Toward Reliable Biodiversity Data References

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

1 Advanced Computing and Information Systems Laboratory (ACIS), University of Florida, Gainesville, Florida, USA

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

*** jhpoelen@jhpoelen.nl**

Abstract

Scientific discovery increasingly relies on digital datasets to capture measurements and outcomes. However, no systematic approach has been adopted to reliably reference and provide access to our digital datasets. Our existing data infrastructures have grown accustomed to using location-based identifiers such as URLs in an attempt to retain our digital knowledge. We hypothesize that URLs are not sufficient to ensure long-term data access, then propose a method for evaluating long-term URL reliability.

After taking periodic inventories from March through October 2019 of the data served by major biodiversity aggregators, including GBIF, iDigBio, DataONE, and BHL, we found that, for each network, 5%-44% of registered URLs were intermittently or consistently unresponsive, 0%-64% produced unstable content, and 13%-76% became either unresponsive or unstable over the period of observation. We propose to use content-based identifiers to reliably track and reference datasets while enabling decentralized archiving schemes. We propose a method for properly tracking and archiving datasets that can be used to guarantee fixed content and encourage long-term accessibility by leveraging content- rather than location-based identifiers.

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 [Hortal et al., 2015, Davis and Schmidt, 1996]. 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 work toward Charles Elton's realization that ecosystems can only be fully understood when we "provide conceptions which can link 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 to provide comprehensive aggregate views from previously scattered natural history record siloes [Rinaldo and Norton, 2009, Michener et al., 2011, Edwards, 2000, Matsunaga et al., 2013, Facility, 2019]. 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 keep track of changes in digital networked resources, such as the use of version numbers and last modified dates [Wieczorek et al., 2012, Robertson et al., 2014] and periodic archival [Costello et al., 2013], we are not aware of the adoption of any systematic approach to preserve the accessibility as well as longevity of our digital natural history record and derived datasets. We have collected evidence that, despite hundreds of years of experience in preserving our physical natural history records, we are currently faced with a growing body of digital data that changes daily and can disappear with the push of a button. Our scholarly record is stitched together by an intricate web of associations between scientific publications. These associations are made explicit using citations. These citations point to related scientific works and are assumed to provide enough identifying information to allow the reader to retrieve the unaltered referenced work regardless of the time at which the reader chooses to do so [Garfield et al., 1964]. In the pre-internet era, the lookup of these references required access to one of the many academic libraries in the world. With the rise of internet accessible scientific publications, authors and readers access these references using a networked device by downloading content from publication websites. This means that researchers are increasingly citing online works to support their claims. Because the citation format of online works 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 [GBIF.org, 2019, iDigBio.org, 2016, DataONE, 2012], the future reader expects the web accessed resource to remain accessible and unaltered via this single web location. Future readers may attempt to find a version of the works referenced by searching online data networks 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 future readers to find the referenced work fails to satisfy the FAIR principle of findability: "F1. (meta)data are assigned a globally unique and eternally persistent identifier." [Wilkinson et al., 2016]. Our study is not alone in providing evidence that suggests that networked, location-based access to digital objects is an unreliable mechanism for providing continued access to the unaltered original work [Vision, 2010, Klein et al., 2014]. Unless we change the way we preserve and cite our digital scholarly works, our physical records stored in libraries and museums around the world are likely to outlast our digital ones.

Problem Characterization

We show that the current practice of using Uniform Resource Locators (URLs) [Berners-Lee et al., 1994] to reference online biodiversity datasets provides no guarantee of long-term data accessibility. Readers who encounter references that use URLs as dataset identifiers cannot be certain that the referenced data will continue to be accessible and in its exact original form. This uncertainty might be cause for alarm for researchers because, over time, the integrity of the scholarly record itself is damaged when existing references become reliable due to the loss of access to the data they reference. When data access is lost, it is possible that documented research results may become impossible to reproduce and the justification for any conclusions or hypotheses that relied on lost results may be undermined. If the use of error-prone referencing techniques is not addressed, we expect that any resulting gaps in the biodiversity data record will only become more severe.

The current practice of relying on URLs to locate and identify referenced data is hazardous due to their demonstrated 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, access to any particular version of the data is likely to be short-lived. We show that, in the event of link rot or content drift, any existing references that relied affected URL may become unreliable.

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 articles in journal of Science, Technology, and Medicine provide references that exhibit either link rot and content drift and refer to the existence of either as "reference rot". Since existing biodiversity references largely rely on URLs to locate datasets, it is reasonable to expect that biodiversity data networks are also at risk of providing unreliable dataset references as a result of reference rot. The information systems used by major biodiversity data networks, such as

DataONE, GBIF, and iDigBio, rely on data curators, such as institutional repositories, to maintain active dataset URLs, and aggregate the data found at those URLs for distribution in response to user queries. If a data curator modifies, relocates, or stops serving a particular dataset, it may become impossible to retrieve the original dataset and the integrity of the data network will suffer as a result.

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 both link rot and content drift across all of the biodiversity data networks we considered, including 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

Although it has been demonstrated that reference rot does occur when URLs are used for referencing scientific works [Vision, 2010, Klein et al., 2014], we are not aware of any prior studies that provide quantitative evidence that reference rot occurs specifically in biodiversity data networks. We set out to quantify the extent of reference rot in biodiversity data 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) address in the Domain Name System. If a web server exists at the resolved IP address, a query 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 it, 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 repeatedly 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 how often 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 how often 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 of 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

Before we can determine the reliability of any given URL, we must first monitor its behavior over time by documenting how it responds to periodic queries. For the context of biodiversity, we consider the case when the 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 (Fig. 1) [Poelen, 2019g]. If datasets are considered analogous to specimens, then the URLs that locate datasets 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, well-known repository, a record is kept that documents what the specimen is in addition to when, where, and by whom it was collected.

(insert figure 1 / see appendix)

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. Additionally, the dataset itself should be stored in a safe, well-known dataset archive. The final step in the collection process is to link the actual preserved specimen to its corresponding record (the "specimen history" in Fig. 1) via an assigned unique identifier. For digital datasets, we use cryptographic hashes of the data as unique content-based identifiers.

Data Collection Over Time

By establishing a dedicated data observatory that follows the collection process we have described, we can build a history for each observed URL to capture its long-term reliability. Such an observatory should periodically query the URLs listed in data network's URL registry, producing 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. The use of a content-based data identifier is crucial; it allows us to reliably link each acquired dataset to its provenance record without the need for an intermediate index. Successive provenance records can be aggregated to construct comprehensive histories for both datasets (when and where they were found) and URLs (which datasets they produced 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 it) that a URL produced each time it was queried. Any change in the content identifier from 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.

Data Network Reliability

Now that we have outlined a method for observing and documenting the behavior of URLs over an extended period of time, we can apply our method to observe all of URLs registered by biodiversity data networks. We also extend the idea of URL reliability to entire data networks and propose that the overall reliability of a data network can be evaluated by monitoring the long-term reliability of each individual URL in the network exposes. Whereas we rigidly label individual URLs with binary indicators of responsiveness, stability, and reliability, we grade data networks according to the percentage of registered 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 the data network is 67% reliable.

Experiment

The Preston biodiversity dataset tracker [Poelen et al., 2018] implements mechanisms for monitoring data networks as we have described. It allows users to deploy a data network observatory which systematically observes the entire set of URLs registered by the network, queries each URL for data, then documents data collection and archives the results. All crawl activities, the queries they issue, and the results they produce are meticulously recorded in a string of provenance logs.

We deployed several Preston observatories which periodically queried the registered dataset URLs listed by Biodiversity Heritage Library (BHL), Data Observation Network for Earth (DataONE), Global Biodiversity Information Facility (GBIF), and Integrated Digitized Bio Collections (iDigBio). Each of these networks provides online registries of URLs that locate the data in the network. The registered URLs for DataONE, GBIF, and iDigBio were queried monthly from March 2019 through October 2019. BHL was queried monthly from May 2019 through October 2019. The logs taken by each of these observatories describe the URL queries and their results, which were processed to produce the results that follow. A sixth observatory was constructed by aggregating the queries of the five data network observatories.

(insert figure 2, see appendix)

Results

Breakdowns of the overall reliabilities of the data networks are provided in Table 1. Results are listed as percentages and total counts of URLs in the data network that were assigned each reliability indicator. When analyzing the recorded results of queries to URLs in each data network over a period of seven months, we found that, for each individual network, 5%-44% of registered URLs were intermittently or consistently unresponsive, 0%-64% produced unstable content, and 13%-76% became either unresponsive or unstable over the period of observation.

Overall, 30% of URLs observed across the five networks became unreliable at some point over the period of March 2019 through October 2019. Of those unreliable URLs, 48% were unstable, 22% became consistently unresponsive, and 70% 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 the next time it responded when queried.

The changes in reliability over time for each network are visualized in Fig. 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 Fig. 1, Fig. 2, and Fig. 3 visualizes the cumulative growth of biodiversity data networks during their periods of observation. This growth is illustrated with two metrics: the total number of unique URLs ever registered by each network and the total number of unique contents that had been downloaded from the network at each sampled point in time.

(insert figure 3, see appendix)

The behaviors of the distributions over time of responsive, stable, and reliable URLs vary notably between data networks.

(insert table 1, see appendix) Some reasons for these differences can be inferred when cross-examining the table and figures. For example, although BHL scored relatively low in responsiveness due to frequent link rot, the content that it does provide is more stable than all other networks because content drift within BHL is relatively rare. Conversely, although iDigBio is relatively responsive, it has low stability because the network's near-constant content growth far outpaces its URL growth. GBIF's behavior 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. Even the most reliable network, DataONE, shows a clear downward trend in all three categories, with 13% of URLs becoming unreliable over a period of just seven months. Additionally, DataONE's growth curves indicate that there are far fewer unique contents than unique URLs; this evokes two possibilities: either much of DataONE's URL population is unresponsive, or DataONE lists multiple URLs for many of its datasets. Because DataONE 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 actually 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 query URLs that the data networks list in their dataset registries; this means that, after URL was removed from a network's registry, we could not detect subsequent instances of reference rot. Therefore, our results represent a very optimistic upper bound on URL and network reliabilities.

The results for DataONE and GBIF in 2 are sometimes skewed due to the pagination method that the networks use to supply users with their dataset registries. Registry pages contained set amounts (e.g. 20) of URLs and represent small slices of the actual data network registry. For registries that use pagination, the observatory would keep querying for registry pages until reaching the page or failing to respond. For instance, GBIF's URL and dataset totals in March 2019 (2.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 data network.

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.

Discussion

We have shown that the reliability of URLs decreases over time in all of the major biodiversity data networks that we monitored. If current trends continue, the extent of reference rot will only worsen. Systematic changes in the way we preserve and reference data are needed to reverse these trends and 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 fixed dataset versions. If biodiversity data providers were uniformly committed to allocating one URL per dataset version, then content drift might indeed become far 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.

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 structured according to the Domain Name System specification and inherently contain some minimum amount of intelligence: the domain that the URL belongs to [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 (i.e. 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 which guarantee a single unique hash, representable as text, for any given dataset [NIST, 2001]. Because the hash itself is deterministically derived from the content it identifies, we say that it is a content-based identifier. Because hashes are deterministic, anyone interested in identifying a dataset can simply compute its hash without the need for some 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 [NIST, 2001]. Because hash identifiers can only identify the exact content that was referenced, content drift is impossible; a content hash will never match with either a different version of the content any other content. Additionally, the chance of link rot is diminished due to the lack of a single point of failure in the form of a central authority that is solely responsible for making content available. The shift from location-based to content-based identifiers allows for the decoupling of future dataset accessibility from the original point of access. As long as there exists some well-known and accessible data repository that has archived the desired content, it can always be retrieved. Even if one repository becomes inaccessible, another may be available to retrieve the content. If a repository changes location, the reference is still reliable; it is the interested user's responsibility to find either the repository's new location or another repository that hosts the desired dataset. Additionally, it is worth noting that duplication of content across different information platforms does not lead to ambiguous references, but rather to distributed copies of the same reliably addressed content. Figure 4 demonstrates the differences in referenced dataset retrieval when using location- versus content-based identifiers.

(insert figure 4, see appendix)

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.org, 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 11, 2018. However, at the time of writing, the URL https://doi.org/10.15468/aomfnb redirects to a GBIF internal reference page which states that 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 eBirds dataset adds a SHA-256 content hash [NIST, 2001]:

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 2018-09-02.

The content hash is captured in a content address URI 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 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.

Enhancing Dataset References with Provenance

A dataset reference can be given enhanced context by also referencing 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 2018-09-02 with provenance hash://sha256/b83cf099449dae3f633af618b19d05013953e7a1d7d97bc5ac01afd7bd9abe5d.

As was the case for the dataset, the provenance itself can be retrieved by querying a well-known 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 also retrieve the same context that the original researcher who referenced the dataset had access to. 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.

(insert figure 5, see appendix)

The use cases for the included 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 at some point in time by a provenance-generating data observatory, properly archived, then made publicly accessible.

Our proposal to use Trask's content-addressed URIs to reliably reference data is similar to, and was 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 respectively.

Dataset Retrieval Using Hash References

The dataset and provenance hashes referenced in the sample references above were produced by our Preston observatories which were set up to monitor the four data networks. Both the referenced dataset and its provenance are available online at zenodo.org [Poelen, 2019f, Poelen, 2019e, Poelen, 2019c] and archive.org [Poelen, 2019d]. A query for the provenance hash in the search bar at zenodo.org or hash-archive.org should direct the user to an archived repository of Preston observations that contains both the dataset and its provenance (5). Given Zenodo's long-term guarantee for data persistence and version availability [Zenodo, 2019], the dataset reference is now reliable; it is effectively immune to both link rot and content drift. Future readers can trust that the dataset will stay available and, when downloaded, identically match the exact version of the eBird dataset we referenced. Note that, to comply with Zenodo's limitations on user uploads [Zenodo, 2019], we only exposed the set of provenance hashes collected by each deployed Preston observatory for search indexing, which are far fewer in number than the dataset hashes. Thus, a query to zenodo.org for the dataset hash above should not produce any results. This is an artificial limitation; ideally, an information system would index the dataset hashes as well. Note that our Zenodo publication for the GBIF/iDigBio/BioCASe observatory [Poelen, 2019c] contains only provenance, although the Internet Archive publication [Poelen, 2019d] contains the content as well as provenance. Our Zenodo and Internet Archive publications for BHL [Poelen, 2019e, Poelen, 2019a] and DataONE [Poelen, 2019f, Poelen, 2019b] contain both content and provenance.

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. A citation file lists the URLs of the datasets (among other things, such as attributions and retrieval dates) that were retrieved by the issued 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 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 Handle System [Paskin, 2009] associates DOIs with 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 as referencing mechanisms is just as potentially 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 URLs to IP addresses, the Handle System resolves DOIs to data. When these data are URLs, they must then be resolved through the Domain Name System in order to retrieve the referenced content. 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, GBIF, and iDigBio. These registrars are then responsible, and indeed the central authorities 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 [Paskin, 2009, Foundation, 2018].

The ability of biodiversity data networks to change the URL associated with a DOI is good for reference reliability in the sense that networks 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 a data network 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 poses an existential threat to published biodiversity datasets. We've 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 that prevents data loss: 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 rather than discarding it; and 4) well-known search indexes should be available to resolve hash identifiers to dataset locations within such repositories. For the purposes of archival, it is important that the recorded provenance records do not necessarily describe the datasets themselves, but rather the activities that led to the procurement of those datasets; the primary purpose of provenance in the context of an archive is to document the fact that evidence, i.e. the dataset itself, 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 a well-known, publicly available search index to map content-based identifiers to datasets located in the repository. In practice, repositories and search indexes (and potentially software agents such as Preston deployments) can be co-located; 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 [Zenodo, 2019, Archive, 2019]. These existing information systems may serve as models for long-term biodiversity information systems.

These 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 which thoroughly describe what datasets are and where they originated from. The amenability of the content-based identification paradigm to the operation of independent distributed repositories strengthens accessibility by preventing the failure of a single data repository from inhibiting future data access (4). Content-based identification also allows for interoperability 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. 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.

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 online biodiversity data networks can be made far more resilient to link rot if distributed observation and archival 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. If long-term data observatories for biodiversity data networks are established, their collected data routinely deposited to well-known publicly available archives, and the archived data sufficiently indexed, then researchers and data curators will be able to have certainty that the datasets they contribute and reference will maintain reliability in the midst of an ever-changing digital ecosystem.

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

We thank ...

References

Archive, 2019. Archive, I. (2019). Uploading - a basic guide. Accessed: 2019-12-04.

Bearman, 1995. Bearman, D. (1995). Archival strategies. *The American Archivist*, 58(4):380–413.

Berners-Lee et al., 2005. Berners-Lee, T., Fielding, R. T., and Masinter, L. M. (2005). Uniform Resource Identifier (URI): Generic Syntax. RFC 3986.

Berners-Lee et al., 1994. Berners-Lee, T., Masinter, L. M., and McCahill, M. P. (1994). Uniform Resource Locators (URL). RFC 1738.

Costello et al., 2013. Costello, M. J., Bouchet, P., Boxshall, G., Fauchald, K., Gordon, D., Hoeksema, B. W., Poore, G. C. B., van Soest, R. W. M., Stöhr, S., Walter, T. C., Vanhoorne, B., Decock, W., and Appeltans, W. (2013). Global coordination and standardisation in marine biodiversity through the world register of marine species (WoRMS) and related databases. *PLoS ONE*, 8(1):e51629.

DataONE, 2012. DataONE (2012). Dataone citation guidelines. Accessed: 2019-12-04.

Davis and Schmidt, 1996. Davis, E. B. and 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, 2000. Edwards, J. L. (2000). Interoperability of biodiversity databases: Biodiversity information on every desktop. *Science*, 289(5488):2312–2314.

Elton, 1927. Elton, C. S. (1927). *Animal ecology*. Macmillan Co.,.

Facility, 2019. Facility, G. T. G. B. I. (2019). What is gbif? Accessed: 2019-12-04.

Foundation, 2018. Foundation, I. D. (2018). Doi handbook. Accessed: 2019-12-04.

Garfield et al., 1964. Garfield, E., Sher, I. H., and Torpie, R. J. (1964). *The Use of Citation Data in Writing the History of Science*. Institute for Scientific Information Inc Philadelphia PA.

GBIF.org, 2019. GBIF.org (2019). Gbif citation guidelines. Accessed: 2019-12-04.

Güntsch et al., 2017. 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, V. S., and Triebel, D. (2017). Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. *Database*, 2017.

Hortal et al., 2015. Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M., and Ladle, R. J. (2015). Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, 46(1):523–549.

iDigBio.org, 2016. iDigBio.org (2016). idigbio citation guidelines. Accessed: 2019-12-04.

Klein et al., 2014. Klein, M., de Sompel, H. V., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., and Tobin, R. (2014). Scholarly context not found: One in five articles suffers from reference rot. *PLoS ONE*, 9(12):e115253.

Kuhn and Dumontier, 2015. Kuhn, T. and Dumontier, M. (2015). Making digital artifacts on the web verifiable and reliable. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2390–2400.

Matsunaga et al., 2013. Matsunaga, A., Figueiredo, R., Thompson, A., Traub, G., Beaman, R., and Fortes, J. A. (2013). Integrated digitized biocollections (idigbio) cyberinfrastructure status and futures. In *TDWG 2013 ANNUAL CONFERENCE*.

Michener et al., 2011. Michener, W., Vieglais, D., Vision, T., Kunze, J., Cruse, P., and Janée, G. (2011). DataONE: Data observation network for earth: Preserving data and enabling innovation in the biological and environmental sciences. *D-Lib Magazine*, 17(1/2).

Mockapetris, 1987. Mockapetris, P. (1987). Domain names - concepts and facilities. RFC 1034.

NIST, 2001. NIST (2001). Descriptions of sha-256, sha-384, and sha-512. Accessed: 2019-12-04.

Page et al., 2015. Page, L. M., MacFadden, B. J., Fortes, J. A., Soltis, P. S., and Riccardi, G. (2015). Digitization of biodiversity collections reveals biggest data on biodiversity. *BioScience*, 65(9):841–842.

Paskin, 1999. Paskin, N. (1999). Toward unique identifiers. *Proceedings of the IEEE*, 87(7):1208–1227.

Paskin, 2009. Paskin, N. (2009). Digital object identifier (DOI®) system. In *Encyclopedia of Library and Information Sciences, Third Edition*, pages 1586–1592. CRC Press.

Poelen et al., 2018. Poelen, J., Elliott, M., Alzuru, I., and Patel, P. (2018). Preston: a biodiversity dataset tracker.

Poelen, 2019a. Poelen, J. H. (2019a). A biodiversity dataset graph: Biodiversity Heritage Library (BHL).

Poelen, 2019b. Poelen, J. H. (2019b). A biodiversity dataset graph: DataONE.

Poelen, 2019c. Poelen, J. H. (2019c). A biodiversity dataset graph: GBIF, iDigBio, BioCASe.

Poelen, 2019d. Poelen, J. H. (2019d). Biodiversity Dataset Archive.

Poelen, 2019e. Poelen, J. H. (2019e). A biodiversity dataset graph: Bhl.

Poelen, 2019f. Poelen, J. H. (2019f). A biodiversity dataset graph: Dataone.

Poelen, 2019g. Poelen, J. H. (2019g). To connect is to preserve: on frugal data integration and preservation solutions.

Rinaldo and Norton, 2009. Rinaldo, C. and Norton, C. (2009). BHL, the biodiversity heritage library: An expanding international collaboration. *Nature Precedings*.

Robertson et al., 2014. Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., Otegui, J., Russell, L., and Desmet, P. (2014). The GBIF integrated publishing toolkit: Facilitating the efficient publishing of biodiversity data on the internet. *PLoS ONE*, 9(8):e102623.

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

Vision, 2010. Vision, T. J. (2010). Open data and the social contract of scientific publishing. *BioScience*, 60(5):330–331.

Weissberg, 2008. Weissberg, A. (2008). The identification of digital book content. *Publishing Research Quarterly*, 24(4):255–260.

Wieczorek et al., 2012. Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., and Vieglais, D. (2012). Darwin core: An evolving community-developed biodiversity data standard. *PLoS ONE*, 7(1):e29715.

Wilkinson et al., 2016. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1).

Zenodo, 2019. Zenodo (2019). General policies. Accessed: 2019-12-04.

Figures

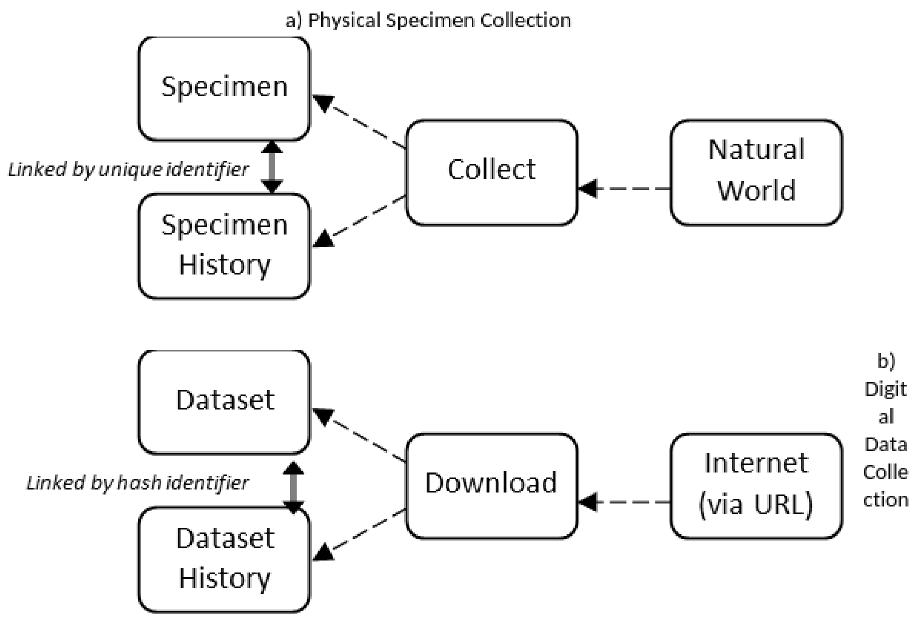

Figure 1. Reliable record keeping for digital datasets (b) can be achieved in an analogous way to current practices in record keeping for physical specimens (a). Biologists collect physical specimens from the natural world, thoroughly document the process, then store the specimens in facilities equipped 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 history records, the downloading of digital datasets can also be recorded and identified in dataset history records.

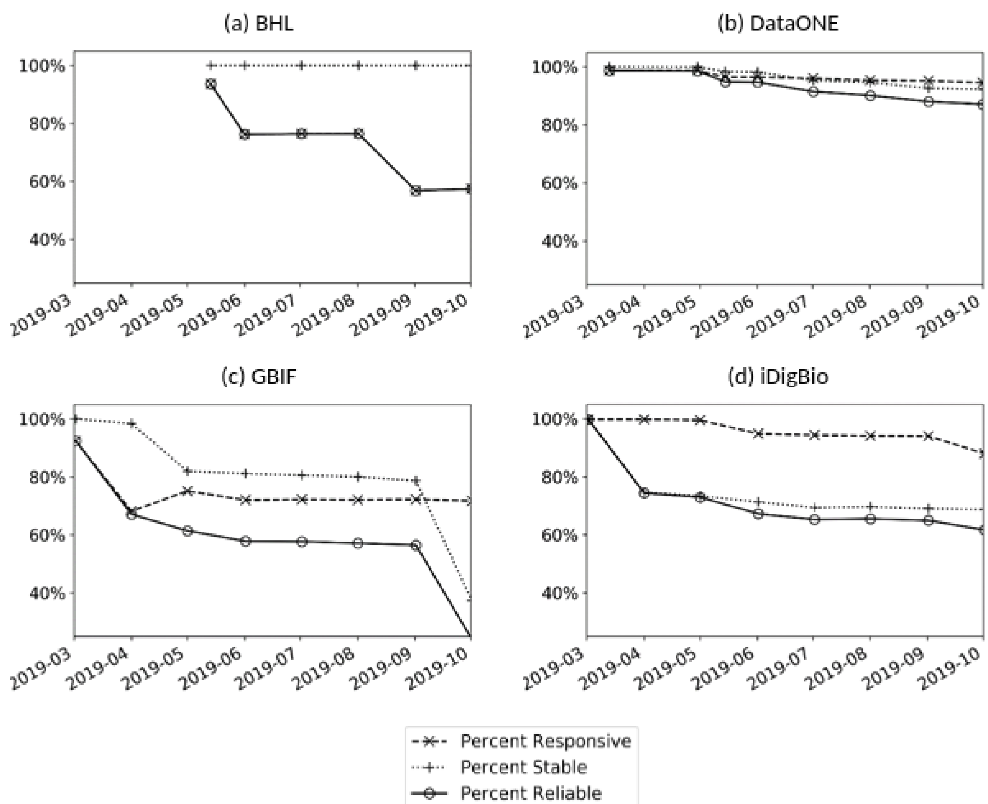

Figure 2. Overall responsiveness, stability, and reliability from March 2019 to October 2019 as a percentage of URLs that exhibit each indicator in a) BHL, b) DataONE, c) GBIF, and d) iDigBio.

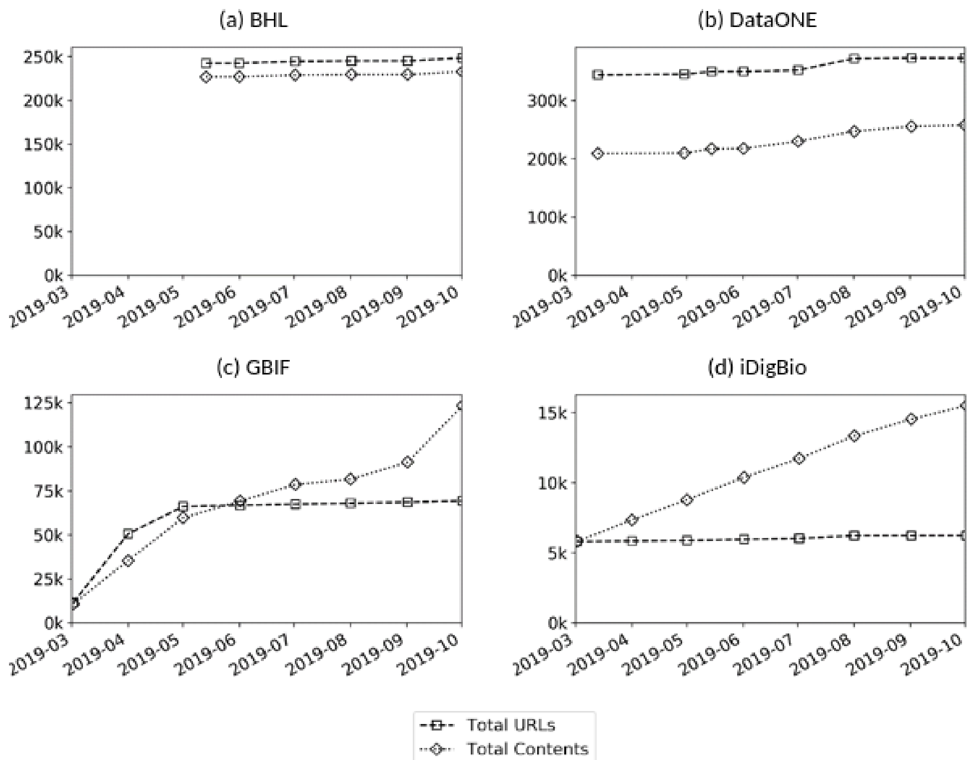

Figure 3. Total number of URLs and unique contents observed from March 2019 to October 2019 for a) BHL, b) DataONE, c) GBIF, and d) iDigBio.

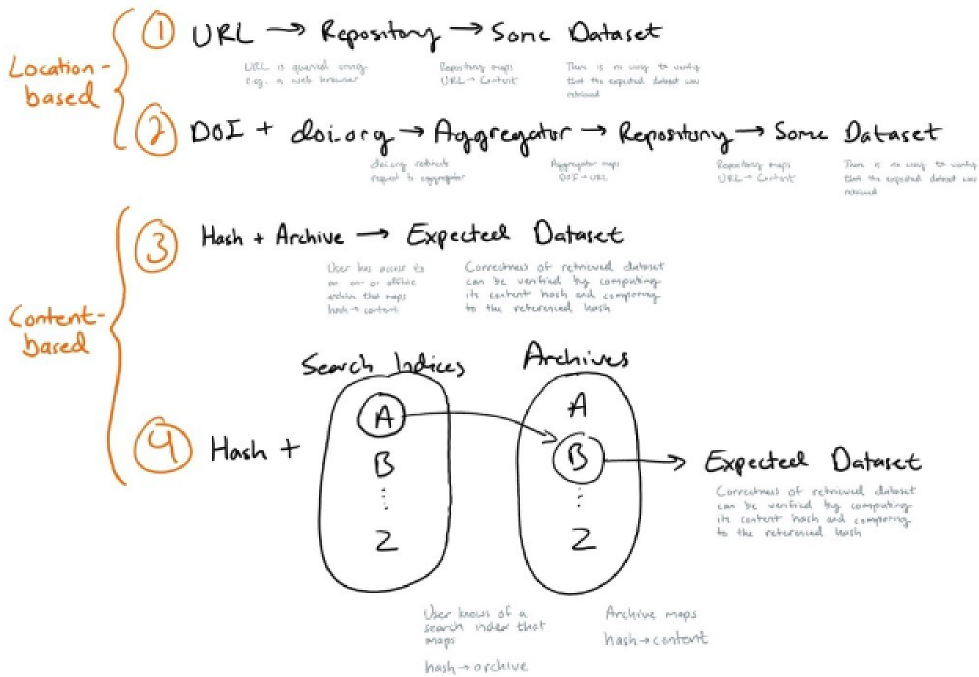

Figure 4. Visualization of content resolution for location- versus content-based identifiers. 1) URLs point to a known location of a dataset, but do not guarantee either the presence or authenticity of the retrieved dataset; 2) the use of a DOI that resolves to a URL adds a layer of redirection; 3) A content-addressed dataset can be found by matching against recomputed hashes of available datasets in an archive; 4) well-known (online) hash indices can be used to facilitate discovery of dataset locations associated with a specific content hash.

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.

Tables

Data Network	Responsive URLs	Stable URLs*	Reliable URLs
BHL	57.41% (142,672)	99.97% (232,996)	57.39% (142,633)
DataONE	94.55% (352,438)	92.27% (339,109)	87.09% (324,641)
GBIF	71.72% (49,707)	37.35% (20,094)	24.05% (16,669)
iDigBio	88.04% (5,477)	68.69% (4,251)	61.68% (3,837)
All observed URLs	78.94% (546,645)	90.43% (593,469)	70.07% (485,203)

Table 1. Overall responsiveness, stability, and reliability for URLs observed in each biodiversity data network and for all observed URLs as of October 2019. * URLs that never provided content were omitted from the divisor when calculating Stable URLs percentages.