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 publication[23]. 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 ledgers[13], 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 in[5], 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 work[5] 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-256[6])

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