IJDC | *Brief Report*

TROV - A Model and Vocabulary for Describing Transparent Research Objects

Meng Li, Timothy M. McPhillips, Craig Willis, Nikolaus Parulian, and Bertram Ludäscher
School of Information Sciences, University of Illinois at Urbana-Champaign

Kacper Kowalik
NCSA, University of Illinois at Urbana-Champaign

Lars Vilhuber
Labor Dynamics Institute, Cornell University

Thu-Mai Lewis and Mandy Gooch
Odum Institute for Research in Social Science, University of North Carolina, Chapel Hill

Abstract

The Transparent Research Object Vocabulary (TROV) is a key element of the *Transparency Certified* (TRACE) approach to ensuring research trustworthiness. In contrast with methods that entail repeating computations in part or in full to verify that the descriptions of methods included in a publication are sufficient to reproduce reported results, the TRACE approach depends on a controlled computing environment termed a *Transparent Research System* (TRS) to guarantee that accurate, sufficiently complete, and otherwise trustworthy records are captured when results are obtained in the first place. Records identifying (1) the digital artifacts and computations that yielded a research result, (2) the TRS that witnessed the artifacts and supervised the computations, and (3) the specific conditions enforced by the TRS that warrant trust in these records, together constitute a *Transparent Research Object* (TRO). Digital signatures provided by the TRS and by a trusted third-party timestamp authority (TSA) guarantee the integrity and authenticity of the TRO. The controlled vocabulary TROV provides means to declare and query the properties of a TRO, to enumerate the dimensions of trustworthiness the TRS asserts for a TRO, and to verify that each such assertion is warranted by the documented capabilities of the TRS. Our approach for describing, publishing, and working with TROs imposes no restrictions on how computational artifacts are packaged or otherwise shared, and aims to be interoperable with, rather than to replace, current and future Research Object standards, archival formats, and repository layouts.

Submitted 18 June 2024 ~ *Accepted* 9 July 2024

Correspondence should be addressed to Timothy M. McPhillips, 501 E Daniel St 493, Champaign, IL 61820. Email: tmcphill@illinois.edu

This report was presented at International Digital Curation Conference IDCC24, Edinburgh, 19-21 February 2024

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v19i1.1019
DOI: 10.2218/ijdc.v19i1.1019

Introduction

Research communities across the sciences increasingly require that authors of research publications make the methods yielding computational results transparent and thereby subject to scholarly review, namely by sharing the data, code, and computational workflows used to obtain those results (Willis & Stodden, 2020). Verifying that the computational artifacts provided by authors in fact represent those that were employed to produce reported results generally has proved troublesome, however. A common approach—used for example by the journals of the American Economic Association[1] and the Odum Institute[2] in support of journal data and code policies, and supported by platforms such as Binder (Jupyter et al., 2018) and Whole Tale (Brinckman et al., 2019)—is to attempt to reproduce results using the provided artifacts. This method of enforcing transparency has two fundamental limitations. First, it is not always possible to obtain access to the necessary datasets (Hrynaszkiewicz, Harney, & Cadwallader, 2021). Digital humanities research may rely on copyrighted data hosted by the HathiTrust Research Center (HTRC) (Murdock et al., 2017); health data employed in biomedical research may be subject to protection under HIPAA[3] (Nass, Levit, & Gostin, 2009); and other kinds of data may similarly be prevented from being openly or easily shared. Second, it is not always practical or even possible to gain access to the requisite computational resources. Examples of this are common in AI-related fields that require significant computational power for training and inference (Pineau et al., 2021), but also in disciplines that employ geoinformatics methods (Kray, Pebesma, Konkol, & Nüst, 2019) or rely on bootstrap or other probabilistic inference procedures requiring substantial computing resources. The alternative to verifying research transparency via explicit reproduction of computational results is to ensure (1) that sufficiently complete and accurate records are collected when results are originally obtained, and (2) that these records are subsequently shared with publishers in a reliable manner. The *Transparency Certified* (TRACE) model[4] formalizes this approach.

TRACE

The TRACE model and Transparent Research Object Vocabulary (TROV) together provide the technical means both to identify the computations and artifacts employed in deriving a reported result, and to justify why these records should be considered trustworthy. TRACE depends crucially on the concept of a *Transparent Research System* (TRS), a computing environment that can reliably (1) record the identities and arrangements of computational artifacts—including datasets, custom scripts, and required third-party software—employed during the computations it hosts; (2) enforce one or more specific conditions on computations to ensure that records captured are both accurate and complete; and (3) package records of supervised computations and identities of associated computational artifacts in the form of a *Transparent Research Object* (TRO) that can be readily examined and programmatically queried.

The TRACE model further adopts public key encryption technologies[5] and associated security practices to ensure, for example, that a TRO was in fact created by the TRS it claims produced it. A TRS digitally signs each TRO it produces using a certificate that describes the capabilities of the TRS and the specific conditions it can enforce on the computations it supervises. Moreover, a TRO is considered valid only if the associated TRS signature is itself signed by a trusted third-party timestamp authority (TSA)[6]. These two digital signatures enable

[1] AEA data and code policies guidance: https://www.aeaweb.org/journals/data

[2] Journal verification framework: https://odum.unc.edu/archive/journal-verification

[3] HIPAA: Health Insurance Portability and Accountability Act of 1996

[4] TRACE project: https://transparency-certified.github.io

[5] Public-key cryptography: https://csrc.nist.gov/glossary/term/public_key_cryptography

[6] RFC 3161 Time-Stamp Protocol: https://www.ietf.org/rfc/rfc3161.txt

anyone with access to the TRO, the TRS certificate, and the public key of the TSA to verify that the TRO is authentic and trustworthy, i.e. that the TRO was in fact signed by the TRS, that the TRO was not modified since it was signed, and that the TRS certificate was valid at the time the TRO was produced.

Enumerating within each TRS certificate the capabilities that a TRS provides for ensuring research transparency facilitates tailoring of the TRACE model to the requirements of different research communities. These requirements, and the corresponding TRS capabilities addressing them, may be expected to vary significantly depending on the research domain; the type, scale, and management of computing resources; size of data, methods of data access, and the practicality of persisting input and intermediate data streams; and a multitude of other variables. Our design allows TRS capabilities to vary significantly as needed. Moreover, by providing means to declare which of the numerous, diverse dimensions of research trustworthiness the TRS ensured during the production of a TRO and the artifacts the TRO describes (McPhillips, Willis, Gryk, Nuñez-Corrales, & Ludäscher, 2019), TRACE sidesteps the complexities faced by attempts to standardize the meanings of terms such as *reproducible* and *replicable* across diverse research domains (National Academies of Sciences, Engineering, and Medicine, 2019).

The descriptions of computations and artifacts that comprise TROs and the declarations of TRS capabilities included in TRS certificates both are articulated using the Transparent Research Object Vocabulary (TROV), a representation of the conceptual model illustrated in Figure 1 and described more fully in the next section. TROV generally is expressed using RDF[7] and is readily serialized in JSON-LD format, queried using SPARQL[8], and validated using the shapes constraint language (SHACL[9]) (Gayo, Prud'hommeaux, Boneva, & Kontokostas, 2018; Pareti & Konstantinidis, 2022). TROV complements the W3C PROV-O[10] ontology for describing general provenance relationships; TROs produced by a TRS with provenance capture capabilities likely will employ both the TROV and the PROV vocabularies. Finally, TROV aims to be easy to integrate with existing RDF-based Research Object standards (Bechhofer, De Roure, Gamble, Goble, & Buchan, 2010; Soiland-Reyes et al., 2022; Ton That, Fils, Yuan, & Malik, 2017).

Conceptual Model

The *Transparent Research System* (TRS) and *Transparent Research Object* (TRO) concepts were introduced in the preceding section. Definitions of the remaining concepts that may be expressed using TROV are as follows. Figure 1 highlights the relationships between these concepts, while Figure 2 in the following section depicts a concrete instantiation of this model.

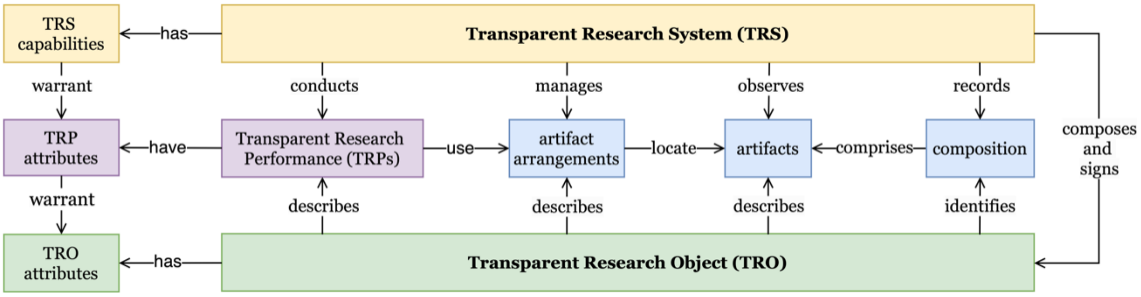

Figure 1. TRACE concepts that may be expressed using TROV. See text for definitions of each concept and the meanings of the relationships between them.

[7] RDF proposed as a W3C Recommendation: https://www.w3.org/RDF/

[8] SPARQL11 proposed as a W3C Recommendation: https://www.w3.org/TR/sparql11-query/

[9] SHACL proposed as a W3C Recommendation: https://www.w3.org/TR/shacl/

[10] W3C PROV-O: https://www.w3.org/TR/2013/REC-prov-o-20130430/

The centerpiece of the TRACE conceptual model is the *Transparent Research Performance* (TRP). A TRP enacts one or more research activities, typically a set of interrelated computations, under the supervision of a TRS. Activities comprising a TRP may access one or more *artifacts* (e.g. data files, scripts, parameter files, program binaries) organized according to one or more *artifact arrangements*, and may update or create new artifacts within these arrangements. An essential (i.e. mandatory) supervisory function of any TRS is to observe the artifacts present in the arrangements accessible to a TRP before and after the TRP is conducted. In combination with (optional) TRS capabilities ensuring that the artifacts in arrangements available to TRPs are not modified (and new artifacts not added) by activities external to the TRP, observation of the initial and final state of each arrangement suffices to establish which artifacts represent products of the TRP and to identify those artifacts from which these products might have been derived. Because representing a multistep computational workflow as a sequence of distinct TRPs reduces the number of dependencies compatible with the before and after states of the arrangements accessible to each TRP, we expect that a typical TRO will describe a sequence of TRPs that access artifacts and arrangements introduced or updated by prior TRPs.

Support for multiple artifact arrangements additionally facilitates recording the multiple locations (distinct resource paths) associated with key artifacts at different stages of the overall workflow. An artifact representing an input dataset might have one resource path representing the public URL from which it is downloaded, a second resource path representing the destination of the downloaded artifact on a local computer, and a third resource path representing the location of the artifact in an archive (e.g. zip) file that eventually will be shared with a publisher for review. Associating these different locations within distinct artifact arrangements permits verifying that each script included in a TRO employs the appropriate path when accessing the artifact; simply associating multiple paths with each artifact on an individual basis (i.e. in the absence of explicit artifact arrangements selectively made available to particular TRPs) would not accomplish this.

The TRACE model does not require that computational artifacts are packaged with a TRO declaration that refers to them. Because each artifact may be associated with multiple resource paths, artifacts are identified on the basis of (e.g. SHA256 [11]) digests of the bits comprising the artifacts. In this way, artifacts associated with a TRO declaration but not bundled with it (e.g. datasets that are very large or that cannot be shared publicly) may be recognized later by those who gain access to those artifacts. The conglomeration of all digital artifacts described by a TRO is termed the *TRO composition*. The *composition fingerprint*, a digest of the sorted digests of each of the individual artifacts in the composition, is a mandatory element of all TROs. Composition fingerprints facilitate identifying TROs that describe the same composition.

The relationships between the*TRS capabilities*, *TRP attributes,* and *TRO attributes* occupying the left side of Figure 1 illustrate how claims about the trustworthiness of a TRO are declared, justified, and verified. TRO attributes represent declarations of the dimensions of trustworthiness asserted by the TRS about the TRO. The hypothetical TRO attribute *IncludesAllInputData* might indicate that the TRO composition encompassess all of the data ultimately needed to derive the results represented by the TRO. No TRO attribute can be asserted without justification, however. One or more attributes of the TRPs described by the TRO must warrant each TRO attribute. A hypothetical TRP attribute *IsolatedFromInternet* might, in part, justify the *IncludesAllInputData* TRO attribute (by precluding access to data outside the artifact arrangements available to the TRP). But in turn, each TRP attribute asserted in a TRO must be justified by a TRS capability. A TRS capability such as *CanProvideInternetIsolation* would justify the *IsolatedFromInternet* TRP attribute. The chain of justifications would stop here; a simple query of a TRS certificate is sufficient to verify that the TRS is justified in claiming any particular capability.

Note again that different TRS implementations, deployments, and configurations may enforce a wide variety of very different conditions on the computations and data management operations performed under their supervision. They also may employ very different means to enforce a particular condition. One TRS may enforce the condition of network isolation via

[11] RFC 4868 SHA256: https://datatracker.ietf.org/doc/html/rfc4868

low-level system observability techniques that verify that no network connections were established during enactment of a TRP; a second TRS may enforce the same condition by fully isolating computations and artifact arrangements from the network automatically; while a third may simply confirm that the network interface was disabled by the user for the duration of the TRP.

Demonstration

Figure 2 represents elements of a TRO featured in the demonstration repository associated with this paper[12]. The repository includes TROV-expressed JSON-LD representations of the TRO declaration and TRS certificate; representative SPARQL queries and results; and scripts for validating the RDF artifacts according to rules expressed in SHACL, executing the SPARQL queries using RDFLib[13], and generating standardized reports using the Geist[14] templating toolkit.

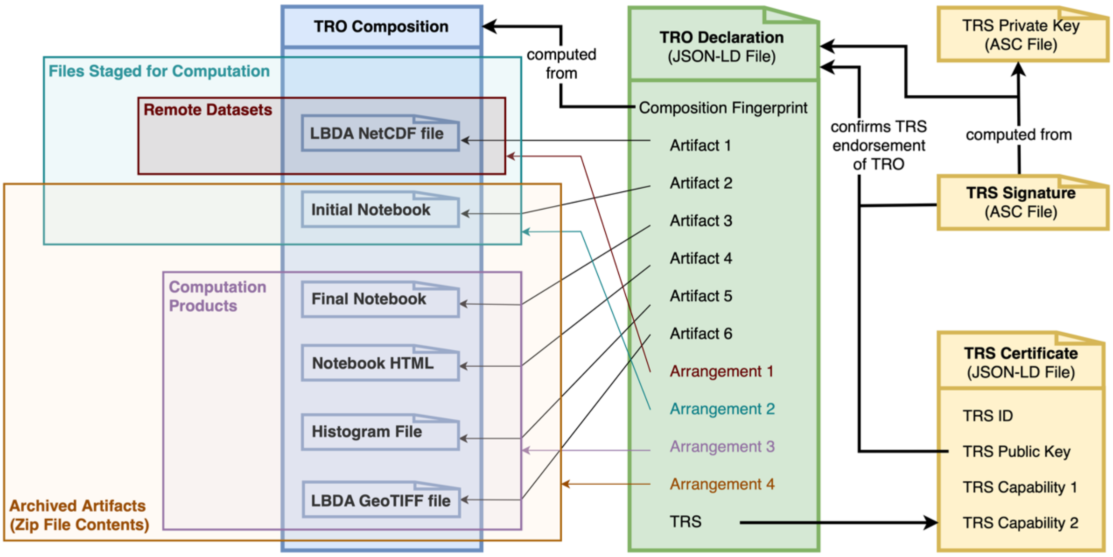

Figure 2. Instance diagram illustrating elements of a TRO, the associated TRS certificate, and the digital signature provided by the TRS. All artifacts are included in the demonstration repository.

This TRO describes a computational workflow comprising three TRPs. The first downloads the Living Blended Drought Atlas (LBDA) dataset[15] from the NOAA web server; the second analyses the downloaded dataset by invoking a Jupyter notebook; and the third packages a subset of the artifacts in a zip file. Each of the six digital artifacts that comprise the TRO composition are included in at least two different artifact arrangements. Note that while the LBDA NetCDF file is not included in the final zip file, the SHA256 digest of this file included in the TRO declaration enables anyone downloading the dataset from the NOAA web site to confirm that it is identical to the file described in the TRO.

[12] Demonstration: https://github.com/transparency-certified/trov-demos

[13] RDFLib package: https://github.com/RDFLib/rdflib

[14] Geist documentation: https://cirss.github.io/geist-p

[15] LBDA dataset: https://www1.ncdc.noaa.gov/pub/data/paleo/drought/LBDP-v2

Acknowledgements

This work was supported by NSF Awards 2209628, 2209629, and 2209630.

References

Bechhofer, S., De Roure, D., Gamble, M., Goble, C., & Buchan, I. (2010). Research Objects: Towards Exchange and Reuse of Digital Knowledge. *Nature Precedings*, 1–1. https://doi.org/10.1038/npre.2010.4626.1

Brinckman, A., Chard, K., Gaffney, N., Hategan, M., Jones, M. B., Kowalik, K., … Turner, K. (2019). Computing environments for reproducibility: Capturing the "Whole Tale." *Future Generation Computer Systems*, *94*, 854–867. https://doi.org/10.1016/j.future.2017.12.029

Gayo, J. E. L., Prud'hommeaux, E., Boneva, I., & Kontokostas, D. (2018). *Validating RDF Data*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-79478-0

Hrynaszkiewicz, I., Harney, J., & Cadwallader, L. (2021). A Survey of Researchers' Needs and Priorities for Data Sharing. *Data Science Journal*, *20*(1). https://doi.org/10.5334/dsj-2021-031

Jupyter, P., Bussonnier, M., Forde, J., Freeman, J., Granger, B., Head, T., … Willing, C. (2018). *Binder 2.0—Reproducible, interactive, sharable environments for science at scale*. 113–120. Austin, Texas. https://doi.org/10.25080/Majora-4af1f417-011

Kray, C., Pebesma, E., Konkol, M., & Nüst, D. (2019). Reproducible Research in Geoinformatics: Concepts, Challenges and Benefits. *DROPS-IDN/v2/Document/10.4230/LIPIcs.COSIT.2019.8*. Presented at the 14th International Conference on Spatial Information Theory (COSIT 2019). Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPIcs.COSIT.2019.8

McPhillips, T., Willis, C., Gryk, M. R., Nuñez-Corrales, S., & Ludäscher, B. (2019). Reproducibility by Other Means: Transparent Research Objects. *2019 15th International Conference on eScience (eScience)*, 502–509. https://doi.org/10.1109/eScience.2019.00066

Murdock, J., Jett, J., Cole, T., Ma, Y., Downie, J. S., & Plale, B. (2017). Towards Publishing Secure Capsule-Based Analysis. *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 1–4. Toronto, ON, Canada: IEEE. https://doi.org/10.1109/JCDL.2017.7991585

Nass, S. J., Levit, L. A., & Gostin, L. O. (Eds.). (2009). *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*. Washington, D.C.: National Academies Press. https://doi.org/10.17226/12458

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and Replicability in Science*. Washington, D.C.: National Academies Press. https://doi.org/10.17226/25303

Pareti, P., & Konstantinidis, G. (2022). A Review of SHACL: From Data Validation to Schema Reasoning for RDF Graphs. In M. Šimkus & I. Varzinczak (Eds.), *Reasoning Web. Declarative Artificial Intelligence: 17th International Summer School 2021, Leuven, Belgium, September 8–15, 2021, Tutorial Lectures* (pp. 115–144). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-95481-9_6

Pineau, J., Vincent-Lamarre, P., Sinha, K., Lariviere, V., Beygelzimer, A., d'Alche-Buc, F., … Larochelle, H. (2021). Improving Reproducibility in Machine Learning Research(A Report from the NeurIPS 2019 Reproducibility Program). *Journal of Machine Learning Research*, *22*(164), 1–20.

Soiland-Reyes, S., Sefton, P., Crosas, M., Castro, L. J., Coppens, F., Fernández, J. M., … Goble, C. (2022). Packaging research artefacts with RO-Crate. *Data Science*, *5*(2), 97–138. https://doi.org/10.3233/DS-210053

Ton That, D. H., Fils, G., Yuan, Z., & Malik, T. (2017). Sciunits: Reusable Research Objects. *2017 IEEE 13th International Conference on E-Science (e-Science)*, 374–383. Auckland: IEEE. https://doi.org/10.1109/eScience.2017.51

Willis, C., & Stodden, V. (2020). Trust but Verify: How to Leverage Policies, Workflows, and Infrastructure to Ensure Computational Reproducibility in Publication. *Harvard Data Science Review*, *2*(4). https://doi.org/10.1162/99608f92.25982dcf