Report on NIH Workshop on Trustworthy Data Repositories (TDR) for Biomedical Sciences

This Report highlights challenges identified during the NIH Workshop on Trustworthy Data Repositories held April 8-9, 2019. An Executive Summary is also available. During the breakout sessions, the participants reviewed certification documents for a range of individual CoreTrustSeal certified repositories. The CoreTrustSeal Repository Extended Guidance includes information on the 16 requirements, which are grouped into four areas: 1) Context, 2) Organizational Infrastructure, 3) Digital Object Management, and 4) Technology. A summarized list of the requirements is provided in Appendix 1 of this report. Each group addressed the value and sufficiency of the requirements for biomedical repositories.

One general concern expressed by participants was a lack of a clear definition of "data repository." The terms database, databank, dataset, data repository, data archive, digital archive, and knowledgebase are related terms that also lack clarity. Without universal understanding of these terms, there is confusion on whether a resource is appropriate for certification as a Trustworthy Data Repository.

During breakout sessions, participants addressed the three questions below in relation to the requirements in each of the four CoreTrustSeal themes, i.e., Context, Organizational Infrastructure, Digital Object Management, and Technology.

1. Do the existing trustworthy 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?

Participants acknowledged each of the 16 CoreTrustSeal requirements were important and relevant to biomedical data repositories. One area that was not felt to be highlighted in the CoreTrustSeal requirements, but highly relevant to clinical data repositories, was Human Subjects Protections. However, most felt that regulations and standards to handle the unique requirement for data from human subjects such as privacy, confidentiality, and security were already in place.

Several areas generated more discussion:

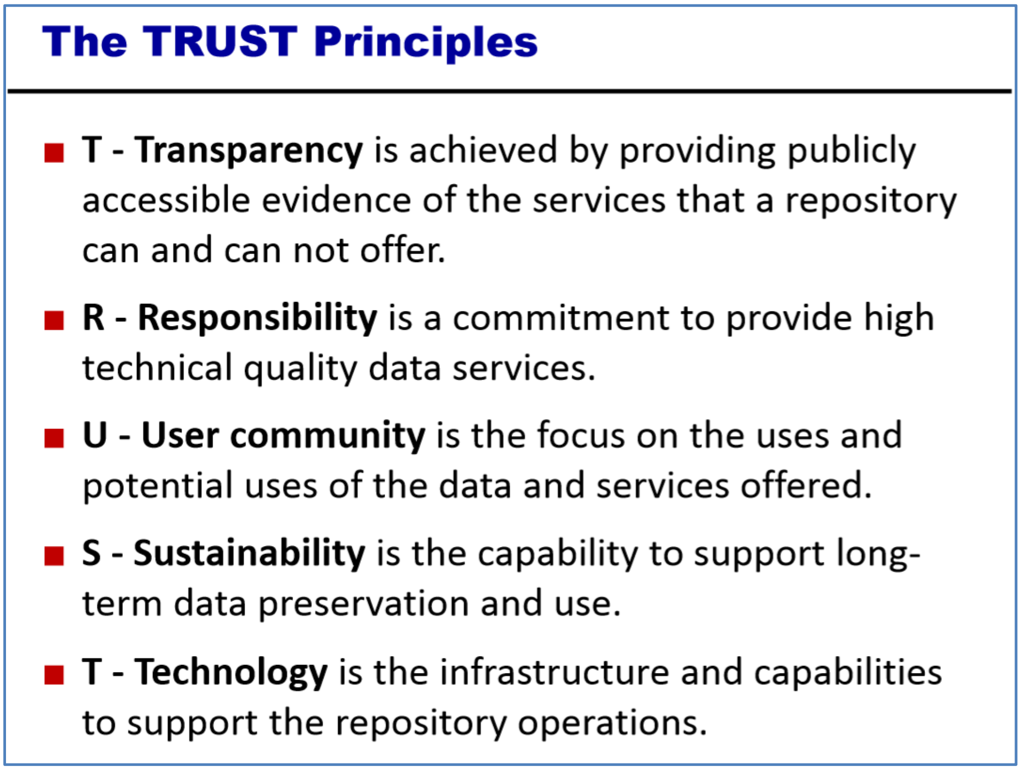
1. Many expressed a desire for more clarity related to the Context requirement. In particular, the user community could be very diverse for some biomedical repositories.
2. Several participants were unclear if licenses were applicable to their repositories.
3. Several expressed concerns about the qualification of reviewers from other repository types to review biomedical repository applications.
4. The Preservation requirement (within the Digital Object Management area) generated discussion since many biomedical repositories are funded for limited time frames with

grant or contract funds. The Preservation plan should also include when data will not be preserved as some biomedical data may lose utility. Policy and regulatory issues may also impact preservation of data.

5. Biomedical data repositories may be part of a data resource that includes specimens and tools. Participants suggested that the CoreTrustSeal requirements only apply to digital data repositories. Certification of biospecimen repositories many need to be addressed separately.
6. Addition of guidance specific for biomedical repositories could be helpful, including making easily accessible examples of how biomedical repositories with CoreTrustSeal certification addressed each of the requirements. Further outreach could also be beneficial to understand the TRUST principles, CoreTrustSeal requirements, and how they apply to biomedical data repositories.

The TRUST Principles

Feedback on the TRUST principles was obtained from the workshop participants using Mentimeter, an interactive audience engagement software tool.



The TRUST Principles

- **T - Transparency** is achieved by providing publicly accessible evidence of the services that a repository can and can not offer.
- **R - Responsibility** is a commitment to provide high technical quality data services.
- **U - User community** is the focus on the uses and potential uses of the data and services offered.
- **S - Sustainability** is the capability to support long-term data preservation and use.
- **T - Technology** is the infrastructure and capabilities to support the repository operations.

Transparency

Components of transparency in trustworthy data repositories include not only organizational transparency, but also data transparency. Areas identified that require transparency in biomedical repositories are data provenance, data curation, and documentation of process and policy.

Responsibility

The responsibility principle was felt to be shared among data producers, institutions, funders, data repositories, publishers, and data users. Data validation, data confirmation, reproducibility and replication were raised as possible components of the responsibility principle. Education, communication, and stewardship by all partners were recognized as components of responsibility. Building a culture of shared responsibility was thought critical to this principle as well.

User community

Engagement, communication, collaboration, and service were identified as critical components of the User Community principle of TRUST for data repositories because engaging and being responsive to user community needs is built on good communication and collaboration and reliable, responsive, and consistent service. User community includes data producers, funders, data users, and scientific community.

Sustainability

The sustainability principle was recognized as critical for TRUST, but challenges were acknowledged. Funders, institutions, data users, data providers, and government were identified as partners in achieving sustainability. The sustainability principle rests on data as a shared resource requiring all to support that resource. Key issues were seen as long-term planning, funding models, business models, and prioritization.

Technology

The technology principle requires commitment to infrastructure that incorporates reliability, flexibility, scalability, transferability, security and agility. Evolving a resource as technology advances requires strong technologic expertise.

Conclusions

Trustworthiness data repository standards

- The existing standards for trustworthy data repositories are applicable and relevant to biomedical repositories.
- Specific biomedical data related needs can be incorporated into current trustworthiness standards (e.g., human subjects protection and reviewer qualifications).

TRUST principles

- **T**ransparency – organizational and data transparency are critical; data provenance, data curation process, and documentation of process and policy
- **R**esponsibility – shared by repository, data producers, data users, publishers, institutions, funders, government; build culture of responsibility
- **U**ser community – all stakeholders engaged with the data and its use; requires engagement, communication, collaboration, and commitment to service
- **S**ustainability – critical, but challenges in achieving; multiple partners and long-term planning required; novel business and funding models likely required
- **T**echnology – infrastructure that incorporates reliability, flexibility, scalability, security and agility; technology expertise required to evolve with new developments in technology, policy, and community needs

Appendix 1: Summary of CoreTrustSeal Requirements

Theme	Key words	Core Trustworthy Data Repositories Requirement
Context	Context	R0. Provide context of the repository: repository type, designated community, Level of curation, outsourcing partner and other.
Organizational Infrastructure	I. Mission/Scope	R1. The repository has an explicit mission to provide access to and preserve data in its domain.
	II. Licenses	R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.
	III. Continuity of access	R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.
	IV. Confidentiality/Ethics	R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.
	V. Organizational infrastructure	R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.
	VI. Expert guidance	R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either inhouse, or external, including scientific guidance, if relevant).
Digital Object Management	VII. Data integrity and authenticity	R7. The repository guarantees the integrity and authenticity of the data.
	VIII. Appraisal	R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.
	IX. Documented storage procedures	R9. The repository applies documented processes and procedures in managing archival storage of the data.
	X. Preservation plan	R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.
	XI. Data quality	R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.
	XII. Workflows	R12. Archiving takes place according to defined workflows from ingest to dissemination.
	XIII. Data discovery and identification	R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

	XIV. Data reuse	R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.
Technology	XV. Technical infrastructure	R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.
	XVI. Security	R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

source: *https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf*