The TRUST Principles for Trustworthy Data Repositories – An Update

Dr. Dawei Lin, Ph.D.

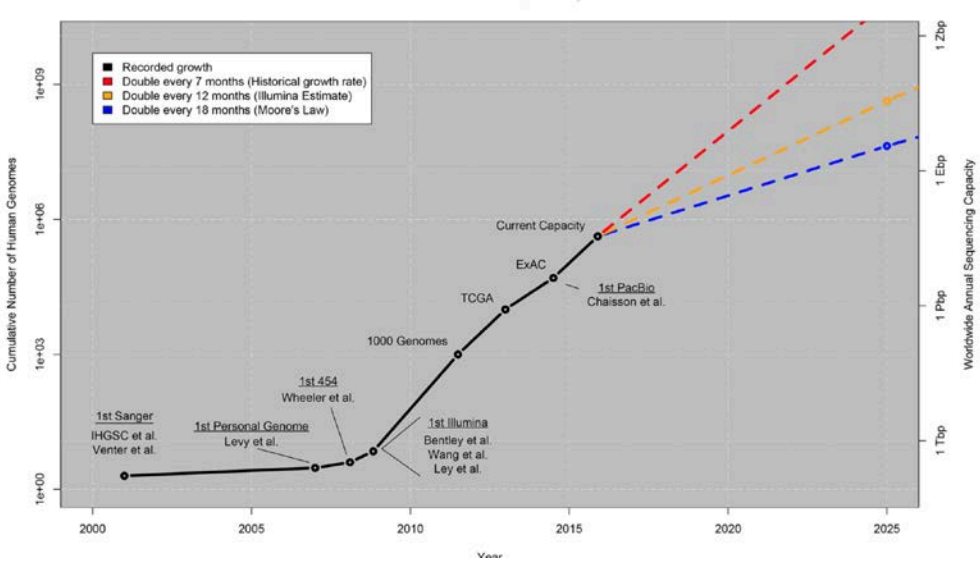
Division of Allergy, Immunology, and Transplantation, NIAID, NIH
dawei.lin@nih.gov

RDA/WDS Repository Certification IG, September 12, 2019

National Institutes of Health
Turning Discovery Into Health

Data is Value & Biological Data is Large

Growth of DNA Sequencing



Storage needs for Big Data in 2025:

Astronomy – 1 ZB/year

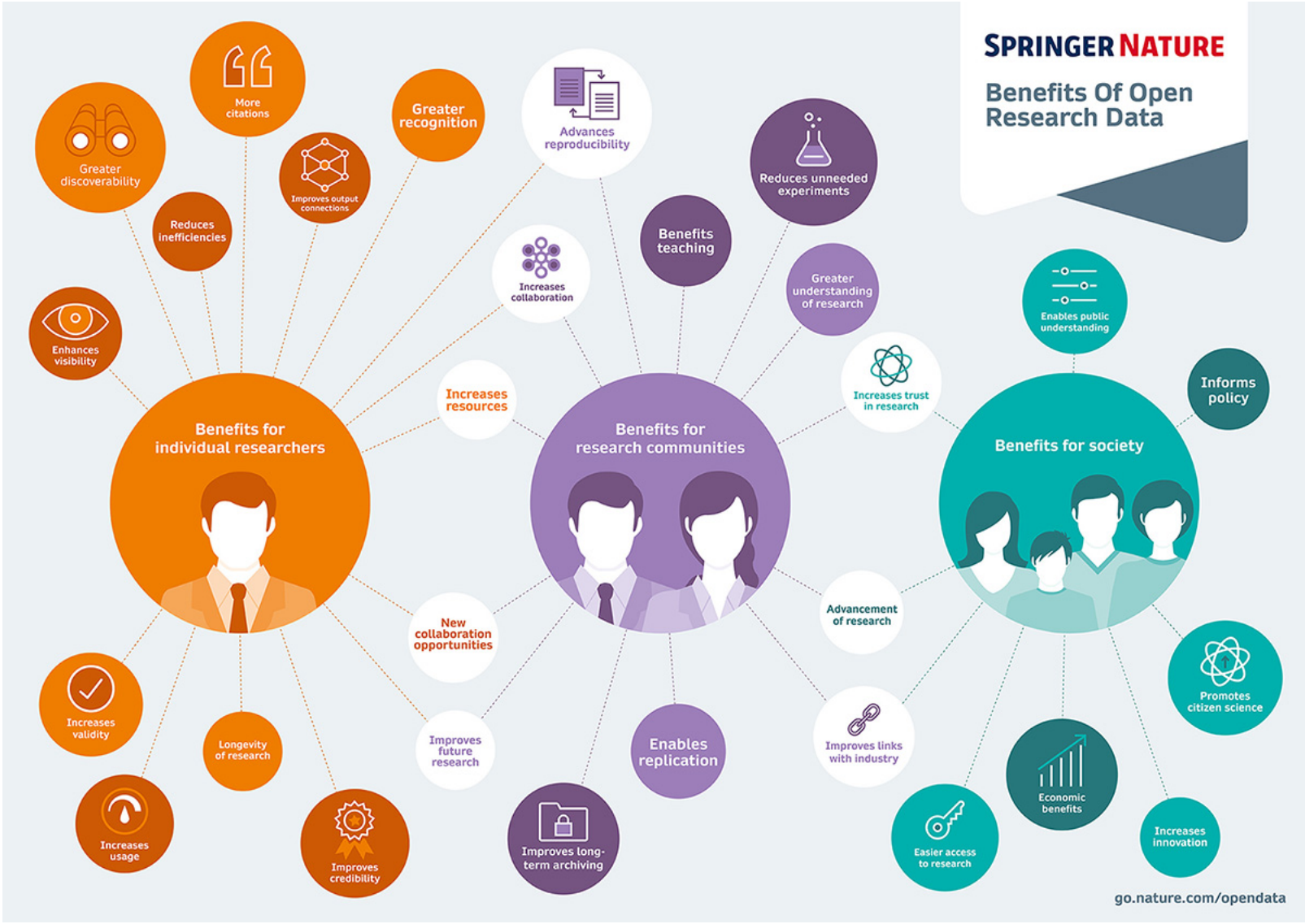
Twitter – 1 to 17 PB/year

YouTube – 1 to 2 ZB/year

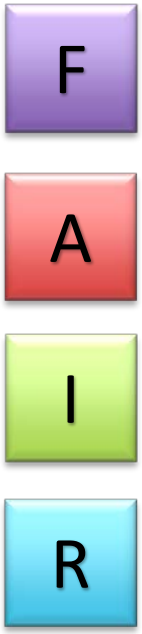
Genomics – 2 to 40 ZB/year

PLOS Biology | DOI:10.1371/journal.pbio.1002195 July 7, 2015

Sharing Makes Data More Valuable



FAIR principles: Trusting *data* management and stewardship



Findable

Accessible

Interoperable

Reusable

Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship [published correction appears in Sci Data. 2019 Mar 19;6(1):6]. Sci Data. 2016;3:160018. Published 2016 Mar 15. doi:10.1038/sdata.2016.18

Measuring the Impact of the Digital Repositories

Interagency Workshop (2017) to identify current assessment metrics, tools, and methodologies that are effective in measuring the impact of digital data repositories

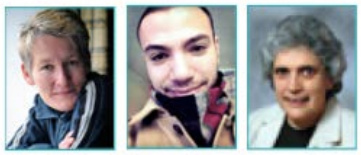
Key Takeaways:

- A group with broad expertise and experience able to formulate and recommend best practices for data sharing and reuse.
- A data citation system that treats data as first-class objects comparable to publications in the research life cycle.
- **Data repository certification that is understandable and usable across a broad range of repositories.**
- New methods to assess economic impacts and opportunity costs when a repository is maintained or eliminated.
- A suite of strategies that repositories can use to achieve financial sustainability.

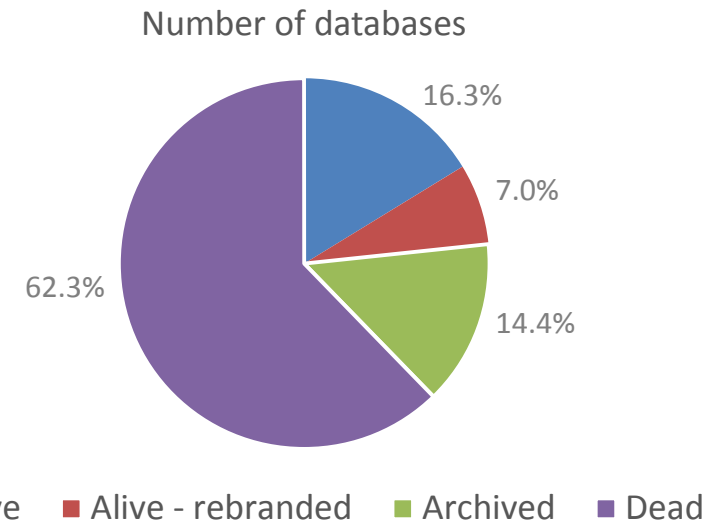
https://www.nitrd.gov/pubs/BD-IWG-Digital-Repository-Workshop-Report-2018.pdf

Longevity of 328 Biomedical Databases over 18 years

Category	N	Percent
Alive	53	16.3%
Alive - rebranded	23	7.0%
Archived	47	14.4%
Dead	203	62.3%
TOTAL	326	100%



Teresa K. Attwood [1] ✉, Bora Agit [1], Lynda B.M. Ellis [2]

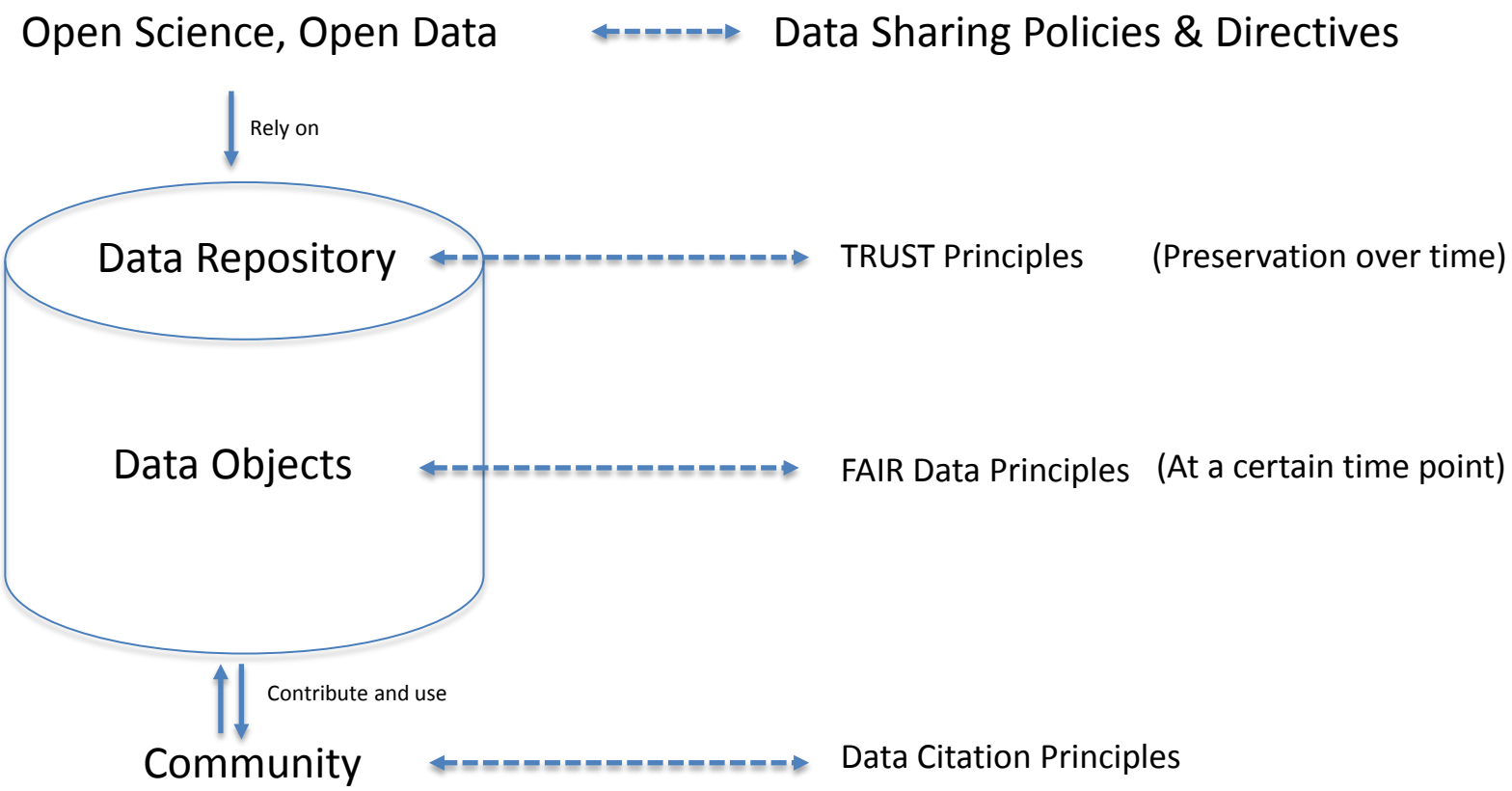


http://journal.embnet.org/index.php/embnetjournal/article/view/803/1209 (Attwood et al, 2015)

Motivation

- Develop concise and measurable approaches to achieve Trustworthiness of Digital Repositories (TDR)
- Form an open forum to define TRUST with continued community feedback and consensus
- **Not** to replace the existing standards and best practices
- Provide a conceptual starting point for thinking about data life-cycle management and preservation

The Data Repository Ecosystem



FAIR principles apply to the *data objects*.
TRUST principles apply to the *data repositories*.

The TRUST Principles

- **T - Transparency** is achieved by providing publicly accessible evidence of the services that a repository does and does not offer.
- **R - Responsibility** is a commitment to provide reliable data services.
- **U - User community** is a commitment to implement and enforce the standards and norm of the user community.
- **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.

T- Transparency

Transparency: Publicly accessible data curation policies, capabilities and services

- T - Transparency is achieved by providing publicly accessible evidence of the services that a repository does and does not offer.

R – Responsibility

Responsibility: Accountable provision of data services for the user community

- R - Responsibility is a commitment to provide reliable data services.

U – User Community

Users: Enable current and future use of data in line with the norms of the community served

- **U - User community focused** is a commitment to implement and enforce the standards and norms of the user community.

S – Sustainability

Sustainability: Capability to continually facilitate long-term stewardship and use of data

- S - Sustainability is the capability to support long-term data preservation and use.

T – Technology

Technology: Ongoing advancement of secure, reliable tools, services and infrastructure

- T - Technology is the infrastructure and capabilities to support the repository operations.

Advantages of the TRUST Principles

- Offers common framework for evaluation of repositories.
- Like OAIS, generalizes trustworthiness beyond disciplinary data repositories
- Work in concert with other principles, such as the FAIR principles.
- Easily understandable guidance to communicating

Future Directions

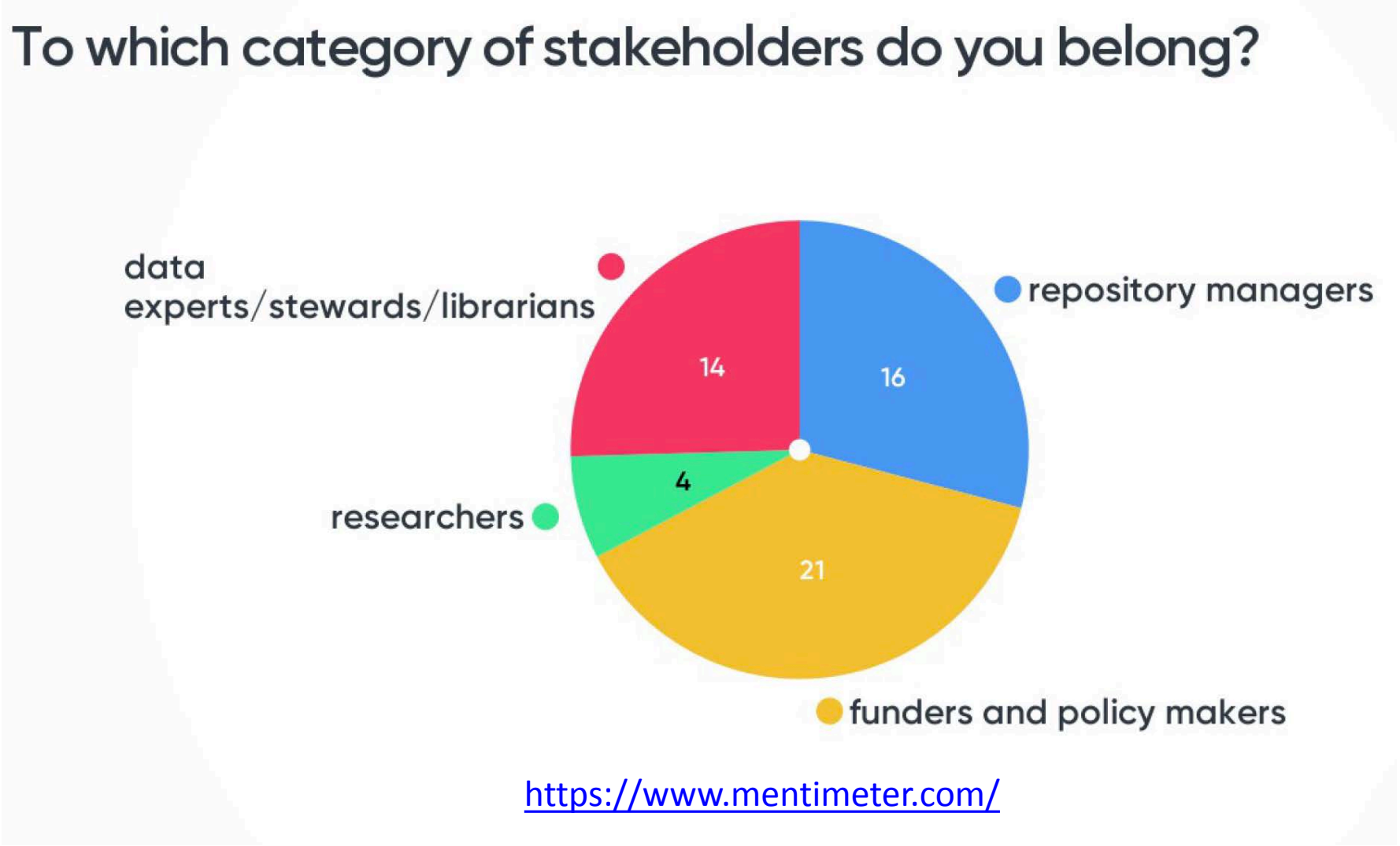
- Mappings of TRUST principles to a range of relevant standards and practices, such as OAIS-ISO14721, TRAC/ISO16363, and CoreTrustSeal.
- Examples of trustworthy digital repositories
- Some form of logical elaboration of basic tenets of psychology.

Trustworthy Data Repositories for Biomedical Sciences

- April 8-9, 2019
- Chairs: Ingrid Dillo, John Westbrook
- Participants included:
  - Trustworthy certification experts
  - Repository managers
  - NIH staff
  - Online: government agencies, universities, industry, international
- Reach: 54 in person, 70 online

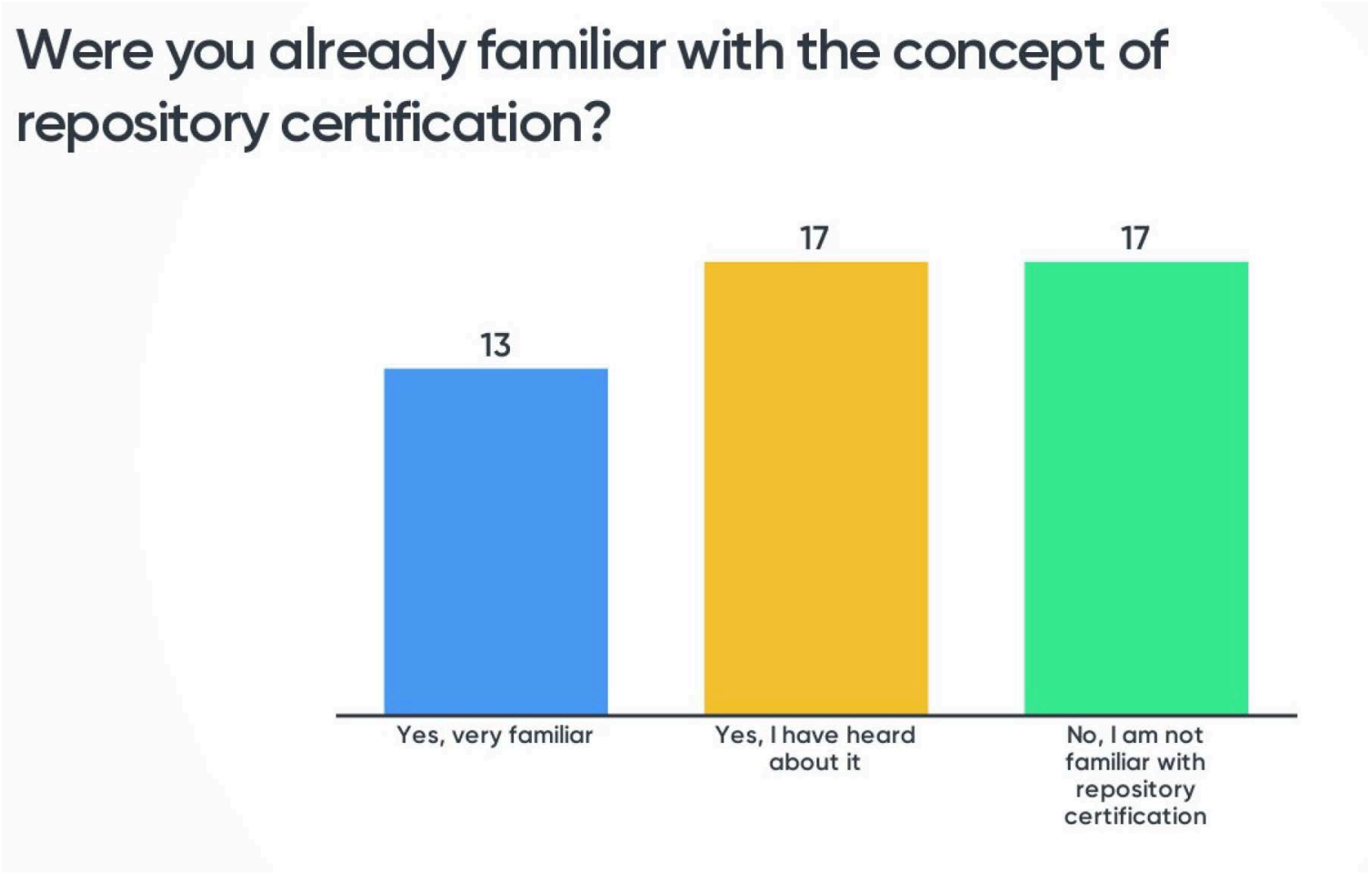
https://datascience.nih.gov/data-ecosystem/trustworthy-data-repositories-workshop

Workshop chairs used Mentimeter to engage participants



https://www.mentimeter.com/

Pre-discussion: assess concept familiarity



Assess essential TRUST concepts

What are the three keywords that would be important for you in the context of certification?

T – Transparency

Transparency is achieved by providing publicly accessible evidence of the services that a repository does and does not offer.

Key concepts discussed by workshop participants

- Data provenance
- Data curation
- Documentation of process and policy.

R - Responsibility

Responsibility is a commitment to provide reliable data services.

Key concepts discussed

- Responsibility is shared (data producers, institutions, funders, repositories, publishers, and data users).
- Data quality (validation, reproducibility).
- Education, communication, and stewardship.
- Building a culture of shared responsibility was thought to be critical to responsibility principle of TRUST.

U – User Community

User community is a commitment to provide and enforce the standards and norm of the user community.

Key concepts discussed

- Engagement, communication, collaboration, and service were identified as critical components this principle.
- Engaging and being responsive to user community needs is built on good communication and collaboration and reliable, responsive, and consistent service.
- User community is broad (data producers, funders, data users, and scientific community).

S – Sustainability

Sustainability is the capability to support long-term data preservation and use.

Key concepts discussed

- Funders, institutions, data users, data providers, and government, were identified as partners in achieving sustainability.
- Data is a shared resource requiring all partners to support that resource.
- Key issues were seen as long-term planning, funding models, business models, and prioritization.

T – Technology

Technology is the infrastructure and capabilities to support the repository operations.

Key concepts discussed

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

Acknowledgements

- **Version 0.02 TRUST principles White Paper co-authors:** Dawei Lin, Jonathan Crabtree, Ingrid Dillo, Robert R. Downs, Rorie Edmunds, Marisa De Giusti, Hervé L'Hours,Wim Hugo, Reyna Jenkyns, Varsha Khodiyar, Maryann Martone, Mustapha Mokrane, Vivek Navale, Jonathan Petters, Barbara Sieman, Dina V. Sokolova, Martina Stockhause
- The suggestions for improving this work that were offered by several members of the CoreTrustSeal Standards and Certification Board and by participants of the Research Data Alliance Plenary 13 session, “Build TRUST to be FAIR - Emerging Needs of Certification in Life Sciences, Geosciences and Humanities”, which was convened by the RDA/WDS Certification of Digital Repositories Interest Group. We are grateful for thoughtful discussions from Shelley Stall, Maryann Martone (Professor Emeritus of Neuroscience UCSD) to write the version 0.01 of the White Paper.
- The generous and insightful comments offered for version 0.01 from Robert S. Chen, Mark Conrad, Peter Doorn, Eliane Fankhauser, Elizabeth Hull, Siri Jodha Singh Khalsa, Micky Lindlar, and Limor Peer.