

cited using a single fixed-length content signature. Trusty URIs use this feature of content signatures to build and identify provenance graphs for publications, encoded in RDF, which conveys structured descriptions of the properties of, and relationships between, uniquely identified resources. Then, Trusty URIs that are resolvable (e.g., embedded in URLs) can be followed to discover the origins of a publication²³. Such provenance graphs are robust because the content signatures used to link content can only ever identify those exact contents, such that the no content referenced in the graph (including the RDF that forms the graph) can be altered (to the extent that the identifier distinguishes between objects, e.g., adding or removing whitespace does not alter a dataset's Trusty URI) without changing the hash of the entire graph. This property is also fundamental to the design of blockchains, most notably used by cryptocurrencies such as Bitcoin to create and evolve robust transaction ledgers¹³, and can be leveraged in scientific citations to allow the origins of scientific findings to be more reliably traced as well as allow the authors of published datasets to be more readily credited for their contributions.

Contributions and origin of work. Evidence for the reliability of content signatures as references has been described in⁵, where they were used to detect and quantify link rot and content drift over time for biodiversity dataset URLs registered with several data aggregators. We found that 20% to 75% of dataset URLs exhibited link rot or content drift over a span of two years. Specifically, link rot was detected in 5% to 70% of URLs, and content drift in 0.05% to 66%. Although we did not investigate the frequencies of link rot and content drift for persistent identifiers such as DOIs, we realized that such identifiers did not generally provide any means of verifying the identity of associated data, whereas content signatures could. In this paper, we build upon our earlier work⁵ by formalizing the role of content hashes in content signatures used to construct signed citations and documenting real-world usages of content signatures to create repeatable, reproducible workflows and robust provenance graphs. We discuss how content signatures enable verifiable content retrieval and discovery using decentralized content registries, repositories, and search indexes, with the potential for leveraging existing infrastructures.

Methods

Signed citations are robust. A content signature is a content identifier that contains the following two components: a cryptographic hash of the identified content, and a description of which cryptographic hashing algorithm was used to compute the hash. A signed citation is simply a data citation that includes a unique and verifiable content signature of the cited data. To create a signed citation, the following two criteria need to be fulfilled:

1. the referenced content is digital
2. means are available to calculate a content hash of that digital content (e.g., via cryptographic hashing algorithms such as SHA-256⁶)

A signed citation can then be formed by including all the elements of a traditional citation (e.g., author, title, publisher, and publication date) as well as the content signature for the cited data. In example 1, below, we illustrate the construction of a signed citation for an image of a bee specimen.

Example 1. In this example, we verifiably cite a digital image of the bee specimen MCZ:Ent:17219 held at the Museum of Comparative Zoology, Harvard University. A rendering of the image is shown in Fig. 1, below. Because the image is digital, its content hash can be computed using a cryptographic hashing algorithm. We chose to use the SHA-256 algorithm to compute a hash of the image and formed the following content signature:

```
hash://sha256/edde5b2b45961e356f27b81a3aa51584de4761ad9fa678c4b9fa3230808ea356
```

which can be listed alongside the author, date, title, and location the content was retrieved from to form the following signed citation:

```
Museum of Comparative Zoology, Harvard University. 2021. Head Frontal View of MCZ:ENT:17219 Nomadopsis puellae (Cockerell, 1933) hash://sha256/edde5b2b45961e356f27b81a3aa51584de4761ad9fa678c4b9fa3230808ea356 Accessed at http://mczbase.mcz.harvard.edu/specimen\_images/entomology/large/MCZ-ENT00017219\_Spinoliella\_puellae\_hef.jpg on 2021-12-07.
```

A future reader who finds the cited content may verify its identity by recomputing the content hash listed in the signed citation.

The signed citation in example 1 begins as a traditional citation, attributing the source of the image and its publication date, along with descriptive text. We then supplement this information with a content signature of the image. For convenience, we also cite the Internet location from which we retrieved the image, as well as the date of retrieval. Note that hash://sha256/edde5b2b45961e356f27b81a3aa51584de4761ad9fa678c4b9fa3230808ea356 is the content signature of the image, expressed as a content-hash URI, which has the general form of hash://[hash function]/[hash value] (e.g., hash://sha256/abc...). The notation provides an intuitive way to indicate that the hexadecimal sequence (where each character can be a digit or one of the letters in the set {a, b, c, d, e, f}) is a content hash and which hash function was used to generate the hash (e.g., sha256 indicates the SHA-256 algorithm). The hash function takes binary content as input and computes a content hash. In the above content signature, the binary content is the digital



Fig. 1 Headshot of *Nomadopsis puellae* (Cockerell, 1933) specimen MCZ:Ent:17219, used with permission from the Museum of Comparative Zoology, Harvard University, ©President and Fellows of Harvard CC-BY-NC-SA 4.0. The image was not modified from its original form. The image has the SHA-256 content signature hash: // sha256/edde5b2b45961e356f27b81a3aa51584de4761ad9fa678c4b9fa3230808ea356. A signed citation for this image is provided in example 1.

image from example 1, and the SHA-256 algorithm was used to generate the 256-bit hash that is encoded in the content signature as a 64-character hexadecimal sequence (each character representing 4 bits).

Hash functions are deterministic; given a specific digital content as input, a hash function will always produce the same hash, regardless of when, where, or by whom the hash is calculated. This means that the association between a hash and digital content can be verified by re-running the hash function to reproduce the hash. Furthermore, secure hash functions such as SHA-256 provide statistical guarantees that no two inputs will be assigned the same hash⁶. In other words, if a hash function produces different hashes for two pieces of digital content, their contents must be non-identical. Therefore, a hash is uniquely associated with exactly one digital content. By using content signatures that specify both a hash and a hash function, the association between a content signature and digital content is both unique and verifiable.

By including content signatures that are unique, verifiable, and location-agnostic, signed citations are resistant to both link rot and content drift in the sense that both can be detected and potentially repaired. If locations (e.g., URLs) listed in a signed citation become inaccessible, the reader may consult a content signature registry (if one is available) to look up alternative locations of the content identified by the content signature. If content is retrieved, content drift can be detected by recomputing its hash and checking whether it differs from the cited hash, in which case the reader can attempt to find an alternative location. Because the uniqueness and persistence²⁷ of content signatures as identifiers cannot be corrupted by link rot and content drift, we say that signed citations are robust.

Recursive signed citations form robust citation graphs. Now that we've established and exemplified a standards-based (e.g., using SHA-256⁶, URI²⁸, and hexadecimal notation²⁹) way to verifiably cite a single digital image, we will show by example how to cite a collection of digital data containing multiple images and associated metadata.

Example 2. A basic form of a collection description may look like a reference list:

Museum of Comparative Zoology, Harvard University. 2021. Head Frontal View of MCZ:ENT:17219 *Nomadopsis puellae* (Cockerell, 1933) hash://sha256/edde5b2b45961e356f27b81a3aa51584de4761ad9fa678c4b9fa3230808ea356 Accessed at http://mczbase.mcz.harvard.edu/specimen_images/entomology/large/MCZ-ENT00017219_Spinoliella_puellae_hef.jpg on 2021-12-07.

Museum of Comparative Zoology, Harvard University. 2021. Habitus Lateral View of MCZ:ENT:17219 *Nomadopsis puellae* (Cockerell, 1933)

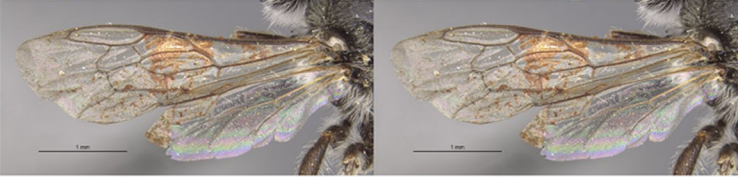
Specimen Record	Media Record
urn:uuid:16f6de61-541c-46b1-b7e7-d8303e6f53db@jhpoelen.nl urn:uuid:16f6de61-541c-46b1-b7e7-d8303e6f53db@idigbio.org <pre>{ "dwc:identificationRemarks": "Labels: Bodega Calif Apl 3; type 17219; Spinoliella puellae Kll cotype.", "dwc:specificEpithet": "puellae", "dwc:countryCode": "US", "dwc:county": "Sonoma", "dwc:recordedBy": "[no agent data]", "dwc:georeferenceSources": "GEOLocate", "dwc:order": "Hymenoptera", "dcterms:references": "http://mczbase.mcz.harvard.edu/guid/MCZ:Ent:17219", { "dwc:individualCount": "1",</pre>	urn:uuid:6e0122c9-28b7-4bd2-9a0c-46bb3346c713@jhpoelen.nl urn:uuid:6e0122c9-28b7-4bd2-9a0c-46bb3346c713@idigbio.org  sha256:05b41ea4707614b20... sha256:05b41ea4707614b20... location of clone or original location of clone or original register clone or original register clone or original find more copies find more copies <pre>"xmpRights:WebStatement": "http://creativecommons.org/licences/by-nc-sa/3.0/",</pre>

Fig. 2 A screenshot of a machine-generated interactive web page visualizing the collection description cited in example 3. The web page displays 10 specimen records and their 39 associated media records, as well as information about where the data came from, their content signatures, and where they can be accessed. The web page was accessed at <https://jhpoelen.nl/bees>. The depicted photograph is included with permission from the Museum of Comparative Zoology, Harvard University, ©President and Fellows of Harvard CC-BY-NC-SA 4.0. The image was not modified from its original form.

hash://sha256/8d49bd24f6ba300b4de44fd218b53294f4cc0106cd9631018ef819b38345c75d Accessed at http://mczbase.mcz.harvard.edu/specimen_images/entomology/large/MCZ-ENT00017219_Spinoliella_puellae_hal.jpg on 2021-12-07.

A signed citation can be generated for the entire collection description by computing the SHA-256 hash of an ASCII encoding³⁰ of the text, then including a content signature (in this case, a content-hash URI) that specifies the hash and the hashing algorithm:

Some author, 2021. A collection of signed citations for various images. Hash://sha256/fe21dbf7e3ac1f9f82afa303a927015ada16ff84571e1fe21914c7053f00fb59 Accessed at <https://example.org/citations.txt> on 2021-12-08.

Because the collection description in example 2 robustly cites the two images using signed citations, the signed citation of the collection description also robustly cites the images. After retrieving the cited collection description and verifying its content signature, correct retrieval of the cited images can be similarly verified using the content signatures listed in the collection description. This demonstrates that signed citations can recursively cite other signed citations to form robust, traversable citation graphs. The process demonstrated in the example can be used to robustly cite any number of contents using a single signed citation.

Robust citation graphs can be annotated to form evolving knowledge graphs. A more elaborate version of the digital collection description in example 2 can be considered. Citation graphs formed by recursive signed citations can be annotated with descriptions of cited objects and their semantic relationships with other cited objects to form robust knowledge graphs. Example 3, below, cites a collection description that includes a semantic representation of the provenance (or origination) of bee images using the PROV Ontology³¹. This representation includes a description of the image origins and the software tools used to discover and collect them, namely a tool called Preston³² and the iDigBio Web Application Programming Interface (i.e., iDigBio's Web API, <https://search.idigbio.org>). The collection description cited in example 3 includes the same references as the one in example 2, but expresses the observed origins (e.g., the discovery process, including location, date of observation, and intermediate data) of the images in the machine-readable Resource Definition Framework³³ (RDF). Whereas example 2 is geared toward human readers, example 3 is easy for machines to read so that, for example, it may be used to re-generate example 2 using programmatic text manipulation or RDF query languages such as SPARQL³⁴, or even generate an interactive web page (Fig. 2) containing renderings of the images and descriptions of how they were discovered.

Example 3. The following is a signed citation of a machine-readable collection of 39 digital bee images³⁵:

A biodiversity dataset graph: <https://jhpoelen.nl/bees>. 2020. hash://sha256/85138e506a29fb73099fb050372d8a379794ab57fe4bdf141743db0de2b985c

The collection uses PROV Ontology to describe how and from where the images were collected. This information is expressed using machine-readable RDF triples (i.e., a list of statements, each expressing a relationship between two objects) encoded as N-Quads³³. A sample from the collection text is provided below.