Integration, Attribution, and Value in the Web of Natural History Museum Data

In 2012 the community articulated a plan for a Network Integrated Biocollections Alliance (NIBA) (https://www.nsf.gov/bio/pubs/reports/niba_implementation_plan.pdf). Goal 2 of the NIBA Implementation plan expresses a plan to "Advance engineering of the US Biocollections cyberinfrastructure". Within this goal, a number of subgoals are presented:

2.1. Create a national database of all digitized specimen records from US institutions and agencies.
2.2. Establish a research and development environment to deliver new specimen digitization workflow methods, tools, and techniques.
2.3 Complete development of required standards and protocols.
2.4. Promote a consensus for the adoption of standards.
2.5. Anticipate the future of biodiversity specimen data integration.
2.6. Develop a strategy for long-term data archiving of specimen information, including 2D and 3D images, text information, and metadata about digitization processes.
2.7. Support the development of a robust, Web-services-based architecture for handling taxonomic names applied to specimens as determinations and annotations.

Although much work remains, numerous groups have made significant progress on a number of these individual goals. Thanks to the efforts of iDigBio (https://www.idigbio.org/), ADBC (https://www.nsf.gov/pubs/2013/nsf13569/nsf13569.htm), the TCN's and individual collections, more specimens are being digitized and imaged. Data is being published at an accelerated rate due to advances in collections management systems (CMS's) data models and tools, publishing tools and the availability of numerous data aggregation portals. iDigBio along with SPNHC (http://www.spnhc.org/) and TDWG (http://www.tdwg.org/) are leading the charge in standards, best practices, digitization protocols and the dissemination thereof through publications, workshops, webinars, and wikis. Groups like Cyverse (http://www.cyverse.org/), Zenodo (https://zenodo.org/), and DataONE (https://www.DataONE.org/) are developing a strategy for long-term data archiving while BCoN (https://bcon.aibs.org/), SPNHC, NSCA (http://www.nscalliance.org/) and iDigBio are involved in galvanizing the community around a common purpose.

While projects such as Wholetale (http://wholetale.org/) and BiotaPhy (http://wiki.biotaphy.org/) have begun to look at data integration, this represents the major remaining challenge in collections and research data. Data needs a pipeline beginning with collection of specimen data in the field, to incorporation of this data into collections databases and subsequent aggregators, to facilitating use of this data by the wider research and end user community with as little human intervention as possible using technological bioinformatics products. There are aspects of all interactions in the data pipeline that require further development and functionality by the various actors involved.

Fig 1: Graphical representation of the data landscape and all actors in the data integration and attribution pipeline.

All members of the data pipeline are unfortunately primarily placing emphasis on the outward flow of data and resources, and as such, attribution of data products to the source collections for the purposes of collections advocacy is a major issue (Pyke & Ehrlich, 2010; Arbelaez-Cortes et al., 2017; Rouhan et al., 2017). Collections are at the heart of specimen-based research but currently are not able to adequately highlight their contribution to this research or receive credit for their contribution due to the currently inconsistent mechanisms for citing products of these research endeavors.

Collections rely on a variety of metrics of collections use to advocate for the utility and future funding of collections and their infrastructure. A number of these are metrics generated through internal usage by the collections, such as number of loans, research visitors, data requests and internal publications. However, it is the external collection specimens and data research use that are important to administrators and funding agencies, and therefore of greatest impact on collections advocacy. Collections use of this kind falls into two main categories:

- Use of specimens in primarily taxonomic and systematic research endeavors (specimen pathways)
- Use of collections data where no specimens are examined e.g. niche modelling for species distribution, climate change, invasive species (data pathways)

These uses are usually highlighted through publication, as well as other primary research products e.g. GenBank or Isobank sequences. There are also secondary products produced during the research process that may, or may not, be published such as images, CT scans, x-rays etc. These products need to be connected back to the original specimens to augment these records and provide these resources to the research and other external user communities. Unfortunately, tracking use of collections in publication and through lodging of products of these research endeavors is becoming increasingly difficult. Attribution and collections advocacy through specimen and data usage is also complicated by the convoluted web of access points including individual collections and the ever-increasing number of aggregators now providing collections data. The fact that there is no standardization across this landscape is severely hampering both integration and attribution. The benefits to collections from other players such as aggregators is not being adequately highlighted or addressed. There are numerous benefits that aggregators could be providing to the collections community that are only possible at the aggregator level – measures of uniqueness, data cleanup tools as well as annotation functionality. The bi-directional nature of this data pipeline is essential to ensure the viability of collections infrastructure and continued growth of collections for the sustainability of the remainder of the pipeline. Collections are an essential element in the data pathway but are reliant on the remaining elements (aggregators, researchers, repositories, publishers, funders, etc.) to provide the necessary structures to ensure this attribution. At present, it is extremely difficult for collections to accumulate data usage statistics from the various aggregators because there is no standardization of statistics or they are just not being provided. Similarly, it is equally difficult to link data products such as publications, GenBank sequences, and other research products back to specimen records, the primary mechanism employed by collections to show use of collections for advocacy purposes. This is due to the community's unwillingness or inability to do so but also due to the ever-increasing variety of collections data uses by an ever-broadening user community. This user community also requires scalable data of high quality to effectively utilize this data to its full potential. It is thus important in this context to also address issues of scalability of data with other products outside of our realm as well as data quality or fitness for use.

In an attempt to address these issues, a BCoN Data integration and Attribution Needs Assessment workshop was held in Lawrence, KS, February 13-14, 2018 (https://bcon.aibs.org/working-groups/data-

attribution-and-integration/). During two days of presentations and discussions, a number of conclusions were reached regarding the integration and attribution pathways and possible solutions. Thirty-five attendees representing a selection of the actors involved in these pathways (collections, collection management databases, aggregators, repositories, researchers and the library community) presented their perspectives on data integration and attribution on the first day and these oral presentations informed discussion on the second day. Initially, smaller groups discussions were convened with the aim of discussing the integration and attribution data cycles individually. Groups were formed with representation of all actors in the data pipeline and each group reported to the attendees after each session. The day culminated in a joint discussion of the interactions between the two cycles and potential solutions to the attribution issue.

It was clear from the discussions that solving the attribution issue is achievable, a pressing need, and of vital importance to the sustainability of the collections infrastructure necessary for the effective use of these collections and their future. Additionally, as the two cycles are integrally linked, in solving the attribution pathway, we can effectively solve the integration pathway simultaneously which will inform and promote reproducible science and the delivery of research products to a wider community of users.

It was evident that a functional attribution cycle requires technological and social solutions that provide both the bioinformatics tools necessary to link data across the wide variety of pathways in the pipeline as well as the development of best practices and standards necessary to guide the individual actors in the pipeline to standardize their practices.

The landscape is a complex web of interactions between multiple actors both within and between levels with multiple possible access points. The ultimate aim is to create a system with the ability for users to access the pathway at any point and still be able to navigate the data pathway in either direction to gain access to all the content related to a single record. For instance, a researcher accessing a GenBank record should be able to link to the original specimen voucher record, publications citing the specimen or GenBank sequence as well as any other products associated with the specimen (images, CT scans etc.). Likewise, a researcher accessing the specimen record either from an aggregator or directly from the collection should be able to traverse that same data cycle in the opposite direction in essence linking all of the disparate sources into a distributed web of information accessible from any space and by any actor. This functionality will create a truly distributed web of information and negate the need for centralization of data products by any one actor. The technological key to making this possible lies in having a standardized set of identifiers for the individual occurrence records, data sets and individual institutions and collections that persist throughout the data cycle. However, it is recognized that the creation and adoption of best practices and standards are also necessary to address the social issues inherent in such a complex, distributed system.

The final discussion period culminated in a proposed solution to the attribution cycle based on the use of a combination of three identification mechanisms for the various elements of the data cycle. This solution is articulated in Figure 2, which highlights the role played by these various identifiers in linking the various actors within the cycle.

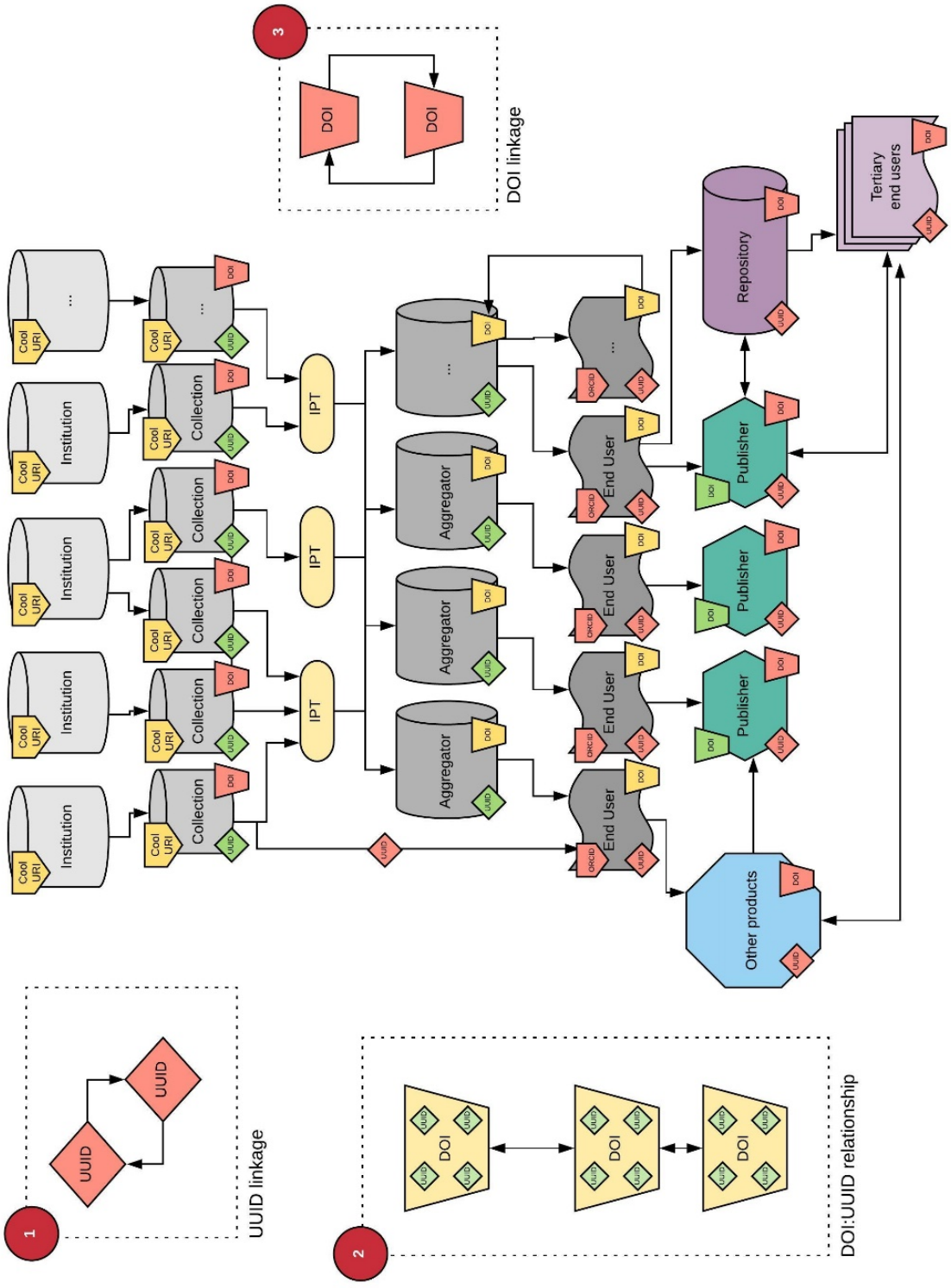

Figure 2: Visual representation of actors and identifier linking mechanisms proposed as a solution to the biodiversity data integration/attribution cycle.

The success of this solution hinges on a number of key properties of these identifiers - the **existence** or **minting** of these identifiers for each element; their **uniqueness** and thus **resolvability**; and their **persistence** not only through time but throughout the data cycle as indicated in Figure 2. The four classes of identifiers proposed are as follows:

- Cool URI's (Uniform Resource Identifiers) for institutions and collections. These URI's would be minted by a central authority such as GRBIO (http://grbio.org/) and would be published as part of the collection dataset through IPT. They would also be used to identify collections contributing to published works.
- UUID's (Universally Unique Identifiers) or GUID's (Globally Unique Identifiers) for individual occurrence records minted either at the collection provider level within the collection management system (CMS) or at the aggregator level for those collections unable to do so themselves. There has been much debate in the literature about what constitutes an identifier and how best to apply these to occurrence records in biodiversity. Numerous solutions have been proposed from Darwin Core triplets, to UUID's, GUIDs or LSIDs (Life Science Identifiers) to URIs and DOIs (Page, 2008, 2009, 2016; Pereira et al., 2009; Guralnick et al., 2014, 2015; Hyam et al., 2012, Nelson et al., 2018; McMurry et al., 2017).
- DOI's (Digital Object Identifiers) provided for downloaded datasets. These DOI's would include the UUID's of the individual records contained within the downloaded dataset. New DOI's would need to be created for any changes made to that dataset - elimination of records deemed unfit for purpose e.g. removal of non-georeferenced records for niche modeling; or augmentation through e.g. georeferencing or taxonomic cleanup.
- ORCID identifiers (Open Research and Contributor IDentifiers) (https://orcid.org/) for the agents involved in the data cycle. As well as attribution for collections and institutions, it is equally important to allow for attribution for the work done by individuals in augmenting, editing, correcting records or datasets in the data cycle (Katz, 2014; McCallum et al., 2013).

This cycle is by no means mature or complete. There are numerous areas where the identifiers or linkages needed are either not present or inconsistently applied or utilized. The hope is that through discussion this system can be refined and improved to lay the groundwork for the production of an attribution and integration platform that will benefit all actors. Below is an outline of the challenges and problems envisaged before this proposed solution can be implemented.

The single most important issue is the consistent application of a single system of identifiers to all occurrence records. Presently our community suffers from a lack of consensus regarding which identification system to use while the systematic application of unique occurrenceIDs to occurrence records is still inconsistent or absent.

Through the use of Darwin Core data standards (http://rs.tdwg.org/dwc/), the IPT publishing toolkit (https://www.gbif.org/ipt) and mature collections management systems, we have a system of UUIDs applied to individual occurrence records as an occurrenceID that is functional from a number of collections or data providers to the aggregators. Most of the ubiquitously used collection management software solutions, e.g. Arctos (https://arctosdb.org/), KE EMu (https://emu.axiell.com/), Specify (http://www.sustain.specifysoftware.org/), Symbiota (http://symbiota.org/docs/) etc., provide functionality to create UUID's for all specimen or lot based records at the source. However, although UUIDs are becoming more ubiquitous, there are still a large number of digital occurrence records with non-standardized occurrenceIDs. Application of UUID's to all records at the collection level is inconsistent due to the inclusion of non-voucher based project and observation records, the limitations

of other collection management software packages used by the community and/or a lack of understanding of need.

Additionally, given the lack of a centralized UUID resolving service in biodiversity data similar to the IGSN (http://www.geosamples.org/igsnabout) and ISBN (https://www.isbn-us.com) used in the geology and library communities respectively, there is no guarantee that these UUIDs are indeed unique. Given the current practice of minting UUIDs at the source collections (or in some cases at the aggregator) with no linkage between disparate collections, they are only guaranteed to be unique in the local context of the individual database or aggregator. In fact, looking at the Darwin Core occurrenceID field in any of the aggregators it is clear that this is not the case. Collections based voucher material has a mismatch of various identifiers (LSIDs, URNs, URLs, triplets, etc.). In some cases UUIDs have been applied at the aggregator level for those collections unable to do so themselves and given the present landscape of multiple aggregators operating in this space, this causes additional problems of uniqueness.

A solution to this uniqueness problem is a vital component to the success of this proposal and highlights the need for a UUID resolving service as well as a re-envisioning of the current aggregator model. The funding and long-term support of such a service would be crucial in the success of any identifier-based solution and yet, no host for such a service has been identified and no funding model envisaged. Currently each aggregator holds its own cache of data records and performs various data manipulations on this cache to effect data standardization and use metrics. This allows for not only the possibility of inconsistent versions of data and identifiers across the landscape but also inconsistency and unnecessary complexity reporting of usage metrics to providers for the purposes of collections advocacy. Although the aggregators have begun to use TDWG inspired standardized data cleanup tools, at present there is little standardization of the data standard practices or application of UUIDs. GBIF is currently working on a proposal that would highlight the need for a centralized data cache used by all aggregators from which all data cleanup tools, data standard practices and data use metrics could be obtained in a standardized fashion. As such, this centralized data cache could simplify and standardize the current aggregator landscape with respect to data identifiers, data uniformity and usage metrics while also functioning as the UUID resolving service needed by the community.

A related issue is that there are two major mechanisms of use of data and specimens from collections -

- **Access through aggregators.** Typically, this involves both specimen based individual record use for taxonomic and systematic research as well as batch record downloads for modeling and other such uses.
- **Access through direct contact with collections by end users.** This use bypasses the aggregator infrastructure altogether and is typically specimen-based use by taxonomic and systematic research users. It does however make up the bulk of published literature linked to physical specimens and tissues in collections.

In order to facilitate the standardized flow of data through the attribution pipeline, both pathways would need to utilize the same application of UUID based records. When access to data or specimens occurs directly from collections, where the practice of using the standard triplet of institutional code, collection code and catalog number (or some lesser combination) is more pervasive, sharing of UUIDs is not typical.

Even with a consistent system of identifiers in place at the collection and aggregator level, at present, identifiers do not persist into publication, other data repositories (GenBank, Isobank, EOL etc.) or to

additional products created from specimen use - images, CT scans, sounds, video. The current practice of using institution: collection: catalog number triplets or some combination of these does not support the rigorous needs of a true identifier infrastructure.

Likewise, DOIs to describe data sets are inconsistently applied by aggregators, repositories, publishers and end users and the community lacks the necessary infrastructure to support resolution and linking of these DOIs. When datasets are used for modeling or other data applications, there is no consistent mechanism of identifying these downloaded datasets or tracking any modifications or augmentations made to them by downstream use. In order to track these augmentations or changes, a system of linking DOIs is required. Similarly, ORCID IDs are needed to identify and attribute the individuals creating these DOIs and subsequently contributing to the augmentation, editing or correction of data records. Attribution for this work will produce positive incentives that will encourage more users to contribute in this way.

Along with these fundamental principles of identifiers necessary to make this plan work, there are also ancillary properties that need addressing as shown in Figure 2:

1. UUID linking to facilitate the expression of relationships between objects - host:parasite, tissue:voucher, herbarium and other duplicate specimens, associated derivative products such as images, CT scans, sounds, videos, field notes, etc.
2. The inherent inclusion of all UUIDs in DOI datasets to allow identification of individual occurrence records associated with dataset use, together with;
3. The similar linkage of DOIs to facilitate the tracking of any subsequent editing, correction or augmentation of records within the dataset by end users.

Although this solution relies heavily on the technological aspects of this system of identifiers, there is also a large social element with community standards, best practices and encouragement required to ensure the implementation and use of these identifiers ubiquitously throughout the data pipeline. The collections community as a whole through individual collections, collections based societies (like SPNHC, NSCA, NatSCA, TDWG etc.) and taxonomic discipline specific societies have a role to play in promoting the creation and use of these identifier systems. Funding agencies such as the National Science Foundation (NSF) also have a role to play in setting policy and standard practices for specimens collected and data lodged from grant funded activities.

So, what needs to be accomplished by the various groups in the data cycle to effect the above solution?

Collections need to use database software that allows for the minting of unique identifiers at the source. It would be ideal if minting of these identifiers occurred in the field during original collection of the specimens, but numerous problems with varying collecting practices and the research: collections interface hamper this possibility. Collections need to utilize and preserve existing unique identifiers when directly sharing specimens or data to end user communities for use, or when exchanging or gifting specimens to other collections. Collections should continue to digitize collection records to increase the number of records in the digital realm that have UUIDs while publishing data to aggregators to allow for data reuse by the end user community. Collections also need to facilitate the creation of linkages between related entities within databases - tissue:voucher, host:parasite, GenBank sequences, publications, or other products etc.

Aggregators should work towards creating a single data cache model to facilitate standardization or, short of this, standardize practices across the aggregator landscape to facilitate integration and attribution. They also need to do a better job of providing reliable, standardized usage metrics of data use by end users and mechanisms that are more effective for users to report potentially erroneous data or data augmentations. Aggregators must preserve and promote the unique identifiers already assigned to aggregated content.

Repositories need to create the mechanisms and infrastructure necessary to allow for the use of occurrence record and dataset identifiers to facilitate their persistence through the data lifecycle.

End Users need to be collections advocacy aware. They need to be incentivized to not only utilize consistent identifiers on all specimen record and dataset use through publication, but also repatriate any additional products created from this use - be it dataset augmentations in the form of georeferencing etc. or taxonomic or other data corrections, or products of research in the form of images, CT scans etc.

Publishers and their editorial boards need to institute policies requiring use of unique identifiers for specimens, datasets and collections in publications to ensure ability to track use. The Pensoft ARPHA writing tool is a good model of how this can be accomplished but more publishers need to adopt or copy this functionality.

This plan can only be realized through participation and inclusion of all actors in the data pipeline and the hope is that through future efforts these actors can be engaged and persuaded as to the merits of this plan.

Next steps

- Distribution of this document to the community for further review, comment and discussion.
- Further engagement of the community through presentation of the plan at symposia at both the 2nd Digital Data in Research meeting in Berkeley in June (https://www.idigbio.org/content/registration-now-open-emerging-innovations-biodiversity-data#) and the joint TDWG/SPNHC meeting in Dunedin, New Zealand in August (http://spnhc-tdwg2018.nz/).
- Journal publication of the final plan to engage all actors in the data cycle in an attempt to put structures in place that will facilitate the use of identifiers in data integration and attribution.

References

Arbeláez-Cortés, E., Acosta-Galvis, A.R., DoNascimiento, C. et al. Scientometrics (2017) 112: 1323. https://doi.org/10.1007/s11192-017-2461-4

Guralnick, R., T. Conlin, J. Deck, B. J. Stucky, and N. Cellinese. 2014. The trouble with triplets in biodiversity informatics: A data-driven case against current identifier practices. PLoS One 9(12): e114069.

Guralnick, R. P., N. Cellinese, J. Deck, R. L. Pyle, J. Kunze, L. Penev, R. Wals, et al. 2015. Community next steps for making globally unique identifiers work for biocollections data. ZooKeys 494: 133–154.

Hagedorn, G. 2013. Beyond Darwin Core–Stable identifiers and then quickly beyond towards linked open data. TDWG 2013, Florence, Italy. https://www.slideshare.net/G.Hagedorn/tdwg-2013-florence-italy-hagedorn-Beyond-dw-c-stableids-linkedopendata.

Hagedorn, G., T. Catapano, A. Güntsch, D. Mietchen, D. Endresen, S. Sierra, and Q. Groom, et al. 2013. Best practices for stable URIs [online]. http://wiki.pro-ibiosphere.eu/wiki/Best_practices_for_stable_URIs

Hyam, R., R. E. Drinkwater, and D. J. Harris. 2012. Stable citations for herbarium specimens on the internet: An illustration from a taxonomic revision of Duboscia (Malvaceae). Phytotaxa 73(1): 17–30.

Katz, D S 2014 Transitive Credit as a Means to Address Social and Technological Concerns Stemming from Citation and Attribution of Digital Products. Journal of Open Research Software, 2(1): e20, pp. 1-4, DOI: http://dx.doi.org/10.5334/jors.be

McCallum, I., H.-P. Plag, S. Fritz and S. Nativi. 2013. Data Citation Standard: A means to Support Data Sharing, Attribution, and Traceability. E3S Web of Conferences, 1 (2013) 28002. DOI: https://doi.org/10.1051/e3sconf/20130128002.

McMurry, J. A., N. Juty, N. Blomberg, T. Burdett, T. Conlin, N. Conte, M. Courtot, et al. 2017. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLOS Biology 15(6): e2001414.

Page, R. D. M. 2008. Biodiversity informatics: The challenge of linking data and the role of shared identifiers. Briefings in Bioinformatics 9(5): 345–354.

Page, R. D. M. 2009. bioGUID: Resolving, discovering, and minting identifiers for biodiversity informatics. BMC Bioinformatics 10(Suppl 14): S5.

Page, R. D. 2016. Towards a biodiversity knowledge graph. Research Ideas and Outcomes 2: e8767.

Pereira, R., K. Richards, D. Hobern, R. Hyam, L. Belbin, and S. Blum. 2009. TDWG Life Sciences Identifiers (LSID) applicability statement, version 2009-09. Biodiversity Information Standards (TDWG) [online]. http://www.tdwg.org/standards/150.

Pyke, G.H., and P.R. Ehrlich. 2010. Biological collections and ecological/environmental research: A review, some observations and a look to the future. Biological Reviews of the Cambridge Philosophical Society. 85:247–266.

Rouhan, Germinal & J. Dorr, Laurence & Gautier, Laurent & Clerc, Philippe & Muller, Serge & Gaudeul, Myriam. (2017). The time has come for Natural History Collections to claim co-authorship of research articles. Taxon. 66. 1014–1016. 10.12705/665.2.

Funding for this Workshop

This workshop and report were made possible by a National Science Foundation grant (DBI #1441785) to the American Institute of Biological Sciences (AIBS).

Acknowledgements

Planning for this workshop was led by Andrew Bentley of the University of Kansas with support and contributions from the Biodiversity Collections Network Advisory Council. The contributions of all workshop participants are greatly appreciated. Syreeta Jones of the American Institute of Biological Sciences provided logistical support for the meeting, as did a number of individuals with the Univeristy of Kansas.

Suggested Citation:

Biodiversity Collections Network. 2018. Integration, Attribution, and Value in the Web of Natural History Museum Data: A Needs Assessment Workshop. February 13-14, 2018. Lawrence, KS. https://bcon.aibs.org