The Digital Extended Specimen will Enable New Science and Applications

Michael S Webster[‡], Jutta Buschbom[§], Alex Hardisty[|], Andrew Bentley[¶]

‡ Cornell University, Ithaca, NY, United States of America
§ Statistical Genetics, Ahrensburg, Germany
| Cardiff University, Cardiff, United Kingdom
¶ University of Kansas, Lawrence, KS, United States of America

Corresponding author: Michael S Webster (msw244@cornell.edu)

Abstract

Specimens have long been viewed as critical to research in the natural sciences because each specimen captures the phenotype (and often the genotype) of a particular individual at a particular point in space and time. In recent years there has been considerable focus on digitizing the many physical specimens currently in the world's natural history research collections. As a result, a growing number of specimens are each now represented by their own "digital specimen", that is, a findable, accessible, interoperable and re-usable (FAIR) digital representation of the physical specimen, which contains data about it. At the same time, there has been growing recognition that each digital specimen can be extended, and made more valuable for research, by linking it to data/samples derived from the curated physical specimen itself (e.g., computed tomography (CT) scan imagery, DNA sequences or tissue samples), directly related specimens or data about the organism's life (e.g., specimens of parasites collected from it, photos or recordings of the organism in life, immediate surrounding ecological community), and the wide range of associated specimen-independent data sets and model-based contextualisations (e.g., taxonomic information, conservation status, bioclimatological region, remote sensing images, environmental-climatological data, traditional knowledge, genome annotations). The resulting connected network of extended digital specimens will enable new research on a number of fronts, and indeed this has already begun. The new types of research enabled fall into four distinct but overlapping categories.

First, because the digital specimen is a surrogate—acting on the Internet for a physical specimen in a natural science collection—it is amenable to analytical approaches that are simply not possible with physical specimens. For example, digital specimens can serve as training, validation and test sets for predictive process-based or machine learning algorithms, which are opening new doors of discovery and forecasting. Such sophisticated and powerful analytical approaches depend on FAIR, and on extended digital specimen data being as open as possible. These analytical approaches are derived from biodiversity monitoring outputs that are critically needed by the biodiversity community because they

are central to conservation efforts at all levels of analysis, from genetics to species to ecosystem diversity.

Second, linking specimens to closely associated specimens (potentially across multiple disparate collections) allows for the coordinated co-analysis of those specimens. For example, linking specimens of parasites/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 multi-trophic interactions.

Third, linking derived data to their associated voucher specimens increases information richness, density, and robustness, thereby allowing for novel types of analyses, strengthening validation through linked independent data and thus, improving confidence levels and risk assessment. For example, digital representations of specimens, which incorporate e.g., images, CT scans, or vocalizations, may capture important information that otherwise is lost during preservation, such as coloration or behavior. In addition, permanently linking genetic and genomic data to the specimen of the individual from which they were derived—something that is currently done inconsistently—allows for detailed studies of the connections between genotype and phenotype. Furthermore, persistent links to physical specimens, of additional information and associated transactions, are the building blocks of documentation and preservation of chains of custody. The links will also facilitate data cleaning, updating, as well as maintenance of digital specimens and their derived and associated datasets, with ever-expanding research questions and applied uses materializing over time. The resulting high-quality data resources are needed for fact-based decision-making and forecasting based on monitoring, forensics and prediction workflows in conservation, sustainable management and policy-making.

Finally, linking specimens to diverse but associated datasets 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), and social vectors in disease transmission.

A network of extended digital specimens will enable new and critically important research and applications in all of these categories, as well as science and uses that we cannot yet envision.

Keywords

infrastructure network, open science, FAIR, derived and associated information, persistent links, transaction documentation, quality, discovery, forecasting, risk assessment

Presenting author

Michael S. Webster

Presented at

TDWG 2021