Guidance for Trustworthy Data Management in Science Projects

https://trustedci.org/2020-trustworthy-data

December 15, 2020

Distribution: Public

Trustworthy Data Working Group

Andrew Adams, Kay Avila, Jim Basney, Laura Christopherson, Melissa Cragin, Jeannette Dopheide, Terry Fleury, Calvin Frye, Florence Hudson, Manisha Kanodia, Jenna Kim, W. John MacMullen, Mats Rynge, Scott Sakai, Sandra Thompson, Karan Vahi, John Zage

About the Trustworthy Data Working Group

The Trustworthy Data Working Group is a collaborative effort of Trusted CI[1], the four NSF Big Data Innovation Hubs[2], the NSF CI CoE Pilot[3], the Ostrom Workshop on Data Management and Information Governance[4], the NSF Engagement and Performance Operations Center[5] (EPOC), the Indiana Geological and Water Survey[6], the Open Storage Network[7], and other interested community members. The goal of the working group is to understand scientific data security concerns and provide guidance on ensuring the trustworthiness of data.

Acknowledgments

The co-authors thank all the other members of the Trustworthy Data Working Group for their help with developing this guidance.

Trusted CI is supported by the National Science Foundation under Grant 1920430. The Cyberinfrastructure Center of Excellence (CI CoE) Pilot project is supported by the National Science Foundation under Grant 1842042. The Engagement and Performance Operations Center (EPOC) is supported by the National Science Foundation under Grant 1826994. The Open Storage Network is supported by Schmidt Futures[8] and the National Science Foundation under Grants 1747483, 1747490, 1747493, 1747507, and 1747552. The NSF Big Data Innovation Hubs are supported in part by the National Science Foundation under awards 1916613, 1916585, 1916589, and 1916573. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

[1] https://www.trustedci.org
[2] https://www.bigdatahubs.org
[3] https://cicoe-pilot.org
[4] https://ostromworkshop.indiana.edu/research/data-management
[5] https://epoc.global
[6] https://igws.indiana.edu
[7] https://www.openstoragenetwork.org
[8] https://schmidtfutures.com

Using & Citing this Work

Cite this work using the following information:
Trustworthy Data Working Group. Guidance for Trustworthy Data Management in Science Projects. December 2020. https://doi.org/10.5281/zenodo.4056241

This work is available on the web at the following URL:
https://trustedci.org/2020-trustworthy-data

Table of Contents

Executive Summary **6**

1 Introduction **7**

2 Stakeholders **8**

3 What is Trustworthy Scientific Data? **10**

4 Key Findings from Survey Report **17**

5 Barriers to Trustworthiness **19**

6 Tools and Technologies for Trustworthy Data **21**

6.1 Strong Coverage: Availability, Integrity, and Authorization 24

6.2 Attributes In Need of Stronger Coverage 24

6.2.1 Accepted Techniques of Creation and Reproducibility 24

6.2.2 Confidentiality 25

6.2.3 Authenticity and Credible Source and Stewardship 26

6.3 Roles and Responsibilities 27

6.3.1 Data Users 27

6.3.2 Data Providers 28

6.3.3 Infrastructure Providers 28

6.3.4 Facilitation and Compliance Professionals 28

7 Communicating Trustworthiness **29**

7.1 Communicating Attributes of Trustworthiness 30

7.2 Sharing Scientific Metadata 31

8 Conclusions **32**

References **33**

Appendix A - List of Terms **36**

List Of Images

- Figure 1. Distribution of Respondents to Question 15 of the Survey
- Figure 2. Stakeholder Roles
- Figure 3. Distribution of Responses to Question 5 of the Survey
- Figure 4. Distribution of Responses to Question 13 of the Survey

List Of Tables

- Table 1. Attributes of Trustworthy Data for Various Stakeholders
- Table 2. Tool Effectiveness for Various Trustworthy Data Attributes

Executive Summary

In April and May of 2020, the Trustworthy Data Working Group[9] conducted a survey of the scientific community about data security concerns and practices. 111 participants from a wide range of positions and roles within their organizations and projects, respectively, completed the survey. The working group analyzed the survey results with an eye for patterns, themes, correlations, and aggregates and produced a report in June 2020 detailing the process, survey methodology, and their analysis [1].

Several key findings emerged from the report, including:

1. Data owners, maintainers, and users are concerned about the trustworthiness of data throughout the lifecycle of the scientific process, especially with regard to the loss of reputation if data trustworthiness isn't preserved.
2. Data owners and maintainers welcome help in securing trustworthy data workflows with encryption, provenance, and regulatory compliance (e.g., FERPA, HIPAA, FISMA).
3. Trustworthiness is not precisely defined in the scientific community.

Based on what the Trustworthy Data Working Group learned from its research in preparation for the survey,[10] the survey findings, subsequent group discussions around the data needs and uses uncovered in the survey, and knowledge of best practices around security and privacy, this guidance report was written as a follow-up to the analysis report mentioned above. The report delves further into key survey findings, explores concerns regarding trustworthy data uncovered in the survey, and provides recommendations to address those concerns. Topics are organized around the following sections: stakeholders of trustworthy data (Section 2), definition of trustworthiness (Section 3), findings from the survey report (Section 4), barriers to trustworthiness (Section 5), tools and technologies for trustworthy data (Section 6), and communication of trustworthiness (Section 7).

[9] https://www.trustedci.org/2020-trustworthy-data

[10] A literature review was conducted that explored recent and current efforts to characterize trustworthiness in data, and efforts to assess and support the protection of research data.

1 Introduction

In April and May of 2020, the Trustworthy Data Working Group conducted a survey on scientific data security concerns and practices in the academic research community. The survey was distributed widely via email and listservs that included both technical and research computing groups, as well as some domain science groups. There were a total of 111 respondents, predominantly from research and computing services, infrastructure providers, science software developers, and educators.

The survey and subsequent report [1] met the first goal of the working group: to understand scientific data security concerns based on a survey of researchers and cyberinfrastructure professionals working in a variety of roles to produce domain science. The purpose of this report is to meet our second goal: to provide guidance, shaped by the results of the survey, on ensuring the trustworthiness of data.

To gauge the community's desire for guidance, and to confirm we are not offering guidance where it is not welcomed, we asked the following question in the survey. "If you were provided with additional guidance, resources, or support, would you apply additional tools or technologies to help maintain the trustworthiness of the research data that you use/produce/curate?" An overwhelming majority of respondents selected Yes (52%) or Maybe (41%), and very few selected No (6%). This demonstrates a clear receptiveness to guidance from the working group and justification for the working group's effort in this report.

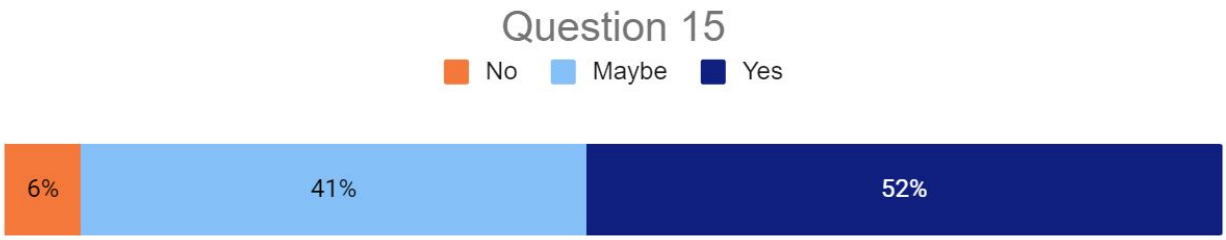

Figure 1. Distribution of Respondents to Question 15 of the Survey

This report is organized as follows. In Section 2, we provide an analysis of the stakeholders of trustworthy data. In Section 3, we explore what "trustworthy scientific data" means. We review key findings from the survey report in Section 4, then discuss barriers to trustworthiness in Section 5. Section 6 describes tools and technologies for trustworthy data, and Section 7 discusses communicating trustworthiness. We conclude in Section 8.

This December 2020 (final) version of the report incorporates community input and the results of additional working group discussions since our September 2020 (initial) publication, including:

- Additional tools and technologies added to Section 6.
- Revisions to Section 7 to match the attributes specified in Section 3.
- Minor edits for clarity.

2 Stakeholders

Multiple actors engage in the research and data management process. They may create data, conduct research with it, curate it, share it, protect it, store it, transmit it, etc. They each have a stake in ensuring its trustworthiness throughout its life cycle. In the 2020 Trustworthy Data survey, respondents were asked to select the roles they fill as a part of their jobs. Roles are defined here as sets of responsibilities or duties that shape a particular way of interacting with data. The 110 responses below show a range of roles occupied by respondents. Eighty-one percent of respondents indicated filling multiple roles as a part of their job. For instance, a scientific data user or researcher may also use data for research. During the course of the analysis, the data user or researcher is also responsible for the data's stewardship, thus positioning the individual as a data manager or curator. Should the user/researcher wish to share the data with others, s/he becomes a data provider (spanning multiple roles below). A data provider may also provide the infrastructure within which the data is housed, making the data provider an infrastructure provider as well.

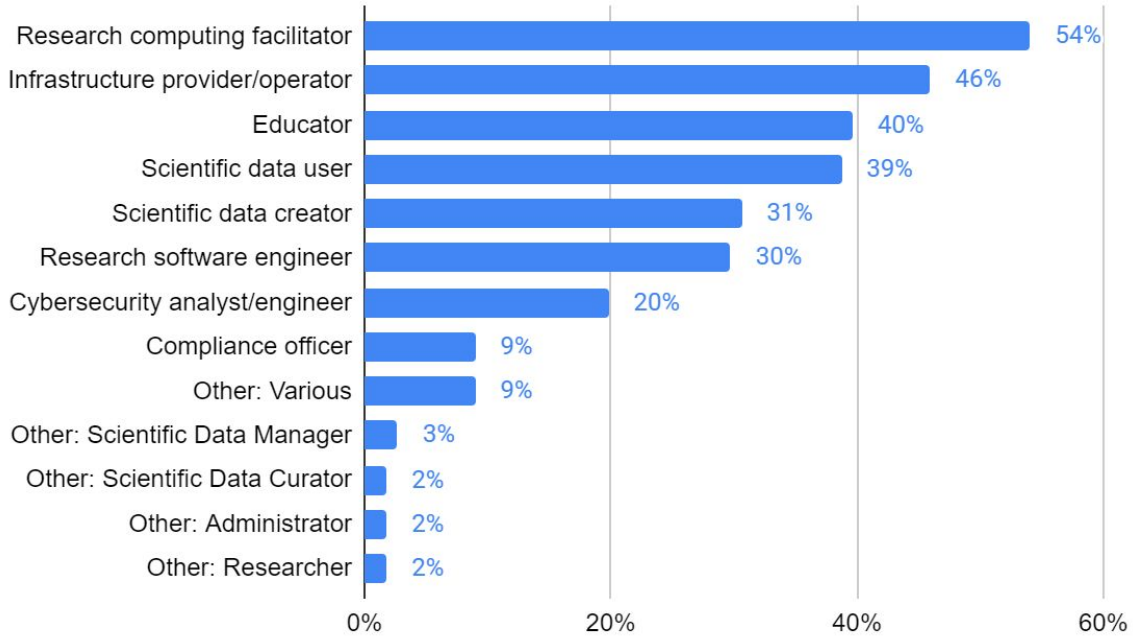

Figure 2. Stakeholder Roles

For the purposes of this report, the different roles listed above have been grouped into four broader roles:

- Data User *(e.g., educator, researcher)* - A data user is someone who works with data for use in research. A data user may need to search for data, review and assess the data for potential use, download or obtain the data in some way, conduct analysis on the data, and publish any findings made as a result of that analysis. A data user may also use data for instruction or educational purposes, particularly when teaching students how to conduct or evaluate research. A data user may also find themselves in a role of a data provider, sharing data created or used in prior studies.
- Data Provider *(e.g., data creator, scientific data curator, scientific data manager)* - A data provider is someone who has data; can speak authoritatively about its integrity, authenticity, and means of creation; and has the authority to make it available to others. This person or organization may be the original creator of the data or they may have been authorized by the creator to distribute the data. In either case, this entity is responsible for defining acceptable use of the data by others. The person may be a researcher him/herself and may have conducted their own analyses on the data and thus also consider themselves a user of the data. Individuals or groups that create data repositories for sharing data or make data publicly available through other means would be considered data providers.

- Secure Infrastructure Provider *(e.g., infrastructure provider/operator, research software engineer, cybersecurity analyst/engineer)* - An infrastructure provider is someone who supplies one or more services around the data's use. For example, an infrastructure provider may house the data, provide a way to search or browse the data, be responsible for its transmission, or enable its use and sharing. Infrastructure providers offer these services via the development, adaptation, provision, or secure configuration and management of cyberinfrastructure, specifically, hardware, software, networks, or other technological tools. Staff at campus computing facilities, high performance computing centers, information security offices, libraries, or commercial service providers that are responsible for the infrastructure or services around data use may fit this role.
- Facilitation and Compliance Professional *(e.g., research computing facilitator, compliance officer*) - A facilitation and compliance professional is someone who supports and consults with others on the use of technology with data and ethical treatment of data. Research facilitators consult with data users about computational methods that may be employed in the analysis of data, services and tools that can help facilitate the analysis, and institutional resources available to support research endeavors. They may also engage in distilling best practices around sharing and use of data. Compliance professionals include individuals or groups responsible for ethical use and care of data, compliance with regulations and policies on data sharing and use, and review and assessment of research and data use activities.

In the following sections, we provide guidance applicable to the above four stakeholders.

3 What is Trustworthy Scientific Data?

Data trustworthiness can be difficult to define as different organizations and roles have varying sets of attributes they consider important and other attributes that they consider unimportant. In this report, we focus mainly on academia and the stakeholders described above. However, it is still important to be able to convey data trustworthiness in many situations. For example, trustworthiness must be communicated to the public when publishing science findings, or to industry partners in a requirements document for purchasing new systems.

One of the illuminating outcomes of the survey was that when given a list of attributes to define trustworthiness, few of the participants defined it similarly. This should not be surprising, though, as several cyber-related associations have also defined 'trustworthy'

differently (e.g., NIST [2]). The basis of the attribute set utilized in these Guidelines comes from the responses from question 5 in the survey.

5. Which attributes do you believe scientific data must have in order to be trustworthy?

- ❑ Accuracy - The data is free from error.
- ❑ Integrity - The data has not been altered.
- ❑ Methodology - The processes and inputs used to create the data are well-established and accepted by the community.
- ❑ Provenance - The data's origin and lineage can be readily established.
- ❑ Reproducibility - The data can be re-created, or the associated scientific results are replicable.
- ❑ Reputation - The data was generated by a credible or trusted source.
- ❑ Responsible stewardship - The ownership of the data is well managed and can be transferred.
- ❑ Significance - The data enables future research directions (with associated funding/support).
- ❑ Other: _____

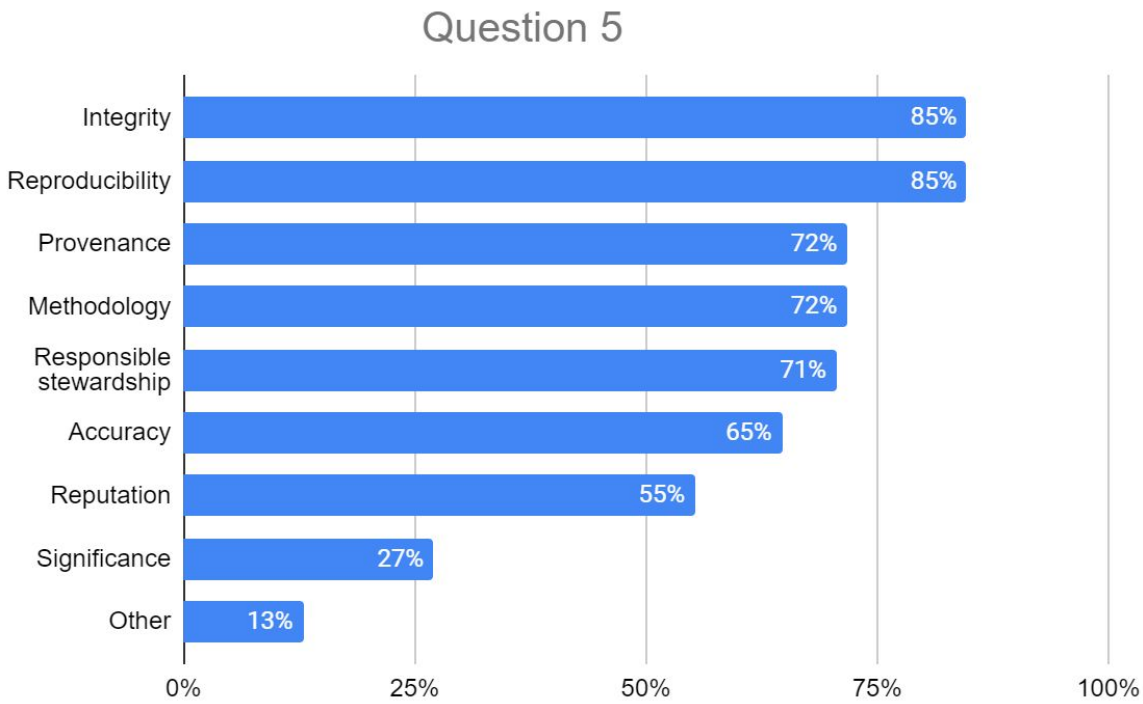

Figure 3. Distribution of Responses to Question 5 of the Survey

Each attribute was selected by one or more respondents, with the "Other" write-ins adding "transparency", "documentation", and "methodology". These attributes were then compared

to the TRUST (Transparency, Responsibility, User focus, Sustainability and Technology) Principles for digital repositories [3], and the FAIR Data Principles (Findable, Accessible, Interoperable, and Reusable) [4]. The working group discussed coverage and importance in relation to the identified stakeholders, as well as the relationship of trustworthiness to the traditional CIA triad (Confidentiality, Integrity, and Availability), and derived the following updated set of attributes for trustworthy data.

Availability: The data is accessible to authorized users, served from a reliable system, during the agreed-to timeline, and according to the data provider's usage policies.

Integrity: The data has not been damaged by faulty systems or altered by malicious actors. Controls can include checksums, logging transactions and actors, replication, and backups.

Authenticity: The data has a clear provenance (origin and lineage) which can be verified. Provisions for metadata describing the data and versioning are included.

Accepted Techniques of Creation: The processes and inputs used to create the data are well-established and accepted by the community.

Authorization: Access controls should be properly in place. For example, open access data will have controls in place to allow anyone to read the data, but limit uploading/updating to the team maintaining the data set.

Confidentiality: The data repository hides/masks personally identified information (PII) or other sensitive information from those not granted access.

Credible Source and Stewardship: The ownership and oversight of the data is well managed and can be transferred securely.

Reproducibility: The data can be re-created, or the associated scientific results are replicable.

In the following table, which spans the next four pages, we describe the attributes of trustworthiness in the context of different stakeholders, highlighting the importance of why the attributes were included in the trustworthy data definition.

Table 1. Attributes of Trustworthy Data for Various Stakeholders

	Data User	**Data Provider**	**Secure Infrastructure Provider**	**Facilitation and Compliance Professional**
Attribute	I want to ensure that the data I download is available to me when I need it, has integrity, is authentic, was created in a proven manner, does not disclose information I have no rights to, has been credibly and trustworthily tended, and is reproducible.	I want to ensure my data is protected and made accessible to only those I have had the opportunity to vet and grant approval to after they've agreed to any stipulations I might have on the data's use.	I want to guarantee that the data we serve in the infrastructure we provide abides by the Data Provider's data policies and satisfies the Data User's needs for verification of its trustworthiness.	We want to ensure all parties concerned have the support they need to ensure the data is trustworthy and the research is sound. We want to ensure all parties are honoring any agreements or policies on data sharing and use.
Availability	Data is available when needed.	The infrastructure I choose to hold and serve and allow use of the data should make the data available on a timeline I have agreed to.	My infrastructure enables the secure and timely access and use of the data as defined in any of the Data Provider's usage policies.	Facilitation: We may help data users find data by making them aware of repositories and other resources, and helping them understand policies on sharing and use. We may help data providers find appropriate repositories or places to list and share their data.
Integrity	The data is complete, unaltered from its promised form, undamaged (not corrupt), and accurate (free from errors). I can rely on it to be in the condition the Data Provider or Infrastructure Provider promises.	The integrity of the data will be maintained no matter where the data lives and how the data is served or transferred before the user comes into its possession.	My infrastructure is designed and maintained to guarantee the data's integrity. It protects the data from alteration, corruption, loss of completeness, or loss of accuracy.	Facilitation: We make data users and providers aware of and help them employ the methods, services, tools, and other resources that are available to assess and ensure the integrity and accuracy of the data.

	Data User	Data Provider	Secure Infrastructure Provider	Facilitation and Compliance Professional
Authenticity	Information about the provenance of the data (i.e., its origin and lineage) is established and available for my verification.	There is a means by which I can make the data's origins and lineage transparent to any potential users of the data so they can review it for authenticity. I may even have a way to stamp the data with my seal of approval and guarantee that the data is mine and that I stand by the data.	The data housed/served from my infrastructure is authentic. It is the same data that was originally given to us by the Data Provider and is as promised to the user. It's provenance can be verified.	Facilitation: We consult with data users on how to assess data authenticity and make them aware of any tools or services available to guide them in this task.
Accepted Techniques of Creation	The data was created in a manner proven and trusted by the larger scientific community. The means of creation are available for my review.	I created the data in a manner proven and trusted by the larger scientific community. The means of creation have been detailed and made available for review by potential users.	The infrastructure allows data owners to document the means of creation when storing data.	Facilitation: We help data users and providers with data cleaning and preparation, documentation, methodology, analysis techniques, and statistics so that the results of any study can be later used by others as a new and trusted data set.

	Data User	Data Provider	Secure Infrastructure Provider	Facilitation and Compliance Professional
Authorization	Access to the data is granted after a vetting process where I agree to provisions related to its use/handling, as laid out by the Data Provider to ensure its protection and care.	I have a way to vet potential users of my data and grant them access based on my assessment of their trustworthiness.	The infrastructure was built with access controls that enable only those users the Data Provider has deemed acceptable access to data. Users must provide their credentials to access the data. Their identity must be verified and communications between the user, Data Provider, and the infrastructure are encrypted and secure.	Compliance: We advise all parties on and/or may monitor the release and transmission of the data to ensure it complies with any security policies or data use policies. We may review the design of the infrastructure to ensure it keeps the data safe and secure and does not allow it to be released to someone that has not been approved.
Confidentiality	No personally-identifiable or sensitive information is revealed in the data other than that which I have made specific promises to protect that have been vetted and approved by the Data Provider or Infrastructure Provider.	Any personally-identifiable or sensitive information contained in the data can be masked from those I choose to share the data with if I do not wish them to have access to this information. Anyone I do not grant access to will not be able to see this information under any circumstances.	The infrastructure supports tools and technologies for maintaining data confidentiality.	Facilitation: We advise data users and providers on any techniques for de-identifying data and protecting the individuals that the data describes. Compliance: We review and advise on any proposed processes on data sharing and use to ensure the identities of any of the people the data describes cannot be deductively disclosed.

	Data User	Data Provider	Secure Infrastructure Provider	Facilitation and Compliance Professional
Credibility	The data was created by a credible, trusted source which can be verified. The data is managed and tended by a credible, trusted source; this source can be trusted to protect, house, and serve the data securely.	Whatever Infrastructure Provider houses, serves, or grants use of my data, must protect it with the security measures necessary to ensure the data is not exfiltrated or transmitted without my express permission. If the infrastructure enables use of the data in its care, such use must also be protected and granted to only those I have approved of.	The infrastructure will protect the data from unwanted, unapproved exfiltration, transmission, or use (if use occurs within the infrastructure).	Facilitation: We may assist data users with assessing data sources for credibility. Compliance: We may review the design of the infrastructure to ensure it is secure and protects the data as agreed upon by various parties.
Reproducibility	The results of any study in which the data has been used should be replicable using the same methods and data parameters.	The data should not undergo any transformations during its storage and transmission to others that might render different study findings than those I arrived at through my investigations (if I am the data creator) or the investigations of those who created the data and for whom I am managing the data.	The infrastructure must protect the data from unwanted transformations that would result in different findings from those of the original study.	Facilitation: We help data users and providers with data cleaning and preparation, documentation, methodology, analysis techniques, and statistics so that the results of any study can be replicated later.

4 Key Findings from Survey Report

The survey produced both expected and unanticipated findings. The primary goal was to gauge the communities' temperature regarding trustworthy data, i.e., is the current state of data used in the scientific mission trustworthy? Answers to some questions affirmed that the majority of participants believe data trustworthiness is important.[11] The most common theme from those questions was that trustworthy data is a cornerstone of the scientific process. Many participants cited the scientific process/method specifically, and others used language such as, *"The very mission of science as a whole is to produce data that is processed, analyzed, and published with integrity,"* or *"results based on improperly collected, maintained, or understood data are inherently untrustworthy*."

Although the responses suggest that most are satisfied with the level of trustworthiness in the data they produce/use/curate,[12] roughly a quarter of respondents are concerned about reputational risk, both personal (e.g., scholarly rebuke, people unwilling to collaborate, loss of position), and to their institutions if their data were to be considered untrustworthy.[13]

A second finding was that data owners and maintainers welcome help in securing trustworthy data workflows with encryption, provenance, and managing regulatory compliance (e.g., FERPA, HIPAA, FISMA). Responses revealed that a third of the respondents were not satisfied with the current state of tools/technologies in achieving sufficient assurance of trustworthiness.[14] When asked about this, half of the respondents stated that they did desire additional help.[15] A few participants were very specific in sharing their specific needs. For example:

- *"QA (quality assurance) software for acquisition and processing (provenance generation through manual and automated entry)"*
- *"Data management software that ties QA metadata to data streams and points"*
- *"QC (quality control) software that allows repeatable QC/analysis and generates QA information for those processes"*
- *"Offsite data repositories with metadata indexing"*
- *"Encryption"*
- *"Tooling to validate integrity of container images"*

[11] See Question #6 (*Importance of Protecting Trustworthiness*) and its free form response follow-up, Question #7.
[12] See Question #10 (*Confidence in the Data*).
[13] See Question #9 (*Potential Consequences*).
[14] See Question #14 (*Sufficiency of Tools and Technologies*).
[15] See Question #15 and its free form response follow-up, Question #16.

- *"Tooling and methodology to perform 'security posture checking' of end user devices that does not limit end users' control over their devices nor their access to the network"*
- *"Tools/technologies around CUI"*
- *"ERPID (enhanced robust persistent identifiers of data) to generate PIDs (persistent identifiers of data) and track the data products that way. But other tools would be of interest as well."*

Some participants answered in regards to future concerns revolving around how data will be saved, where it will be saved/stored, maintaining its readability, sustaining it through lack of funding, and updating systems to fulfill new local and federal requirements. Specific comments include:

- "*There is a lot of social media data harvesting that is used in research today. There are no guidelines, it is up to every individual researcher."*
- *"Having lived through numerous storage media revolutions, I am mildly concerned with future-proofing the readability of data I save. I have CDs full of data that I can no longer open, and I worry USB drives and even cloud storage will be the same."*
- *"Long term preservation on public cloud platforms while avoiding vendor lock-in."*

Finally, the last and perhaps most profound finding from the survey responses was that there is no clear agreement on what precisely describes 'trustworthy data' or on which roles should own responsibility for ensuring it. One multiple response question provided the participants with a list of potential attributes that might be associated with trust and asked them to select which ones scientific data must have in order to be considered trustworthy.[16] Few participants, if any, used the same set of attributes to define trustworthy data. Perhaps more surprising was that the attribute "integrity" was not selected by all (15% did not select it), and that "reproducibility" was significantly higher (85%) than either "provenance" (72%), "accuracy" (65%) or "reputation" (55%). Hence, a component governing the concern of trustworthiness is that it appears the scientific community has yet to agree upon what trustworthiness means.

While 90% of respondents either agreed or strongly agreed that protecting the trustworthiness of data is important, only 69% believed that establishing or maintaining trustworthiness fell within their job duties. This indicates potential confusion over who holds the responsibility for this function. Interestingly, 90% of the ten respondents who self-identified as compliance

[16] See Question #5 (*Attributes of Scientific Data*). See also Section 3 of this document for full results to this question. Possible attributes included accuracy, integrity, methodology, provenance, reproducibility, reputation, responsible stewardship, and significance, as well as a free-form "Other" option.

officers either agreed or strongly agreed with this statement, but only 73% of the twenty-two respondents who self-identified as cybersecurity analysts or engineers agreed or strongly agreed that establishing or maintaining trustworthiness fell within their job duties.

5 Barriers to Trustworthiness

If a consensus on the definition of trustworthiness has yet to be established within the scientific community, then identifying the barriers to establishing trustworthiness poses its own set of challenges. Any one of the attributes from Section 3 could have related barriers expounded upon in depth. However, in lieu of that extensive examination, we take a more generalized approach here by outlining potential difficulties applicable to one or more possible attributes.

As discussed in Section 4, two potential barriers were identified in the survey:

- Lack of a common definition for trustworthiness
- Lack of consensus on which role(s) should own responsibility for trustworthiness

Additionally, four additional barriers emerged as part of internal discussions and a literature review conducted by the working group:

- Conflicts between cybersecurity and reproducibility
- Accidental failures through systems or human error
- Lack of funding/incentives for making supporting data and provenance information available
- Perceived data quality issues

Lack of a common definition. First, as mentioned in the previous section, no widely accepted definition of trustworthiness exists, making it difficult to hold a discussion with common understanding or to assign responsibility for ensuring it to a particular role. The NIST definition of trustworthiness focuses primarily on integrity, one of the three components of the CIA triad (Confidentiality, Availability, and Integrity) traditionally associated with security and under the purview of cybersecurity positions [2]. However, not all respondents in the survey believe data integrity is a critical part of trustworthiness, and few cybersecurity engineers or analysts believe trustworthiness should fall under their purview.

Lack of consensus on which role(s) should own responsibility for trustworthiness. The cybersecurity perspective reflected in the survey responses did not show any disagreement with NIST's focus on integrity, which would seem to be a clear fit for cybersecurity analysts and engineers. All twenty-two of the cybersecurity analyst/engineer respondents indicated that integrity is an important attribute of trustworthiness. Yet (as discussed in Section 4), almost a fourth of these individuals felt that trustworthiness does not belong under their purview. One hypothesis for this discrepancy related to integrity could be the distinction between malicious and non-malicious threats to integrity. It may be that cybersecurity engineers, analysts, and professionals in related roles view themselves as primarily preventing unauthorized or malicious alteration of data rather than alteration introduced by human or computer error.

Conflicts between cybersecurity and reproducibility. Further complicating cybersecurity's relationship with trustworthiness, protecting the security of data, including its integrity, can come into direct conflict with other facets of trustworthiness. As Deelman, et. al. note, patching against security flaws is critical for protecting systems and data, yet can adversely affect reproducibility by harming performance or by creating incompatibilities with older code [5]. Potential conflicts arise when older software or equipment is necessary for reproducing findings, yet can no longer receive security updates. Additionally noted in this reference, publishing the code used to generate or manipulate data increases reproducibility, as well as illuminating provenance; both are components of trustworthiness. However, an argument can be made from a security standpoint that publicly publishing code makes it more prone to having vulnerabilities discovered and exploited. This view is admittedly more commonly seen in industry or in the private sector.

Accidental failures through systems or human error. Accidental failures, whether due to system or human error, also pose a problem to establishing or maintaining the trustworthiness of data. As one article notes, "Confidence in data depends upon trust in the entire data life cycle," [6] from the initial gathering or computation to how the data is archived [7]. Bit flips can occur through natural or accidental means [8]. Researchers depend upon the ability to move data confidently within computational resources, yet there are well-known cases where significant problems have occurred. When the Scientific Workflow Integrity with Pegasus project began testing production workflows, previously undetected changes in checksums were observed with data being processed on the Open Science Grid [9]. Additionally, failures in trustworthiness can occur through incorrect algorithms, whether those are developed by hand or through machine learning. In one instance, a commonly used set of python scripts were found to give different results depending on the underlying operating system. Those scripts had been cited over 130 times between 2014 and 2019 [10].

Lack of funding/incentives for making supporting data and provenance information available. These issues can be exacerbated by political and cultural issues surrounding scientific publishing. Due to the competitive nature of research, publishing early is incentivized, and carefully curating the underlying data to make it available may not be part of the process. Despite occasional and often unfunded mandates to make the corresponding data available publicly, "data donation is not acknowledged as a contribution to scientific research," and "[a]ctivities such as data donation and participation in data curation are not currently rewarded within the academic system. Therefore, many scientists who run large laboratories and are responsible for their scientific success perceive these activities as an inexcusable waste of time, despite being aware of their scientific importance" [11]. A report generated after a workshop on data preservation stated it bluntly: "data's default state is being in a state of risk" [12]. This is even more true when underlying methodology and provenance data are not published, in which case the results may default to a state of untrustworthiness.

Perceived data quality issues. Finally, in the era of "Big Data" and citizen science, many people may be involved in creating and using a data set. At one point, control and unbroken provenance were seen as necessary components for data integrity, yet this is changing as funding agencies push for open data, science continues to become more collaborative, mashups of data continue to be used, and data is used from "messy" sources like social networks [13]. These can call the trustworthiness of the data into doubt if the data quality seems questionable. New ways of tracing and evaluating the trustworthiness of these different kinds of data will need to be developed and disseminated into the community.

6 Tools and Technologies for Trustworthy Data

Having explored and defined the attributes of Trustworthy Data, the selection of tools and technologies to help assure those attributes becomes a *gap-analysis* exercise. Starting with a survey of the tools and technologies that the respondents indicated a strong familiarity with, we evaluate their effectiveness at providing the eight attributes identified in Section 3. The remainder of the gap analysis consists of identifying any uncovered attributes and suggesting approaches to strengthen their assertion.

As a preface to suggested tools or technologies, it is the intent of this paper and its authors to offer general ideas, concepts, or capabilities; rather than endorse or promote specific products as a "best" solution, lest anyone be misled into believing that it is wrong to use anything but the products that would otherwise be mentioned here. Additionally, all stakeholders are

encouraged to seek out contemporary, unembellished, well-proven products or solutions that fit these generalizations (and their requirements) before designing a product on their own.

Question 13 of the survey shows general awareness of technologies for providing a subset of our eight attributes. These serve as the basis for the gap analysis.

Which of the following tools and technologies (if any) help to secure the research data that you use/produce/curate?

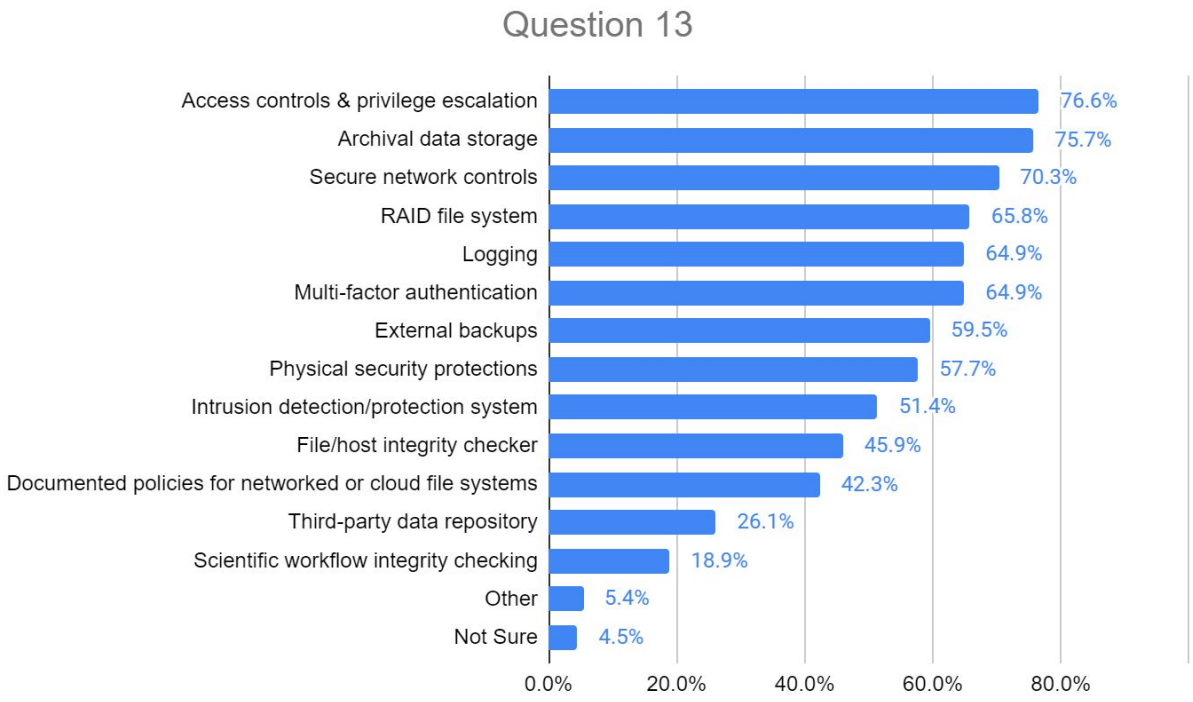

Figure 4. Distribution of Responses to Question 13 of the Survey

In the following table, we quantify the effectiveness of each tool/technology from Question 13 towards asserting one of the desired attributes.

Table 2. Tool Effectiveness for Various Trustworthy Data Attributes

Tool / Attribute	Availability	Integrity	Authenticity	Accepted Techniques	Authorization	Confidentiality	Credible Source	Reproducibility
3rd party data repo	O	O	O				O	O
Policy for network/cloud storage	N	O			N		N	
Archival storage	N	O					O	O
Workflow integrity checking		S						N
Access controls	N	N			S	N		
Physical security protections	N	N			N	N		
Network controls	N	N			N			
Logging	O	O			O			
Multifactor Authentication	O	O			O			
Intrusion detection/protection	O	O			O			
File/host integrity check	O	N			O			
RAID file system	N	N						
External backups	N	O						

Symbol	Meaning
S	Sufficient: This tool/technology alone can establish an assertion of the desired attribute, however weak it may be.
N	Necessary: This tool/technology is required to provide a stronger, credible assertion of the desired attribute.
O	Optional: This tool/technology can help strengthen the assertion of the desired attribute, however that is not its design intent.

To arrive at these subjective scorings, we consider the meaning of each tool, then discuss how each attribute can be asserted using these tools. Unfortunately, as definitions for each tool were not provided in the survey, our definitions may vary from what any given respondent had in mind, and thus may change the intent of their response. It is our intent to minimize the effect of this deviation by being transparent about our adopted meaning of these terms, listed in Appendix A.

6.1 Strong Coverage: Availability, Integrity, and Authorization

The attributes, "availability," "integrity," and "authorization" are clearly well-covered by tools that are staples for a typical enterprise information-technology (IT) deployment, and many environments using hardware and software designed for server (as opposed to desktop) roles. This should come as no surprise, as asserting those attributes are general IT concerns, and general IT tools are designed to address them. However, the effective use of these tools may be a responsibility shared amongst multiple roles.

6.2 Attributes In Need of Stronger Coverage

As seen in our gap analysis, we find that the initial set of tools and technologies leaves several remaining attributes poorly covered, in particular, "authenticity," "accepted techniques of creation," "confidentiality," "credible source and stewardship," and "reproducibility." These attributes are more closely related to the science domain rather than the IT domain, and also, as no surprise, are not directly addressed by IT tools. IT can help protect data from undesired forms of exposure, deletion, and modification, but has no knowledge about the data's significance or how it was created. Yet that sort of information, or *metadata*, is required to address these attributes.

6.2.1 Accepted Techniques of Creation and Reproducibility

As mentioned in the previous section, Barriers to Trustworthiness, the IT practice of regular patching can be at odds with reproducibility. Containerization technologies, especially those designed to run containers without elevated privileges, can help strike a middle-ground, where older software may be used in the container while the system running the container may be regularly patched by its system administrators. As an added benefit, most containerization technologies describe the programmatic contents of the container and its execution as an *image*, often a single computer file. This feature enables publishing the code in a manner that is likely to execute identically on another system with minimal extra effort from the Infrastructure Provider or Data User.

A second concern touched on in Section 5 relates to the increased chance of vulnerability discovery and exploitation if programs and their source code are made public. While such a concern may be appreciable in the domain of general IT, one should note that the discovery of any latent software flaws in the science domain is *desirable*, especially when such flaws can result in the injection of improperly-derived data into follow-on research or decision-making

processes. Making source code, workflows, and processes transparently available to the public will facilitate any such discovery and bolster trust in the resulting data.

6.2.2 Confidentiality

The attribute of confidentiality does not map directly onto the IT concept of confidentiality, and thus has weak coverage by IT tools used to assure confidentiality. IT tools to protect confidentiality tend to work at the granularity of a file (e.g., password-protected zip, encrypted document, file permissions), or storage media (e.g., BitLocker, LUKS). Our attribute of confidentiality involves finer-grained protection, with the ability to selectively control access to parts of a file, or fields in a record, possibly in an ad-hoc structure or one specific to a domain of science.

Encryption is the technology designed to protect confidentiality when other mechanisms, such as access controls, fail or are insufficient. Encryption is also easy to use improperly, such that it looks like it works, but fails to maintain confidentiality [14, 15, 16]. Because of the danger of this property, we feel that it is not sufficient to discourage writing one's own encryption routines or directly using cryptographic encryption primitives (e.g., AES encrypt/decrypt). If encryption is selected as a technology to help assure confidentiality in a custom software product, the designers should instead use an *abstraction* (i.e., software package) written, maintained, and/or reviewed by individuals or groups credible in the field of cryptography. Such an abstraction removes the most risky aspects of encryption from the software developer; exposing opaque elements rather than cryptographic nuances such as key generation, initialization vectors, random number generation, cipher modes, etc.[17]

In cases where data is structured and stored in a relational database management system (RDBMS), many popular RDBMSs implement column-level access permissions, encrypted tables, and sometimes encrypted columns. If the need to publish some data fields but not others exists, these features may help satisfy that need. A process utilizing RDBMS-provided encrypted columns and column-level permissions offers a stronger assurance of confidentiality than one using an RDBMS that only supports column-level access permissions.

In any case, one should keep in mind that an encryption key or an equivalent is still a secret, and needs to be kept confidential as well.

[17] One such abstraction is NaCl with ports to popular programming languages under the name libsodium. http://nacl.cr.yp.to/, https://en.wikipedia.org/wiki/NaCl_(software)

While traditional approaches to confidentiality require some trust in the Infrastructure Provider to protect (decrypted) data from unwanted exposure while the data is being processed, there is a growing area of research to alleviate the importance of this trust, or eliminate it completely. Those interested may look at the technologies of Data-Oblivious Computation, Secure Multiparty Computation. As an exercise in management of expectations, one should note that while these technologies are promising, they may still be waiting to be incorporated in a friendly, practical application.

However, a more effective assurance for confidentiality is to reduce the data that must be kept confidential. For example, if a dataset must have certain fields masked out during publishing, the Data Provider could publish a copy with those fields removed, rather than rely on the Infrastructure Provider to provide a solution to perform masking on-the-fly. Or, in cases where sensitive information is involved and the values in sensitive fields are not necessary for analysis, the Data User could remove or anonymize them from the dataset before putting a copy in resources provided by an Infrastructure Provider. Another option is the use of Differential Privacy technology, which enables publishing a dataset for statistical analysis without exposing individual data. Data Providers exercising these options would be keen to note the deviation from the original and provide a rationale of how the change does not impact the results.

6.2.3 Authenticity and Credible Source and Stewardship

Our definitions for credible source and stewardship, and authenticity have some overlap with confidentiality and attributes covered by general IT tools. Those overlapped components are not concerns in this section. Rather, we are concerned with the remainder, the need to establish ownership and provenance.

Journal archives and domain-specific data repositories, along with identifiers such as the Digital Object Identifier (DOI), allow one to see with limited scope how a dataset relates to other research and publications. However, what is often missing is a means to ensure a dataset is identical to the one the Data Provider published. At worst, the dataset exists in shared storage and one Data User trusts another Data User's word that the contents are authentic. At best, the Data Provider includes, along with references to the repository, some kind of authoritative metadata, such as a cryptographic hash or PGP signature. Although it does not go so far as including cryptographic hashes, ReproZip provides a high degree of reproducibility by bundling all necessary data files, libraries, environment variables and options in order to re-create the research into a self-contained package.[18]

[18] https://www.reprozip.org

Possibly one of the most impactful changes to the current practices regarding data publication would be for Data Providers to augment the metadata with instructions and information to independently authenticate the dataset. While not universally feasible due to the varying structure of datasets, adoption can be expedited by using, where applicable, a standard process used in IT, namely bundling a list of cryptographic hashes with the links to file downloads, as well as in metadata and in the references of any journal articles or papers using the dataset.

Providing a means of independent authentication not only helps bridge the gap in the chain of trust, but also aids in identifying situations where random data corruption has disrupted integrity. As pointed out by Gentz and Peisert [17], there are many sources of corruption that have nothing to do with malicious actors, and some have a high probability in introducing at least one error in larger datasets. A nice feature of this approach is that detection of corrupted or inauthentic data works regardless of how the inconsistency occurred.

6.3 Roles and Responsibilities

As suggested by the gap analysis at the beginning of this section, building and maintaining trustworthiness is a shared responsibility, not a delegated one. All roles have a part to play.

6.3.1 Data Users

With respect to the other roles, Data Users are in a position analogous to a customer, and in fact, are sometimes referred to as such. This places them in a position to request the availability of the tools and technologies discussed in this section. It is incumbent upon the Data User to make use of such tools when they are available. A Data User concerned with trustworthiness should:

- Work with Infrastructure Providers to set access controls that meet the requirements of Data Providers and their local Facilitation and Compliance Professionals.
- Request their Data Providers to publish and enable means of independently establishing the authenticity of their datasets.
- Independently establish the authenticity and integrity of the datasets they use, as close to processing as possible, preferably integral to processing.
- Work with local Facilitation and Compliance Professionals, and Infrastructure Providers to verify that the appropriate IT tools and technologies mentioned here are being used, and used effectively.

6.3.2 Data Providers

Similar to Data Users, Data Providers have a similar relation to the other roles, and thus similar set of responsibilities, with the addition of stewardship of their data. A Data Provider concerned with trustworthiness should:

- Adopt the same responsibilities as a Data User.
- Enable Data Users, who need to independently establish the authenticity of the Data Provider's datasets, in the immediate, as well as distant future.
- Work with local Facilitation and Compliance Professionals, and Infrastructure Providers to ensure that data remains available, accessible, and relatively free from error for the duration of the data's anticipated lifetime.
- Transparently publish their dataset, along with any software used to create or refine it, subject to contractual or legal restrictions.
- Use container images to produce ready-to-run distributions when publishing software.
- Work with local Facilitation and Compliance Professionals to identify sensitive information and define a process to protect it.

6.3.3 Infrastructure Providers

This role primarily supports the efforts of Data Users and Data Providers. However, it should not be thought of as oblivious to the goals of either of those roles, especially with respect to trustworthiness. Infrastructure Providers are the subject-matter experts for the tools and technologies mentioned here, and thus bear the brunt of the responsibility for laying the foundations of trustworthiness. An Infrastructure Provider interested in helping to lay this foundation should:

- Develop analysis software, workflow frameworks, and repositories that integrate and abstract the necessary components for independently authenticating datasets.
- Explore new and existing tools and technologies to help Data Users and Data Providers meet their responsibilities.
- Work with Data Users, Data Providers, and Facilitation and Compliance professionals to ensure that their selection of tools and technologies for building trustworthiness is neither deficient nor overzealous in meeting the needs of those other roles. For example, this can include meeting privacy policies, data use agreements, and intellectual property protections.

6.3.4 Facilitation and Compliance Professionals

While not a strictly technical role, Facilitation and Compliance Professionals help orchestrate the relationship between the other roles, define the bounds within which the other roles may operate, and importantly, are most likely to be situated in the organizational structure to

advocate for the needs of those other roles. A Facilitation or Compliance Professional concerned with trustworthiness should:

- Work with Data Users, Data Providers, and Infrastructure Providers to develop guidance (policies, guidelines) for *frustration-free* use of infrastructure, data publishing, protection, authentication, and stewardship.
- Advocate for changes to policy and resource allocation, within their organization and to funding bodies, to enable the use of tools and technologies for building trustworthiness in a manner that is cost-free and low-friction to Data Users and Data Providers.

7 Communicating Trustworthiness

Our survey results indicated some general themes about "data trustworthiness":

- Untrustworthy scientific results can cause reputational harm to researchers and organizations. This can reduce use of other data from the same source.
- Data quality is important, and this should be the focus of data management efforts (rather than specifically tied to trust) (e.g., Survey 2.24).
- Security guidance for regulated and sensitive data is important.

As discussed in Section 5, barriers to trustworthiness are context- and community-dependent. In general, perceived trustworthiness of data is improved by communication about the data per se, describing its source(s), the methods of generation, processing, transformations, and corrections. Collectively, this is often referred to as provenance information.

Data service organizations should provide clear policies and procedures about their data stewardship for the resources they provide, which can indicate trustworthiness of the organization or service. Further, providing rich and accurate metadata is essential to enabling stakeholders, either users or content distributors, to verify and independently assess the data provided.

Each of the stakeholder roles discussed in Section 2 have some need to assess trustworthiness of data in resources, both to communicate from providers to users, and for users to know what to expect from providers. Put most simply: the burden of proof falls to those who have data and either provide it directly or communicate results based upon it. Consumers of the data or results need to understand what has been done to establish and preserve trustworthiness in order to conduct their own evaluations. This section provides guidance for these stakeholders on how to communicate the trustworthiness of their data or resource to users and other stakeholders.

7.1 Communicating Attributes of Trustworthiness

In addition to the general practice of communicating provenance information, another way of addressing trustworthiness involves starting from the ground up, and communicating *how* data meets the requisite attributes for trustworthiness. These include the list from Section 3: *Availability, Integrity, Authenticity, Accepted techniques of creation, Authorization, Confidentiality, Credible source and stewardship, Reproducibility*. Below, we address methods for communicating trustworthiness in relation to these attributes.

Availability - Data providers should document and make available their processes for maintaining data and securing the data, as well as creating and publishing their long term preservation plans.

Integrity - Data users must be assured that the specific copy of the data in the resource has not been altered from the original source beyond those ways documented and communicated. This can be shown via checksums and by data repositories publishing their data integrity methods, i.e., for preserving the data against potential alterations and remediation for unauthorized or unintended modifications.

Authenticity - Beyond the techniques used to create and process the data, understanding the origin and lineage of the data reassures end users that it comes from a known source, to provide assurance that the resource is what it purports to be..

Accepted Techniques of Creation - Provide metadata on the processes and inputs used to create or modify the data, particularly when they are methods that are well-established and accepted by the community, indicating trustworthiness.

Authorization - Data providers should publish their policy on technologies used for authorization for access and modifications.

Confidentiality - Confidentiality for trustworthiness relates to whether personally identified information (PII) or other sensitive data is not available to those who are unauthorized to see it. Data providers should publish their policies on how data are restricted, what portions of the data are considered confidential, and what processes are required to allow users access to it.

Credible Source and Stewardship - The net result of providing details on a data set's original creation demonstrates that the data creator can be trusted. Responsible stewardship can be communicated through both formal and informal approaches. Data use communities may have informal assessments about which data resources are trustworthy, based on the preceding attributes. Resource providers may also wish to formalize community recognition through the use of third-party certification processes such as CoreTrustSeal.[19]

Reproducibility - Similar to using accepted techniques of creation, reproducibility focuses on other components provided alongside the data, such as information about processes/algorithms used to gather and analyze data, with enough detail that a subject matter expert in the same field could recreate the results.

7.2 Sharing Scientific Metadata

A significant aspect of responsible data sharing is the creation of information about the data, which describes their source attributes, processing, bias and error, who can access the data, how it is to be used, and the people and software involved in the data production. All of this kind of information is included in the general concept of "metadata," providing this content in an easily-accessible form is one key to communicating the trustworthiness of data. Another approach that is growing in the Open Science arena is to make Data Management Plans (DMPs) public. For some disciplines, the DMP is a tool for capturing the data handling routines, and processes for maintaining the integrity of the data across the research lifecycle.

Returning to metadata specifically, the scientific community is perhaps most aware of the metadata elements required for submitting "published" materials to a repository. While general repositories (such as academic Institutional Repositories) tend to require fairly minimal records, disciplinary repositories often extend metadata requirements to address specific domain standards that might include, for example, instrument types and software versions.

For example, the Institutional Repository at UIUC - *IDEALS* - provides a fill-in page for people who are depositing materials. In addition to accessible policies, they provide "Metadata Best Practices" guidance on completing the deposit form [18]. Other general repositories, such as Zenodo and Dryad, provide similar guidance. Domain specific repositories will have additional requirements to support data sharing and publishing requirements agreed to by a particular scientific community.

[19] https://www.coretrustseal.org/

There are many sources available (e.g. the California Digital Library [19], Cornell University Libraries [20], and the USGS materials on research data management [21]) to provide best practice recommendations on the creation of metadata for research data.

8 Conclusions

In recognition of the importance of trustworthy data to the scientific process, the working group members have developed this report to provide guidance to the community toward achieving the attributes that constitute data trustworthiness. Members of the scientific community fill multiple roles in the data management process, including data user, data provider, secure infrastructure provider, facilitator, and compliance officer. Our goal for this report is to provide to practitioners an understanding of the different attributes of data trustworthiness to assist in the selection, use, and support of tools and technologies to achieve those attributes.

For additional information about our working group, please visit https://www.trustedci.org/2020-trustworthy-data.

References

[1] Adams, Andrew; Avila, Kay; Basney, Jim; Cragin, Melissa; Dopheide, Jeannette; Fleury, Terry; ... Zage, John. (2020, June 30). Scientific Data Security Concerns and Practices: A survey of the community by the Trustworthy Data Working Group. Zenodo. http://doi.org/10.5281/zenodo.3924533

[2] Anon., "Trustworthiness," Glossary, Computer Security Resource Center, National Institutes of Standards and Technology, Accessed September 29, 2020, https://csrc.nist.gov/glossary/term/trustworthiness

[3] Lin, D., Crabtree, J., Dillo, I. et al. The TRUST Principles for digital repositories. Sci Data 7, 144 (2020). https://doi.org/10.1038/s41597-020-0486-7

[4] FAIR: https://www.go-fair.org/fair-principles/

[5] E. Deelman, V. Stodden, M. Taufer, and V. Welch, "Initial Thoughts on Cybersecurity And Reproducibility," in *P-RECS '19: Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer System*, Phoenix, AZ, 2019, pp. 13–15, doi: 10.1145/3322790.3330593.

[6] J. C. Wallis, C. L. Borgman, M. S. Mayernik, A. Pepe, N. Ramanathan, and M. Hansen, "Know Thy Sensor: Trust, Data Quality, and Data Integrity in Scientific Digital Libraries," in *Research and Advanced Technology for Digital Libraries*, Budapest, Hungary, 2007, pp. 380–391, doi: 10.1007/978-3-540-74851-9_32.

[7] L. Christopherson, A. Mandal, E. Scott, and I. Baldin, "Toward a Data Lifecycle Model for NSF Large Facilities," in *Practice and Experience in Advanced Research Computing*, Portland OR USA, Jul. 2020, pp. 168–175, doi: 10.1145/3311790.3396636.

[8] Reinhard Gentz and Sean Peisert, "An Examination and Survey of Random Bit Flips and Scientific Computing," Trusted CI Technical Report, December 20, 2019. https://hdl.handle.net/2022/24910

[9] M. Rynge *et al.*, "Integrity Protection for Scientific Workflow Data: Motivation and Initial Experiences," in *Proceedings of the Practice and Experience in Advanced Research Computing*

on Rise of the Machines - PEARC '19, Chicago, IL, USA, 2019, pp. 1–8. https://doi.org/10.1145/3332186.3332222

[10] J. Bhandari Neupane, R. P. Neupane, Y. Luo, W. Y. Yoshida, R. Sun, and P. G. Williams, "Characterization of Leptazolines A–D, Polar Oxazolines from the Cyanobacterium Leptolyngbya sp., Reveals a Glitch with the 'Willoughby–Hoye' Scripts for Calculating NMR Chemical Shifts," Org. Lett., vol. 21, no. 20, pp. 8449–8453, Oct. 2019. https://doi.org/10.1021/acs.orglett.9b03216

[11] S. Leonelli, "What difference does quantity make? On the epistemology of Big Data in biology," *Big Data & Society*, vol. 1, no. 1, p. 205395171453439, Jul. 2014, doi: 10.1177/2053951714534395.

[12] Mayernik, Matthew S., Kelsey Breseman, Robert R. Downs, Ruth Duerr, Alexis Garretson, Chung-Yi (Sophie) Hou, Environmental Data Governance Initiative (EDGI), and Earth Science Information Partners (ESIP) Data Stewardship Committee. 2020. "Risk Assessment for Scientific Data." Data Science Journal 19 (10): 1–15. https://doi.org/10.5334/dsj-2020-010

[13] C. Lagoze, "Big Data, data integrity, and the fracturing of the control zone," *Big Data & Society*, vol. 1, no. 2, pp. 1–11, Jul. 2014, doi: 10.1177/2053951714558281.

[14] 27th Chaos Communication Congress, Console Hacking 2010, PS3 Epic Fail, https://fahrplan.events.ccc.de/congress/2010/Fahrplan/events/4087.en.html

[15] The Heartbleed Bug, https://heartbleed.com

[16] CVE-2008-0166 [predictable random number generator], https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2008-0166

[17] Reinhard Gentz and Sean Peisert, "An Examination and Survey of Random Bit Flips and Scientific Computing," Trusted CI Technical Report, December 20, 2019. https://hdl.handle.net/2022/24910

[18] Anon., Bran Eveland, "Metadata Best Practices," IDEALS Resources and Information, University of Illinois at Urbana-Champaign, August 25, 2020, https://wiki.illinois.edu/wiki/display/IDEALS/Metadata+Best+Practices

[19] Strasser, C., Cook, R., Michener, W., & Budden, A. (2012). Primer on Data Management: What you always wanted to know. UC Office of the President: California Digital Library. doi:doi:10.5060/D2251G48 Retrieved from https://escholarship.org/uc/item/7tf5q7n3

[20] Anon. "Best Practices," Research Data Management Service Group, Cornell University Accessed September 29, 2020, https://data.research.cornell.edu/content/best-practices

[21] Anon. "Metadata Creation," Data Management, United States Geological Survey, Accessed September 29, 2020, https://www.usgs.gov/products/data-and-tools/data-management/metadata-creation

[22] R. Shirey, "Internet Security Glossary, Version 2," RFC 4949, Request for Comments, Networking Working Group, Internet Engineering Task Force, August 2007, https://tools.ietf.org/html/rfc4949

Appendix A - List of Terms

Access Control(s)
A process by which use of system resources is regulated according to a security policy and is permitted only by authorized entities (users, programs, processes, or other systems) according to that policy. [22, p.10, access control]

Archival Storage
A system designed for storing data for long periods of time. An archival storage system may or may not permit instantaneous modification or read-access to data (e.g., tape), however it must be designed with the intent to protect the integrity and availability of the data stored in it for the lifetime of the archive.

Network Control(s)
Access controls applied to a computer network. Enforcement of policy is generally accomplished through the use of a combination of dedicated network firewall devices, host-based firewall software, and judicious network service configuration.

RAID [backed] Filesystem
A filesystem built upon a redundant array of independent disks, configured in a manner that preserves the availability of data stored in the filesystem during a disk failure and while the failed disk's replacement is undergoing integration to the array.

Logging
An infrastructure for generating, collecting, and analyzing security events -- occurrences in a system that are relevant to the security of the system. [22, p. 268, security event]

Multi-Factor Authentication
An authentication process requiring proof of more than one of: something you know, something you have, something you are. A common example of multifactor authentication is the use of a password in conjunction with a proof of possession of a registered mobile phone by interacting with an application on the phone.

External Backups
The creation of a reserve copy of data [22, p. 31, back up] and storing that copy such that it is not accessible from (i.e., is external to) the computer system it was made from. Note that

backup copies, unlike archives, are meant as a contingency measure, and should be expected to have a short, finite lifespan.

Physical Security Protections
Tangible means of preventing unauthorized physical access to a system. Examples: Fences, cages, walls, and other barriers; locks, safes, and vaults; dogs and armed guards; sensors and alarm bells. [22, p. 222, physical security]

Intrusion Detection / Prevention
A process or subsystem, implemented in software or hardware, that automates the tasks of (a) monitoring events that occur in a computer network and (b) analyzing them for signs of security problems. [22, p. 165, intrusion detection system] An intrusion detection system may be augmented with an automated-response capability to interrupt detected security problems, creating an intrusion *prevention* system.

File/Host Integrity Check
The application of a (cryptographic) hash function to determine if the contents of files (data and/or program and/or configuration) have unexpectedly changed. A hash function is a (mathematical) function which maps values from a large (possibly very large) domain into a smaller range. [22, p. 139, hash function] Common examples of hash functions include sha256 and md5.

Policy for Network / Cloud Storage
Guidance and/or requirements that specify how data is to be stored and transported outside an organization's trust boundary. For example, confidential data should be encrypted when transferred over the Internet, or Amazon S3 buckets should not be publicly accessible.

3rd Party Data Repository
A system for storing and retrieving data, operated by an entity external to the Data User and Data Provider. We assume that 3rd-party data repositories are capable of ensuring the security of the data stored in them, subject to the limitations in their service agreement or use policy. We also assume that the Data Provider finds these limitations acceptable.

Workflow Integrity Checking
The integration of cryptographic hash functions within a scientific workflow to ensure that the data being operated on is the same as the data when it was originally collected, generated, or published.