NIH Workshop on Trustworthy Data Repositories (TDR) for Biomedical Science Content

Executive Summary

Purpose of the Workshop

The NIH STRATEGIC PLAN FOR DATA SCIENCE outlines five overarching modernization goals: (1) Data Infrastructure, (2) Data Ecosystem, (3) Analytics and Tools, (4) Workforce Development and (5) Data Stewardship and Sustainability. This workshop, *NIH Workshop on Trustworthy Data Repositories (TDR) for Biomedical Science,* supports several of these goals and was organized by a working group aligned with goal #2, Data Ecosystem. An essential feature of an effective data repository is that it is *trusted* as a reliable data resource by its user community. In July 2018 the White House National Science and Technology Council's Big Data Interagency Working Group issued a summary of a workshop titled "MEASURING THE IMPACT OF DIGITAL REPOSITORIES." One of the five key improvements recommended to enhance the impact of digital repositories is related to trustworthiness: "Data repository certification that is understandable and usable across a broad range of repositories."

The current workshop focused on the concept of "trustworthiness" for NIH data resources, including what trustworthiness means to key stakeholders (NIH, data repository managers, and the research community), and how some existing repository certification standards address the needs of these key stakeholders in assessing trustworthiness. The workshop held April 8-9, 2019, on the NIH campus was sponsored by the NIH Office of Data Science Strategy (ODSS). Participants included 20 data repository representatives, 29 NIH program officers and staff, and five international experts in data repository trustworthiness certification. An additional 70 people attended the presentations via webcast.

Workshop Content Overview

Session 1 TRUST Concepts and Standards: This session introduced the five characteristics of TRUST (**T**ransparency, **R**esponsibility, **U**ser Community, **S**ustainability and **T**echnology) and provided overviews on data preservation, the development of the Reference Model for Open Archival Information System (OAIS), and information on international activities associated with repository certification. Presentations by Jonathan Crabtree, Robert Downs, and Ingrid Dillo provided an overview on data preservation perspectives, the International Organization for Standardization standard ISO 16363, and the CoreTrustSeal (CTS), respectively. CoreTrustSeal (CTS) resulted from an international collaboration fostered by the Research Data Alliance to combine two earlier repository certification standards by the World Data Systems and the Data Seal of Approval. The ISO 16363 has 107 requirements and requires a larger financial and time investment for certification and an audit. The workshop breakout groups used CTS as an exemplar. The CTS has 16 requirements and requires a smaller financial and time investment for certification. CTS requirements are grouped into three themes: organizational, digital object

management, and technology. CTS documents including the principles and worksheets are freely assessible. Several biomedical repositories have already received CTS certification. In the first breakout, eight groups examined different CTS requirements in terms of the relevance, feasibility, and utility for biomedical data repositories then reconvened to share findings. Most requirements were deemed relevant and helpful for managing biomedical data repositories.

Session 2 - TRUST Examples: This session explored the experiences of two different repositories seeking trustworthiness certifications. Jared Lyle from the **I**nter-university **C**onsortium for **P**olitical and **S**ocial **R**esearch (ICPSR) described their experiences since 2005 with several certification sources, and John Westbrook from the RCSB **P**rotein **D**ata **B**ank (PDB) described their recent experience applying for CTS certification. Both speakers value their certifications and shared that the process of thinking through the requirements was beneficial.

Session 3 – **TRUST Challenges and Opportunities:** Breakout groups focused on the distinct needs of biomedical data repositories for trustworthiness assessments. Each group focused on identifying fitness (for biomedical repositories) versus gaps in the CTS requirements for one of three themes: organization, digital object management, and technical infrastructure. Participants felt that the existing requirements are appropriate for biomedical repositories and identified requirement gaps specific to human data and data life cycle, e.g., should data generated using early less-accurate methods ever be removed or de-referenced.

Session 4 – TRUST Community: On the second day, the workshop focused on the TRUST principles and their alignment and relevance to biomedical repositories. This session stimulated active discussion and creative thinking on how to align the biomedical repository community with the five characteristics of TRUST. Themes arising from this discussion included how TRUST applies to repositories holding distinct types of data (such as sequences, images, or phenotypes), the relevance of trust certification to data types, how to build community engagement for TRUST principles, and the ease/burden of the certification process.

Highlights

The two-day workshop stimulated discussions on challenges, barriers, and opportunities. Workshop attendees were generally in agreement that TRUST principles should be used by data repository managers, and that the principles of certification, if not actually obtaining certification, will benefit repository management practices. Four main themes emerged:

- **Repository definition:** The workshop identified the need to clarify the distinction between repositories, databases, knowledgebase, databanks, data centers, and archives and questioned if TRUST principles should be applied differently based on the distinction.
- **Value proposition:** Participants indicated that it would be good to have NIH consider when trustworthiness certification would bring value and ensure that it was encouraged or required in those situations. In addition, participants indicated it would be useful to understand the priority of pursuing TRUST principles, certification, how that may be included in program descriptions and review criteria, and how it may impact evaluation, which typically is based on feature delivery.

- **Self-assessment vs. certification:** Many participants indicated interest in doing self-assessment activities, as they felt it to be a useful and informative process that may help improve repository management practices. They were less certain of the benefits of formal certification outweighing the associated costs (in terms of both staff time and money).
- **Trustworthiness vs. sustainability**: Participants perceived tension between trustworthiness and longevity/sustainability in that databases/repositories should strive to be trustworthy (and possibly get certification) while at the same time are not guaranteed long-term funding. Participants wished to bring this conflict to the attention of NIH.

Outcomes from the workshop

- Five repositories that attended the workshop indicated they plan to go through the CTS certification process in the next year as a result of their participation. Another 13 repositories are considering obtaining CTS certification at a later date.
- Several of the participants who represent data repositories are planning to remain in contact to share information and support each other in obtaining trustworthiness certification.
- Workshop participants proposed to establish an NIH Trustworthy Data Repository Interest Group. A mailing list has been set up to facilitate the formation of such as a group to foster continued engagement and education. The list is open to the pubic but requires registration. It is accessible at https://list.nih.gov/cgi-bin/wa.exe?A0=TDR.
- Participants representing repositories expressed interest in having access to a white paper or other form of guidance on the import and value of TRUST principles and trustworthiness certification.
- Participants representing NIH expressed interest in developing guidance on whether and how TRUST principles may be of value to grant review panels.