Trust Concerns for Data Repositories: The COVID-19 Data Domain

Technical Report

Glenda Amaral[1], Tiago Prince Sales[1], Giancarlo Guizzardi[1], João Paulo A. Almeida[2], and Daniele Porello[3]

[1]Conceptual and Cognitive Modeling Research Group (CORE), Free University of Bozen-Bolzano, Bolzano, Italy
{gmouraamaral,tiago.princesales,giancarlo.guizzardi}@unibz.it
[3]Ontology & Conceptual Modeling Research Group (NEMO), Federal University of Espírito Santo, Vitória, Brazil
jpalmeida@ieee.org
[3]ISTC-CNR Laboratory for Applied Ontology, Trento, Italy
daniele.porello@loa.istc.cnr.it

July, 2020

Contents

1 Introduction

The COVID-19 pandemic created a global human crisis, which can ultimately be seen as a data-driven crisis. Over the past months, a huge number of datasets on COVID-19 have sprung up, yet not all of them are reliable, leading to potentially poor decision making with consequences for millions. As argued in [1] "as the international community responds to an outbreak of coronavirus [...] early and open data sharing — which are vital for its control — depend on the trust that the data will not be used without proper attribution to those who generated it". Furthermore, according to the Research Data Alliance COVID-19 Working Group [4], "during a pandemic like COVID-19, it is important to concentrate efforts on scrutinising reliable data sources that provide data and metadata of high quality and guarantee the authenticity and integrity of the information". The RDA Group argues that in addition to having a certification (such as the CoreTrusSeal [6]) or accreditation, trustworthy repositories should consider a wide range of community-based standardised quality criteria, best practices, and principles (e.g. TRUST Principles [3]).

The importance of understanding what composes trust in a particular COVID-19 data repository, has motivated us to conduct an investigation of the characteristics a COVID-19 repository should have in order to be held in a position of trust by the communities they intend to serve. We address this issue by making a comparison between the trust-related concerns about data repositories in the domain of COVID-19 raised in [5] and three acknowledged approaches to trustworthy information sources, namely the TRUST Principles for Digital Data Repositories [3], the CoreTrustSeal [6] and the Requirements of Trustworthy AI defined by the European High-Level Expert Group on Artificial Intelligence (AIHLEG) [2].

This report is structured as follows: Section 2 presents the TRUST Principles for Digital Data Repositories [3], the CoreTrustSeal [6] and the AIHLEG Requirements of Trustworthy AI [2]. In Section 3, we correlate the trust-related concerns about COVID-19 data repositories presented in [5] to the three approaches discussed in Section 2, focusing on in identifying similarities and verifying to which extend they are aligned with well established initiatives on the trustworthiness of information sources. We conclude in Section 4, with some final considerations.

2 Approaches to Trustworthy Information Sources

In this section, we present three approaches to trustworthy information sources, namely the TRUST Principles for Digital Data Repositories [3] (Section 2.1), the CoreTrustSeal [6] (Section 2.2) and the Requirements of Trustworthy AI defined by the European High-Level Expert Group on Artificial Intelligence (AIHLEG) [2] (Section 2.3). Whilst a comprehensive review of all initiatives on this subject is beyond the scope of this report, we here attempt to provide some of the most prominent approaches.

2.1 The TRUST Principles for Digital Data Repositories

The TRUST Principles for Digital Data Repositories [3] stand for "a set of guiding principles to demonstrate digital repository trustworthiness", collaboratively developed and endorsed by several stakeholders, representing various segments of the digital repository community. These principles were conceived in the context of the Open Science discourse, aiming at providing "a common framework to facilitate discussion and implementation of best practice in digital preservation by all stakeholders" [3]. Table 1 presents the five TRUST Principles for Digital Data Repositories. For a complete description of the TRUST Principles approach the reader should refer to [3].

Table 1: The TRUST Principles for Digital Data Repositories [3]

P01	Transparency	To be transparent about specific repository services and data holdings that are verifiable by publicly accessible evidence.
P02	Responsibility	To be responsible for ensuring the authenticity and integrity of data holdings and for the reliability and persistence of its service.
P03	User focus	To ensure that the data management norms and expectations of target user communities are met.
P04	Sustainability	To sustain services and preserve data holdings for the long-term.
P05	Technology	To provide infrastructure and capabilities to support secure, persistent, and reliable services.

2.2 The CoreTrustSeal

The *CoreTrustSeal Trustworthy Data Repository Requirements* [6], were developed by the DSA–WDS Partnership Working Group on Repository Audit and Certification, a Working Group of the Research Data Alliance, aiming at creating a set of harmonized common trustworthy requirements for certification of repositories at the core level. Table 2 presents the CoreTrustSeal requirements. For a detailed description of these requirements reader should refer to [6].

Table 2: CoreTrustSeal Trustworthy Data Repositories Requirements [6]

C01	Provide access to and preserve data in its domain.
C02	Maintains applicable licenses for data access and use and monitors compliance.
C03	Has a continuity plan to ensure ongoing access to and preservation of its holdings.
C04	Ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.
C05	Has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.
C06	Adopts mechanism(s) to secure ongoing expert guidance and feedback.
C07	Guarantees the integrity and authenticity of the data.
C08	Accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.
C09	Applies documented processes and procedures to manage data storage.
C10	Assumes responsibility for long-term preservation and manages this function in a planned and documented way.
C11	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.
C12	Archiving is made according to defined workflows from ingest to dissemination.
C13	Enables users to discover the data and refer to them in a persistent way through proper citation.
C14	Enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.
C15	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.

2.3 AIHLEG Requirements of Trustworthy AI

In the context of AI systems the High-Level Expert Group on Artificial Intelligence (AIHLEG), elaborated a set of ethics guidelines for trustworthy AI, as part of the European Strategy on Artificial Intelligence [2]. In its framework for achieving trustworthy AI, the AIHLEG identifies seven requirements for Trustworthy AI that should be met and offers guidance on the potential methods that can be used to realize it. They are presented in Table 3. The reader interested in thorough description of the AIHLEG Requirements of Trustworthy AIUFO is referred to [2].

Table 3: AIHLEG Requirements of Trustworthy AI [2]

T01	Human agency and oversight	Fundamental rights, human agency and human oversight
T02	Technical robustness and safety	Resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
T03	Privacy and data governance	Respect for privacy, quality and integrity of data, and access to data
T04	Transparency	Traceability, explainability and communication
T05	Diversity, non-discrimination and fairness	Avoidance of unfair bias, accessibility and universal design, and stakeholder participation
T06	Societal and environmental wellbeing	Sustainability and environmental friendliness, social impact, society and democracy
T07	Accountability	Auditability, minimisation and reporting of negative impact, trade-offs and redress.

3 Trust Concerns for COVID-19 Data

The COVID-19 pandemic is a global problem that requires cross-disciplinary strategies, combining scientific advances in areas such as virology, pharmacology, epidemiology, information technology, among others. The virus outbreak has led to a proliferation of data sharing initiatives to encourage research collaboration through digital platforms. Relying on digital data sources is of particular value, however there are concerns about the quality of data and publications performed under enormous time pressure and provided in near real-time.

In order to shed some light on how to separate the wheat from the chaff, in [], the authors raise some questions that should be taken into consideration when evaluating the trustworthiness of a particular COVID-19 information source. Table 4 presents a compilation of these trust concerns.

Table 4: Trust concerns for Covid-19 data [5]

Data representativeness transparency
Analytic methods transparency
Algorithms transparency
Data provenance transparency
Data semantics transparency
Data quality level transparency
Share interpretations uncertainties
Data is provided in a meaningful, interpretable format
Respect for privacy
Respect for human rights
Based on acknowledged scientific facts
Fairness
Follows ethical guidelines and is compliant with local and international law
Data and code are shared in safe and responsible ways
Staff Expertise
Quality of data fits the purpose

We compare the trust concerns for COVID-19 data presented in [5] to the TRUST principles [3], the CoreTrustSeal requirements [6] and the AIHLEG requirements [2], in order to find a common ground. Table 5 shows the comparisons results and correlations found.

Table 5: Trust concerns for Covid-19 data [5], TRUST Principles, CoreTrustSeal, and AIHLEG Requirements similarities

Trust concerns for Covid-19 data	**TRUST Principle**	**CoreTrustSeal Requirement**	**AIHLEG Requirement**
Data representativeness transparency	(P01)Transparency	(C08)Ensures relevance and understandability (C11)Has appropriate expertise to address technical data and metadata quality and sufficient information is available for end users	(T04) Transparency
Analytic methods transparency			
Algorithms transparency			
Data provenance transparency			
Data semantics transparency			
Data quality level transparency			
Share interpretations uncertainties			
Data is provided in a meaningful, interpretable format			
Respect for privacy	(P02)Responsibility	(C02)Maintains applicable licenses for data access and use	(T03)Privacy and data governance
Respect for human rights	(P02)Responsibility	(C04)Compliance with disciplinary and ethical norms	(T01)Human agency and oversigh
Based on acknowledged scientific facts	(P02)Responsibility	(C07)Guarantees the integrity and authenticity of the data; (C08)Ensure relevance and understandability	(T02)Technical robustness and safety
Fairness	(P02)Responsibility; (P03)User Focus	(C04)Compliance with disciplinary and ethical norms	(T05)Diversity,non-discrimination and fairness
Follows ethical guidelines and is compliant with local and international law	(P02)Responsibility; (P03)User focus	(C02)Maintains applicable licenses for data access and use; (C04)Compliance with disciplinary and ethical norms	(T03)Privacy and data governance; (T05)Diversity, non-discr., fairness; (T06)Societal and environmental well-being
Data and code are shared in safe and responsible ways	(P02)Responsibility; (P04)Sustainability; (P05)Technology	(C01)Provide access to and preserve data in its domain (C02)Maintains applicable licenses for data access and use (C03)Has a continuity plan to ensure ongoing access to data (C10)Assumes responsibility for long-term preservation in a planned and documented way	(T02)Technical robustness and safety; (T03)Privacy and data governance; (T06)Societal and environmental well-being
Staff Expertise	(P05)Technology	(C05) Has adequate funding and qualified staff	(T02)Technical robustness and safety
Quality of data fits the purpose	(P05)Technology	(C07)Guarantees the integrity and authenticity of the data; (C11)Has appropriate expertise to address technical data and metadata quality and sufficient information is available for end users	(T02)Technical robustness and safety

4 Conclusions

In this report we presented a comparison that highlights the resemblance among different approaches to the trustworthiness of information sources. Moreover, we correlated these approaches to a set of trust concerns about COVID-19 data, in order to to find a common ground.

Due to the heterogeneous nature that data may have, just one set of specific principles or requirements may not be enough to deal with the particularities of different contexts. So, finding correspondences among them can be helpful in adding new evaluation guidelines.

Observing comparisons on Table 5, each of the sixteen trust concerns extracted from [5] is related to at least one principle or requirement of the other three approaches. Moreover, by similarities perspective, it is possible to identify that some concerns are a common ground among the four sets of principles, such as transparency, privacy of data, compliance to ethical norms, respect for human rights and data quality. In the future, we plan to expand our analysis to investigate the correlation between the initiatives presented here and further approaches, focused on the COVID-19 data domain.

References

[1] Rapid outbreak response requires trust. Nature Microbiology 5(2), 227–228 (Jan 2020)

[2] Hleg, A.I.: Ethics Guidelines for Trustworthy AI. B-1049 Brussels (2019)

[3] Lin, D., Crabtree, J., Dillo, I., Downs, R.R., Edmunds, R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., et al.: The trust principles for digital repositories. Scientific Data 7(1), 1–5 (2020)

[4] Research Data Alliance: Rda covid-19 guidelines and recommendations (2020)

[5] Satchit Balsari, C.B., Khanna, T.: Which covid-19 data can you trust? Harvard Business Review (2020)

[6] Standards, C., Board, C.: CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022 (Nov 2019)