Toward Reliable Biodiversity Dataset References

Michael J. Elliott[1†], Jorrit H. Poelen[2†*], José A.B. Fortes[1]

[1] Advanced Computing and Information Systems Laboratory (ACIS)

Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL

339 Larson Hall, PO Box 116200, Gainesville, Florida 32611-6200, USA

[2] 400 Perkins St Apt 104, Oakland, California, USA

[†] These authors contributed equally to this work

[*] Corresponding author

Toward Reliable Biodiversity Dataset References

Abstract

No systematic approach has yet been adopted to reliably reference and provide access to digital biodiversity datasets. Based on accumulated evidence, we argue that location-based identifiers such as URLs are not sufficient to ensure long-term data access. We introduce a method that uses dedicated data observatories to evaluate long-term URL reliability.

From March 2019 through May 2020, we took periodic inventories of the data provided to major biodiversity aggregators, including GBIF, iDigBio, DataONE, and BHL by accessing the URL-based dataset references from which the aggregators retrieve data. Over the period of observation, we found that, for the URL-based dataset references available in each of the aggregators' data provider registries, 5% to 70% of URLs were intermittently or consistently unresponsive, 0% to 66% produced unstable content, and 20% to 75% became either unresponsive or unstable.

We propose the use of cryptographic hashing to generate content-based identifiers that can reliably reference datasets. We show that content-based identifiers facilitate decentralized archival and reliable distribution of biodiversity datasets to enable long-term accessibility of the referenced datasets.

Keywords— Biodiversity, Ecological Informatics, Information Systems, Information Retrieval

Introduction

Over the course of hundreds of years, naturalists and biologists have systematically collected physical evidence from an ever-changing natural world. Through well-established protocols and institutional support, many of these natural history collections have withstood the ravages of time (Davis and Schmidt 1996, Hortal et al. 2015). Records that describe these carefully collected specimens are now made available digitally through online search indices, registries, and data archives (Page et al. 2015). The increased availability of digital natural history records helps realize Charles Elton's vision of "[linking] up into some complete scheme the colossal store of facts about natural history which has accumulated up to date in this rather haphazard manner" (Elton 1927). So far, various initiatives have succeeded in providing comprehensive aggregate views from previously scattered natural history record siloes (Edwards 2000, GBIF 2019, Matsunaga et al. 2013, Michener et al. 2011, Rinaldo and Norton 2009). However, we show that these aggregate views are subject to change as their underlying digital source data changes or becomes inaccessible. Although efforts have been made to track changes in datasets with versioning, last-modified dates (Robertson et al. 2014, Wieczorek et al. 2012), and periodic archiving (Costello et al. 2013), no systematic approach has been adopted to keep our digital natural history record accessible. Despite centuries of expertise in preserving our physical natural history records, biologists currently struggle to maintain a growing body of digital data that can change or disappear with the push of a button.

Our scholarly record consists of an intricate web of associations between scientific studies and the datasets on which they are based. These associations are made explicit through citations that can be used to reconstruct a study's context and provide the chain of evidence that supports its claims (Garfield et al. 1964). In the pre-Internet era, the lookup of cited references required access to one or more of the many academic libraries in the world. With the rise of Internet-accessible scientific publications, authors and readers access these references by using a networked device to download content from publication websites. This means that researchers are increasingly citing online works to support their claims. Because the citation format of online works typically documents only when (e.g., 2019-10-01) and where (e.g., https://doi.org/10.123/456) the referenced work was accessed by the author (DataONE 2012, GBIF 2019, iDigBio 2016), the reader expects the web-accessed resource to remain accessible and unaltered via this single web location. Readers may attempt to find a version of the works referenced by searching online data repositories for the matching author and title, but there is no guarantee that information found this way will be exactly the same as what was originally referenced. Any reference that does not allow readers to find the referenced work fails to satisfy the first FAIR principle of findability: "F1. (meta)data are assigned a globally unique and eternally persistent identifier" (Wilkinson et al. 2016). Our study supports Klein's and Vision's findings that networked, location-based access to digital objects is an unreliable mechanism for providing continued access to the unaltered original work (Klein et al. 2014, Vision 2010). Unless we change the way we preserve and cite our digital scholarly works, the web of knowledge that forms the basis of our scientific record will degrade.

Problem Characterization

The current practice of using Uniform Resource Locators (URLs) (Berners-Lee et al. 1994) to reference online biodiversity datasets provides no guarantee of continued data accessibility. This uncertainty jeopardizes the integrity of the scholarly record. When data access is lost, documented research results may become impossible to reproduce and the justification for conclusions or hypotheses that rely on lost results may be undermined.

Biodiversity data aggregators, such as DataONE, GBIF, and iDigBio, rely on data providers such as data curators and institutional repositories to maintain active dataset URLs, and aggregate the data found at those URLs for distribution in response to user queries. From here on, we use the term "data network" to refer to a collection of URLs that are discoverable through some central URL registry, and the term "provider network" to refer to the subset of URLs in a biodiversity aggregator's data network from which the aggregator retrieves data.

Relying on URLs to locate and identify referenced data carries the risk of link rot and content drift (Klein et al. 2014). Link rot occurs when a URL, or link, that had previously responded to queries can no longer be reached. This can happen, for example, due to temporary outages, URL retirement, or URL migration. A link exhibits content drift when a query to the link provides content that is different from the content it provided in the past. The extent of content drift can vary; content may have received only minor edits with no changes in semantics, or it may reference a different entity altogether. When a single URL is used to locate data that may change over time, a particular data version may become inaccessible over time. In one study on the *Genetics* journal, it was reported that 40% of links (URLs) to supplemental materials became unavailable due to link rot within one year of publication (Vision 2010). Another study (Klein et al. 2014) confirmed that as many as one in five Science, Technology, and Medicine articles contained references that exhibit "reference rot," which includes either link rot or content drift.

In this paper, we propose a methodology for measuring the existence of link rot and content drift in online data networks, then provide experimental results that confirm the existence of link rot and content drift in the provider networks of BHL, DataONE, iDigBio, and GBIF. Finally, we propose a method for referencing and serving biodiversity data in a way that works toward satisfying the Findable, Accessible, Interoperable, and Reusable (FAIR) principles (Wilkinson et al. 2016).

Methodology

While previous studies focus more generally on reference rot of URLs cited in scientific works (Klein et al. 2014, Vision 2010), our study provides quantitative evidence that reference rot occurs in biodiversity provider networks. Because reference rot occurs in the scope of individual data references, and references to digital datasets rely on URLs to locate the data, we begin by introducing terminology for characterizing the reliability of a URL according to how often it exhibits link rot and content drift.

URL Reliability

We assume that the URLs used to reference biodiversity datasets are expected to resolve to an Internet Protocol (IP) (Postel 1981) address via the Domain Name System (Mockapetris 1987). If a web server is accessible at the resolved IP address, a query (i.e., HTTP get request) to that address over the Hypertext Transfer Protocol (HTTP) will return a response code and, in some cases, associated content (Berners-Lee et al. 2005). We classify the reliability of a URL according to the content, or lack of content, that it provides over successive queries. If a query to a URL is unsuccessful, we say that link rot has occurred. However, if a successful response is received but the retrieved content is different from the content retrieved by previous query, we say that content drift has occurred. Monitoring URLs in this way allows us not only to determine whether link rot and content drift occur, but also to capture their long-term behaviors. For example, one URL that has exhibited link rot might have failed to respond only once, whereas another might have become consistently unresponsive. Likewise, one URL might exhibit content drift less frequently than another whose contents change rapidly. Furthermore, various combinations of link rot and content drift behavior may indicate that one URL is more reliable than another, even though both exhibit reference rot.

We label URLs with sets of reliability indicators according to their link rot and content drift behaviors. The defined reliability indicators are differentiated by the degree of link rot and content drift observed over a series of queries to the URL at different points in time. We characterize the responsiveness of a URL according to whether it exhibits link rot:

- Unresponsive: the link has failed to respond to one or more queries
- Responsive: the link has responded to all recorded queries

We characterize the stability of a URL according to whether it produces different content from one query to the next:

- Unstable: the content that the link points to sometimes changes
- Stable: the content that the link points to never changes

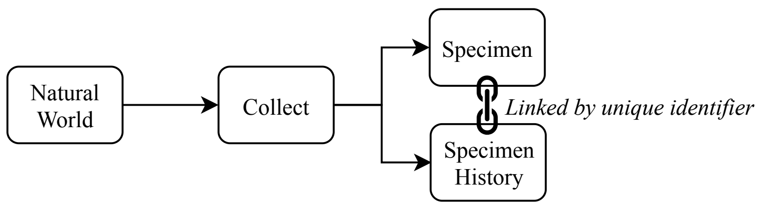
We characterize the overall reliability of a URL according to both its responsiveness and stability:

- Unreliable: the link does not always provide the expected content; it is either unresponsive, unstable, or both
- Reliable: the link always provides the expected content; it is both responsive and stable

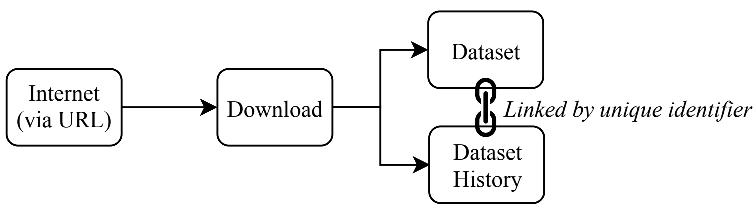
In order to determine the reliability of any given URL over time, we must monitor its behavior by documenting how it responds to periodic queries. We propose a method for monitoring URL behavior in the Data Collection Over Time section of this paper. First, however, we must propose a method for documenting a URL's response to a single query. For the context of biodiversity, we consider the case in which any content that a URL produces is a dataset.

The Data Collection Process

We suggest that digital dataset collection practices have some analogies to well-established physical specimen collection procedures (see figure 1) (Poelen 2019d). If datasets are considered analogous to specimens, then the URLs that locate datasets on the Internet are analogous to the physical locations of specimens in the natural world; they are where digital datasets were originally found, but not where they should be preserved. Once found, physical specimens are collected by hand; similarly, digital datasets are downloaded by querying their URLs. Once a specimen is collected and deposited to a safe, accessible repository, a record is kept that documents what the specimen is in addition to when, where, and by whom it was collected.



(a) Physical specimen collection



(b) Digital data collection

Figure 1. Reliable record keeping for digital datasets (b) can be achieved in an analogous way to current practices in record keeping for physical specimens (a). Biologists collect physical specimens from the natural world, thoroughly document the process, then store the specimens in facilities equipped for long-term preservation. Analogously, digital datasets that are downloaded from the internet can be thoroughly documented and archived in dedicated repositories for long-term preservation. Just as the collection of physical specimens is recorded and identified in specimen information records, the downloading of digital datasets can also be recorded and identified in dataset provenance records.

The same can be done for downloaded datasets. When a dataset is downloaded, a record can be kept that details the URL that was queried, the time of query, and who (e.g., a human or software agent) issued the query that initiated the download event; we refer to this record as the dataset's provenance record (Pasquier et al. 2017). Additionally, the dataset itself should be stored in a safe, accessible dataset archive so that it may be retrieved at a later date if needed. The final step in the collection process is to link the preserved specimen to its corresponding record (see figure 1(a)) via an assigned unique identifier.

The identifiers assigned to datasets must differ only if the contents of their datasets differ. This can be achieved by deriving the identifiers from the contents of their datasets. Furthermore, the identifier must be unique to the dataset; a dataset will always be assigned the same identifier and no two datasets (including different versions of a dataset) can share an identifier. Cryptographic hashing is one such method for producing content-based identifiers which are both content-derived and unique. A variety of cryptographic hashing algorithms exist that receive some digital file as input and uniquely encode its contents into a fixed-length series of bits called a "hash." We use hashes generated by the SHA-256 algorithm (NIST 2001) as unique content-based identifiers. For example, given two different bits of text, "first example" and "second example", their computed SHA-256 hashes (in hexadecimal format) are b84283f1f4cb997eae b28dce84466678ea611824ac97978749b158d2cd3886ac and c64eee387ccc1d04 38765129a8c423dab0b67d094710e395ac3193c52591a3ba, respectively. These hashes are the only ones that can possibly be computed from the example texts using the SHA-256 algorithm, and no other input to the SHA-256 algorithm can produce either of these specific hashes (NIST 2013). One benefit of the SHA-256 algorithm is that its computation time and space requirements scale linearly and remain constant, respectively, with the amount of data being hashed (NIST 2001). That is, computing a hash for a dataset that is twice as big as another dataset should take twice as long but use the same amount of memory. This is important for the biodiversity domain, where large media files such as computed tomography (CT) scans may consist of terabytes of data (Keklikoglou et al. 2019). Another benefit is that all SHA-256 hashes have the same length, regardless of the amount of data being hashed; a hash computed for a terabyte-sized CT scan is no longer than the hash computed for "first example".

Content-based identifiers that meet the requirements we have described are reliable references; they are not susceptible to either link rot or content drift. Additionally, the derivation of the content-based identifier for a given dataset can be performed by anyone, anywhere, and at any time. There is no need for some central authority to generate and assign identifiers, as is the case for non-content-based identification schemes (Paskin 1999). Therefore, dataset provenance can be collected in a decentralized manner; if two agents collect provenance for the same dataset acquired from potentially different locations, they can both reference the dataset using the same content-based identifier without any need for coordination. In this scenario, the two provenance records produced by the two agents can also be uniquely identified by using content-based identifiers in the same manner as we identify and reference datasets. We elaborate on uses for identifying and referencing provenance records in the discussion section of this paper.

Data Collection Over Time

By establishing a dedicated data observatory, we can build a history for each observed URL to capture its reliability over time. Such an observatory periodically queries URLs in a data network and produces for each URL two complementary parts: 1) an archived copy of the response to the corresponding query, whether it was a dataset, an error code, or no reply at all, and 2) a record of its provenance, including the URL itself, the current date, and a content-based identifier of any dataset received. Successive provenance records can be aggregated to construct comprehensive histories for both datasets (when and where they were found) and URLs (which datasets they located over a series of queries over time).

The constructed URL histories can be analyzed to determine whether a link was ever broken, when it was broken, and whether it became responsive again. The logs also identify the content (or lack of content) that a URL located each time it was queried. Any change in the content identifier from one query to the next indicates a change in the content of the dataset. These link breakages and content changes correlate to link rot and content drift, respectively, and allow us to determine the responsiveness, stability, and reliability of each URL over time.

URL Reliability in Data Networks

Our method for monitoring the behavior of a single URL over time can be applied to monitor all URLs in a data network. We also extend the idea of URL reliability to data networks and propose that the overall reliability of a set of URLs in a data network can be evaluated by monitoring the reliability of each URL over time. First, we label individual URLs with binary indicators of responsiveness, stability, and reliability at each time they were queried. Next, we characterize data networks according to the percentages of URLs that are assigned each of the reliability indicators. For example, if a data network contains three distinct URLs and we find that only two out of the three are reliable, then we say 67% of the URLs in the data network are reliable.

Experiment

The Preston biodiversity dataset tracker (Poelen et al. 2018) implements mechanisms for monitoring URLs in provider networks. It allows users to deploy a data observatory that discovers URLs in the provider network of a biodiversity aggregator, queries each URL for data, documents the data collection process, then archives the results. All crawl activities, the queries they issue, and the results they produce are recorded in a string of provenance logs. It is important to note that the URLs in provider networks are the sources of the datasets ingested by aggregators, not necessarily the datasets served by the aggregators, which may have been altered to, for example, to add alternate taxonomic information ([GBIF] Global Biodiversity Information Facility 2019b).

We deployed several Preston observatories to monitor the provider network URLs registered in Biodiversity Heritage Library (BHL), Data Observation Network for Earth (DataONE), Global Biodiversity Information Facility (GBIF), and Integrated Digitized Biocollections (iDigBio). The provider network URLs for DataONE, GBIF, and iDigBio were queried monthly from March 2019 through May 2020. The BHL provider network was queried monthly from May 2019 through May 2020.

The logs taken by each of these observatories describe the URL queries and their results, which were processed to produce the results that follow. To analyze the full set of URLs observed across all four provider networks, an fifth observatory was constructed by aggregating the provenance records produced by the four provider network observatories. In an effort to minimize artificial link rot due to Internet access issues in our local network, we deployed the Preston observatories in a large commercial data center in Germany.

Results

Breakdowns of the overall reliabilities of the sets of URLs observed within the provider networks are provided in table 1. Results are listed as percentages and total counts of URLs in the provider network that were assigned each reliability indicator. When analyzing the recorded results of queries to URLs in each provider network, we found that, for each individual network, 5% to 70% of registered URLs were intermittently or consistently unresponsive, 0% to 66% produced unstable content, and 20% to 75% became either unresponsive or unstable over the period of observation.

We found that 43% of URLs observed across the four provider networks became unreliable at some point over the period of observation. Of those unreliable URLs, 41% were unstable, 11% became consistently unresponsive, and 71% were at best only intermittently responsive. For 5% of successful queries, the URL failed to respond to the next query. For 4% of successful queries, the URL provided different content in response to the next successful query.

Provider Network	Responsive URLs	Stable URLs*	Reliable URLs
BHL[a]	29.99% (77,040)	99.95% (241,243)	29.97% (76,998)
DataONE[b]	92.54% (394,568)	87.11% (367,957)	80.30% (342,363)
GBIF[c]	73.93% (60,564)	33.93% (22,491)	24.53% (20,093)
iDigBio[c]	86.80% (5,988)	61.99% (4,265)	54.41% (3,754)
All observed URLs**	69.62% (534,107)	86.46% (632,879)	57.43% (440,606)

Table 1. Overall responsiveness, stability, and reliability for URLs observed in each aggregator's provider network and for all observed provider network URLs as of May 2020. Numbers in brackets indicate total URL counts. *URLs that never provided content were omitted from the denominator when calculating Stable URLs percentages. **Because URLs may be registered in more than one provider network, the total number of observed URLs is expected to be less than the sum of the URL counts for each network. [a]Poelen (2020a) [b]Poelen (2020b) [c]Poelen (2020c)

The changes in reliability over time for each provider network are visualized in figure 2. Note that because we have defined reliable URLs to be those considered both responsive and stable, they always represent the smallest fraction of URLs in table 1, figure 2, and figure 3. Figure 3 visualizes the cumulative growth of biodiversity provider networks during their periods of observation. This growth is illustrated with two metrics: the cumulative total number of unique URLs observed in each network and the cumulative total number of unique contents that were downloaded from the network at each monthly sampling.

The behaviors of the distributions over time of responsive, stable, and reliable URLs vary notably between provider networks. Reasons for these differences might be inferred when cross-examining table 1 and figures 2 and 3. For example, although the set of URLs observed in the BHL provider network scored relatively low in responsiveness due to frequent

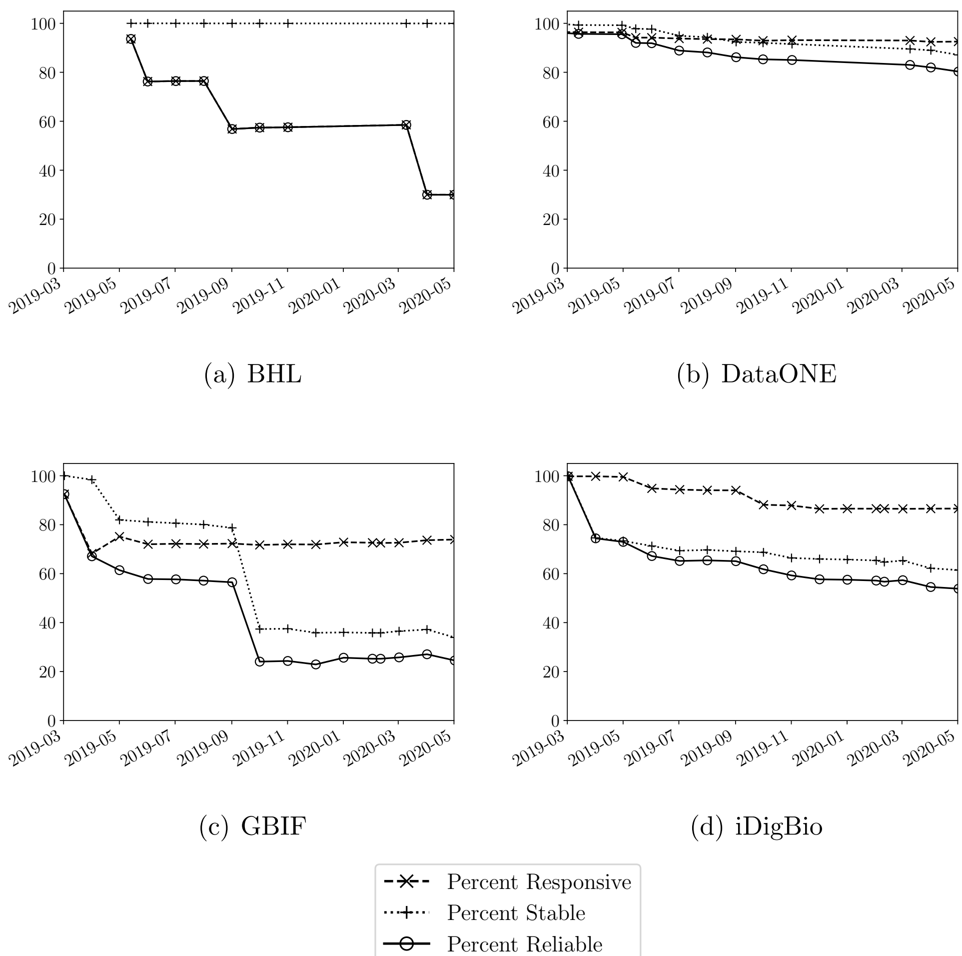
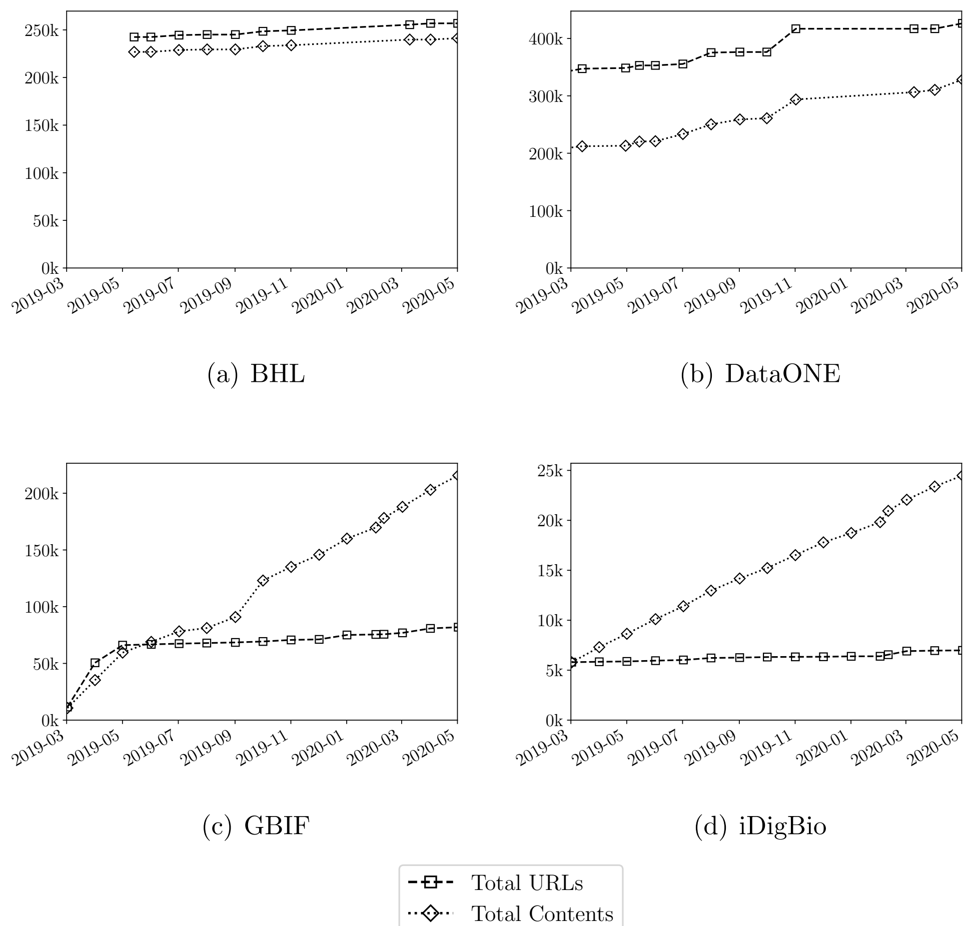


Figure 2. Overall responsiveness, stability, and reliability from March 2019 through May 2020 as percentages of URLs that exhibit each indicator in the provider networks of (a) BHL, (b) DataONE, (c) GBIF, and (d) iDigBio.

link rot, they were more stable than the provider network URLs of other aggregators because content drift within the BHL provider network is relatively rare. Conversely, although URLs observed in the iDigBio provider network were relatively responsive, they scored low in stability because the network's near-constant content growth far outpaces its URL



(a) BHL (b) DataONE (c) GBIF (d) iDigBio

Figure 3. Total number of URLs and unique contents observed from March 2019 through May 2020 in the provider networks of (a) BHL, (b) DataONE, (c) GBIF, and (d) iDigBio.

growth. The behavior of the GBIF provider network was characterized by large sporadic swings; a mass URL migration of over 14,000 Plazi-hosted datasets occurred in May, introducing thousands of new URLs over a short period of time, while over 31,000 URLs (60% of URLs that responded to queries that month) suddenly changed contents in October 2019. Even the most reliable set of URLs, observed in the DataONE provider network, shows a clear downward trend in all three categories, with 13% of URLs becoming unreliable over a period of fourteen months. Additionally, the DataONE provider network's growth curves indicate that there are far fewer unique contents than unique URLs. This mismatch suggests two possibilities: either much of the provider network's URL population is unresponsive, or DataONE lists multiple provider URLs for many of its datasets. Because the DataONE provider network has been shown to be highly responsive, it could be the case that many distinct URLs refer to the same datasets. It's also worth noting that the June and September spikes in BHL's unresponsiveness were largely due to URLs that failed to respond in those particular months but did respond to future queries.

Sources of Potential Numerical Error

We expect that the URL reliability counts generated for the figures and tables are lower than their actual values. When we qualified URLs as being reliable, responsive, and stable, we could not be certain that links did not briefly become unresponsive or change content during the month-long periods between queries. It is therefore likely that some cases of link rot and content drift were not reflected in the results. Additionally, we only queried provider network URLs that the aggregators list in their dataset URL registries; this means that, if a URL were removed from an aggregator's registry, we would not be able to detect subsequent instances of reference rot. Therefore, our results represent an optimistic upper bound on provider network URL reliabilities.

The results for the DataONE and GBIF provider networks in figure 2 are sometimes skewed due to Preston's interactions with the pagination method that the aggregators use to supply users with their dataset registries. Registry pages contained set amounts (e.g., 20) of URLs and represent small slices of the registry. For registries that use pagination, the observatory would keep querying for registry pages until reaching the page or failing to get a response. For instance, GBIF's URL and dataset totals in March 2019 (see figure 3(c)) are low because an early query to a GBIF registry page was not answered and, consequently, the URLs of registry pages that should have followed were not discovered. Similar events happened for both the GBIF and DataONE observatories at later points in time, potentially overestimating the reliability of the URLs in their provider networks.

For the iDigBio provider network, an issue with Preston's parsing of the iDigBio URL registry prevented the discovery and querying of a subset of URLs before February 2020, when the issue was detected and fixed. This likely accounts for the surge in the total number of contents and URLs in early February 2020.

The observatories for DataONE and BHL failed to save new provenance records for December 2019 through February 2020 due to a technical error on their shared server. Therefore, no new contents or URLs were reported for the provider networks of these aggregators during this time frame.

Discussion

We note that our experiment did not consider datasets other than those in the provider networks, i.e., those referenced in the aggregators' registries of data providers. For example, datasets that are retrieved from iDigBio or GBIF via portal/API queries or download events were not included. These datasets also have URL-based references and, unlike provider datasets, are hosted by the aggregators. These URLs are used to reference biodiversity datasets according to existing biodiversity network citation guidelines (DataONE 2012, GBIF 2019, iDigBio 2016). However, while we do not have quantitative measurements of stability for these URLs, content drift can take place. This is because datasets correspond to specific queries which over time produce different content depending on the changes in the data aggregated from the providers. Similarly, link rot can happen when the aggregator systems are down or storage limitations dictate the deletion of datasets. The architecture and policies used for storing and referencing these datasets differ among aggregators and are outside the scope of this paper.

We have shown that the reliability of URLs decreases over time in all of the provider networks that we monitored. If current trends continue, their reliabilities will continue to worsen. Systematic changes in the way we preserve and reference data are needed to improve the longevity and long-term integrity of the biodiversity data record. Before we propose such changes, it's necessary to first understand why URLs are proving to be ill-suited for referencing data in the long term.

Unreliability of Location-Based Identifiers

The problems related to using URLs for referencing datasets are largely due to the fact that they are location-based identifiers: they describe where the data is but not necessarily what it is. Also, by definition, data accessed via URLs must be mediated by a central authority, such as the institutional repositories that serve biodiversity datasets, who can match location-based identifiers with data. Interested users are expected to trust the central authority to guarantee long-term access to the referenced data in its original form.

The use of URLs as identifiers violates the requirements of uniqueness and persistence (Paskin 1999). An identifier must only ever identify one entity (uniqueness) and must persist longer than the entity it identifies (persistence) (Paskin 1999). However, as we have shown in our experiments, many URLs do not possess both uniqueness and persistence; unstable URLs forfeit uniqueness in the event of content drift, while unresponsive URLs do not persist as long as the datasets they identify.

At the core of URL instability is the current practice of using URLs to identify evolving datasets rather than using content-based identifiers to identify fixed dataset versions. If biodiversity data providers were uniformly committed to allocating one URL per dataset version, then content drift might become less common, improving overall URL stability; however, widespread social adoption of such a commitment from all data providers may be unrealistic. Additionally, such a commitment would not address link rot and URL unresponsiveness. Even if a similar commitment were made by data providers to guarantee the long-term responsiveness of URLs, it could not address the case where a data provider either loses authority over a domain name or migrates to another. For example, our deployed Preston observatories recorded the sudden migration of over 14,000 Plazi datasets from the http://plazi.cs.umb.edu/ domain to http://tb.plazi.org/, an event which invalidated any references to URLs within the first domain.

The instability that we have observed across the URLs in provider networks is to be expected, and is not a measure of the quality of either the provider networks or their aggregators. In fact, regular updates to datasets (i.e., URL instability) might indicate continued growth, maintenance, and refinement of those datasets. One might even argue that a stable dataset URL would indicate that the dataset is no longer being maintained or is potentially outdated. Therefore, the issues resulting from the use of URLs as references are not due to poor management on the part of data aggregators or curators, but rather due to the fact that URLs are inherently unreliable.

Paskin proposed that “the best way to ‘future proof’ an identifier scheme is to forego any intelligence within the identifier itself” (Paskin 1999), where the notion of intelligence refers to the inclusion of meaningful information in the textual representation of the identifier. URLs are typically structured according to the Domain Name System specification (though URLs may include an IP address instead of a domain name) and inherently contain some minimum amount of intelligence, namely the domain to which the URL belongs (Mockapetris 1987). Thus, it is necessary to look to another identification scheme to allow for proper identification and reliable referencing.

An Alternative: Unique Content-Based Identifiers

Instead of identifying digital datasets by location (e.g., a URL), we can identify datasets by their content. One way to achieve this is to use algorithmically generated content-based identifiers. A variety of cryptographic hashing algorithms are available that guarantee a unique hash, representable as text, for any given dataset (NIST 2001). Because the hash is deterministically derived from the content it identifies, we say that it is a content-based identifier. These content-based identifiers can be generated for a dataset using openly available algorithms, without a mediating central authority (Paskin 1999). If a change is made to the dataset, then the hash computed from the modified dataset will be different from that of the original. Therefore, if the hash of a dataset is the same as the referenced hash, it must be the originally referenced dataset (figure 4(c)) (NIST 2001). Using hash identifiers eliminates the possibility of content drift.

The shift from location-based to content-based identifiers decouples future dataset accessibility from the original point of access. As long as there exists some discoverable and accessible data repository that serves the desired content, that content can always be retrieved. Such data repositories can be made discoverable through content hash registries such as hash-archive.org (Trask 2015). In response to a user query for a content hash, these content hash registries would provide a list of locators (e.g., URLs), if any, that direct users to the referenced data (e.g., a registry would provide URLs that retrieve data when queried). Even if one repository becomes inaccessible due to either a temporary outage or permanent retirement, another may be available to provide the referenced data. When several repositories serve referenced datasets, there is no single point of failure for content hash lookups; if a referenced dataset is redundantly located across and within data repositories, access to the dataset will only be lost if all associated locations exhibit link rot. Even if access to a dataset is lost, it can be restored as long as the referenced dataset still exists somewhere and can be made discoverable and accessible.

If a dataset version were identified with a content-based hash, its duplication across different platforms would not lead to ambiguous references, but rather to distributed copies of the same reliably addressed content.

Transitioning to Reliable References

Although we propose a change in the fundamental mechanisms used to reference datasets, existing references can be made reliable with only minor modifications. Consider the following citation generated by GBIF according to their citation guidelines (GBIF 2019):

> Levatich T, Padilla F (2017). EOD - eBird Observation Dataset. Cornell Lab of Ornithology. Occurrence dataset https://doi.org/10.15468/aomfnb accessed via GBIF.org on 2018-09-02.

The citation references the eBird dataset hosted at gbif.org as it was retrieved on September 2, 2018. However, at the time of writing, the URL https://doi.org/10.15468/aomfnb redirects to a GBIF internal reference page that states the eBird dataset was last updated in March of 2019. The dataset made available through the listed URL is different from what was originally referenced in the citation, but it is impossible to determine the extent of the changes without having access to previous versions of the data.

Fortunately, references like the example above can be made more reliable by augmenting them with a content-based identifier for the dataset. Consider the following enriched citation for the eBird dataset that adds a SHA-256 content hash (NIST 2001):

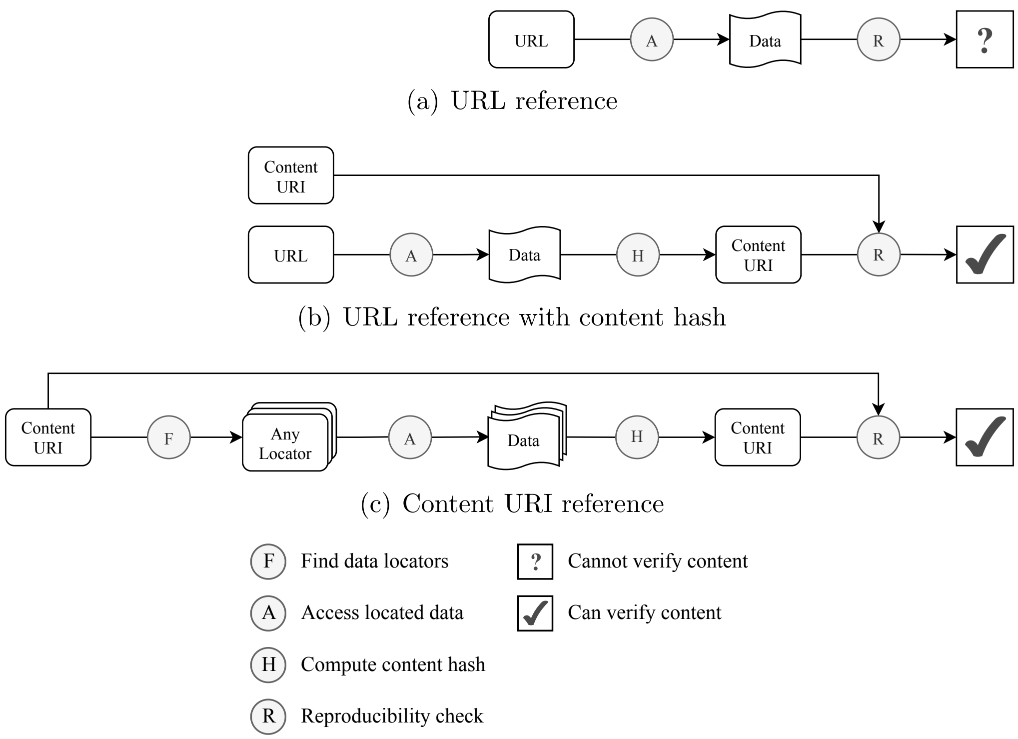


Figure 4. Content resolution and verification for references that use location- versus content-based identifiers. (a) Location-based identifiers (e.g. URLs) cannot verify the authenticity of retrieved content and are vulnerable to link rot due to the use of a fixed locator. (b) If the content hash of the referenced data is known, the authenticity of retrieved data can be verified by comparing the hash of the retrieved data with the provided content hash. However, the fixed locator is still vulnerable to link rot. (c) Content-based identifiers (e.g. Content URIs) can be used to find several locators for the referenced data and contain a content hash to verify the authenticity of retrieved data. The decoupling of the reference from a fixed locator makes the reference resistant to link rot.

> Levatich T, Padilla F (2017). EOD - eBird Observation Dataset. Cornell Lab of Ornithology. Occurrence dataset hash: //sha256/29d30b566f924355a383b13cd48c3aa239d42cba0a55f4 ccfc2930289b88b43c accessed at https://doi.org/10.15468/aomfnb via GBIF.org on 2018-09-02.

The content hash is captured in a content address Uniform Resource Identifier (URI) (Berners-Lee et al. 2005) in the form of hash://algo/hash-string proposed by (Trask 2015), where "algo" is a hashing algorithm (e.g., "sha256") and "hash-string" is the content hash generated by the algorithm in hexadecimal format. In the example above, the hashing algorithm is SHA256 and the hash string starts with "29d3." The added content hash was derived from and uniquely identifies the exact version of the eBird dataset that was originally referenced. If an interested user knows of and has access to an information retrieval system that has indexed the dataset, finding the desired dataset is as simple as querying for its content hash. With the addition of a content hash, the URL becomes superfluous and is included merely to demonstrate that the URL and content hash are not mutually exclusive (see figure 4(b)).

Other cryptographic hashing algorithms besides SHA-256 can be used to generate content-based identifiers with the same uniqueness guarantees (NIST 2013). However, note that different hashing algorithms will generate different content hashes from the same data. We use a URI rather than the content hash itself because it allows us to specify the hashing algorithm. If the hashing algorithm is not specified, one might mistakenly conclude that a dataset does not match a reference if the wrong hashing algorithm is used to verify the dataset's authenticity. Our proposal to use

Trask's content-addressed URIs to reliably reference data is inspired by Kuhn & Dumontier's method to make digital content verifiable and permanent using Trusty URIs (Kuhn and Dumontier 2015). We chose to use Trask's content hash URIs because they are location- and content-agnostic and easy to read. However, we recognize that Trusty URIs can help facilitate content retrieval and processing using a location-based URI prefix and an (optional) extension suffix.

Other content-based identification schemes exist that resist changes in format in digital content. For example, the universal numeric fingerprint (UNF) (Altman and King 2007) resists such changes by first processing the input data before generating a content hash. Among other preprocessing techniques used when generating UNFs, numerical data may be rounded to a certain precision before generating a content hash, with the understanding that a dataset may undergo such format changes when translated, for example, between different computing environments or hardware configurations. Indeed, on manual examination of the changes between successive versions of the biodiversity datasets we observed, we found some cases in which two versions of a dataset (determined to be different because they resulted in different content hashes) differed only in formatting, such as the amount of whitespace and the sequential ordering of observational records. However, for biodiversity data, we expect that such format-specific content-based identification schemes would only prove detrimental in practice. Standard cryptographic hashing algorithms, such as SHA-256, are included in most modern software environments and enjoy widespread use across different digital applications, whereas non-standard algorithms, such as UNF, would first need to be installed and may be unknown to most users, presenting a hurdle to their widespread adoption. Additionally, it may be unrealistic to expect preprocessing efforts to filter out non-informative data effectively enough to be able to trust that semantically identical datasets will always result in the same content-based identifiers. This is especially relevant to biodiversity datasets because they consist mostly of text data, which may be altered in a number of ways without changing the content's meaning.

Enhancing Dataset References with Provenance

A dataset reference can also be enhanced by pointing to the record that describes its provenance. The following citation further augments the eBird dataset reference with the content hash of an associated provenance record:

> Levatich T, Padilla F (2017). EOD - eBird Observation Dataset. Cornell Lab of Ornithology. Occurrence dataset hash://sha256/29d30b566f924355a383b13cd48c3aa239d42cba0a55f4ccfc2930289b88b43c accessed at https://doi.org/10.15468/aomfnb via GBIF.org on 2018-09-04 with provenance hash://sha256/b83cf099449dae3f633af618b19d05013953e7a1d7d97bc5ac01afd7bd9abe5d.

As was the case for the dataset, the provenance itself can be retrieved by querying an information system that has indexed the hash of the referenced provenance record. Note that the provenance hash is not strictly necessary to make a dataset reference reliable; the dataset hash alone is sufficient. However, explicitly referencing the provenance of the dataset is useful because it allows future readers to retrieve the same context to which the researcher referencing the dataset had access. More generally, the provenance describes the context of the retrieval of any type of content (e.g., datasets, metadata, citation files, etc.). The types of information in the provenance depend on the implementation of the data observatory, but at a minimum include the URLs that were queried to produce the content, the dates of the queries, the format of the content, and the data registries that were searched to find the content.

A provenance record relates to a dataset the way that a map relates to a location: a provenance record provides a context to understand the origin and relations of a dataset. This provenance context may be limited to few metadata elements related to a single dataset (e.g., web location, data format, author, license), but can also include a comprehensive description of a biodiversity provider network consisting of thousands of datasets and their associations. Also, because provenance records are datasets themselves, they can be reliably referenced and embedded in other provenance records using their content URIs. We used such a composition of content URIs and provenance records as part of our monitoring scheme (Poelen et al. 2018) to track the reliability of URLs in biodiversity provider networks over time (see table 1 and figures 2 and 3). The following citation references the history of the entire DataONE provider network over the period of observation by one of our Preston observatories:

> Poelen JH. 2019d. A biodiversity dataset graph: DataONE. doi:10.5281/zenodo.3483218 . hash://sha256/2b5c445f0b7b918c14a50de36e29a32854ed55f00d8639e09f58f049b85e50e3

The use cases for including the provenance hash are many. For example, if the provenance record of a dataset is found, it may be possible to traverse the provenance and find newer versions of the dataset. This requires that the various versions of the dataset were observed by a provenance-generating data observatory, properly archived, then made publicly accessible. Provenance can also be used for attribution purposes; a detailed record is kept of the life of each dataset, including when and where it was found, as well as snapshots of aggregator URL registries, which may provide information such as the publisher, authors, and contact information for each dataset. One study found that 88% of publications that cite biodiversity datasets do not provide enough information to identify the original source of the dataset (Escribano et al. 2018). Even in such cases, it may be possible to determine the dataset's publisher by looking up identifying information, such as the dataset's content hash, URL, or DOI, in available provenance records.

Dataset Retrieval Using Hash References

The dataset and provenance hashes referenced in the example references above were produced by our Preston observatories, which were set up to monitor the four provider networks. At the time of writing, both the referenced dataset and its provenance are available online (Poelen 2019a,b,c, 2020a,b,c). A query for the provenance hash in the search bar at hash-archive.org should direct the user to an archived repository of Preston observations that contains both the dataset and its provenance (see figure 5). The dataset reference is now reliable; it is effectively immune to both link rot and content drift. Given that Zenodo and Internet Archive serve as online digital archives (Internet Archive 2020, Zenodo 2019), future readers can expect that the URLs registered as locations for the referenced dataset

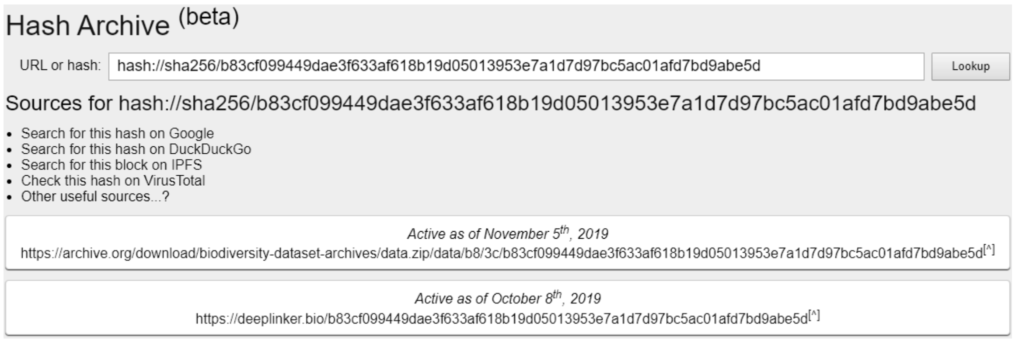


Figure 5. An example of a search index mapping hashes to archives. A search for a content or provenance hash at hash-archive.org will find any associated URLs that have been registered at hash-archive.org.

and provenance will serve the correct version of the eBird dataset we referenced. When archives and their URLs are eventually retired, datasets and provenance can be copied to other archives without compromising existing references, as long as their new locations are made available in an openly accessible hash registry such as hash-archive.org. Note that our Internet Archive publications (Poelen 2019a,b,c) contain data collected only from March 2019 through October 2019, whereas our Zenodo publications (Poelen 2020a,b,c) contain data collected from March 2019 through May 2020. Due to Zenodo's limit on total data size (Zenodo 2019), the Zenodo publication for the combined GBIF and iDigBio observatories (Poelen 2020c) contains only provenance, not biodiversity datasets.

Several biodiversity data aggregators, such as GBIF and iDigBio, produce a citation file for each user query to allow researchers to simply reference a single citation file rather than each individual dataset (GBIF 2019, iDigBio 2016). A citation file lists the URLs, attributions, and retrieval dates of the datasets that were returned by a query. We have demonstrated that dataset URLs are unreliable references; thus, citation files that rely on URLs as references are also unreliable. Citation files could be made reliable if they were augmented with the hashes of the retrieved datasets and, optionally, their provenance records. In fact, citation files themselves can be referenced by hash, along with accompanying provenance hashes, as long as they are archived and made accessible.

DOIs for Datasets and Queries

Biodiversity data aggregators often assign each dataset or query a Digital Object Identifier (DOI) (Paskin 2009) (e.g., 10.123/456) wrapped as a URL (e.g., https://doi.org/10.123/456) and advise researchers to reference the generated DOI rather than a URL. Unfortunately, this abstraction does little to enhance the reliability of the reference.

The DOI System (Paskin 2009) uses the Handle System (Sun et al. 2003) to resolve DOIs to online resources. However, it does not enforce any constraint on type of resource associated with a DOI. When DOIs are used to reference biodiversity datasets, the associated resources are often URLs, and therefore the use of such DOIs can be as unreliable as using URLs. In practice, these DOIs identify the evolving dataset (or set of datasets in the case of a query) rather than a fixed version, as demonstrated in the example references above. It is possible that an author would wish to make such a reference to an evolving online digital object. For example, an author promoting use of a published dataset might want future users to be directed to the most up-to-date content. However, such a fluid reference is not appropriate for making published results reproducible.

The Handle System allows for a complex web of redirection and distributed responsibilities. Just as the Domain Name System resolves domain names in URLs to IP addresses, the Handle System allows "handles" such as DOIs to be resolved to URLs. However, the responsibility for resolving DOIs to URLs is divided between the Handle System and DOI registrars. The Handle System serves as the central authority that maps DOI prefixes to DOI registrars, examples of which include BHL, DataONE, and GBIF. These registrars are responsible for associating DOIs that match their designated prefix with URLs, and are free to change the URL associated with any given DOI under their jurisdiction (IDF 2018, Paskin 2009).

The ability of biodiversity aggregators and providers to change the URL associated with a DOI is good for reference reliability in the sense that they can account for dataset migration without compromising existing references. However, the use of DOIs addresses neither the instability of the URLs they redirect to nor cases of link rot in which no URLs remain responsive to serve the referenced dataset. Additionally, as the number of datasets identified online continues to grow, proper maintenance of all of the DOIs an aggregator or provider administrates might become more unsustainable over time, potentially increasing the risk of unreliable URLs going undetected.

In an article proposing HTTP-URI-based stable identifiers (e.g., URLs that are resolvable over HTTP) for biological collection objects, Güntsch et al. admit that the use of DOIs does not solve the problem of unreliable referencing but merely deflects the burden of URL maintenance onto institutional repositories (Güntsch et al. 2017). In contrast, we propose a dataset referencing scheme that is reliable and can be supported by existing infrastructures and workflows. If existing workflows require references to be in the form of DOIs, it could be convenient to embed content hashes into DOIs. Such an approach has already been established for ISBNs through the creation of actionable ISBNs, or ISBN-As (Weissberg 2008), which may serve as a model for actionable content hashes.

What It Means to Preserve Data

Our results indicate that reference rot threatens the integrity of published biodiversity datasets. We have seen that the use of content-based identifiers can effectively address the issue of reference rot. However, identifiers are of little use in a vacuum. An identifier can only be useful for data retrieval when combined with a resolver to associate identifiers with locations and a database to retrieve the dataset at the associated location (Paskin 1999). Thus, we need to address how resolvers and databases might be organized to accommodate content-based identifiers in order to fully realize long-term data preservation. In this context, we define data preservation as the continued capacity for datasets to be reliably referenced and retrieved in their original form even as the global digital biodiversity network evolves over time.

We propose four requirements that must be met to ensure proper data preservation: 1) datasets must be addressable and retrievable using content-based rather than location-based identifiers; 2) an agent must exist to collect datasets, record their provenance, and deposit both to a dedicated repository; 3) these repositories should archive data that could be used in the future; and 4) content hash registries should be openly accessible to resolve hash identifiers to dataset locations within such repositories. Although openly accessible registries should make archived data discoverable, access to those data can still be restricted. Additionally, for the purposes of archiving, it is important that the recorded provenance records do not describe the datasets themselves, but rather the activities that led to the procurement of those datasets; the primary purposes of provenance in the context of an archive are to document the fact that evidence (i.e., an observation of a dataset) does exist and to make it discoverable for interested users (Bearman 1995).

We have shown that software agents such as Preston can be used to collect datasets and their provenance over time while maintaining content-addressability; all that is needed to ensure proper data preservation are a dedicated repository and an openly accessible content hash registry to map content-based identifiers to datasets located in the repository. In practice, repositories and registries (and potentially software agents such as Preston deployments) can be colocated; examples include Zenodo and the Internet Archive, although they impose some limitations that may restrict file size, number of files, and the amount of information that can be indexed (Internet Archive 2019, Zenodo 2019). Zenodo and the Internet Archive may serve as models for long-term biodiversity information systems.

These four requirements help to ensure that biodiversity data remain FAIR (Findable, Accessible, Interoperable, and Reusable) (Wilkinson et al. 2016). Findability is achieved through the publishing of provenance logs that thoroughly describe what datasets are and where they were retrieved. The amenability of the content-based identification paradigm to the operation of independent decentralized repositories strengthens accessibility by preventing the failure of a single data repository from inhibiting future data access (see figure 4). Content-based identification also contributes to interoperability across data networks due to the absence of any central authority to administrate data access; a content hash computed from a dataset is guaranteed to match the hash computed by any other agent using the same dataset. Furthermore, content-based identifiers can be embedded in or referenced by DOIs to maintain compatibility with systems that use DOIs as identifiers. Finally, and particularly relevant to this paper's purpose, reusability is strengthened by enhancing the retrievability of referenced datasets and allowing users to verify that a retrieved dataset exactly matches that which was referenced.

Future Work

The fourteen-month span of our experimental results might not be considered long-term in the context of biodiversity data. To evaluate the long-term reliability of provider network URLs in the aggregators, continued monitoring is needed.

Although we only monitored the provider networks of each aggregator, the same methods used in this paper to monitor URLs, collect datasets, and record provenance could be used for any of the URLs in biodiversity data networks.

In this study, we only monitored URLs that locate datasets. However, datasets may internally contain references to other data, such as media, literature, and genetic sequence information (Wieczorek et al. 2012). Such references are often URLs and therefore potentially unreliable. For datasets that contain links to other data, a recursive approach could be considered where those links are themselves queried for content and tracked through provenance records. This is the subject of future work and beyond the scope of this paper.

Conclusions

Although reference rot is resulting in a steady decline in the reliability of our digital biodiversity record, realistic solutions are available to address the root causes of the issue. Content drift can be eliminated altogether by changing the way we reference datasets from using location-based identifiers to ones that are content-based. Meanwhile, the biodiversity provider networks can be made more resilient to link rot if decentralized observation, archiving, and distribution techniques are used to capture incremental changes to the data record so that references can remain valid even when online datasets are updated, removed, or relocated. The use of content-based identifiers should be considered by biodiversity data aggregators in order to increase the reliability of references to the data they aggregate.

We have demonstrated that data observatories can be deployed to track the growing digital biodiversity data record. Using the dataset provenance collected over a period of fourteen months, we were able to quantify the change in reliability over time in terms of link rot and content drift exhibited by the provider network URLs registered in major biodiversity data aggregators. Even if aggregators and providers uniformly adopted content-based identification of datasets and maintained versioned datasets, our method of quantifying link rot and content drift in data networks could be used to monitor whether either of these issues persist in practice due to implementation flaws or nontechnical issues.

Biodiversity data observatories can also be used to increase the longevity of the biodiversity data record. Such observatories can be used to form reliable dataset references as well as recover datasets that would otherwise become inaccessible due to link rot and content drift. Additionally, the dataset provenance captured by such observatories serves as evidence of the evolution and distribution of the digital biodiversity data record. The combination of archived datasets and provenance can ensure the long-term reproducibility of scholarly works that reference ever-evolving biodiversity datasets.

Furthermore, the establishment of dedicated data repositories and publicly accessible content hash registries are beneficial for making content-addressed biodiversity data discoverable, distributable, and long-lived, by securely archiving the datasets and provenance captured by biodiversity data observatories and making them publicly available.

Great care has been taken to establish rigorous preservation guidelines for physical specimens, yet there is much that can be done to increase the longevity of our digital data. Our method is not only suited for tracking datasets in biodiversity data networks, but also provides a resilient and reliable way to publish, reference, and preserve scientific digital datasets without having to abandon our existing infrastructures. The method provides a much-needed foundation for constructing digital provenance graphs from an accessible, verifiable, and citable digital scholarly record.

Acknowledgments

The research reported in this paper was funded in part by a grant (NSF OAC 1839201) from the National Science Foundation and the AT&T

Foundation. We acknowledge early exchanges with Matt Collins, Anne Thessen, Jen Hammock, Katja Seltmann, Carl Boettiger, and Deborah Paul. Also, we thank Pepper Luboff for proofreading our manuscript.

References

[IDF] International DOI Foundation. 2018. Doi handbook. Technical report. International DOI Foundation. doi:10.1000/182. Accessed: 2019-12-04.

Altman M, King G. 2007. A proposed standard for the scholarly citation of quantitative data. D-Lib Magazine 13.

Bearman D. 1995. Archival strategies. The American Archivist 58:380–413. doi:10.17723/aarc.58.4.pq71240520j31798.

Berners-Lee T, Fielding RT, Masinter L. 2005. Uniform resource identifier (uri): Generic syntax. STD 66. RFC Editor. `http://www.rfc-editor.org/rfc/rfc3986.txt`. Accessed: 2020-02-03.

Berners-Lee T, Masinter L, McCahill M. 1994. Uniform resource locators (url). RFC 1738. RFC Editor. `http://www.rfc-editor.org/rfc/rfc1738.txt`. Accessed: 2020-02-03.

Costello MJ, Bouchet P, Boxshall G, Fauchald K, Gordon D, Hoeksema BW, Poore GCB, van Soest RWM, Stöhr S, Walter TC, Vanhoorne B, Decock W, Appeltans W. 2013. Global coordination and standardisation in marine biodiversity through the world register of marine species

(WoRMS) and related databases. PLoS ONE 8:e51629. doi:10.1371/journal.pone.0051629.

[DataONE] Data Observation Network for Earth. 2012. DataONE citation guidelines. https://www.dataone.org/citing-dataone. Accessed: 2019-12-04.

Davis EB, Schmidt D. 1996. Guide to Information Sources in the Botanical Sciences. Vol. 2nd ed. Reference Sources in Science and Technology. Englewood, Colo: Libraries Unlimited.

Edwards JL. 2000. Interoperability of biodiversity databases: Biodiversity information on every desktop. Science 289:2312–2314. doi:10.1126/science.289.5488.2312.

Elton CS. 1927. Animal ecology. Macmillan Co. doi:10.5962/bhl.title.7435.

Escribano N, Galicia D, Ariño AH. 2018. The tragedy of the biodiversity data commons: a data impediment creeping nigher? Database: the journal of biological databases and curation 2018:bay033. doi:10.1093/database/bay033.

Garfield E, Sher IH, Torpie RJ. 1964. The Use of Citation Data in Writing the History of Science. Institute for Scientific Information Inc Philadelphia PA.

[GBIF] Global Biodiversity Information Facility. 2019a. GBIF citation guidelines. https://www.gbif.org/citation-guidelines. Accessed: 2019-12-04.

[GBIF] Global Biodiversity Information Facility. 2019b. Gbif secretariat:

Gbif backbone taxonomy. https://doi.org/10.15468/39omei. doi:10.15468/39omei. Accessed: 2020-05-04.

[GBIF] Global Biodiversity Information Facility. 2019c. What is the GBIF? https://www.gbif.org/what-is-gbif. Accessed: 2019-12-04.

Güntsch A, Hyam R, Hagedorn G, Chagnoux S, Röpert D, Casino A, Droege G, Glöckler F, Gödderz K, Groom Q, Hoffmann J, Holleman A, Kempa M, Koivula H, Marhold K, Nicolson N, Smith VS, Triebel D. 2017. Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. Database 2017. doi:10.1093/database/bax003.

Hortal J, de Bello F, Diniz-Filho JAF, Lewinsohn TM, Lobo JM, Ladle RJ. 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. Annual Review of Ecology, Evolution, and Systematics 46:523–549. doi:10.1146/annurev-ecolsys-112414-054400.

[iDigBio] Integrated Digitized Biocollections. 2016. iDigBio citation guidelines. https://www.idigbio.org/content/idigbio-terms-use-policy. Accessed: 2019-12-04.

Internet Archive. 2019. Uploading - a basic guide. https://help.archive.org/hc/en-us/articles/360002360111-Uploading-A-Basic-Guide. Accessed: 2019-12-04.

Internet Archive. 2020. About the internet archive. https://archive.org/about. Accessed: 2020-05-25.

Keklikoglou K, Faulwetter S, Chatzinikolaou E, Wils P, Brecko J, Kvaček J, Metscher B, Arvanitidis C. 2019. Micro-computed tomography for natural history specimens: a handbook of best practice protocols. European Journal of Taxonomy 0. doi:10.5852/ejt.2019.522.

Klein M, de Sompel HV, Sanderson R, Shankar H, Balakireva L, Zhou K, Tobin R. 2014. Scholarly context not found: One in five articles suffers from reference rot. PLoS ONE 9:e115253. doi:10.1371/journal.pone.0115253.

Kuhn T, Dumontier M. 2015. Making digital artifacts on the web verifiable and reliable. IEEE Transactions on Knowledge and Data Engineering 27:2390–2400. doi:10.1109/tkde.2015.2419657.

Matsunaga A, Thompson A, Figueiredo RJ, Germain-Aubrey CC, Collins M, Beaman RS, MacFadden BJ, Riccardi G, Soltis PS, Page LM, Fortes JAB. 2013. A computational- and storage-cloud for integration of biodiversity collections. In: 2013 IEEE 9th International Conference on e-Science. p. 78–87. doi:10.1109/eScience.2013.48. Accessed: 2020-05-20.

Michener W, Vieglais D, Vision T, Kunze J, Cruse P, Janée G. 2011. DataONE: Data observation network for earth: Preserving data and enabling innovation in the biological and environmental sciences. D-Lib Magazine 17. doi:10.1045/january2011-michener.

Mockapetris P. 1987. Domain names - concepts and facilities. STD 13. RFC Editor. `http://www.rfc-editor.org/rfc/rfc1034.txt`. Accessed: 2020-02-03.

[NIST] National Institute for Standards and Technology. 2001.

Descriptions of sha-256, sha-384, and sha-512. https://web.archive.org/web/20130526224224/http://csrc.nist.gov/groups/STM/cavp/documents/shs/sha256-384-512.pdf. Accessed: 2019-12-04.

[NIST] National Institute for Standards and Technology. 2013. Digital signature standard (dss). doi:10.6028/NIST.FIPS.186-4. Accessed: 2020-05-04.

Page LM, MacFadden BJ, Fortes JA, Soltis PS, Riccardi G. 2015. Digitization of biodiversity collections reveals biggest data on biodiversity. BioScience 65:841–842. doi:10.1093/biosci/biv104.

Paskin N. 1999. Toward unique identifiers. Proceedings of the IEEE 87:1208–1227. doi:10.1109/5.771073.

Paskin N. 2009. Digital object identifier (DOI®) system. In: Encyclopedia of Library and Information Sciences, Third Edition. CRC Press. p. 1586–1592. doi:10.1081/e-elis3-120044418.

Pasquier T, Lau MK, Trisovic A, Boose ER, Couturier B, Crosas M, Ellison AM, Gibson V, Jones CR, Seltzer M. 2017. If these data could talk. Scientific Data 4. doi:10.1038/sdata.2017.114.

Poelen J, Elliott M, Alzuru I, Patel P. 2018. Preston: a biodiversity dataset tracker. doi:10.5281/zenodo.1410543.

Poelen JH. 2019a. A biodiversity dataset graph: Biodiversity Heritage Library (BHL). hash://sha256/34ccd7cf7f4a1ea35ac6ae26a458bb603b2f6ee8ad36e1a58aa0261105d630b1. https://archive.org/details/preston-bhl. Accessed: 2019-12-04.

Poelen JH. 2019b. Biodiversity Dataset Archive. hash://sha256/8aacce084 62b87a345d271081783bdd999663ef90099212c8831db399fc0831b. https://archive.org/details/biodiversity-dataset-archives. Accessed: 2019-12-04.

Poelen JH. 2019c. A biodiversity dataset graph: DataONE. hash://sha256 /2b5c445f0b7b918c14a50de36e29a32854ed55f00d8639e09f58f049b85e50e 3. https://archive.org/details/preston-dataone. Accessed: 2019-12-04.

Poelen JH. 2019d. To connect is to preserve: on frugal data integration and preservation solutions. doi:10.17605/OSF.IO/A2V8G.

Poelen JH. 2020a. A biodiversity dataset graph: BHL. hash://sha256/34cc d7cf7f4a1ea35ac6ae26a458bb603b2f6ee8ad36e1a58aa0261105d630b1. doi:10.5281/zenodo.3849560.

Poelen JH. 2020b. A biodiversity dataset graph: DataONE. hash://sha256 /2b5c445f0b7b918c14a50de36e29a32854ed55f00d8639e09f58f049b85e50e 3. doi:10.5281/zenodo.3849494.

Poelen JH. 2020c. A biodiversity dataset graph: GBIF, iDigBio, BioCASe. hash://sha256/8aacce08462b87a345d271081783bdd999663ef90099212c8 831db399fc0831b. doi:10.5281/zenodo.3852671.

Postel J. 1981. Internet protocol. STD 5. RFC Editor. http://www.rfc-editor.org/rfc/rfc791.txt. Accessed: 2020-02-03.

Rinaldo C, Norton C. 2009. BHL, the biodiversity heritage library: An expanding international collaboration. Nature Precedings doi:10.1038/npre.2009.3620.1.

Robertson T, Döring M, Guralnick R, Bloom D, Wieczorek J, Braak K, Otegui J, Russell L, Desmet P. 2014. The GBIF integrated publishing toolkit: Facilitating the efficient publishing of biodiversity data on the internet. PLoS ONE 9:e102623. doi:10.1371/journal.pone.0102623.

Sun S, Lannom L, Boesch B. 2003. Handle system overview. RFC 3650. RFC Editor. `https://www.rfc-editor.org/info/rfc3650`. Accessed: 2020-05-25.

Trask B. 2015. Principles of content addressing. https://bentrask.com/?q=hash://sha256/98493caa8b37eaa26343bbf73f232597a3ccda20498563327a4c3713821df892. Accessed: 2019-12-04.

Vision TJ. 2010. Open data and the social contract of scientific publishing. BioScience 60:330–331. doi:10.1525/bio.2010.60.5.2.

Weissberg A. 2008. The identification of digital book content. Publishing Research Quarterly 24:255–260. doi:10.1007/s12109-008-9093-8.

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D. 2012. Darwin core: An evolving community-developed biodiversity data standard. PLoS ONE 7:e29715. doi:10.1371/journal.pone.0029715.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, 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. doi:10.1038/sdata.2016.18.

Zenodo. 2019. General policies. `https://about.zenodo.org/policies/`. Accessed: 2019-12-04.