Mass Digitizing Biodiversity Collections of the United States

Nicholas D. Pyenson

January 23, 2020

DAY ONE PROJECT

dayoneproject.org

Imagine the year is 2046, the Smithsonian Institution's bicentennial. Then, like now, a group of children scamper loudly to the far corners of the museum halls. But unlike today, these children do far more than gaze at an object. Instead, they ***live*** *its context. They point devices that look like pencils at an object and unfurl everything about it: the children step into a map of the exact place where the object was collected; they see inside it, rotate it, and pull it apart; and then they spool the map forward in time, to see what their future worlds might be.*

Summary

Global climate change is already proving to be a major driver for species extinction and for shifting species' geographic distribution—trends that are almost certain to continue for the rest of the 21st century and beyond. The consequences on ecosystem structure, function and biodiversity are complex, but predictable.[1] Reliable predictions depend in part on the nation's biodiversity collections, especially those housed at museums, universities, and other learning centers. This key scientific infrastructure, consisting of hundreds of millions of objects, records the biology of many past eras on this planet. But without detailed, digitized information on these objects, we cannot come close to unlocking the insights contained within this massive body of knowledge. Mass digitization of U.S. biodiversity specimens will increase their accessibility, visibility, and value for researchers, students, and citizens for generations to come.

Mass digitization of U.S. biodiversity collections would position the nation to achieve massive advances in the life sciences—a leap forward on par with the way that DNA technology transformed genomics at the start of the 21st century. This heritage consists of hundreds of millions of dry, wet, and otherwise preserved specimens in U.S. museums and other collections, including plant germplasm, microbial cultures, non-human biomedical samples (e.g., parasites), fossils, and other plant and animal samples. The next administration can catalyze this advance by hosting a White House Summit on Biodiversity Digitization that paves the way for a sustained, coordinated effort to mass digitize the physical specimens in U.S. biodiversity collections (and their associated metadata).

The Summit would include representatives of citizen-science initiatives, academic institutions, private companies with the capabilities and technology (*e.g.*, 3D object capture, machine-learning networks) for mass digitization, and federal science agencies that have previously convened workshops on digitization standards with the White House Office of Science and Technology Policy's (OSTP) Interagency Working Group on

[1] IPBES, *Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*, Sandra Díaz, *et al.* [Eds.] (April–May, 2019).

Scientific Collections (IWGSC). Using the Trump Administration's "Summit on America's Bioeconomy"[2] as a template, the Summit on Biodiversity Digitization would solicit input from stakeholders while also building on Obama Administration-era actions to make federal science data publicly available.[3] The Summit would integrate existing digitization efforts that are distributed and siloed across the private and public sectors. The Summit would also generate a set of actions through the IWGSC, both in the near term and over longer time scales (i.e., decades to generations) that would illuminate this mostly invisible, but vital scientific infrastructure. The effort to mass digitize the Nation's biodiversity heritage will safeguard an irreplaceable investment about our biological past, preserve our rapidly changing present world, and answer questions we have not yet asked about the origin and fate of biodiversity, human health, and our food security and national security.

1. Challenge

Biodiversity collections (*i.e.*, biological and Earth-science collections) provide the fundamental physical basis for understanding any scientific question about the origin and diversity of life on our planet. For example, museum eggshell collections from the 19th and early 20th centuries provided pivotal evidence of the effect of DDT (a chemical found in pesticides) on biological health. By examining baseline eggshell thickness prior to the widespread application of DDT-containing pesticides in the mid-20th century, scientists were able to show that DDT detrimentally thinned bird eggs, among other widespread impacts.[4]

Biodiversity collections are an irreplaceable type of scientific infrastructure because the specimens within them cannot be collected again. The power of any biodiversity specimen derives from its context: the where, when, and how that specimen was collected. This information—also known as metadata—enables specimens to reflect the actual presence of an organism on the planet in the past. Maintaining the association between a physical object and its metadata is essential to good stewardship of these collections. Thus, institutions that house and protect these specimens (and their metadata) quite literally save past worlds.

The United States possesses a large number of biodiversity collections. The U.S. government (dominated by the Smithsonian Institution) is the largest single holder of

[2] Office of Science and Technology Policy, *Summary of the 2019 White House Summit on America's Bioeconomy*, The White House (October 2019).
[3] U.S. Government Accountability Office, *Federal Research: Additional Actions Needed to Improve Public Access to Research Results*, GAO-20-81 (November 2019).
[4] Andrew V. Suarez and Neil D. Tsutsui, "The value of museum collections for research and society", *BioScience* 54, no. 1 (2004): 66–74.

these collections. Other collections are held by governments at the sub-federal level, academic institutions, and private entities. Physical specimens include everything from pinned insects to bird feathers to pressed plants and much more. Specimens range in size, complexity, and condition (*i.e.*, fragile to stable), requiring careful and expert custody. The total number of specimens across all U.S. collections likely reaches close to half a billion objects—a significant percentage of the biodiversity specimens in collections worldwide.[5]

However, our ability to realize the full value of U.S. biodiversity collections is impeded by two challenges.[6] First, collections are at risk of neglect, damage, or total loss. Inadequate budgets for many biodiversity institutions results in inadequate care for the collections they house. If institutions lack the resources to protect biological specimens from exposure to light, air, vibration, or changes in humidity, specimens will slowly decay. Moreover, the effects of climate change—including rising sea levels and increased frequency of flooding—jeopardize the long-term security of certain U.S. collections, including the Smithsonian Institution's complex in Washington, DC; the American Museum of Natural History in New York; Harvard University's Museum of Comparative Zoology; the California Academy of Sciences in San Francisco; and other institutions located along coastlines or at sea level. Even the most remote and secure long-term storage facilities (*e.g.*, the Svalbard Global Seed Vault[7]) are not immune to climate threats. The risk of loss applies not only to physical specimens themselves, but also associated metadata and other artifacts (such as field notes or maps).

Second, many specimens in U.S. biodiversity collections remain unstudied, misidentified or unidentified, or unlabeled. Also, the metadata for many specimens that have been examined exists only in analog format (*e.g.*, hand-written notes on labels), inaccessible to all but a few researchers. Illuminating these "dark data" through mass digitization[8] could help researchers combat emerging pandemics; provide baselines for past, ongoing, and future environmental catastrophes; and answer questions we do not yet know to ask.

The loss or damage of biodiversity collections is irreparable. Individual specimens cannot be recollected, and the increasing destruction of the natural world means that some

[5] Ian Owens and Kirk Johnson, "One World Collection: The state of the world's natural history collections", *Biodiversity Information Science and Standards* 3 (2019).
[6] Christopher Kemp, "Museums: The endangered dead", *Nature* 518 (2015): 292–294.
[7] Kayla Epstein, "The Doomsday Vault's home is already altered by climate change. A report says it could get worse.", *Washington Post*, March 27, 2019, https://www.washingtonpost.com/climate-environment/2019/03/27/doomsday-vaults-home-is-already-altered-by-climate-change-report-says-it-could-get-worse/.
[8] C.R. Marshall, *et al.*, "Quantifying the dark data in museum fossil collections as palaeontology undergoes a second digital revolution", *Biology Letters* 14, no. 9 (2018).

species and geographic habitats can no longer be sampled at all because they are extinct or destroyed, respectively. Securing U.S. biodiversity collections for the long term is therefore essential to preserving our ability to learn from these collections for generations to come.

2. Opportunity

Digitization provides an affordable, scalable, and durable solution for capturing information that might be completely lost or substantially degraded. Digitization is thus an effective and compelling set of solutions to the challenges outlined above. Digitization also raises the visibility and value of their associated physical specimens, even those that are relatively secure.

In the United States, even dark data are relatively well organized. Management systems for U.S. biodiversity collections are traceable and validated, meaning that objects can be located in a repository even if they are not fully visible to researchers and the public. Digitization makes specimen information accessible to anyone in the world. Digital data may include:

- Information on the specimen itself (*e.g.*, qualitative description, such as its texture; quantitative information, such as dimensions and weight).
- Specimen metadata (*e.g.*, information on when, where, and how the specimen was collected).
- Associated data (*e.g.*, maps and field notes relevant to the specimen).
- Digital facsimiles (created through techniques such as photographic imaging and 3D modeling) that replicate physical attributes.

Digitization cannot prevent the destruction of the natural world. But digitization can ensure that information on past natural worlds is maintained even in the face of accelerating species loss, destruction of natural habitats, and urbanization of whole ecosystems. Just as there is an urgent need for action to protect the parts of our natural world that remain, there is an urgent need for action to protect the natural specimens under our custody.

Example of the Smithsonian's digitization efforts

As the holder of over 145 million biodiversity specimens, the Smithsonian Institution (SI) possesses the largest biodiversity collection in the world—outstripping the next-largest collection (London's Natural History Museum) by over twofold. The SI's collections include biological specimens from every ocean body, continent, and biome; earth-science specimens (*e.g.*, fossils) from every geological time period; and even geological samples from the Moon, meteorites, and other sources older than our planet. The SI has been the premier federal science institution for managing U.S. biodiversity collections since 1846. While the sheer size of the SI's collections means that not every single specimen is ready nor available for digitization, the SI's data-management policies have been widely adopted by other U.S. science to manage large specimen collections.

SI's digitization portfolio illustrates the challenges of designing digitization workflows that capture, render, and organize this information: the time and effort to 3D digitize large objects (*e.g.*, dinosaur or whale skeletons) is greater than small objects (*e.g.*, insects), but the potential public interest for the former may be greater than the latter[1]. Only about 2.6 million of the 145 million biodiversity objects at the Smithsonian have been digitized (*i.e.*, a specimen with a digital record and an image)[1], but technological and workflow improvements in FY 2018 have allowed nearly 0.5 million biodiversity objects to be digitized in 12 months[1]. This increase, which focused on a specific collection (i.e., fossil mollusks), demonstrates the conditions needed to successfully mass digitize biodiversity collections at a large scale.

Overall, across the entire Smithsonian, SI's Mass Digitization Program has digitized over 3 million objects (including collections in arts, culture, and history) since 2014. SI has voluntarily adopted the OSTP public access policy[1] for scientific research publications and data (including those related to SI's biodiversity collections); long-term compliance will be enhanced by public-private partnerships that leverage unique technological innovations for digitizing millions of complex, irregular objects that vary in size, stability, condition, and accessibility.

3. Proposed action

The next administration should catalyze a mass effort to digitize U.S. biodiversity collections by hosting, through OSTP, a White House Summit on Biodiversity Digitization. The Summit would include representatives of the U.S. government, academia, the private sector, and citizen-scientist groups, and would aim to (1) integrate and align existing digitization efforts (2) establish and implement data standards for digitization (*e.g.*, consistent file formats), and (3) identify key infrastructure needs (*e.g.*, data repositories, and common platforms for data access). The Summit would also encourage institutions that house biodiversity collections to commit to digitizing 25% of their collections by 2025, and would encourage private partners to assist in the workflows (*e.g.*, technological support of capturing, rendering, and visualizing objects) to support mass digitization. Most importantly, the Summit would provide visibility and compelling leadership on a major effort that currently lacks centralized coordination.

Summit priorities should be informed by digitization efforts from the past 20 years. Digitization programs for U.S. biodiversity collections have been implemented with mixed success[9]. The mass digitization techniques pioneered by industry in the late 2000s for 2D objects (*e.g.*, books by Google[10]) have proven useful for certain types of biodiversity collections, but not others. For instance, these methods are highly successful when applied to pressed plants of near-uniform size but are not well-suited for more other collections of skeletal elements. Similarly, 3D surface-scanning techniques can successfully capture objects of roughly shoebox size but encounter limits on larger objects. Surface-scanning techniques also cannot capture internal structure. CT scanning technology has enabled many institutions to capture high-resolution internal data of biodiversity collections at large scale. The oVert project, funded by the National Science Foundation (NSF), is using CT technology to capture 3D digital representations of every branch of vertebrate life (*e.g.*, ~20,000 species)[11]. By reviewing the successes and shortcomings of previous and current digitization initiatives, the Summit will help identify what is needed accelerate and expand digitization of U.S. biodiversity collections.

The next administration should also direct OSTP to continue supporting IWGSC's work with other federal agency partners on long-term planning for digitizing U.S. biodiversity collections. In tandem with the Summit, the IWGSC should:

- Begin long-term planning that includes community partners inside and outside the IWGSC (*e.g.*, biodiversity collections at U.S. universities and other institutions, key private sector partners).
- Identify strategic investments (*e.g.*, equipment, personnel) needed to advance digitization efforts.
- Identify cross-linkages that coordinate and exploit urgent needs among the community of partners, as well as needs over longer timeframes (*i.e.*, decadal-centennial scales).
- Engage with the network and capacity building of the One World Collection project[12] to ensure that the United States, as the largest contributor to this project, can lead by modeling a cohesive strategy for making biodiversity collections visible and secure.

---

[9] Biodiversity Collections Network, "Extending U.S. Biodiversity Collections to Promote Research and Education," American Institute of Biological Sciences (2019), https://bcon.aibs.org/wp-content/uploads/2019/04/BCoN_March2019_FINAL.pdf
[10] James Somers, "Torching the Modern-Day Library of Alexandria", *The Atlantic*, April 20, 2017, https://www.theatlantic.com/technology/archive/2017/04/the-tragedy-of-google-books/523320/.
[11] Ryan Cross, "New 3D scanning campaign will reveal 20,000 animals in stunning detail", *Science*, August 24, 2017, available at https://www.sciencemag.org/news/2017/08/new-3d-scanning-campaign-will-reveal-20000-animals-stunning-detail.
[12] Vince Smith, *et al.*, "Shining a New Light on the World's Collections", iDigBio, August 14, 2018, https://www.idigbio.org/content/shining-new-light-world%E2%80%99s-collections.

- Update the IWGSC report on scientific collections, originally (and last) published in 2009.

These activities build on U.S. scientific leadership in biodiversity, and are intended to help catalyze the next revolution in biological sciences.

The next administration should also use the Summit to discuss other challenges and solutions related to the long-term security of U.S. biodiversity collections. As stated above, many biodiversity collections are threatened by inadequate stewardship and the effects of climate change. The Summit is an opportunity to discuss how U.S. science leadership in science can support a concerted effort to secure or rebuild infrastructure and facilities for U.S. biodiversity collections. Other priority discussion topics stem from additional concerns:

- What the shared concerns and unique challenges for legal custody of digital biodiversity collections at different institutions?
- What standards for metadata preservation should be used?
- What are the timescales, expectations, and needs for future-proofing the specimens and their digital records?
- Which other science or non-science U.S. federal agencies (*e.g.*, those with human health portfolios[13]) could be stakeholders for digitizing biodiversity collections?

4. Key partners and stakeholders

OSTP's IWGSC is well situated to oversee and organize the Summit. The IWGSC already engages with U.S. federal science agencies on scientific collections, and group membership includes traditional research agencies (*e.g.*, the SI and the Departments of Agriculture, Commerce, and the Interior), as well as agencies whose domains include national security (*e.g.*, the Department of Defense). The IWGSC also maintains strong partnerships with international consortia that guide the interoperability of large-scale biodiversity databases. These include the Global Biodiversity Information Facility (GBIF), which implements Darwin Core (a set of community standards intended to facilitate sharing of information about biological diversity).

Key private-sector partners include technology companies (*e.g.*, Google, Amazon) that have created digitization workflows and pipelines that may be useful for biodiversity collections. Companies involved in developing augmented- and virtual-reality platforms are also important partners, as the immense narrative potential of digital content pinned

[13] Diane DiEuliis, *et al.* "Opinion: Specimen collections should have a much bigger role in infectious disease research and response", *Proceedings of the National Academy of Sciences* 113, no. 1 (2016): 4–7.

to historically or scientifically significant biodiversity specimens may have a home with such platforms. In engaging with private-sector partners, the next administration should emphasize ways in which digitized biodiversity collections could be transformative for the private sector and spur additional growth. For example, the emergence of so-called digital twin or mirror world visualization platforms, where digital data are overlain on real world objects[14], would make explicit and necessary use of digital biodiversity collections datasets. The digital story-telling potential of hundreds of millions of biodiversity objects is untapped.

The many and diverse types of institutions that house biodiversity collections in the United States are also important stakeholders. These institutions include public museums (federal, state, and county repositories), private museums (*e.g.*, the American Museum of Natural History in New York City, the Field Museum in Chicago), museums at universities, and smaller research centers (including some maintained by federal agencies). One robust biodiversity-digitization effort that already exists at the university level is the Advancing Digitization of Biodiversity Collections (ADBC) national resource, which is funded by NSF, and operates through the Integrated Digitized Biocollections (iDigBio) program at the University of Florida. A final stakeholder group comprises the multiple U.S. citizen-science programs that employ digitization methods.

All of these institutions are strong partners and advocates, and should be involved in the Summit, which stands on solid consensus from the professional and scientific communities. The United States is a visible and dominant participant of these communities, and a large portion of the world's biodiversity collections are under the custody and stewardship of the U.S. government. Also, as U.S. science agencies are increasingly closing or moving their scientific collections, digitization provides a key process for managing the workflows associated with this trend that places greater strain on flatlined investments in infrastructure and staffing.

4. Templates and other tools

The Trump Administration's "White House Summit on America's Bioeconomy" would serve as a template for the proposed Summit on Biodiversity Digitization. The Summit on Biodiversity Digitization could emulate the Summit on America's Bioeconomy by using a Request for Information to solicit input from the biodiversity and digitization communities, including researchers in both private and public sectors. The proposed

[14] Kevin Kelly, "AR Will Spark the Next Big Tech Platform—Call It Mirrorworld", *WIRED*, February 2, 2019, https://www.wired.com/story/mirrorworld-ar-next-big-tech-platform/.

Summit also builds on the Obama Administration's Presidential Memorandum[15] and Executive Order[16] focusing on open access and making federal science data publicly available. The linkages to private sector interests would be strong, especially for the rapidly emerging market of 3D visualization. Mass digitization of biodiversity collections has the potential to generate new markets and yield additional, unforeseen benefits. Most importantly, the Summit is an easy but critical step to integrate separate efforts that are running in parallel at different rates and scales. The need for cohesion and long-range planning is critical because now is the time for coordination. Notably, the Summit requires no Congressional funding nor action.

5. Conclusion

U.S. biodiversity collections are made up of nearly half a billion specimens that are irreplaceable and mostly unknown. A large-scale digitization program will illuminate these data, making them visible, accessible, and searchable. Digitization will also protect these data in perpetuity so that future researchers and citizens can answer questions not yet asked. To catalyze this mass digitization effort, the next administration should host a White House Summit on Biodiversity Digitization that (1) capitalizes on U.S. science leadership and (2) results in a coordinated digitization plan that will help secure a precious scientific resource—the biodiversity holdings of our nation—for generations.

[15] Office of Science and Technology Policy, *Increasing Access to the Results of Federally Funded Scientific Research*, The White House (February 2013), available at https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[16] "Executive Order 13642 of May 9, 2013 on Making Open and Machine Readable the New Default for Government Information", *Code of Federal Regulations*, title 3 (2013): 244–246, https://www.govinfo.gov/content/pkg/CFR-2014-title3-vol1/pdf/CFR-2014-title3-vol1-eo13642.pdf.

About the author

Nicholas Pyenson is the curator of fossil marine mammals at the Smithsonian Institution's National Museum of Natural History in Washington, DC. As a paleontologist, his expeditions have taken him to every continent, and the results of his team's discoveries have been published extensively, including cover articles in the journals *Science* and *Nature*. Along with his collaborators, he has named nearly a dozen new species of fossil vertebrates, discovered the richest fossil whale graveyard on the planet, and described an entirely new sensory organ in living whales. His research has received the highest awards from the Smithsonian, and he has also received a Presidential Early Career Award for Scientists and Engineers from the White House. His book describing his work, *Spying on Whales*, was featured on national television and radio, and is included in many best science book compilations. He completed his postdoctoral fellowship at the University of British Columbia, received his doctoral degree from the University of California, Berkeley, and has a bachelor's degree from Emory University. Pyenson is also a member of the Young Scientists community at the World Economic Forum.

About the Day One Project

The Day One Project is dedicated to democratizing the policymaking process by working with new and expert voices across the science and technology community, helping to develop actionable policies that can improve the lives of all Americans, and readying them for Day One of a future presidential term. For more about the Day One Project, visit dayoneproject.org.