Opening Statement for the First NIH Trustworthy Data Repository

Dawei Lin, Ph.D.
Division of Allergy, Immunology, and Transplantation
National Institute of Allergy and Infectious Diseases
National Institutes of Health
April 8, 2019

My name is Dawei Lin. On behalf of the organizing committee and logistic staff, it is my greatest pleasure to welcome you to participate in the first NIH Trustworthy Data Repository Workshop in person or online.

May I ask the people who have helped organize the workshop to stand up from both NIH and CoreTrustSeal. And give them a big round of applause.

It is amazing how hard they have been working to put this important workshop together. The NIH committee has a weekly call for the past a few months.

I also want to welcome you to NIAID, the National Institute of Allergy and Infectious Diseases, which is one of 27 institutes and centers of NIH (ICs). Sometimes we call them ICs. Today you will see colleagues from many of the ICs.

The Webinar is open to the public. Please feel free to tweet using the hashtag #NIHTDR.

My role is to explain the purposes of this workshop and what we are going to work on for the next two days.

To start off, why do we talk about “Trustworthy Data Repository”?

Data repositories are the crown jewels of the NIH data ecosystem. I was in awe of the diversity and the depth of knowledge that reflect in both the Bios of more than 20 repositories and those of the people who represent them. The question becomes that - Why did not we have this workshop earlier?

To the repository representatives as well as the NIH Program staff to work them over the years, thank you again for taking a trip to come here and contributing your expertise.

For these 20 repositories and other ones who are not here today, they not only preserve invaluable datasets from life-saving research for the generations to come but also provide the critical services and technical infrastructure to accommodate rapidly changing research needs from the biomedical user community. They are the hubs where communities form. They are the resources that empower people to make discoveries, generate new hypotheses, and develop treatments and cures. They help the community to maximize the value of data beyond their original purposes to generate them.

While we all understand the data repository is a fundamental infrastructure. The priority of building and sustaining such a critical resource is not always aligned with the urgency of the need to have robust and reliable operations. We’ve often lost valuable data without even knowing that we are losing them. I know many repository representatives will agree with me that the support for data repositories does not always commensurate with the level of impact they have upon the community. For the Program Officers, it is often not easy to find an appropriate funding mechanism to support data repositories. These are the challenges we are facing.

With all these challenges, how can we trust that our investments are protected? how can we trust that we are well-equipped in the digital age to accomplish our missions? What we need is not just data repositories but trustworthy data repositories. Trustworthy is the element that will give researchers a peace of mind to deposit their data, share their data and provide many uses for the data.

The first step to having a trustworthy data repository is to define what they are and how to assess them based on the definitions. This workshop is about this first step. To define trustworthiness in the context of biomedical sciences and more importantly to define requirements on how to assess the trustworthiness in a standardized manner. From this workshop, we hope to generate actionable recommendations in a report that we will share with the biomedical community for further comments.

Before we talk about trustworthiness further, it is important to distinguish a closely related and well-known concept, which is FAIR principles. Phil Bourne, who was the former Associate Director of Data Science of NIH, brought the concepts to the NIH community. We are fortunate to have a couple of the co-authors of the FAIR Principles in the room, Maryann Martone and Ingrid Dillo. So when the questions arise, we have experts to answer them.

FAIR has revolutionized the way of how we think about research data. The catchy acronym of FAIR, Findable, Accessible, Interoperable and Reusable, make it easy to understand what properties of data people need. The power of “making it understandable” is that it can make it easy to motivate people to act. However, we think that FAIR is about the properties of datasets. This workshop is not to talk about having these properties. Instead, the focus of the workshop is on how to get these properties into data. It is to focus on repositories, who enable data FAIR. We think that a proper word to describe the characteristics of repositories is TRUST. TRUST is the focus of the workshop.

What does TRUST mean:

TRUST is about having transparent policies, organizational capabilities. It is about transparency.

TRUST represents a commitment to transparently fulfill the services promised to support the continuing use of data. It is about responsibility.

TRUST is about the people who are behind the websites, infrastructure, and databases, who understand deeply what FAIR means to the users of their designated community. It is about the user and the community.

TRUST is about sustaining infrastructures that are needed to support sustainable operations and long-term data and knowledge preservation. It is about sustainability.

TRUST is about providing services for archiving and distributing data. It is about maintaining reliable and secure operations through technology and data stewardship procedures. It is about technology.

TRUST is our catchy name for trustworthiness. The TRUST spell as Transparency, Responsibility, User Community, Sustainability and Technology. Last week, the White Paper on TRUST Principles is released at the RDA in Philadelphia.

Ingrid Dillo will talk about it a little bit more in detail later and point you to the White Paper.

The good news about Trustworthy Data Repositories is that there exists a couple of standards for defining and assessing trustworthiness already, so the Biomedical community does not have to start from scratch.

The goal of this workshop is to rely on your expert opinions to answer three key questions:

1. Do the existing trustworthiness standards cover all the essential aspects needed for Biomedical Data Repositories?

2. Are there aspects that the existing trustworthiness standards cover that seem irrelevant to Biomedical Data Repositories?

3. Are there essential aspects of trustworthiness needed for Biomedical Data Repositories that are not touched on by existing trustworthiness standards?

These three questions: the standards good to use, any irrelevant ones, anything missing, have been summarized by Jennie Larkin will be our guide throughout the workshop. You should have a copy of it in your handout.

At this workshop, we will use CoreTrustSeal (CTS) as the standard of choice. The CTS has been used by more than 140 certifications and including using CTS and its former versions are known as Data Seal of Approval (DSA) and World Data Systems (WDS). The reason for its wide adoption is simplicity and lightweight. It has 16 requirements instead of 30+ to 100+ requirements. So it is a standard that we can realistically evaluate in two days.

CTS is a standard that has been adopted by multiple disciplines including social sciences and earth sciences. One feature I like the most is that once a repository is certified, all the certification materials are online, which can serve as the templates for other repositories. The openness and transparency are important to form a trusted community and help each other to improve trustworthiness. These are the main justification we use CTS as our first experiment to assess trustworthiness.

The game plan for the workshop. In the morning and early afternoon, we will listen and learn the definitions of trust requirements and examples of building trustworthiness and get peer-reviewed. In the late afternoon, we will dive into the requirements and answer the three questions. We want to have a deep discussion so we divide into four groups to talk a couple of requirements by each group. We will explain the details of the breakout sessions off-line since the online people will not participate in the activities. The discussion output of the breakout sessions will be the basis for the discussion of recommendations tomorrow morning on defining trustworthiness and how to assess them. After that, we will have a session on how to form a trusted community, which is necessary to have a long lasting trust.

That is our workshop. I want to pause to answer a couple of clarifying questions.