Digital Extended Specimens: Enabling an Extensible Network of Biodiversity Data Records as Integrated Digital Objects on the Internet

ALEX R. HARDISTY, ELIZABETH R. ELLWOOD, GIL NELSON, BREDA ZIMKUS, JUTTA BUSCHBOM, WOUTER ADDINK, RICHARD K. RABELER, JOHN BATES, ANDREW BENTLEY, JOSÉ A. B. FORTES, SARA HANSEN, JAMES A. MACKLIN, AUSTIN R. MAST, JOSEPH T. MILLER, ANNA K. MONFILS, DEBORAH L. PAUL, ELYCIA WALLIS, AND MICHAEL WEBSTER

The early twenty-first century has witnessed massive expansions in availability and accessibility of digital data in virtually all domains of the biodiversity sciences. Led by an array of asynchronous digitization activities spanning ecological, environmental, climatological, and biological collections data, these initiatives have resulted in a plethora of mostly disconnected and siloed data, leaving to researchers the tedious and time-consuming manual task of finding and connecting them in usable ways, integrating them into coherent data sets, and making them interoperable. The focus to date has been on elevating analog and physical records to digital replicas in local databases prior to elevating them to ever-growing aggregations of essentially disconnected discipline-specific information. In the present article, we propose a new interconnected network of digital objects on the Internet—the Digital Extended Specimen (DES) network—that transcends existing aggregator technology, augments the DES with third-party data through machine algorithms, and provides a platform for more efficient research and robust interdisciplinary discovery.

Keywords: digital specimen, extended specimen, Digital Extended Specimen, natural history, biodiversity collections

Fundamental and applied biodiversity research depends on making full use of the enormous potential of physical specimens preserved within the world's estimated 6500 natural science collections (Meineke and Davies 2019, Meineke et al. 2019), along with an impressive quantity of related data. Physical specimens are persistent reservoirs of data about the planet's spectacular diversity of plants, animals, fungi, and other organisms, as well as its geology. Specimens provide enduring time capsules for discovery and verification in existing and previously unexplored avenues of research, even long after populations and species might disappear in nature. Therefore, natural science collections represent and inform biological, environmental, paleontological, geological, and anthropogenic patterns and processes at varying ranges of spatial, temporal, and functional distribution and resolution. For example, the study of the origins of the chytrid fungus *Batrachochytrium dendrobatidis* Longcore, Pessier and D.K. Nichols in amphibians at the beginning of the twenty-first century used whole-genome sequencing to solve the spatiotemporal origins of the most devastating panzootic disease to date (O'Hanlon et al. 2018). The analysis could not have been done without well-preserved specimens. Specimens have also been used in human health research (Suarez and Tsutsui 2004, Thompson et al. 2021), climate science (Johnson et al. 2011), conservation and natural resource management (Drew 2011), ecology (Pyke and Ehrlich 2010, Lister 2011), education (Monfils et al. 2017), law enforcement (NatSCA 2005, Rivers and Dahlem 2014), and policy (Abrahamse et al. 2021).

When linked with other sources of biodiversity and environmental information, specimen-based and specimen-derived data have the capacity to catalyze transformative science and interdisciplinary applications needed to address global change, biodiversity loss, zoonotic diseases, food and

BioScience 72: 978–987.
https://doi.org//10.1093/biosci/biac060 Advance Access publication 3 August 2022

water security, and the sustainable management of natural resources (NASEM 2020). Persistent links with other in-house data (e.g., trait measurements, DNA sequence data, geographic locality, collector's field notes), relevant third-party data (e.g., habitat characteristics, ecosystem functions, environmental factors, and pharmaceutical, agricultural, human health uses), and relevant traditional knowledge add to the research value of curated specimens.

In the present article, we propose the Digital Extended Specimen (DES) as the appropriate paradigm for digitally linking specimen data from the world's natural science collections to relevant ecological, environmental, and related data from numerous domains as a mechanism to support a quantum leap in enhancing the knowledge to be derived from such collections. Transforming scientific collections' primary specimen data into widely applicable *digital objects* on the Internet, with a single common representation (structure, format, etc.) and common interpretation of the subparts (fields) of that structure, enables integration of divergent data types and easier sharing among institutions, research projects, information systems, and software tools. This provides the basis for combining distributed data resources into robust data sets ready for large-scale processing by computers and sophisticated, multifactor analysis. Such digital resources increase the usability of collections' holdings and allow data providers and users, including researchers, conservation managers, and commercial enterprises, to leverage powerful, high-quality data sets at unprecedented size, scale, and versatility (Heberling et al. 2021).

We envision the DES as an essential building block in the evolution of present-day biodiversity informatics infrastructures. Individual DESes become data products to be accessed, updated, and processed as a new public domain level of trusted, open data as infrastructure to be used as a commons. People and—increasingly, in the future—computers will manipulate and apply this high-value information in research and as empirical evidence for applied predictive modeling, problem solving, management decisions, and decision-making. We show how the DES introduces expanded capabilities and capacities to the biodiversity research community and associated disciplines, as well as support for knowledge-based approaches for achieving the substantial changes needed to avert anthropogenic catastrophes in the medium and long term (Bak-Coleman et al. 2021).

Digitally linking specimen data

Openly available DESes combining specimen data with related data make it possible for the wider community of collections experts to enrich and curate biodiversity data more effectively and more comprehensively. Although digital specimens foster extended, increased, and nontraditional use by new audiences of data users and scientists, it is presently difficult to find and expose provenance and thematic connections, especially in a manner that is easy for computers to process. One example involves the International Nucleotide Sequence Database Collaboration (INSDC; e.g., the DNA Data Bank of Japan, the European Nucleotide Archive, GenBank), which specializes in genetic sequence data. INSDC includes many sequences derived from or represented by physical specimens in collections. However, in the majority of cases, INSDC databases do not list voucher specimen or specimens, nor do they include links to tissue or voucher samples from which the DNA were extracted and the sequence data derived (Buckner et al. 2021). Similarly, collection management systems (CMS) in collection-holding institutions such as museums and universities are still improving their ability to store details of or point to genomic data derived from the specimens they catalog (Krimmel et al. 2020). This highlights that different kinds of data demand custom approaches to finding and making connections between domain-specific data and the digital anchors representing the specimens from which such data were derived.

Extending and linking the scientific components of digital specimen records beyond the proprietary CMS of individual collection-holding institutions leads us into a new era of dynamic data management and maintenance. Digital representations on the Internet of physical specimens in collections can act as focused anchoring points for networks of derived and associated data. This extends the research value of single specimen records considerably, and promotes more FAIR (findable, accessible, interoperable, and reusable) data objects (Wilkinson et al. 2016).

Digitizing and mobilizing collections data

Over the past three decades, natural science collections have continued to transcribe the label data and to record images of their specimens, aggregate these into databases, and make specimen text and image data available online. Many institutions have mobilized such data directly through their own data portals and through national aggregators such as iDigBio (in the United States), e-ReColNat (France), the Finnish Biodiversity Information Facility (Finland), the Atlas of Living Australia, Reflora by the National Council for Scientific and Technological Development and the Centro de Referência em Informação Ambiental (both Brazil); through transnational or international data infrastructures such as BioCASe and the Global Biodiversity Information Facility (GBIF); and through thematic infrastructures such as JACQ and SEINet (both virtual herbaria), VertNet (vertebrates), and GeoCASe (geoscience collections), to list just a small number. To improve discoverability of such digital records on the Internet in an unambiguous way, persistent identifiers (PID) such as URLs, DOIs, ARKs, IGSN Generic Sample Numbers, CETAF Stable Identifiers, and others have been assigned by multiple initiatives (Kunze 2003, Güntsch et al. 2017, DOI Foundation 2019, Klump et al. 2021). This has had the effect of making large volumes of biodiversity and geosciences specimen data openly accessible and available on the Internet but these have largely remained disconnected from other data, including that derived from the same specimens.

Between 2003 and 2019, more than 4000 published studies have made use of data delivered through GBIF to address questions at taxonomic, temporal, and spatial scales that would otherwise have been impossible, indicating the far reach mobilized data can have in enabling and catalyzing new studies (Heberling et al. 2021). Heberling and colleagues (2021, p. 6) noted that "though promising, this work is far from a culmination." Their review highlighted "the need for continued development to facilitate a new era of data-intensive biodiversity science." They selected several topics as themes to develop further: digitization and publishing, a more complete and unbiased view of biodiversity, efficient routes for providing feedback to improve data quality, and (most important) new initiatives for linking and integrating data from multiple sources.

The US National Science Foundation's Advancing Digitization of Biodiversity Collections (ADBC) program, launched in 2010 and followed by the Biodiversity Collections Network (BCoN), launched in 2014, proposed a national agenda for creating and leveraging digital biodiversity collections data for new uses. Building on the extended specimen concept (Webster 2017), the BCoN vision for the future of collections includes transparent annotation of the underlying data and the linking of disparate specimen records; ecological and environmental data; and derivative items such as gene sequences and isotope readings, through voucher–tissue, predator–prey, and host–parasite type relationships—all enabled by continued growth of physical and digital collections. The European program for a Distributed System of Scientific Collections (DiSSCo), which began planning in 2018 under the European Commission Horizon 2020 project on innovation and consolidation for large scale digitization of natural heritage (ICEDIG; Hardisty et al. 2020), aims for digital unification of all European natural science assets under common policies for curation, access, and use. As such, DiSSCo enables the fragmented landscape of the extensive European natural science collections to be transformed into an integrated digital specimen knowledge base that provides interconnected hard evidence about the natural world.

The DES builds on the previously established ADBC, BCoN, the Atlas of Living Australia (ALA), and DiSSCo concepts by recognizing the data import limitations constraining integration with local CMSes while leveraging the DES network to release the community data of thousands of institutions. In the present article, we describe efforts to effect common global understanding of DESes, and shared approaches for future work. Although the US- and European-led initiatives are largely compatible with one another, we are aware of the potential for divergence over time, especially given funding sources with limited geographic scope.

Converging to a common understanding

Extending the scope of the specimen to include more than merely biological (or geological) materials but also audio recording, video, photographs, and a wide range of other environmental and ecological data types, directly derived or indirectly related, coalesces a much richer and more complete source of biodiversity information than a single disconnected specimen can ever achieve. Likewise, a DES can be placed in multiple classifying or taxonomic structures simultaneously, to be discoverable alongside additional categorical data based on, for example, media type, locality, or taxon. This potential has been recognized by the biodiversity community for some time, although the necessary framework, such as it was implemented through ADBC and ALA, was insufficient for full execution (e.g., NSF 2012, Hardisty and Roberts 2013, Hobern et al. 2013, Belbin et al. 2021).

Since Webster's (2017) publication, growing conversations have recognized the exciting possibilities of enhancing and digitally representing the billions of specimens currently held in the world's natural science collections. Two concepts in particular have emerged that have advanced the dialogue and provided the foundation for actualizing Webster's (2017) extended specimen.

In Europe, preparatory work has developed a design blueprint for DiSSCo's digitization infrastructure following the concept of digital specimens (Hardisty et al. 2019, Addink and Hardisty 2020, Hardisty 2020, DiSSCo 2021) that act as technical nodes for interconnecting specimens and collection-based data into a digital specimen network. In the United States, the BcoN, in conjunction with representatives of the collections and research communities, has presented its strategy for the Extended Specimen Network (Thiers et al. 2019, Lendemer et al. 2020). These efforts align with the shared vision of connecting all information related to a specimen. These ideas, and the sensible notion to unify efforts, have spurred a wide range of discussions, notably through the annual Biodiversity Information Standards (TDWG) conferences (Hardisty and Bentley 2021), as well as by numerous smaller groups of international stakeholders with significant interest in achieving a fully integrated and interoperable digital data infrastructure.

Following from TDWG 2020, more than 35 organizations worldwide and many individuals signed a letter of intent under the umbrella of the Alliance for Biodiversity Knowledge (ABK 2021), expressing interest to work collaboratively toward a global specification and interoperability for the digital specimen and extended specimen concepts. The present article is a product of that intent. It is based on outcomes of a worldwide consensus-forming discussion that took place February through August 2021 (GBIF 2021), resulting in a big step forward to converge the two ideas and the international communities behind them. The newly accepted term *Digital Extended Specimen* circumscribes this convergence into one technical concept as is illustrated in figure 1. It is the term we use in the present article.

Defining the Digital Extended Specimen

We define the DES as the collective representation on the Internet of all digital assets referring to a physical specimen (which can include physical evidence of related observations),

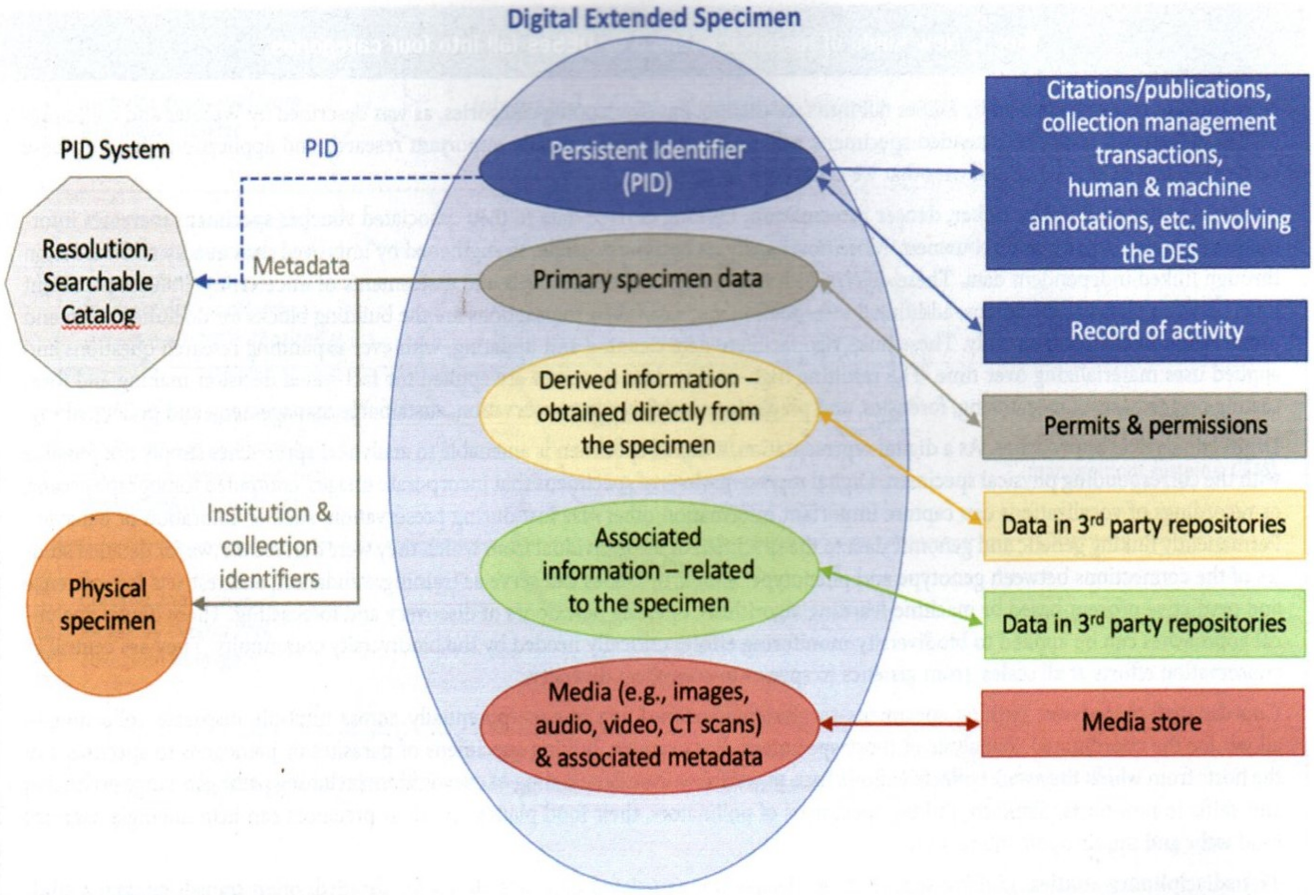

Figure 1. The Digital Extended Specimen (DES) as an enclosure for digital information about a physical specimen. Data about a specimen and pointers to data derived from a specimen, to data associated with a specimen, and to images or other media of the specimen and related objects are collated together as a DES digital object. A persistent identifier (PID) is assigned to uniquely identify the DES. A PID and metadata (the dashed or dotted lines) allow the DES to be indexed in searchable catalogs such that the specimen can be found, accessed, and used in line with relevant terms and conditions. This digital specimen on the Internet is tightly bound to the physical specimen it represents through institution, collection, and specimen identifiers included as part of the authoritative information about the specimen. A DES can be the focus for different kinds of transactions (e.g., annotations, citations, exchanges, loans, visits) and can have an associated long-lasting record of provenance (what was done, when, and by whom). Omitted from the figure for clarity, operations can be invoked against the DES to fetch data from it, to modify or update data, and to perform processing using the data. Also not shown, access control lists determine who has what permissions in relation to the DES content.

that meets the FAIR principles and that is distinguished and linked using globally unique persistent and resolvable identifiers to create an extensive online network of knowledge regarding life and related natural science objects.

A DES is a FAIR digital representation of a specimen. A DES can be linked to data or samples derived from the curated specimen itself (e.g., computed tomography scan imagery, DNA sequences, tissue samples), to DESes for other related specimens, or to data about the organism's life (e.g., parasite specimens collected from it, photos or recordings of the organism in life, data on the immediate surrounding ecological community). A DES can also be linked to the wide range of associated specimen-independent data sets and model-based contextualization (e.g., assignment via identification to a specific systematic classification framework, conservation status, bioclimatic region, remote sensing images, environmental–climatological data, traditional knowledge, genome annotations). The resulting connected network of DESes will enable new research on many fronts (box 1), and indeed, this has already begun.

A DES exists as a digital representation on the Internet—a resource, a data construct, a surrogate—acting for a curated specimen in a natural science collection and its associated data within the custodial institution, as well as related data outside the institution. Therefore, a DES represents the sum of the digital information about the physical specimen, data derived from the specimen, and data associated with the specimen regardless of source, as is illustrated in figure 1.

A DES encloses several principal classes of information related to the corresponding specimen:

Authoritative information. It includes data about the physical specimen, typically captured from its label, card catalog or ledger entry, field notes, or the collection to which it belongs.

Box 1. New kinds of research enabled by DESes fall into four categories.

New kinds of research enabled by DESes fall into four distinct but overlapping categories, as was described by Webster and colleagues (2021). A network of digital extended specimens will enable novel and critically important research and applications in all of these categories, as well as science and uses that we cannot yet imagine.

Analyses better enabled by richer, denser information. Linking derived data to their associated voucher specimens increases information richness, density, and robustness. More novel analyses become possible, strengthened by improved data quality and validation through linked independent data. These deliver with improving confidence levels and assessments of uncertainty. Indeed, persistent links between physical specimens, additional information, and associated transactions are the building blocks for documentation and preservation of chains of custody. These links also facilitate data cleaning and updating, with ever-expanding research questions and applied uses materializing over time. The resulting high-quality data resources are applied for fact-based decision-making and forecasting on the basis of monitoring, forensics, and prediction workflows in conservation, sustainable management, and policymaking.

Digital analytical approaches. As a digital representation, a digital specimen is amenable to analytical approaches simply not possible with the corresponding physical specimen. Digital representations of specimens that incorporate images, computed tomography scans, or recordings of vocalizations can capture important information otherwise lost during preservation, such as coloration or behavior. Permanently linking genetic and genomic data to the specimen of the individual from which they were derived allows for detailed studies of the connections between genotype and phenotype. Similarly, DESes can serve as training, validation, and test sets for inference and predictive process-based or machine-learning algorithms, opening new doors of discovery and forecasting. These digital analytical approaches can be applied to biodiversity monitoring efforts critically needed by the biodiversity community. They are central to conservation efforts at all scales, from genetics to species to ecosystem diversity.

Coordinated coanalysis. Linking specimens to closely associated specimens—potentially across multiple disparate collections—allows for the coordinated coanalysis of those specimens. For example, linking specimens of parasites or pathogens to specimens of the hosts from which they were collected allows for a powerful new understanding of coevolution, including pathogen range expansion and shifts to new hosts. Similarly, linking specimens of pollinators, their food plants, and their predators can help untangle complex food webs and multitrophic interactions.

Transdisciplinary studies. Linking specimens to diverse but associated data sets allows for detailed, often transdisciplinary, studies of topics ranging from local adaptation, through the forces driving range expansion and contraction (critically important to our understanding of the consequences of climate change), vector identification, and the environmental and social factors shaping disease transmission. Furthermore, opportunities arise when those associated data sets are socioeconomic or cultural in nature.

This corresponds mainly to what it is, where it came from, when it was collected, and who collected it. Additional data indicating the institution responsible for curation and the identifier of the physical specimen within that institution ties the DES tightly to its physical counterpart.

Derived information. It includes data that have been obtained directly from the curated specimen by direct observation, measurement, and analysis of it. Examples include morphological descriptions, measurements of traits, chemical analysis, DNA sequencing, and more. Images produced by a variety of different techniques logically fall into this category but are treated separately as a media class because of their special status.

Associated information. It includes data that have not been derived directly from the specific physical specimen but that are associated in some way. Examples include taxon-level behavioral data, habitat data, data about conservation status, literature including species descriptions, and so on.

A DES is identified using a PID, such as a DOI. This allows a DES to be unambiguously cited in literature and for work done on it (such as extending, annotating, and improving the available information) to be attributed to the proper person or people and organization or multiple organizations. By resolving a PID, the DES can be found, accessed, and used on the Internet. Resolving a PID not only locates the digital object but also provides the location of the corresponding physical object (to institutional collection level).

Metadata captured when the PID is assigned allows the DES to be indexed in searchable catalogs such that the specimen can be found and accessed for use in line with relevant terms and conditions. Its use might only involve the digital data, or it can also involve arranging access to the curated specimen through loans or visits. Relevant terms and conditions encoded as part of the digital representation incorporate and support compliance requirements with various legal or regulatory, ethical or moral, and sensitive data obligations—for example, the CARE Principles for Indigenous Data Governance (Carroll et al. 2020).

A DES does not normally contain all the actual data nor even facsimiles of it. A DES is conceived to be mainly a collection of pointers to other locations where the data can be found. A typical example of this is that we would not expect to include all the DNA sequence data derived from a specific specimen but would include one or more pointers (links) to the relevant records in INSDC databases. As we explained earlier, such databases already contain some references to the specimens from which sequence data have been derived,

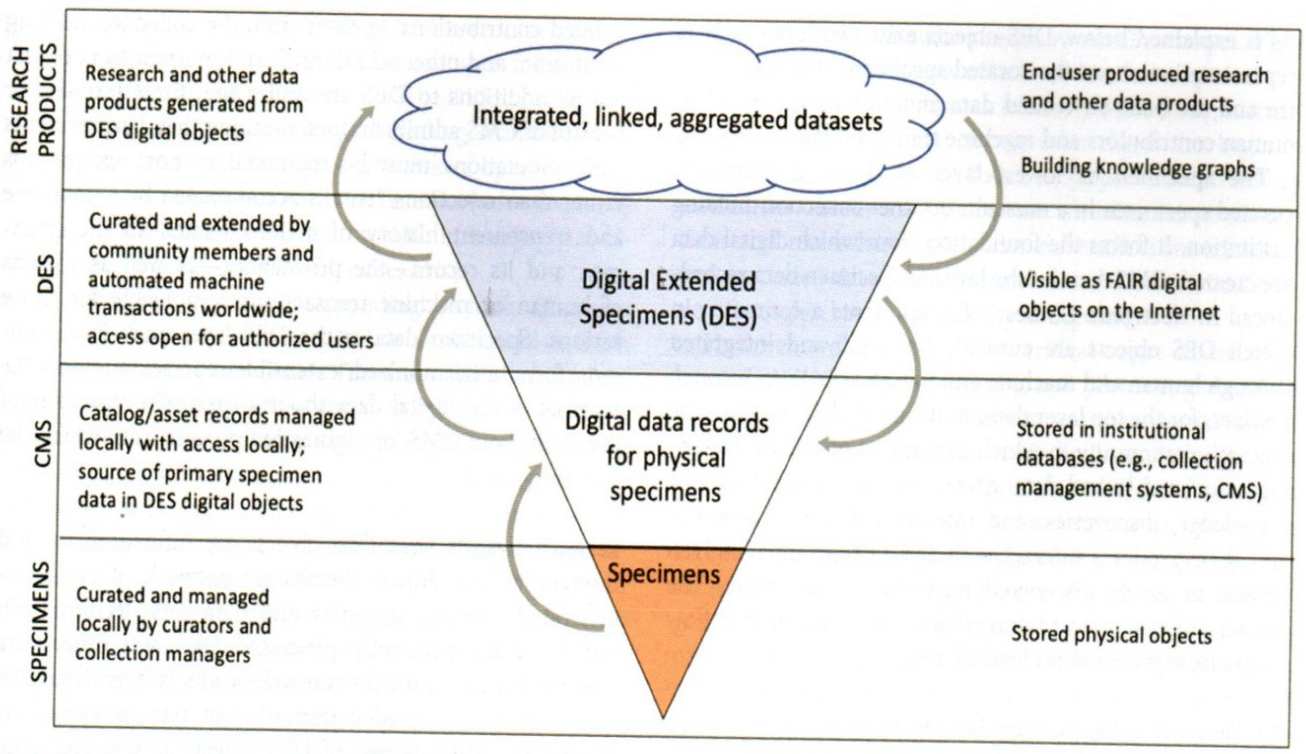

Figure 2. Layers of digital data representation beyond physical specimens and the data flows between them.

but exploiting them today is largely a manual process. In the future, third-party databases should ideally contain reciprocal pointers (links) to the relevant DES, such that an investigator beginning with the DNA sequence or other third-party data can easily locate the specimen and all data associated with it (or vice versa). An important design consideration for ease of use and good performance will be to achieve the right balance between packaging actual data and including pointers to data within the DES.

The role of Digital Extended Specimens

We have suggested that a DES acts on the Internet on behalf of a curated specimen and its data in a collection. This is meant in the sense of an actor standing in for a character in a theatrical script, subject to manipulation by the director, other actors, and the audience. In much the same way, DESes are amenable to manipulation and processing by computer software on behalf of humans. Moreover, as needs dictate, computer software can intermediate the data content of the objects into forms (styles, presentations) that can be comprehended, interpreted, and processed not only by humans but by other computer programs. This will become increasingly important. We say that DESes are machine actionable, meaning that operations can be invoked on them and results returned. As one type of digital object they are editable, interactive, reprogrammable, and distributable (Kallinikos et al. 2013). They can be processed openly or transported transparently between heterogeneous information systems. Not only can generic operations such as create, retrieve, and update be supported, but domain and application-specific operations can be defined. Interoperability difficulties between institutions and systems are much reduced by the definition of object types and operations that underlie the concept. Such objects have the implicit capability to remain FAIR over timescales familiar to collection-holding institutions, which is many decades (more than 100 years). As a specific kind of the more general FAIR digital object (De Smedt et al. 2020), DESes assist to integrate collection data with the data-rich world of (*inter alia*) the life and Earth system sciences.

Focusing on the curatorial role within scientific collections, figure 2 highlights the interconnectedness among physical specimens, digital representations in an institutional CMS, and DES. It illustrates how data at the DES layer benefits from the cognitive surplus of wider, external curatorial and other creative and productive activities of the broader scientific community, expanding and extending the scope and capacities of institutionally based curation and collection management. DESes become channels fostering social connectedness, appealing to our innate and generous human desire to make and share things, especially when such contributions can be widely recognized and attributed to specific persons (Shirky 2011). With DESes acting as anchor points for bringing together available but dispersed information arising from community contributions, this can appeal to individuals and organizations in curatorial and collections-management roles, those in research, teaching, and policy roles, and nature enthusiasts and the engaged public.

A second focus of figure 2, especially of the DES layer, is the linking and use of third-party sources of digital information that form the crux of the extended specimen concept.

As is explained below, DES objects exist two levels of data representation above the curated specimens they characterize and are built on related data annotations generated by human contributors and machine transactions.

The specimen or lowest layer in figure 2 represents curated specimens in a museum or other collection-holding institution. It forms the foundation from which digital data are created. *CMS* denotes the layer of specimen data records stored in electronic CMSes. *DES* represents a commons in which DES objects are curated, extended, and integrated through human and machine annotation. The term *research products* for the top layer denotes the myriad activities in the research community in which processing of DESes, including associated linked data, drives the development of new knowledge, discoveries, and integrations. These activities and their products interact with and further enhance DES objects as newly discovered relationships are made. The arrows indicate forward and reverse directions of data flow from one representation level to another.

Specimen layer. The specimen layer consists of curated objects held by a collection. The specimen may be a complete organism or part of one—for example, tissues in biobanks or specimens that have been divided into preparations such as skin and skeleton. Physical objects are under the care and control of the curators and collections managers who maintain the specimens' labels, ledgers, card catalogs, field notes, storage, physical well-being, and proper environmental conditions, as well as acting as reservoirs of expert knowledge about the collection holdings. Physical specimens are the basis for all electronic transformations and data representations derived from them. They cannot be displaced by digital replicas. They are the anchors for the digital representations and are permanently linked from those. They constitute the material evidence of our natural environment over time and space and provide the substantive matter for continuously and dynamically advancing technological and statistical methodologies and evolving management contexts (Waples et al. 2008). The importance of physical specimens and access to them, therefore, remains high far into the future (Cook et al. 2020, Colella et al. 2021a, 2021b).

CMS Layer. The CMS layer in figure 2 results from the creation of digital specimen records and their inclusion into an institution's CMS. This creates a searchable, electronic inventory of a collection's holdings. The CMS typically includes a wealth of data about and associated with the specimens, such as the object's label data, unique catalog number, original determination, subsequent annotations and determinations, collecting locality, date of collection, and collector name, as well as any local curatorial data used in managing the collection.

DES Layer. The data at the DES layer are initially generated from and anchored in CMS data released from numerous collections and subsequently expanded through community curated contributions, updates from the collection-holding institution, and other additions. It is important to note that not all additions to DES are under the direct purview or control of CMS administrators, meaning that disagreements with annotations must be registered as new annotations rather than rejections. This is accompanied by a complete and transparent history of actions related to the specimen and its record—the provenance—as well as records of human or machine transactions responsible for those actions. Specimen data at the DES layer exist in a common form: a harmonized, extensible representation on the Internet of the digital data about a specimen regardless of the institution, CMS, or digitization process from which the data originated.

Research products layer. DES objects are influenced by and interact with a diffuse distributed network of end-user-generated research products and data sets derived from jointly or independently processed DES and other data sources that constitute the research products layer (figure 2). These products included derived data sets (produced by processing disparate sets of DES objects as was described above); integrated data sets (result of grouping DES objects with other data on the basis of relationships; e.g., bitrophic or multitrophic interactions, genetics, phylogenetic trees, cladistic representations, food webs); linked data sets (which are the same as integrated data sets, but the data sets remain distinct rather than integrated; i.e., transformed into a single data set that contains the desired information); aggregated data sets (data sets brought together for administrative, management, functional purposes, or other reason; e.g., to be available from a single location); literature, including species descriptions and taxonomic treatments; and knowledge graphs (result of identifying relationships between sets of DES objects).

Such data sets and knowledge graphs in turn enhance data at the DES level through annotations adding novel discoveries and contributions to a comprehensive, globally improved DES layer. Given the global nature of contributed annotations, this is also the layer at which language translation services will be appropriate.

Examples and case studies

One example of what could be immediately possible involves the potential for reintegrating results from an experimental clustering algorithm under development by GBIF that identifies potentially related records by matching similar entries in individual fields across different data sets (GBIF 2020). This would allow the GBIF to bring together potentially related biodiversity records across multiple institutions by matching similar entries (e.g., duplicates, exsiccatae, tissue voucher, host–parasite relationships) in individual fields across different data sets mediated at the GBIF.

Needs for specimen-based interlinked data arise in many circumstances, some of which are urgent—for example, when a species is heading toward extinction or during a

global pandemic of zoonotic origins. Three case studies, briefly described below and elaborated more fully in the supplemental materials, illustrate how the DES approach can overcome current shortcomings of finding, accessing, and reusing the largely disconnected network of open data that is presently available. The approach benefits wider society by equipping the scientific community with immediately available, high-value data when such crises arise.

In 2014, the Poweshiek skipperling butterfly (*Oarisma poweshiek* Parker, 1870) was listed as Endangered in both Canada and the United States. This once abundant, prairie-specific butterfly has been reduced to two small Michigan populations in the United States and a metapopulation in Manitoba, Canada. Thousands of specimens in natural science collections play a unique role in addressing the historical landscape-level changes contributing to the precipitous decline of the species. Extending these specimens with associated data from research on population genetics, the bacteria genus *Wolbachia*, environmental contaminants, plant diversity metrics, and climate and weather data extends the ability to ask questions about factors endangering the Poweshiek. Data gathered over the past 30 years integrated like this supports studies and recovery planning by state or province and federal authorities that can inform reintroduction of the Poweshiek into resilient prairie systems across its historical range.

Populations of the koala (*Phascolarctus cinereus* Goldfuss, 1819), an iconic Australian species, are decreasing, in large part because of habitat loss. An important part of understanding the population dynamics of this species and conservation applications lies in distinguishing the three subspecies, the associated literature of each, their current conservation status, and all other available data. Currently, however, even the holotypes and published species descriptions are not digitally linked for two of the three subspecies (*Phascolarctus cinereus victor* and *Phascolarctus cinereus adustus*), and considerable manual effort is required to find, link, and associate each of the relevant elements. This digital disconnect hampers the usefulness of these data, particularly for time-sensitive conservation measures and by data users who may be less fluent in finding and linking the necessary data.

Stimulated by the SARS-CoV-2 pandemic, work to enhance data about the world's specimens of horseshoe bats and related species has revealed that less than 5% of the approximately 90,000 relevant data records published to the GBIF and iDigBio contain references to genetic sequence data. This is even though there are more than 200,000 deposited sequence records for these species in the INSDC databases. More than 1100 bat specimens in collections were "discovered" as the sources of analyzed DNA from among the 200,000 sequence records. These 1100 specimens were not previously linked to DNA sequence data that had been derived from them.

Next steps

DESes offer a striking opportunity for finding, accessing, interacting with, and reusing enriched biodiversity data, especially in the context of computer-assisted research processes. They enhance the connectedness of information and empirical evidence of importance in human health, food, security, sustainability, and environmental change value chains that begin in natural science collections. They can support natural history institutions, ideally in ways that are advantageous to institutions and researchers alike and can benefit from enhanced and expanded data sharing, enrichment, curation, and citation through a broad community of experts.

However, we recognize some critical issues where work is needed. Foremost is maturing and promoting a culture that values DES objects both as a means to advance research and also to support the needs of individuals, helping them advance professionally, whatever their career stage from students to established professionals. Access to appropriate training and continuous professional development opportunities has an essential role to play in this process. Pilot programs for scaling up the transposition of CMS data to DES objects can nurture opportunities for innovating, working, and rewarding in new ways, ultimately showcasing what is desirable in a mature integrated data culture.

Implementing the social and technological aspects of a DES object infrastructure requires a roadmap marking stages of development, with reasonable timeframes and phases of stable and sustainable funding. This roadmap must delineate the social and technical issues to be addressed at each stage, the actors that must act, and what the actors must do. Several must-dos for effective implementation are integral components of such a roadmap.

Core to the necessary social developments will be to build a backbone to support an active, valued, trusted community of curators that addresses divergent perceptions and opinions regarding responsibility for these data and institutional control, the social issues inherent in shared ownership, and the challenges of authentication and authorization of trusted human annotators and machine algorithms required for biodiversity data integration. Likewise we will need to provide opportunities for individuals at various career stages to enter this expanding field of biodiversity science and its wide-ranging intersections with data and computer sciences, biodiversity informatics, and associated disciplines. They must be able to explore broadening the horizon to learn from other disciplines, designing data structures that include other types of specimens–for example, geological. As a community, we will work to ensure accessibility and propagation of updated, annotated, and enriched DES data back to the contributing institution.

Numerous technical advancements will be needed for implementation. It will be important to ensure that a globally unique, persistent, resolvable identifier is assigned to every DES, based on a standard international mechanism for allocating, resolving, and maintaining such identifiers (i.e., PID services) in the domain of natural science collections and that each identified DES is indexed in a searchable global registry. Implementation also calls for the need to establish attribution of DES records through inclusion of collection

date, collector name or names, collecting locality, collecting event details, name and identifier of the curating institution, a PID, and an appropriately formatted normalized citation string. Further, the community will work to develop and promote adoption of standards and specifications for the logical structure and content of DES, the operations permitted on them, general handling rules and behaviors for such operations, and application programming interfaces that can be used by software for harmonized remote operations and the mechanisms of serialization, packaging, and transfer of DES content between different information systems and tools when needed.

Last but not least, DES will need to comply with legal obligations, including constraints imposed by national and international regulations, intellectual property rights, the protection of sensitive information, and the just, equitable, and fair sharing of access, use, and benefits for all people.

Acting with a mandate from the community, an agency committed to managing a program for the implementation of the roadmap must engage both existing and new partners, together identifying existing building blocks and developing advancements that can contribute functionality, options, and strategies for connecting and integrating the building blocks, while periodically evaluating and redirecting as necessary.

Conclusions

Combining the digital specimen and extended specimen concepts to become the DES gives the future foundation for digital information about natural science specimens. It results in physical collections having explicit representation on the Internet. When fully implemented, DES enables research across an array of rapidly evolving scientific fields while providing a nimble architecture capable of supporting the needs of data-intensive research and rapid information extraction. As the biodiversity and specimen data communities work to realize the DES, we invite individuals, organizations, institutions, infrastructures, and representatives from all corners and backgrounds to participate in dialogue, expansions, and applications.

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 871043 (DiSSCo Prepare) and the US National Science Foundation awards no. DBI 2027654 and no. DBI 2033973. Thanks to James Beach, Sharif Islam, Talia Karim, Dimitris Koureas, Nicky Nicolson, Jyotsna Pandey, Barbara Thiers, and Andrew Young for productive discussion and reviews of the manuscript.

Supplemental material

Supplemental data are available at *BIOSCI* online.

References cited

[ABK] Alliance for Biodiversity Knowledge. 2021. Alliance for Biodiversity Knowledge. ABK. www.allianceforbio.org.

Abrahamse T, Andrade-Correa M, Arida C, Galsim R, Häuser C, Price M, Sommerwerk N. 2021. The Global Taxonomy Initiative in Support of the Post-2020 Global Biodiversity Framework. Convention on Biological Diversity. Technical series no. 96.

Addink W, Hardisty A. 2020. "openDS": Progress on the new standard for digital specimens. Biodiversity Information Science and Standards 4: e59338.

Bak-Coleman JB, et al. 2021. Stewardship of global collective behavior. Proceedings of the National Academy of Sciences 118: e2025764118.

Belbin L, Wallis E, Hobern D, Zerger A. 2021. The atlas of living Australia: History, current state and future directions. Biodiversity Data Journal 9: e65023.

Buckner JC, Sanders RC, Faircloth BC, Chakrabarty P. 2021. The critical importance of vouchers in genomics. eLife 10: e68264.

Carroll SR, et al. 2020. The CARE Principles for Indigenous Data Governance. Data Science Journal 19: 43.

Colella JP, et al. 2021a. Leveraging natural history biorepositories as a global, decentralized, pathogen surveillance network. PLOS Pathogens 17: e1009583.

Colella JP, Stephens RB, Campbell ML, Kohli BA, Parsons DJ, Mclean BS. 2021b. The open-specimen movement. BioScience 71: 405–414.

Cook JA, et al. 2020. Integrating biodiversity infrastructure into pathogen discovery and mitigation of emerging infectious diseases. BioScience 70: 531–534.

De Smedt K, Koureas D, Wittenburg P. 2020. FAIR digital objects for science: From data pieces to actionable knowledge units. Publications 8: 21.

[DiSSCo] Distributed System of Scientific Collections. 2021. README.md. DiSSCo.https://github.com/DiSSCo/openDS.

DOI Foundation. 2019. DOI Handbook. DOI Foundation. www.doi.org/hb.html.

Drew J. 2011. The role of natural history institutions and bioinformatics in conservation biology. Conservation Biology 25: 1250–1252.

[GBIF] Global Biodiversity Information Facility. 2020. New data-clustering feature aims to improve data quality and reveal cross-dataset connections. GBIF News (28 July 2020). www.gbif.org/news/4U1dz8LygQvqIywiRIRpAU/new-data-clustering-feature-aims-to-improve-data-quality-and-reveal-cross-dataset-connections.

[GBIF] Global Biodiversity Information Facility. 2021. Converging digital specimens and extended specimens: Towards a global specification for data integration: Phase 1: Digital/extended specimen. GBIF Community Forum. https://discourse.gbif.org/t/converging-digital-specimens-and-extended-specimens-towards-a-global-specification-for-data-integration-phase-1/2394.

Gries C, Gilbert EE, Franz NM. 2014. Symbiota: A virtual platform for creating voucher-based biodiversity information communities. Biodiversity Data Journal 24: e1114.

Güntsch A, et al. 2017. Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. Database 2017: bax003.

Hardisty AR. 2020. What is a digital specimen? DiSSCoTech (31 March 2020). https://dissco.tech/2020/03/31/what-is-a-digital-specimen.

Hardisty A, Bentley A, eds. 2021. Digital Extended Specimens SYM07. Biodiversity Information Science and Standards. https://biss.pensoft.net/collection/302.

Hardisty A, Roberts D. 2013. A decadal view of biodiversity informatics: Challenges and priorities. BMC Ecology 13: 1–23.

Hardisty AR, Ma K, Nelson G, Fortes J. 2019. "openDS": A new standard for digital specimens and other natural science digital object types. Biodiversity Science and Standards 3: 37033.

Hardisty A, Saarenmaa H, Casino A, Dillen M, Görderz K, Groom Q, Hardy H, Koureas D, Nieva De La Hidalga A, Paul DL. 2020. Conceptual design blueprint for the DiSSCo digitization infrastructure—Deliverable D8.1. Research Ideas and Outcomes 6: e54280.

Heberling JM, Miller JT, Noesgaard D, Weingart SB, Schigel D. 2021. Data integration enables global biodiversity synthesis. Proceedings of the National Academy of Sciences 118: e2018093118.