Governance Committee for the Earth Science Information Partners (ESIP) and on the Editorial Board of the CODATA Data Science Journal. He also serves on the Data Archive Interoperability (DAI) working group of the Consultative Committee for Space Data Systems (CCSDS), which is currently reviewing and revising ISO 14721:2012, the standard for the Open Archival Information System (OAIS) Reference Model, and ISO 16363, the standard for Audit and Certification of Trustworthy Digital Repositories. He also is a Senior Member of the Association for Computing Machinery (ACM) and a member of the American Geophysical Union (AGU) and the International Association for Social Science Information Services and Technology (IASSIST).

Dr. John Westbrook

John Westbrook is the lead data and software architect at the RCSB Protein Data Bank (www.rcsb.org). He is active in data standards activities in the field of structural biology including: the International Union of Crystallography (IUCr) Commission on the Maintenance of the CIF Data Standard (COMCIFS), the IUCr Commission on Data (COMMDAT), and the American Crystallographic Association (ACA) SIG on Best Practices for Data Analysis and Archiving (chair).

Mr. Jared Lyle

Jared Lyle is an Archivist at the Inter-university Consortium for Political and Social Research (ICPSR), where he directs the Metadata and Preservation Unit, which is responsible for Metadata, the Bibliography of Data-Related Literature, and Digital Preservation. He also serves as Director of the Data Documentation Initiative (DDI), an international metadata standard for describing survey and other social science data.

Participating Repositories

PDB

Protein Data Bank (PDB) was established as the 1st open access digital data resource in all of biology and medicine. It is today a leading global resource for experimental data central to scientific discovery. Through an internet information portal and downloadable data archive, the PDB provides access to 3D structure data for large biological molecules (proteins, DNA, and RNA). These are the molecules of life, found in all organisms on the planet. Knowing the 3D structure of a biological macromolecule is essential for understanding its role in human and animal health and disease, its function in plants and food and energy production, and its importance to other topics related to global prosperity and sustainability. RCSB PDB operates the US data center for the global PDB archive, and makes PDB data available at no charge to all data consumers without limitations on usage.

Representative: Mr. John Westbrook
John Westbrook is the lead data and software architect at the RCSB Protein Data Bank (www.rcsb.org). He is active in data standards activities in the field of structural biology including: the International Union of Crystallography (IUCr) Commission on the Maintenance of the CIF Data Standard (COMCIFS), the IUCr Commission on Data (COMMDAT), and the American Crystallographic Association (ACA) SIG on Best Practices for Data Analysis and Archiving (chair).

ICPSR

ICPSR advances and expands social and behavioral research, acting as a global leader in data stewardship and providing rich data resources and responsive educational opportunities for present and future generations. As an international consortium of more than 750 academic institutions and research organizations, **Inter-university Consortium for Political and Social Research** (ICPSR) provides leadership and training in data access, curation, and methods of analysis for the social science research community. ICPSR maintains a data archive of more than 250,000 files of research in the social and behavioral sciences. It hosts 21 specialized collections of data in education, aging, criminal justice, substance abuse, terrorism, and other fields. ICPSR collaborates with a number of funders, including U.S. statistical agencies and foundations, to create thematic data collections and data stewardship and research projects. ICPSR's educational activities include the Summer Program in Quantitative Methods of Social Research, a comprehensive curriculum of intensive courses in research design, statistics, data analysis, and social methodology. ICPSR also leads several initiatives that encourage use of data in teaching, particularly in undergraduate instruction. ICPSR-sponsored research focuses on the emerging challenges of digital curation and data science. ICPSR leads or takes part in many policy initiatives and grant-funded activities that result in publications that address issues related to data stewardship. ICPSR researchers also examine substantive issues related to our collections, with an emphasis on historical demography and the environment. ICPSR is a unit within the Institute for Social Research at the University of Michigan and maintains its office in Ann Arbor.

Representative: Mr. Jared Lyle
Jared Lyle is an Archivist at the Inter-university Consortium for Political and Social Research (ICPSR), where he directs the Metadata and Preservation Unit, which is responsible for Metadata, the Bibliography of Data-Related Literature, and Digital Preservation. He also serves as Director of the Data Documentation Initiative (DDI), an international metadata standard for describing survey and other social science data.

LONI Image Data Archive

The Image & Data Archive (IDA) provides tools and resources for de-identifying, integrating, searching, visualizing and sharing a diverse range of neuroscience data, helping facilitate collaborations between scientists worldwide. We are committed to the ideal of fostering open scientific inquiry within a context of reliable data stewardship. The IDA contains data collected for more than 80 studies focused on processes such as development, aging and the progression of specific diseases. Many studies have generous data sharing policies and support online access requests. The Featured Studies section above provides more details.

Representative: Dr. Arthur Toga

TCIA

The Cancer Imaging Archive (TCIA) is a service which de-identifies and hosts a large archive of medical images of cancer accessible for public download. TCIA Increases public availability of high quality cancer imaging datasets for research, supports NIH data sharing requirements for the cancer imaging community, enhances reproducibility in research, and creates a culture of open data sharing and collaboration among cancer imaging researchers. The data are organized as "Collections", typically patients related by a common disease (e.g. lung cancer), image modality (MRI, CT, etc) or research focus. DICOM is the primary file format used by TCIA for image storage. Supporting data related to the images such as patient outcomes, treatment details, genomics, pathology, and expert analyses are also provided when available.

Representative: Dr. John Freymann
Dr. John Freymann directs the Cancer Imaging Informatics Lab as part of the Frederick National Lab for Cancer Research support to the NCI Cancer Imaging Program. He leads efforts to enhance reproducibility in research and to create a culture of open data sharing and collaboration among cancer imaging researchers through The Cancer Imaging Archive and through support to NIH program activities such as the Quantitative Imaging Network (QIN), Informatics Technology in Cancer Research (ITCR), and CPTAC. He leads the imaging collection and analysis efforts for the Moonshot APOLLO project, and directs the accrual and hosting of imaging data generated from NCI supported clinical trials.

NIF/dkNet

The NIDDK Information Network (dkNET) serves the needs of basic and clinical investigators by providing seamless access to large pools of data and research resources relevant to the mission of

The National Institute of Diabetes Digestive and Kidney Diseases (NIDDK). dkNET is hosted at University of California San Diego, and is supported by NIH NIDDK grant 2U24DK097771-06.

Representative: Dr. Maryann Martone

Dr. Maryann Martone received her BA from Wellesley College in Biological Psychology and Ancient Greek and her Ph. D. in Neuroscience from the University of California, San Diego. She is a professor Emerita at UCSD, but still maintains an active laboratory and currently serves as the Chair of the University of California Academic Senate Committee on Academic Computing and Communications. She started her career as a neuroanatomist, specializing in light and electron microscopy, but her main research for the past 15 years focused on informatics for neuroscience, i.e., neuroinformatics. She led the Neuroscience Information Framework (NIF), a national project to establish a uniform resource description framework for neuroscience, and the NIDDK Information Network (dknet), a portal for connecting researchers in digestive, kidney and metabolic disease to data, tools, and materials. She just completed 5 years as Editor-in-Chief of Brain and Behavior, an open access journal, and has just launched a new journal as Editor in Chief, NeuroCommons, with BMC. Dr. Martone is past President of FORCE11, an organization dedicated to advancing scholarly communication and e-scholarship. She completed two years as the chair of the Council on Training, Science and Infrastructure for the International Neuroinformatics Coordinating Facility and is now the chair of the Governing Board. Since retiring, she served as the Director of Biological Sciences for Hypothesis, a technology non-profit developing an open annotation layer for the web (2015-2018) and founded SciCrunch, a technology start up based on technologies developed by NIF and dkNET.

ImmPort

ImmPort, the Immunology Database and Analysis Portal, provides advanced information technology support in the archiving and exchange of scientific data for the diverse community of life science researchers support by the Division of Allergy, Immunology and Transplantation. ImmPort serves as a long-term, sustainable archive of research and clinical data. The core component of ImmPort is an extensive data warehouse containing experimental data and metadata describing the purpose of studies, methods of data generation and result files.

Representative: Ms. Elizabeth Thomson

Ms. Elizabeth Thomson is the Project Manager for the ImmPort Contract awarded by the National Institutes of Health, National Institute of Immunology and Infectious Diseases, Division of Allergy, Immunology and Transplantation to Northrop Grumman and have held this position since June 2017. Northrop Grumman has maintained the ImmPort contract, previously known as Bioinformatics Integration Support Contract (BISC) since it'd inception in 2004. The ImmPort Contract employs 12 staff members comprised of bioinformaticians, applications programmers, database architect/administrator, network administrator, test engineer as well as outreach personnel. The ImmPort team develops and maintains the ImmPort database, associated applications and analysis tools in support of the NIH mission to shared publicly funded data with the scientific community. ImmPort has shared 309 studies to date and has over 100 studies in the data sharing pipeline. n My career has been singularly focused on improving human health through research positions in industry targeting autoimmune disease as well as government contracting positions in clinical/research database development. Leadership positions held include

management of a DNA synthesis core, coordinator of multiple NIH-funded working groups as well as management of the development of a suite of open source analysis tools.

PhysioNet

PhysioNet offers free web access to large collections of recorded physiologic signals (PhysioBank) and related open-source software (PhysioToolkit).
PhysioNetWorks workspaces are available to members of the PhysioNet community for works in progress that will be made publicly available in PhysioBank and PhysioToolkit when complete.

Representative: Dr. Tom Pollard

ZEBrA

The Zebra finch Expression Brain Atlas (a.k.a. ZEBrA; www.zebrafinchatlas.org) is a publically accessed *in situ* hybridization database that documents the constitutive brain-wide expression of >650 genes in the zebra finch (*Taeniopygia guttata*), a vocal learning songbird species. The database, hosted by Dreamhost, consists of >3,200 high resolution digital images (0.42um/pixel; ~3TB data) that can be browsed down to a cellular resolution. Images are also presented in alignment with an annotated histological brain atlas. ZEBrA is built on an extensive relational MySQL database that links gene expression patterns to pertinent information about gene function (based on the NCBI:Gene database), human diseases and communication disorders (based on the Online Mendelian Inheritance in Man - OMIM database), mouse phenotypes (based on the Mouse Genome Informatics – MGI database), and brain expression patterns in mouse (based on the Allen Mouse Brain Atlas – MBA). Still expanding, ZEBrA currently contains brain expression data for ~650 genes in adult male zebra finches, and includes genes covering a range of gene families and pathways of relevance for brain function, development, and behavior. All data were generated by a high-throughput non-radioactive protocol optimized for low background and high sensitivity (Carleton et al., 2014). The database does not currently hold ISO, CoreTrustSeal, or other trustworthiness certifications.

Representative: Peter V. Lovell
Dr. Peter V. Lovell is a Sr. Research Associate at OHSU with extensive experience studying the neurobiology of behavior in a wide-array organisms, including crayfish, rodents, and songbirds. As a grad student at Cornell University he trained under Dr. Carl Hopkins, studying the evolution of electric communication in fish, and later completed my dissertation with Dr. David McCobb, studying the long-term effects of stress on the electrophysiology of large conductance voltage- and calcium-dependent potassium channels. He joined the Mello lab as a post-doctoral research fellow to acquire techniques in molecular genetics and genomics, including PCR, cloning, gene expression analysis, and computational neurogenomics. As one of the founding members of the Songbird Neurogenomics Consortium (SoNG) he have been full committed to the development of modern neurogenomics resources for songbirds, and have participated in the implementation and application of many new technologies, including laser-capture microdissection, cDNA microarray analysis, as well as tools for studying novel genes and gene losses in avian genomes. He have extensive experience with both single and two probe in situ hybridization (ISH) and tract-tracing, and was singularly responsible for the developing the non-radioactive ISH method currently used

by the Mello lab. He is the chief architect (and programmer) of ZEBrA (the Zebra finch Expression Brain Atlas; www.zebrafinchatlas.org), an online public resource that makes available a large collection of in situ images, and includes online tools that allow users to relate the expression of genes to the organization of the songbird brain. This expertise aligns well with the primary goal of the present proposal, to complete an in-depth molecular analysis of >500 genes available on the ZEBrA website. In recent years, his efforts have been focused on understanding how comparative features of genomes (i.e. gains/loss in genes, regulatory elements) give rise to the anatomical, physiological, and molecular properties of brain circuits evolved for the acquisition and production of learned song in hummingbirds, zebra finches, and budgerigars. As the lead PI on a successful F32, he have demonstrated his ability to successfully develop and administer research projects that have resulted in peer-reviewed publications, and have showed that he is capable of managing timelines, budgets, and personnel.

TalkBank: FluencyBank, AphasiaBank, CHILDES, PhonBank and HomeBank

The **TalkBank** system (http://talkbank.org) is the world's largest open-access integrated repository for spoken language data. It provides language corpora and resources to support researchers in psychology, linguistics, education, computer science, and the speech sciences. The National Institutes of Health (NIH) and the National Science Foundation (NSF) have provided support for the construction of five of the components of TalkBank:

1. AphasiaBank, at https://aphasia.talkbank.org, for the study of language in aphasia in six languages;
2. CHILDES, at https://childes.talkbank.org, for the study of child language development in 42 languages, from infancy to age 6;
3. FluencyBank, at https://fluency.talkbank.org, for the study of fluency and disfluency in stuttering, aphasia, second language learning, and normal processing;
4. HomeBank, at https://homebank.talkbank.org, for the application of automatic speech recognition technology to untranscribed daylong recordings in the home and elsewhere; and
5. PhonBank, at https://phonbank.talkbank.org, for the analysis of children's phonological development in 18 languages.

The data in each of these banks involve multiple corpora that were contributed by individual researchers.

Representative: Dr. Brian MacWhinney

Brian MacWhinney is Teresa Heinz Professor of Psychology, Computational Linguistics, and Modern Languages at Carnegie Mellon University. He received his Ph.D. in psycholinguistics in 1974 from the University of California at Berkeley. With Elizabeth Bates, he developed a model of first and second language processing and acquisition based on competition between item-based patterns. In 1984, he and Catherine Snow co-founded the CHILDES (Child Language Data Exchange System) Project for the computational study of child language transcript data. This system has extended to 13 additional research areas such as aphasiology, second language learning, TBI, Conversation Analysis, developmental disfluency and others in the shape of the TalkBank Project. MacWhinney's recent work includes studies of online learning of second language vocabulary and grammar, situationally embedded second language learning, neural

network modeling of lexical development, fMRI studies of children with focal brain lesions, and ERP studies of between-language competition. He is also exploring the role of grammatical constructions in the marking of perspective shifting, the determination of linguistic forms across contrasting time frames, and the construction of mental models in scientific reasoning. Recent edited books include *The Handbook of Language Emergence* (Wiley) and *Competing Motivations in Grammar and Usage* (Oxford).

FITBIR

The Federal Interagency Traumatic Brain Injury Research (FITBIR) informatics system, an instantiation of the BRICS platform, was developed to share data across the entire TBI research field and to facilitate collaboration between laboratories, as well as interconnectivity with other informatics platforms. Sharing data, methodologies, and associated tools, rather than summaries or interpretations of this information, can accelerate research progress by allowing re-analysis of data, as well as re-aggregation, integration, and rigorous comparison with other data, tools, and methods. This community-wide sharing requires common data definitions and standards, as well as comprehensive and coherent informatics approaches.

Representative: Dr. Matthew McAuliffe

Dr. Matthew McAuliffe earned his Bachelor of Science in Electrical Engineering degree from the University of Detroit and PhD in Biomedical Engineering from the University of North Carolina, Chapel Hill, NC. He has been at NIH since 1998 and is currently the Chief of the Biomedical Image Research and Services Section in the Division of Computational Bioscience. He is the original author the Medical Image Processing Analysis and Visualization (MIPAV) application and actively manages the continued development of the application. Dr. McAuliffe also leads the development of the Biomedical Research Informatics Computing System, or BRICS, that supports NIH's strategic goals and enables research data to be Findable, Accessible, Interoperable, and Reusable (FAIR). His current research interests include, biomedical informatics, biomedical imaging analysis and visualization.

WormBase

WormBase (www.wormbase.org) is a free and open-source database and web site dedicated to cataloging and disseminating information about the genes, genome, and biology of the model organism nematode (roundworm) *Caenorhabditis elegans* (*C. elegans*) and related nematode species. *C. elegans* researchers across the globe rely on WormBase on a daily basis as a reference to keep them updated on new information regarding their favorite gene(s) and help them discover information about new genes of interest that emerge from their own experimental studies into health, medicine, and basic biology. Information regarding gene expression (spatio-temporal expression patterns and condition-specific gene expression), gene phenotypes, gene and gene product interactions, human disease models, biological processes, and gene product molecular function are extracted from the literature by a team of expert biocurators, deposited in structured format to our database, and made available through our web site, FTP site, and application programming interface (API) for biologists and bioinformaticians interested in nematode genetics and genomics.

Representative: Dr. Christian A. Grove

Dr. Christian A. Grove currently works as a biocurator for the WormBase database, a repository of genome sequences and gene annotations for the model organism nematode (roundworm), *Caenorhabditis elegans* (*C. elegans*). He curates gene phenotype and interaction annotations from the *C. elegans* research literature, maintains and develops the *C. elegans* phenotype ontology, works on ontology-based data accessibility and display, and plays an active role in data modeling and community curation pipelines for WormBase. Before WormBase, he earned a PhD studying the basic helix-loop-helix (bHLH) family of protein transcription factors in *C. elegans*, determining their protein dimerization patterns, spatio-temporal co-expression, and DNA binding specificities while working in the laboratory of Marian Walhout at UMass Medical School in Worcester, MA. He started working with *C. elegans* while studying the genetics of aging as a laboratory technician with Heidi Tissenbaum at UMass Medical School. He received his bachelor's degree (B.S.) as an interdisciplinary degree in both biology and physics at Worcester Polytechnic Institute (WPI) in Worcester, MA.

UniProt

UniProt (Universal Protein Resource) is a collection of databases that contain comprehensive information on protein sequences. The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. The primary resource is the UniProt KnowledgeBase (UniProtKB), which consists of two sections: UniProtKB/Swiss-Prot containing entries reviewed and annotated from the literature by an expert biocuration team and the unreviewed UniProtKB/TrEMBL with entries annotated by automated systems including rule-based systems. Additionally, UniProt provides the non-redundant UniRef datasets which cluster all sequences at different levels of identity (100, 90 and 50%) as well as sets of Reference Proteomes that provide representative coverage of the tree of life.

UniProt is produced by an international consortium with members from the European Bioinformatics Institute (EMBL-EBI) in the UK, the Swiss Institute of Bioinformatics (SIB) in Switzerland and the Protein Information Resource (PIR) in the USA. The UniProt databases are widely used by scientists around the world and are central to the activities of other resources as a provider of annotation, nomenclature, cross-references as well as sequences. The UniProt website had over 650,000 unique visitors per month in 2018.

Representative: Dr. Peter McGarvey

Dr. Peter McGarvey has a broad background in bioinformatics, software development and molecular biology in both academic and commercial settings. Currently he is actively managing several databases and data repositories as well as initiatives to standardize procedures and formats for genomic and proteomic analysis, annotation, and interoperability. Dr. McGarvey currently serves as: Keystaff for UniProt, coordinating consortium activities in the United States; Co-PI of the CPTAC Data Coordination Center for NCI's Clinical Proteomics and Tumor Analysis Consortium; and, works on several data standards and data integration efforts for ClinGen (Clinical Genome Resource). He has published over 50 peer-reviewed articles,book chapters and patents.

dbSNP

The NCBI **dbSNP** (https://www.ncbi.nlm.nih.gov/snp) was created in 1998 in collaboration with the National Human Genome Research Institute (NHGRI) in response to a need for a general catalog of genome variation to address the large-scale sampling designs required by association studies, gene mapping and evolutionary biology. The primary roles of dbSNP are to process submissions, archive the data, annotate on the genome and NCBI Reference Sequences (RefSeqs), and distribute it worldwide. The database housed variation and frequency data from thousands of submitters including large scale projects including HapMap, 1000Genomes, GO-ESP, ExAC, TOPMED, and HLI to more focused studies such as locus-specific databases (LSDB) and clinical variants in ClinVar. It currently contains more than 1.8 billion Submitted SNP (ss) records and more than 650 million Reference SNP (rs). In addition, more than 580 million of these RS records have frequency data in Build 151. The data is used worldwide and rs accessions are cited in more than 48 thousand publications.

Representative: Dr. Lon Phan

Dr. Lon Phan is the Project Leader for NCBI genetic variation resources that include the Database of Short Genetic Variations, known as the Single Nucleotide Polymorphism Database (dbSNP), and the Database of Genomic Structural Variation (dbVar). He leads a team of scientific curators and developers responsible for the operation and development of the databases.

GEO

The NCBI **Gene Expression Omnibus** (GEO, http://www.ncbi.nlm.nih.gov/geo/) is an international public repository for large-scale, open-access gene expression and epigenomics datasets submitted by the research community. Data are deposited in compliance with journal and grant directives to make research data public. GEO supports archiving of raw data, processed data and metadata which are indexed, cross-linked and searchable. GEO also provides several web-based tools and strategies to assist users to query, analyze and visualize data. The community re-uses GEO data in diverse ways, including finding evidence of novel gene expression patterns, identifying disease predictors, and generally aggregating and analyzing data in ways not anticipated by the original data generators. In this way, GEO represents a foundational resource that helps catalyze basic science, facilitating data-driven discoveries and translation of research results into new knowledge and products that accelerate biological and health sciences.

Representative: Dr. Tanya Barrett

Since 2002, Dr. Tanya Barrett has worked on the NCBI Gene Expression Omnibus (GEO) database project, first as a curator, now as team leader. In that time, GEO has grown to be the largest collection of openaccess gene expression and epigenomics datasets from all branches of life, generated by microarray and next generation sequencing technologies. Her duties include: developing standard operating procedures for curation and data flow; shaping data submission requirement policies; designing new data analysis and visualization tools; writing documentation; integration with related NCBI databases; overseeing tests and troubleshooting GEO functionality; and communicating with the scientific community resolving issues relating to content, data interpretation and site navigation.

DASH

The *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD) established the **Data and Specimen Hub (DASH)** in 2015 as a centralized resource to promote data and biospecimen sharing from completed NICHD funded studies and enable researchers to comply with NIH data sharing policies. By supporting data and biospecimen sharing through DASH, NICHD aims to maximize the value of existing data and biospecimens to accelerate scientific advances and improve human health. DASH was designed with a flexible architecture to accommodate the diverse data types and formats from NICHD's broad scientific portfolio in a manner that promotes FAIR data sharing principles.

Representative: Ms. Rajni Samavedam

Ms. Rajni Samavedam is a Principal/Director with Booz Allen Hamilton's Health and Life Sciences practice. She has worked with HHS and NIH for over 20 years, and specifically has executive oversight for Booz Allen's programs at NIH. As a partner to the federal govern-ment she specializes in research and data sharing programs, and the design, development and implementation of scalable scientific applications. She oversees integrated teams with skillsets ranging from clinical and basic research, portfolio analysis, data sharing policy and technology, software design and development, agile method-ology, cloud implementation, cyber, and researcher communications and management. In addition, she has experience working with other parts of HHS as well as non-governmental organizations in the areas of clinical research, scientific evaluation, and data strategy. Prior to Booz Allen, she carried out research studies both domestically and internationally; worked for state government as the State Health Planner; and stood up a non-governmental organization in India focused on women and children health and the development of health education programs. She holds a master's in public health focused on reproductive epidemiology. And she has a bachelor's in biology and chemistry.

BioLINCC

The NHLBI Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) was established in 2008 to develop the infrastructure needed to link the contents of two related NHLBI Repositories, facilitate access to repository resources, and streamline request processes. The Biologic Specimen Repository, funded under contract, was initially archived biospecimens from population-based blood product safety surveys in 1976 and was later expanded to biospecimens from clinical and epidemiological studies in heart, lung, and blood disorders. The Data Repository was established to store and distribute study data from NHLBI-sponsored research in 2000 and managed by the Epidemiology Branch, Division of Cardiovascular Sciences.

The primary objective of the BioLINCC program is to maximize the scientific value of historical and contemporary NHLBI biospecimen and data collections by facilitating access by qualified researchers to these research resources, and to further enhance utilization by promoting awareness of these resources to the research community. BioLINCC (www.biolincc.nhlbi.nih.gov) makes available online data from over 140 clinical studies and biospecimens from 49 study collections linked at a vial level to the data in the Data Repository.

Representative: Dr. Lucy Hsu
Lucy Hsu serves as a Program Official and co-lead for BioLINCC, one of the Data Repository entities at the National Heart, Lung, and Blood Institute (NHLBI), NIH. Her training is in biostatistics. She joined the NHLBI in 2011. Prior to joining the NHLBI, she had more than 20 years of work experience in the biomedical field, including advanced statistical analyses with various types of data (i.e., disease surveillance, electronic health/medical records, healthcare claims, real world data, etc.), and study designs (i.e., observational, environmental, epidemiological, pharmacoepidemiologic, and experimental studies, clinical trials, complex-design surveys, etc.).

GlyGen

GlyGen is a data integration and dissemination project for carbohydrate and glycoconjugate related data. GlyGen retrieves information from multiple international data sources and integrates and harmonizes this data. This web portal allows exploring this data and performing unique searches that cannot be executed in any of the integrated databases alone.

Representative: Dr. Robel Kahsay
As an Assistant Professor of Biochemistry and Molecular Medicine and while working at DuPont's Central Research and Development as a senior computational scientist, Dr. Robel Kahsay has applied his extensive informatics experience in the management and analysis of biological data to develop various bioinformatics algorithms and robust data analysis pipelines supporting diverse research groups in the field of biological sciences research. His current research focus is on the use of innovative technologies for building analysis engines that feed on high throughput heterogeneous biological datasets to extract actionable knowledge, and making this knowledge available to the research community through advanced interfaces and automated APIs.

OncoMX

OncoMX is a knowledgebase of unified cancer genomics data from integrated mutation, expression, literature, and biomarker databases, accessible through web portal.

Representative: Dr. Jonathon Keeney

eyeGENE

The National Ophthalmic Disease Genotyping and Phenotyping Network (**eyeGENE®**) is a genomic medicine initiative created by the National Eye Institute (NEI), part of the National Institutes of Health (NIH), in partnership with clinics and laboratories across the vision research community. The core mission of eyeGENE® is to facilitate research into the causes and mechanisms of rare inherited eye diseases and accelerate pathways to treatments.

Representative: Dr. Kerry Goetz
Kerry Goetz, MS is a Program Manager and Data Specialist at the National Eye Institute. There she manages an international clinical study of rare, inherited eye conditions called eyeGENE. Her day to day activities at the NEI involve, not only the eyeGENE data, but several other collaborative

studies for which data is being made available for secondary research. Kerry is also actively facilitating advancing the use of common data elements at the NEI.

The National Ophthalmic Disease Genotyping and Phenotyping Network (eyeGENE®) is a genomic medicine initiative created by the National Eye Institute (NEI), part of the National Institutes of Health (NIH), in partnership with clinics and laboratories across the vision research community. The core mission of eyeGENE® is to facilitate research into the causes and mechanisms of rare inherited eye diseases and accelerate pathways to treatments.

ICE

The Integrated Chemical Environment (ICE) provides high-quality, curated data from NICEATM and its partners as well as other data resources and tools to support development of new approaches for assessing chemical safety. ICE allows users to easily query and integrate data streams that can then be explored interactively via table or graphic formats.

Representative: Dr. Shannon Bell