Digitization and the future of natural history collections

Hedrick, B. P.*[a,b], Heberling, J. M.+[c], Meineke, E. K.+[b,d], Turner, K. G.+[e], Grassa, C. J.[d], Park, D. S.[d], Kennedy, J.[d], Clarke, J. A.[f], Cook, J. A.[g], Blackburn, D.C. [h], Edwards, S. V.[b], Davis, C. C.*[d]

[a]Department of Earth Sciences, University of Oxford, Oxford OX1 3AN United Kingdom
[b]Department of Organismal and Evolutionary Biology, Harvard University, Cambridge, MA 02138
[c]Section of Botany, Carnegie Museum of Natural History, Pittsburgh, PA 15213
[d]Harvard University Herbaria, Harvard University, Cambridge, MA 02138
[e]Department of Biology, Pennsylvania State University, University Park, PA 16802
[f]Jackson School of Geosciences, University of Texas Austin, Austin, TX 78712
[g]Department of Biology, University of New Mexico, Albuquerque, NM 87131
[h]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611

*Corresponding authors: bphedrick1@gmail.com, cdavis@oeb.harvard.edu
+Contributed equally

Keywords: digitization, herbaria, natural history collections, specimens, Anthropocene, baselines

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27859v1 | CC BY 4.0 Open Access | rec: 15 Jul 2019, publ: 15 Jul 2019

Abstract

Natural history collections (NHCs) are the foundation of historical baselines for assessing anthropogenic impacts on biodiversity. Along these lines, the online mobilization of specimens via digitization–the conversion of specimen data into accessible digital content–has greatly expanded the use of NHC collections across a diversity of disciplines. We broaden the current vision of digitization (Digitization 1.0)–whereby specimens are digitized within NHCs–to include new approaches that rely on digitized products rather than the physical specimen (Digitization 2.0). Digitization 2.0 builds upon the data, workflows, and infrastructure produced by Digitization 1.0 to create digital-only workflows that facilitate digitization, curation, and data linkages, thus returning value to physical specimens by creating new layers of annotation, empowering a global community, and developing automated approaches to advance biodiversity discovery and conservation. These efforts will transform large-scale biodiversity assessments to address fundamental questions including those pertaining to critical modern issues of global change.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27859v1 | CC BY 4.0 Open Access | rec: 15 Jul 2019, publ: 15 Jul 2019

I. The relevance and importance of digitization

Anthropogenic impacts, including urbanization, globalization, and climate change, are rapidly transforming our world. Despite our best efforts, however, quantifying the biotic impacts of human activity has been challenging, as evidenced by the difficulty of delimiting the onset of the Anthropocene (Lewis and Maslin 2015). Part of this uncertainty stems from a lack of historical data that track biotic change over time. However, natural history collections (NHCs), with their broad taxonomic, geographic, and temporal scope, offer a key solution to this impasse. In the past twenty years, there has been a dramatic increase in the use of NHCs for assessing a wide variety of scientific questions (Suarez and Tsutsui 2004, Pyke and Ehrlich 2010, Park and Potter 2015, Meineke et al. 2018, 2019). Indeed, they have emerged as one of the best resources for establishing biological baselines to understand the impacts of, for example, the origins of agriculture, the industrial revolution, the development of nuclear armaments, and more generally the influence and acceleration of anthropogenic change on biodiversity (Moritz et al. 2008, Johnson et al. 2011, Lister 2011, Funk 2018, Nelson and Ellis 2018).

Most large NHCs provide specimen data to researchers and the public by mobilizing searchable collection databases online. We assert that these mobilized collections are among the most important advances in museum curation in the past century, significantly opening access to NHCs and greatly stimulating large-scale analyses that span novel academic and societal enterprises. These resources are connecting diverse scholarly domains, propelling a new generation of scientists forward, and removing financial, sociological, institutional, and academic obstacles preventing access to these materials

(Drew et al. 2017, Sweeney et al. 2018). In short, digitizing a specimen–translating metadata associated with a physical specimen object into flexible digital data formats–increases the value of the physical specimen exponentially.

Here, we present an ambitious, two-pronged vision for digitization, which we term Digitization 1.0 and Digitization 2.0. Digitization 1.0 represents the ongoing push to create digital images and related content directly from physical voucher specimens; Digitization 2.0, in contrast, relates exclusively to data gathering, tasks, or workflows derived from digitized products of Digitization 1.0 rather than from the physical specimens themselves (figure 1). In addition to the vast expansion and online aggregation of these mobilized collections to create a truly global digital NHC, Digitization 2.0 offers the promise of also shifting and growing the workforce and public who interface with these objects to accelerate the progress of digitization.

II. Digitization 1.0: The Past, Present, and Future

Digitization of NHCs began with the overarching goal of documenting specimen inventory and facilitating research by transcribing label information into centralized, searchable databases as described recently by Nelson and Ellis (2018). These efforts have given rise to Digitization 1.0, which has been widely embraced and continues to be infused with innovation. Digital representations generated through Digitization 1.0 include specimen images and direct transcriptions of specimen metadata from handwritten or printed collection catalogs or labels, including for example details on coloration or measurements. As part of this effort, NHCs have generated millions of digital representations of physical vouchers and have devised numerous technological innovations to facilitate efficient data generation, including conveyor belt and robotic imaging techniques for mass digitization of specimens (Tegelberg et al. 2014, Sweeney et al. 2018). More recent next generation technologies, including photogrammetry, laser-scanning, and computed tomography, create far richer digital representations of specimens than can be visualized by eye or with standard microscopy (figure 2). Given that large portions of most NHCs still remain unavailable in digital format, the innovations and efforts within Digitization 1.0 will continue well into the future, likely for decades. In the subsections below, we outline Digitization 1.0 through the lens of digitization workflows, strategic prioritization, and solutions to impediments.

Digitization workflows and linking data–The practice of digitization is broadly consistent among projects and organismal groups, in so much as each specimen is represented by textual metadata from labels or catalogs and typically digital two-dimensional images, but increasingly also three-dimensional representations and audio or video recordings where relevant. There exists great variation in specimen size, storage conditions (e.g., fluid-preserved, microscope slides, dry storage), dimensionality (2D versus 3D representation), and detail associated with specimens, not to mention widely varying practices in specimen collection and curation across taxonomic domains and institutions. This heterogeneity of collections and institutional policies and priorities thus creates challenges to efficient mass imaging and gathering of metadata. However, at minimum, digitization workflows should attempt to integrate all available specimen metadata into digitization efforts and appropriately link these data to their associated

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27859v1 | CC BY 4.0 Open Access | rec: 15 Jul 2019, publ: 15 Jul 2019

physical voucher specimens. Beyond traditional linkages, non-traditional metadata associated with the specimen include biotic (e.g., mass) and abiotic data (e.g., climate), media (e.g., video and audio recordings), community- and population-level metadata (e.g., abundance), species observations in the field, and genetic samples (i.e., the "extended specimen" *sensu* Webster 2017). Much of these digital data are served in part or in their entirety via online collection databases (e.g., Arctos, Specify, Symbiota, EMu) or in data aggregators (e.g., iDigBio, Global Biodiversity Information Facility–GBIF, Botanical Information and Ecology Network–BIEN). Linking voucher specimens to these new data layers generated post collection is important and has been facilitated by associating URLs, data accession numbers, digital object identifiers (DOI), or ARKs with specimen records in collection databases. In addition, trait data can be incorporated into specimen records using extensions to the Darwin Core Archives (Yost et al. 2018). For the next generation of collections, protocols are under development to expand the digitization workflow to the collecting event itself (Heberling and Issac 2018).

Developing digitization priorities–Given the limited resources available to many NHCs, it is necessary to establish priorities for specimen digitization. Specimens at risk of degradation, such as rare or fragile fossils, and those representing rare or threatened species and habitats are candidates for high priority digitization. Further, efforts should focus on specimens with rich associated metadata from the collection event. A growing number of species are imperiled, and conservation biologists are increasingly reliant on NHCs for baseline data to understand species ranges and climatic tolerances for assessing future changes (Lister 2011). Distributing information for these rare or threatened taxa to conservation biologists is increasingly critical to these species' management and survival (MacDougall et al. 1998, Nualart et al. 2017). Finally, taxa representing a breadth of evolutionary history or unique adaptations are important for research on phenotypic evolution, community ecology, and biologically inspired design. We suggest that such specimens have high priority for digitization.

Owing to the varying effort required by different digitization strategies (e.g., label data, images, 3D reconstructions), data types that serve the largest diversity of use cases should also be prioritized. For instance, key information including taxon name, collection locality, and date can be captured relatively efficiently and can facilitate assessments of species distributions through time. Rapidly expanding areas of research including phenology (e.g., Primack et al. 2004, Willis et al. 2017), large-scale taxonomic inventories (e.g., Cardoso et al. 2017), and morphometric investigations (e.g., Hedrick et al. 2015), rely on such label data and data from post-digitization enhancement (Sweeney et al. 2018). For example, in one of the first studies to demonstrate how historic specimens can be used to quantify the biotic effects of climate change, Primack et al. (2004) used flowering plant specimens collected between 1885 and 2003 in the greater Boston (USA) area to demonstrate that plants were flowering up to eight days earlier in recent years than in the early years of the 20th century. The utility of such diverse data (e.g., geographic location, flowering date, anatomical measurements) is important to a wide array of researchers and should be prioritized. Additionally, we feel it is best to only apply more complex, holistic digitization methods on a key subset of data-rich specimens as has been recently demonstrated in the openVertebrate (oVert) Thematic Collection

Network (Blackburn et al., NSF Abstract #1701714). Increasing the magnitude of the collection of media files (e.g., photogrammetry of bird skins, nuts, etc.) for this subset of data via new pipelines and technological advances will be critical to this effort.

Past impediments and future solutions–Despite the success of Digitization 1.0, this initiative has identified three issues that must be addressed to maximize efficiency of information retention and distribution. First, museums are obligated to manage, store, and steward additional digital data associated with their physical collections. However, the act of digitization entails significant challenges since it requires sustainably curating both the physical objects and rapidly emerging digital datasets. This issue will necessitate the development of new tools, that centralized aggregators assume increasing responsibility, and will require increased funding in the near future (see Digitization 2.0 below).

Second, there is concern that large aggregators aimed at connecting researchers with NHCs (e.g., GBIF, iDigBio) (Edwards 2004) remove NHCs from the attribution chain. NHCs are frequently funded on their research relevance, which is determined both from within and outside institutions. When researchers view specimen images or harvest metadata from aggregators, NHCs that contribute these data often receive little to no credit (Rouhan et al., 2017). A mechanism for referencing these source collections needs to be embedded in the publication process that requires that NHCs be acknowledged and notified when publications incorporate their data. A viable solution to this problem is to mint a digital object identifier (DOI) for a digitized specimen and establish a reporting mechanism for collections to be alerted when their specimens have been cited.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27859v1 | CC BY 4.0 Open Access | rec: 15 Jul 2019, publ: 15 Jul 2019

Automating this attribution pipeline as part of the digitization workflow better ensures that NHCs receive credit for stewarding both voucher specimens and also digitized data, which is likely to stimulate NHCs to embrace open-access policies for their data.

Third, digitized data are inconsistently and redundantly spread across multiple databases at different scales. NHCs often have their own databases, but some data are additionally deposited in regional databases, taxon-specific databases, and national and international data aggregators. This data dispersion causes information to be input/archived redundantly such that each database has a variant of the post-digitization metadata, leading aggregators to archive either inconsistent or duplicated copies of the same primary data. This problem can be partially circumvented by more communication among data aggregators, as well as between NHCs and aggregators. Algorithms linking specimen numbers between aggregators could ensure that post-digitization enhancement metadata are transferred to all aggregators mentioning particular specimens by unique identifiers such as the specimen-based occurrenceID. This is done internally at iDigBio through the iDigBio Record API, which retains current and previous iterations of a specimen's data.

III. Digitization 2.0: charting a road map for the future

Unlike Digitization 1.0, which directly utilizes the physical specimen, Digitization 2.0 instead utilizes the digitized product from Digitization 1.0 for generating additional data and metadata (figure 1). Digitization 2.0 is powered by the online aggregation of these resources and enables digitization to assume new forms and engage vast new workforces.

As we outline below, Digitization 2.0 is already well underway and holds tremendous promise. It includes semi- or fully automated data recording from digitized specimens, which stimulates research and returns value to the physical specimen. Additionally, Digitization 2.0 entails a shift in the workforce engaged in collections science and stewardship. Finally, Digitization 2.0 leverages NHC resources to create trait databases, either from aggregating and better indexing existing metadata or by allowing researchers or citizen scientists to associate trait annotations with images served from NHC databases.

Innovative tools for automating digitization: machine learning and neural networks– Given the massive number of specimen images in digital databases with minimal data, an important first step is to better automate data transcription to augment these skeletal records. The enormity of this task is quickly becoming impossibly large for collections staff to manage without automation, especially considering that funding for NHCs has been decreasing (Thiers 2018). In recent years, machine learning applications utilizing convolutional neural networks have achieved stunning levels of performance in computer vision tasks including image detection and classification (Sudholt and Fink 2016). Neural networks have previously demonstrated promising results for handwriting recognition systems, which could easily be applied to automated label transcription. These forms of innovative technology, which have been applied to medical diagnoses, speech recognition, and driverless cars, are now permeating NHCs (Schuettpelz et al. 2017) and are likely to be enormously useful when trained on existing databases of handwriting samples (Krishnan et al. 2016), as well as those from transcribed labels generated through

Digitization 1.0. These models can be further trained using existing semantic field constraints to much more effectively parse specimen metadata into appropriate database fields. Beyond capturing essential minimal data records in an automated manner, neural networks have recently been implemented to accomplish far more sophisticated tasks than text transcription (Wilf et al. 2016, Schuettpelz et al. 2017, Funk 2018). Wilf et al. (2016), for example, used computer vision to classify fossil leaf images based on leaf shape and venation with high accuracy. This proved not only to be an efficient protocol for classifying images, but also discovered previously unidentified morphological landmarks potentially useful for species identification and for understanding important evolutionary and ecological innovations not previously documented. The community is now ready for deeper exploration of minimal metadata capture using semi- to total automation.

Further, the declining number of taxonomists in the global workforce severely impacts our ability to address key questions concerning biodiversity in the face of global change (Hopkins and Freckleton 2002). Combining taxonomists' expertise (past and present) with student and public training and increased automation will facilitate enhanced specimen curation, and greatly enable biodiversity discovery. Continued robust support for taxonomic research and training is essential. However, given the enormity of the task at hand, and the limited time for this effort, we believe that addressing many taxonomic problems of identification, particularly for well-known groups of organisms, could be greatly facilitated by automation, such as has been demonstrated through Kurator (Dou et al. 2012). Reasonably successful early efforts are underway to machine-learn and automatically identify large sub-collections of insects (e.g., butterflies) (Schermer and Hogeweg 2018). Although simple taxonomic identification may seem rudimentary, it is the foundation of all biological research, and in many groups remains problematic. For example, it is estimated that more than 50% of tropical plant specimens in NHCs are incorrectly identified (Goodwin et al. 2015). Together with the training of more expert taxonomists and organismal biologists, the widespread use of neural networks to identify specimens and target groups that need attention would enhance collection utility for research, teaching, and management and further motivate the discovery and description of new species.

Expansion of the digitization workforce–Expanding digitization to involve a global workforce is now possible and is a major advancement of Digitization 2.0 and is motivated by the increasingly global accessibility of NHCs. These new workforces can be developed to supplement existing NHC staff, especially to include enhanced digitization from the millions of images in databases that have limited associated metadata. One obvious group to engage in this effort are citizen scientists. NHCs associated with museums typically have departments devoted to public outreach, which can easily be tapped for aid, helping collections staff with the task of digitization while simultaneously providing the public with ownership and agency. Using citizen science in this manner has been fruitful in numerous contexts including the transcriptions of label data, georeferencing, and physical specimen annotations (Hill et al. 2012, Ballard et al. 2017, Ellwood et al. 2015, 2017). For example, *CrowdCurio–Thoreau's Field Notes*, an online crowdsourcing platform has successfully facilitated climate change studies from thousands of herbarium specimens utilizing thousands of non-expert crowdsourcers (Willis et al. 2017). Quality control is always a concern in large-scale citizen science projects (Willis et al. 2017, Zhou et al. 2018) and thus an easy-to-use graphical user interface clearly demonstrating to the public how and what to digitize will be necessary (e.g., Notes for Nature), as has been accomplished in several research-based projects (Chang and Alfaro 2016, Cooney et al. 2017, Willis et al. 2017). Increasingly, such citizen science efforts are being supplemented by machine-based learning as well (Unger et al. 2016, Wilf et al. 2016, Schuettpelz et al. 2017). For instance, crowdsourced data can potentially provide reliable and rapid data for training and testing machine learning models, creating a positive feedback loop propelling digitization forward.

Layers of trait annotations–Traits of organisms are fundamental for documenting biodiversity but also for understanding how organisms evolve and respond to changing environments. Building on investments in creating digital NHCs, there is now increasing demand for creating and associating new layers of trait data to these collections. For some taxa, these biological data are already captured in the digitized text of a specimen record (e.g., Darwin Core fields: “organismRemarks”). In mammals and birds, it is common to have measurements on the mass and length of both the whole specimen and parts of the specimen (e.g., testes length, wing length). The aggregation of traits from both the initial collecting event and new annotations will stimulate a wealth of questions and generate a better understanding of global biodiversity through the development of standardized trait vocabularies (Kissling et al. 2018). For example, recently developed data-processing tools for the data aggregator VertNet standardized more than 1.5 million measurements for vertebrates using digital data from collections (Guralnick et al. 2016). Users can now search those specimen records by mass and length, as well as download harmonized trait data associated with individual specimens. The latter allows for new explorations of trait variation within and across species, including spatial and temporal patterns in traits associated with specimens that have collecting dates and georeferenced localities (Riemer et a. 2018). By expanding this framework to annotate traits to specimens and utilizing online platforms for even 3D representations of specimens, NHCs can facilitate the capture of not only simple traits, ranging from specimen length to the presence of a flower, but also more complex traits requiring more sophisticated representation (e.g., virtual automated dissection of the vertebrate nervous system).

IV. Concluding thoughts

Digitization facilitates the democratizing of collections-based research and is essential to establishing and evaluating biological baselines to assess the impacts of climate change, land use changes, species invasions, and the current mass extinction. It allows for the mining of specimen data in much the same way that we explore organismal genomes. The key to further developing Digitization 1.0 and establishing Digitization 2.0 lies in building upon what the research, funding, and policy communities have learned in the several decades since the initiation of this endeavor. Data-rich NHC specimens are useful and provide unique perspectives on the diversity and distribution of a given taxon. However, if a specimen is not searchable, it will likely not be found or studied despite its potential use. We are already witnessing the fruits of the synergy between Digitization 1.0 and 2.0. This will no doubt expand dramatically in the coming decades to involve new domains, new questions, and new audiences that are not yet realized (or even imagined). Only with creativity and improved techniques, including automated and semi-automated methods, a better distributed digitization workload making use of new technologies and workforces, and conscientious attention to the attribution chain, will researchers be best able to track ongoing biodiversity change from all existing data. Moreover, even as new technologies and digitization techniques emerge, we will need to always return to physical specimens, in ways that are unimaginable now, to generate novel data to better understand our changing planet. Although we stress the importance of improved methods and practices for digitization, the active collection and continued curation of physical specimens by expert biologists remains the central pillar supporting advancements in evolutionary biology and conservation represented so importantly by NHCs.

Box 1: ***Estimating the size and scale of a global digitization effort***– Digitization 1.0 has resulted in the mobilization of millions of specimen records and has created the momentum for a massive, global digitization effort. To better establish target goals and evaluate the success of this effort (e.g., estimating the proportion of specimen records that have been digitized and mobilized online), obtaining accurate estimates of the number of specimens housed in NHCs is necessary. Extrapolations from digitized content indicate that roughly 2.5–3 billion specimens are housed in NHCs worldwide (O'Connell et al. 2004, Krishnan et al. 2016). However, more robust assessments of global specimen numbers, including geographic and taxonomic distribution, are required to facilitate thoughtful assessments of collection bias to better target digitization priorities (Meyer et al. 2016). Making robust size estimates are particularly relevant as vended solutions are utilized to achieve digitization milestones, including the mobilization of entire collections like those at the Muséum National D'Histoire Naturelle (France), Naturalis (Netherlands), and the Smithsonian Institution (US) (Rogers 2016, Le Bras et al. 2017). Along these lines, a test case example to illustrate such an effort on a smaller scale comes from the Harvard University Herbaria (HUH), which has been thought to contain 5.5 million specimens. Targeted subsampling of the HUH vascular plant collection facilitated accurate estimates (with confidence intervals) of total specimen collection numbers and their geographic distribution (figure 3A). Once the total number of specimens in NHCs have been accurately quantified, it is necessary to establish the percentage of specimen collection records that have been digitally mobilized.

Because imaging and serving metadata-rich collection information online requires a large financial investment, as well as human labor, its impacts on research should be documented and acknowledged. The most powerful outcomes of digitization would be better characterized by relating these various forms of data usage to one another to explore how digitization increases specimen usage. Along these lines, data relevant to describing the scientific impact of physical specimens (pre-digitization), such as loans and museum visits, remain largely confined to physical collection logbooks, thus limiting assessment of the impact of Digitization 1.0 (figure 3B). Such efforts would allow us to begin to understand the ways that digitization stimulates increased visitation and use of the actual physical versus digital collection (figure 3C). As a community, we must be better prepared to track and assess these questions.

Acknowledgements

We thank the members and organizers of the National Science Foundation Postdoctoral Research Fellows in Biology (NSF-PRFB) 2017 Symposium for discussions contributing to this manuscript. NSF and the Biodiversity Collections Research Coordinating Network supported the meeting (NSF1441785, 1746177). Additional funding from the U.S. National Science Foundation includes awards NSF-DBI: EF1208835 and NSF-DEB 1754584 to CCD. For collections-based data from the MCZ and HUH and for discussion, we thank Mark Omura, Linda Ford, Brendan Haley, Paul Morris, Rachel Hawkins, José Rosado, Stephanie Pierce, and Andrew Williston. We thank Pam Soltis, James Hanken, and Maureen Kearney for providing thoughts and comments on earlier versions of this manuscript.

References Cited

Ballard HL, Robinson LD, Young AN, Pauly GB, Higgins LM, Johnson RF, Tweddle JC. 2017. Contributions to conservation outcomes by natural history museum-led citizen science: Examining evidence and next steps. *Biological Conservation* 208: 87–97.

Cardoso D, Särkinen T, Alexander S, Amorim AM, Bittrich V, Celis M, Daly DC, Fiaschi P, Funk VA, Giacomin LL, Goldenberg R. 2017. Amazon plant diversity revealed by a taxonomically verified species list. *PNAS* 114: 10695–10700.

Chang J, Alfaro ME. 2016. Crowdsourced geometric morphometrics enable rapid large-scale collection and analysis of phenotypic data. *Methods in Ecology and Evolution* 7: 472–482.

Comoglio F, Fracchia L, Rinaldi M. 2013. Bayesian inference from count data using discrete uniform priors. *PLoS One* 8: e74388.

Cooney CR, Bright JA, Capp EJ, Chira AM, Hughes EC, Moody CJ, Nouri LO, Varley ZK, Thomas GH. 2017. Mega-evolutionary dynamics of the adaptive radiation of birds. *Nature* 542: 344–347.

Dou L, Cao G, Morris PJ, Morris RA, Ludäscher B, Macklin JA, Hanken J. 2012. Kurator: A Kepler package for data curation workflows. *Procedia Computer Science* 9: 1614–1619.

Drew JA, Moreau CS, Stiassny MLJ. 2017. Digitization of museum collections holds the potential to enhance researcher diversity. *Nature Ecology & Evolution* 1: 1789.

Edwards JL. 2004. Research and societal benefits of the global biodiversity information facility. *Bioscience* 54: 485–486.

Ellwood ER, Dunckel BA, Flemons P, Guralnick R, Nelson G, Newman G, Newman S, Paul D, Riccardi G, Rios N, Seltmann KC. 2015. Accelerating the digitization of biodiversity research specimens through online public participation. *Bioscience* 4: 383–396.

Ellwood ER, Crimmins TM, Miller-Rushing AJ. 2017. Citizen science and conservation: Recommendations for a rapidly moving field. *Biological Conservation* 208: 1–4.

Funk VA. 2018. Collections-based science in the 21st century. *Journal of Systematics and Evolution* 56: 175–193.

Goodwin ZA, Harris DJ, Filer D, Wood JR, Scotland RW. 2015. Widespread mistaken identity in tropical plant collections. *Current Biology* 25: 1066–1067.

Guralnick RP, Zermoglio PF, Wieczorek J, LaFrance R, Bloom D, Russell L. 2016. The importance of digitized biocollections as a source of trait data and a new VertNet resource. *Database* Volume 2016.

Heberling JM, Isaac BL. 2018. iNaturalist as a tool to expand the research value of museum specimens. *Applications in Plant Sciences* 6: e1193.

Hedrick BP, Manning PL, Lynch ER, Cordero SA, Dodson P. 2015. The geometry of taking flight: Limb morphometrics in Cretaceous theropods. *Journal of Morphology* 276: 152–166.

Hill A, Guralnick R, Smith A, Sallans A, Gillespie R, Denslow M, Gross J, Murrell Z, Conyers T, Oboyski P, Ball J. 2012. The notes from nature tool for unlocking biodiversity records from museum records through citizen science. *ZooKeys* 209: 219.

Hopkins GW, Freckleton RP. 2002. Declines in the numbers of amateur and professional taxonomists: Implications for conservation. *Animal Conservation* 5: 245–249.

Johnson KG, Brooks SJ, Fenberg PB, Glover AG, James KE, Lister AM, Michel E, Spencer M, Todd JA, Valsami-Jones E, Young JR. 2011. Climate change and biosphere response: Unlocking the collections vault. *Bioscience* 61: 148–153.

Kissling WD, Walls R, Bowser A, Jones MO, Kattge J, Agosti D, Amengual J, Basset A, Van Bodegom PM, Cornelissen JH, Denny EG. 2018. Towards global data products of essential biodiversity variables on species traits. *Nature Ecology & Evolution* 2: 1531–1540.

Krishnan P, Dutta K, Jawahar CV. 2016. Deep feature embedding for accurate recognition and retrieval of handwritten text. *Frontiers in Handwriting Recognition (ICFHR)* 15th International Conference: 289–294.

Le Bras G, Pignal M, Jeanson ML, Muller S, Aupic C, Carré B, Flament G, Gaudeul M, Gonçalves C, Invernón VR, Jabbour F. 2017. The French Muséum National D'histoire Naturelle vascular plant herbarium collection dataset. *Scientific Data* 4: 170016.

Lewis SL, Maslin MA. 2015. Defining the Anthropocene. *Nature* 519: 171.

Lister AM. 2011. Natural history collections as sources of long-term datasets. *Trends in Ecology & Evolution* 26: 153–154.

MacDougall AS, Loob JA, Claydenc SR, Goltzd JG, Hindse HR. 1998. Defining conservation priorities for plant taxa in southeastern New Brunswick, Canada using herbarium records. *Biological Conservation* 86: 325–338.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27859v1 | CC BY 4.0 Open Access | rec: 15 Jul 2019, publ: 15 Jul 2019

Meineke EK, Davis CC, Davies TJ. 2018. The unrealized potential of herbaria for global change biology. *Ecological Monographs* 88: 505–525.

Meineke EK, Davies TJ, Daru BH, Davis CC. 2019. Biological collections for understanding biodiversity in the Anthropocene. *Philosophical Transactions of the Royal Society B: Biological Sciences* 374: 20170386.

Meyer C, Weigelt P, Kreft H. 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters* 19: 992–1006.

Moritz C, Patton JL, Conroy CJ, Parra JL, White GC, Beissinger SR. 2008. Impact of a century of climate change on small-mammal communities in Yosemite National Park, USA. *Science* 322: 261–264.

Nelson G, Ellis S. 2018. The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B*. 374: 20170391.

Nualart N, Ibáñez N, Soriano I, López-Pujol J. 2017. Assessing the relevance of herbarium collections as tools for conservation biology. *The Botanical Review* 83: 303–325.

O'Connell AFJ, Gilbert AT, Hatfield JS. 2004. Contribution of natural history collection data to biodiversity assessment in national parks. *Conservation Biology* 18: 1254–1261.

Park DS, Potter D. 2015. Why close relatives make bad neighbours: phylogenetic conservatism in niche preferences and dispersal disproves Darwin's naturalization hypothesis in the thistle tribe. *Molecular Ecology* 24: 3181–3193.

Primack D, Imbres C, Primack RB, Miller-Rushing AJ, del Tredici P. 2004. Herbarium specimens demonstrate earlier flowering times in response to warming in Boston. *American Journal of Botany* 91: 1260–1264.

Pyke GH, Ehrlich PR. 2010. Biological collections and ecological/environmental research: A review, some observations and a look to the future. *Biological Reviews* 85: 247–266.

Riemer K, Guralnick RP, White EP. 2018. No general relationship between mass and temperature in endothermic species. *eLife 7*: e27166.

Rogers N. 2016. Museum drawers go digital. *Science* 352: 762–765.

Rouhan G, Dorr LJ, Gautier L, Clerc P, Muller S, Gaudeul M. 2017. The time has come for natural history collections to claim co-authorship of research articles. *Taxon* 66: 101–1016.

Schermer M, Hogeweg L. 2018. Supporting citizen scientists with automatic species identification using deep learning image recognition models. *Biodiversity Information Science and Standards* 2: e25268.

Schuettpelz E, Frandsen PB, Dikow RB, Brown A, Orli S, Peters M, Metallo A, Funk VA, Dorr LJ. 2017. Applications of deep convolutional neural networks to digitized natural history collections. *Biodiversity Data Journal* 5: e21139.

Suarez AV, Tsutsui ND. 2004. The value of museum collections for research collections. *Bioscience* 54: 66–74.

Sudholt S, Fink GA. 2016. PHOCNet: A deep convolutional neural network for word spotting in handwritten documents. *Frontiers in Handwriting Recognition (ICFHR)* 15th International Conference: 277–282.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27859v1 | CC BY 4.0 Open Access | rec: 15 Jul 2019, publ: 15 Jul 2019

Sweeney PW, Starly B, Morris PJ, Xu Y, Jones A, Radhakrishnan S, Grassa CJ, Davis CC. 2018. Large-scale digitization of herbarium specimens: Development and usage of an automated, high-throughput conveyor system. *Taxon* 67: 165–178.

Tegelberg R, Mononen T, Saarenmaa H. 2014. High-performance digitization of natural history collections: Automated imaging lines for herbarium and insect specimens. *Taxon* 63: 1307–1313.

Thiers B. 2018. Using data from index herbarium to assess threats to the world's herbaria. *Biodiversity Information Science and Standards* 2: e26440.

Unger J, Merhof D, Renner S. 2016. Computer vision applied to herbarium specimens of German trees: Testing the future utility of the millions of herbarium specimen images for automated identification. *BMC Evolutionary Biology* 16: 248.

Webster M. 2017. The extended specimen. In M. Webster [ed.], The extended specimen: Emerging frontiers in collections-based ornithological research, 1–9. CRC Press, Boca Raton, Florida, USA.

Wilf P., Zhang S, Chikkerur S, Little SA, Wing SL, Serre T. 2016. Computer vision cracks the leaf code. *PNAS* 113: 3305–3310.

Willis CG, Law E, Williams AC, Franzone BF, Bernardos R, Bruno L, Hopkins C, Schorn C, Weber E, Park DS, Davis CC. 2017. CrowdCurio: An online crowdsourcing platform to facilitate climate change studies using herbarium specimens. *New Phytologist* 215: 479–488.

Yost JM, Sweeney PW, Gilbert E, Nelson G, Guralnick R, Gallinat AS, Ellwood ER, Rossington N, Willis CG, Blum SD, Walls RL. 2018. Digitization protocol for scoring reproductive phenology from herbarium specimens of seed plants. *Applications in Plant Sciences* 6: e1022.

Zhou N, Siegel ZD, Zarecor S, Lee N, Campbell DA, Andorf CM, Nettleton D, Lawrence-Dill CJ, Ganapathysubramanian B, Kelly JW, Friedberg I. 2018. Crowdsourcing image analysis for plant phenomics to generate ground truth data for machine learning. *PLoS Computational Biology* 14: e1006337.

Figures.

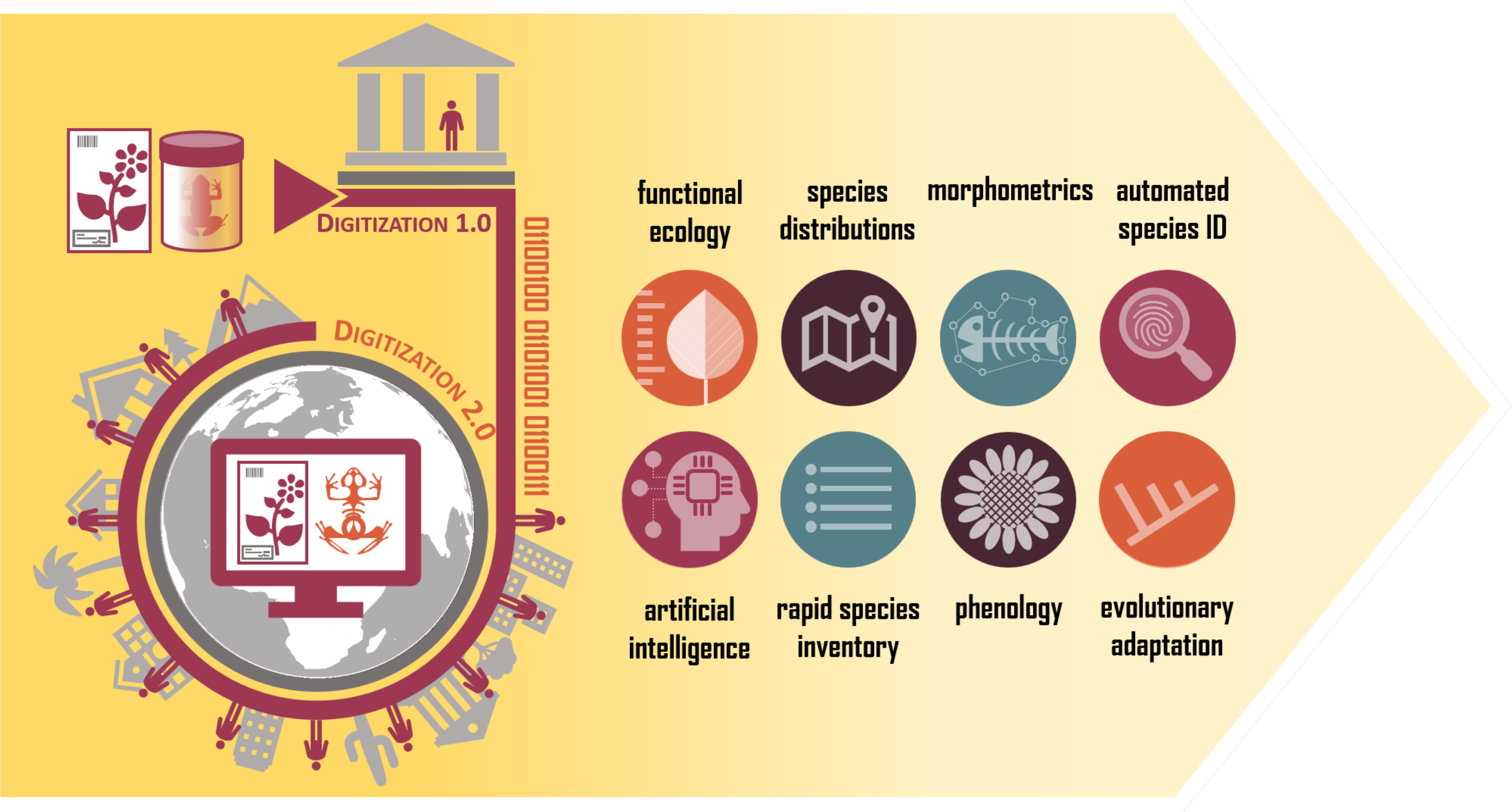
Figure 1. Digitization 1.0 and 2.0. Digitization 1.0 is the creation and online mobilization of digital content derived from physical specimens. This endeavor occurs locally within institutions, most commonly Natural History Museums. Digitization 2.0, in contrast, builds upon the digitized data, workflows, and infrastructure produced by Digitization 1.0 to facilitate enhanced digitization, curation, and data linkages to address increasingly complex questions at a massive global scale not previously imagined. These efforts are stimulating a new work force and connecting diverse scholarly domains, propelling a new generation of scientists forward, and removing financial, sociological, institutional, and academic obstacles restricting access to these materials. Some areas of inquiry that will be greatly stimulated by both Digitization 1.0 and 2.0 are highlighted.

Figure 2. An end-to-end pipeline example to highlight the value and complementarity of Digitization 1.0 and 2.0. The African pig-nosed frog (genus *Hemisus*) shown (A) was collected during recent field research in Angola. In addition to metadata from the collection event, a series of x-ray images (tomograms) were created using diffusible

iodine-based contrast-enhanced computed tomography (diceCT) directly from the voucher specimen. This product of Digitization 1.0 is shown in the black and white x-ray image (B). From these digital x-ray images, a 3D volume was created from the digital data generated during Digitization 1.0 from which students and scientists can digitally dissect and manipulate regions of interest representing the frog's nervous (C), circulatory (D), and muscular (E) systems (Digitization 2.0).

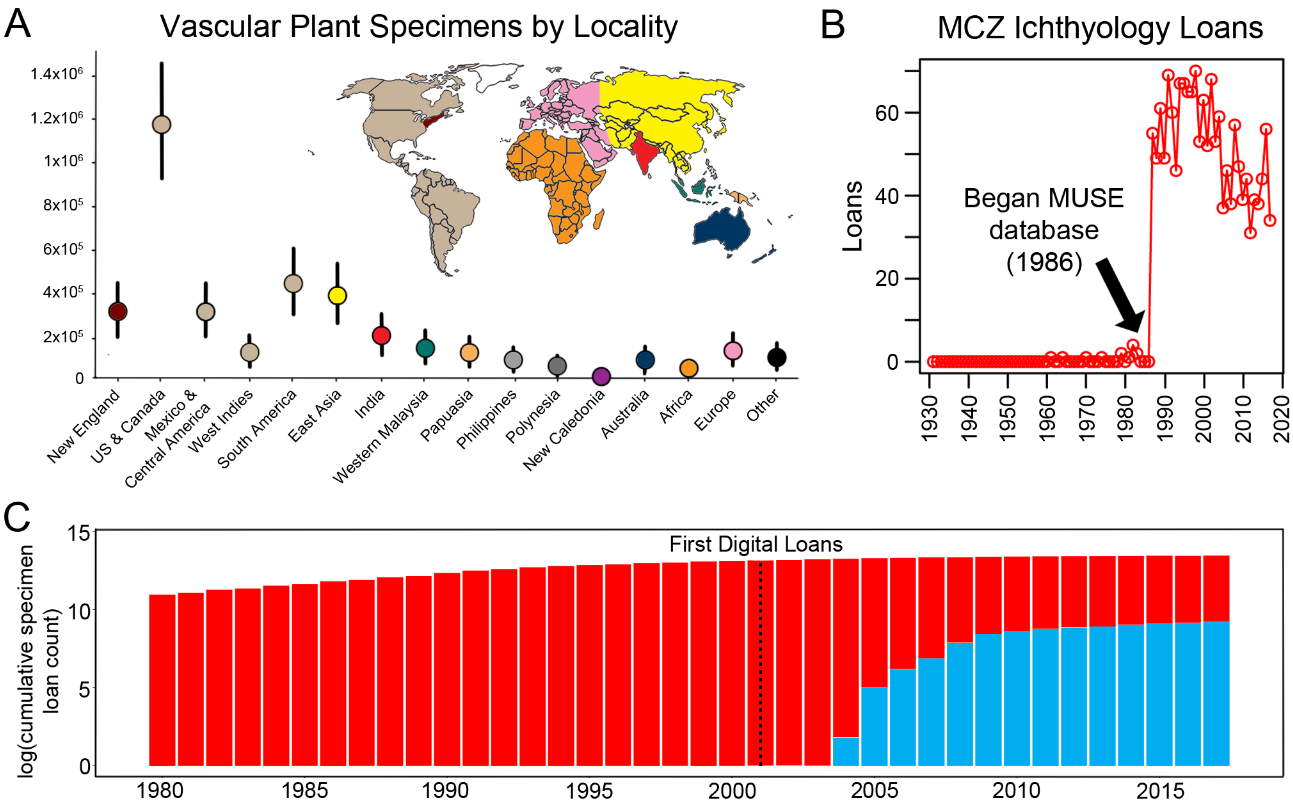
Figure 3. Estimating collection sizes and impact on research. (A) Size and geographical distribution of the vascular plant collection at the Harvard University Herbaria (HUH). To statistically estimate the size of this large collection, the total number of specimens in randomly subsampled cubbies were counted. These data were then used to model a probability distribution of the total number of specimens across the entire collection (Comoglio et al. 2013). Three hundred fifty cubbies were sampled and counted, establishing that the HUH has 3,701,695 vascular plants with a 95% confidence interval spanning 3,644,497 to 3,759,803. A similar approach was applied to further assess geographical distribution of the collection as well. (B) Loan use information for the Harvard Museum of Comparative Zoology ichthyology collection. Digitization greatly enhances the tracking of loan use history post 1980, until which point records are confined to physical logbooks. (C) Cumulative number of HUH specimen loans post 1980. While the total number of physical specimen loans (red) have remained relatively constant in recent years, the number of digital specimen images loaned has grown substantially.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27859v1 | CC BY 4.0 Open Access | rec: 15 Jul 2019, publ: 15 Jul 2019



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27859v1 | CC BY 4.0 Open Access | rec: 15 Jul 2019, publ: 15 Jul 2019

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27859v1 | CC BY 4.0 Open Access | rec: 15 Jul 2019, publ: 15 Jul 2019


A
Vascular Plant Specimens by Locality
1.4x10⁶
1.2x10⁶
1x10⁶
8x10⁵
6x10⁵
4x10⁵
2x10⁵
0
New England
US & Canada
Mexico & Central America
West Indies
South America
East Asia
India
Western Malaysia
Papuasia
Philippines
Polynesia
New Caledonia
Australia
Africa
Europe
Other
B
MCZ Ichthyology Loans
Loans
60
40
20
0
Began MUSE database (1986)
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
C
log(cumulative specimen loan count)
15
10
5
0
First Digital Loans
1980
1985
1990
1995
2000
2005
2010
2015