Signed Citations: Making citations of digital scientific content persistent

Michael J Elliott[‡], Jorrit H Poelen[§,|], Jose AB Fortes[‡]

‡ University of Florida, Gainesville, United States of America
§ Ronin Institute, Montclair, NJ, United States of America
| UC Santa Barbara Cheadle Center for Biodiversity and Ecological Restoration, Santa Barbara, CA, United States of America

Corresponding author: Michael J Elliott (mielliott@ufl.edu)

Abstract

Digital data are a foundation of 21st century science. In order to maintain a stable foundation, the FAIR Guiding Principles (Wilkinson et al. 2016) were proposed to keep data findable, accessible, interoperable, and reusable (FAIR). However, commonly used data citation practices rely on unverifiable retrieval methods that do not always enable access to the cited data. Without verifiability, retrieval methods are susceptible to undetected "content drift", which occurs when the data associated with an identifier have been allowed to change. In the presence of content drift, cited data may lose their findability.

We propose signed citations, i.e., customary data citations extended to include a standards-based, secure, unique, and fixed-length digital content signature. A content signature is a code that is unique to the data it identifies and can be reliably recovered from the data. For example, the signature of a dataset could be the SHA-256 hash (Dang 2015) of its content. We show that the inclusion of content signatures in citations not only enables independent verification of the cited content, but also can improve the reliability and availability of the citation, allowing the cited data to remain findable for longer periods of time and across changing online infrastructures.

If a content signature registry is available which links content signatures to one or more (possibly temporary) known content locations, then content signatures can themselves be used to find identified data. That is, registries make content signatures "resolvable" just like URLs and DOIs. Additionally, signed citations are location- and storage-medium-agnostic, allowing the making of as many copies of cited data as necessary to ensure content persistence across current and future storage media and data networks. As a result, content signatures can be leveraged to help scalably store, locate, access, and independently verify content across new and existing data repositories, search engines, and registries (such as those that exist within services offered by Zenodo, DataOne, and

the Software Heritage archive) without requiring any time-sensitive information (e.g. URLs or references to specific infrastructures) to be baked into the citation.

Signed citations can also be used to reliably identify complex data networks and knowledge graphs. By embedding content signatures inside content and then citing that content with a signed citation, a secure (unforgeable, irrevocable, self-verifying) link is formed between the cited content and those identified by embedded content signatures. Such links create secure data graphs that are annotatable and machine-traversable, acting as a mechanism for manual and automated discovery, which are vital to findability according to the FAIR guidelines (Wilkinson et al. 2016). Additionally, entire knowledge graphs can be similarly securely cited using a single signed citation.

Our proposal originates from our earlier work on reliable dataset identifiers (Elliott et al. 2020). In addition to further discussing signed citations as stated above, we expand upon our previous work by describing real-world examples of the use of content signatures, including signed citations of a corpus of digitized images of bee specimens from natural history collections, datasets which collectively contain over a billion records available through global biodiversity data networks, and a corpus of taxonomic name resources. Our use of signed citations in these real-world examples offers a starting point for the development of community standards on how to build, use, and support independent yet interoperable signature-based services such as content registries, repositories, and search indexes.

Keywords

citation standards, data persistence, verification, provenance

Presenting author

Michael J Elliott

Presented at

TDWG 2022

Funding program

This work was funded by grants from the National Science Foundation (Michael Elliott and Jose Fortes were funded by DBI 202765, Jorrit Poelen was funded by OAC 1839201) and the AT&T Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the AT&T Foundation.

Conflicts of interest

References

- Dang Q (2015) Secure Hash Standard. National Institute of Standards and Technology https://doi.org/10.6028/nist.fips.180-4
- Elliott M, Poelen J, Fortes JB (2020) Toward reliable biodiversity dataset references. Ecological Informatics 59 https://doi.org/10.1016/j.ecoinf.2020.101132
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, ’t Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18