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. This workshop will introduce the concepts of ***trust*** principles for data repositories, which address management aspects of data repositories. The workshop will introduce the CoreTrustSeal Core Trustworthy Data Repositories Requirements (an emerging, international, community-based standard) 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

April 8, 2019

Session 1- TRUST Concepts and Standards

Chair: Dawei Lin (NIH/NIAID)

09:00-09:10 Opening Remarks, *Dawei Lin (NIAID) & Susan Gregurick (NIH/OD)*

09:10-09:30 Introduction of the audience, *Dawei Lin (NIH/NIAID)*

09:30-10:00 Perspectives on Data Preservation, *Jon Crabtree (UNC)*

10:00-10:30 International Standards for Trustworthy Data Repositories, *Bob 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

13:00-14:00 **Lunch NIAID Cafeteria**

Session 2 - TRUST Examples

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

Facilitators: Jon Crabtree, Bob 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

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 on Technology, Jared Lyle on Social Sciences*

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).

Dr. 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).

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).

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

UniProt is funded internationally, with its three main funders being the National Institutes of Health (NIH), the European Molecular Biology Laboratory (EMBL) and the Swiss Secretariat for Education Research and Innovation (SERI). Additional information on the history, content and structure of the UniProt datasets can be found on the UniProt homepage and associated documentation (www.uniprot.org).

dbSNP

dbSNP contains human single nucleotide variations, microsatellites, and small-scale insertions and deletions along with publication, population frequency, molecular consequence, and genomic and RefSeq mapping information for both common variations and clinical mutations.

GEO

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

DASH

NICHD **DASH** is a centralized resource for researchers to store and access data from NICHD-funded research studies to use for secondary research.

BioLINCC

The mission of **BioLINCC** is to facilitate access and maximize the scientific value of the Biorepository and Data Repositories, and to promote the availability and use of other NHLBI-funded population-based biospecimen and data resources.