Persistent Identifiers and Sharing of Digital Information About Scientific Specimens

Cynthia Samar Carr

Museum Studies Digital Curation Certificate Dual Program, Johns Hopkins University AAP

AS.460.674.81.FA21: Research Paper

Prof. Helen Stevens-Martin

December 18, 2021

Abstract

Using persistent identifiers (PIDs) in digital data production and sharing concerning scientific specimens promotes an overarching goal, to allow for creation of relationships. The assignment of unique PIDs is an essential step for enabling findability and accessibility of digital data using the FAIR data model. Implementation of the digital extended specimen links the digital object record with associated and derived specimen parts and research data. Linking to atomized information such as collection event, collector, locality, collection, institutions, taxon (identification), people involved in analyzing and processing the specimen, other related specimens, and many other subsamples and derived and related data can be accomplished with a system incorporating numerous types of unique persistent ids. These many IDs need to be maintained by organizations to prevent broken links and provide redirects for older identifiers. While community development of best practice is influenced by experts in digital data architecture, it must incorporate challenges based on the history of data sharing concerning scientific specimens. The development of identifier systems and normalization around digital object structure and vocabulary needs to accommodate the needs of managers of diverse collections. Most providers are working with a collection management system and with limitations based on past decisions and limited time and finances, so data sharing practices should address these issues to encourage compliance. This paper will use a combination of reviews of the literature and of several interviews with workers in the field to explore community collaboration, persistent ids, and increased mobilization of shared data.

Keywords: Persistent Identifiers, Unique Persistent Identifiers, PIDs, FAIR data sharing, Digital Extended Specimen, DES, Biodiversity Data Pipeline, Data Standards, Darwin Core, Scientific Specimen Metadata Schemes, Required Vocabulary Normalization, Core for Digital Specimens, Community Collaboration in Data Sharing, Persistent Identifier Schemes

Table of Contents

Persistent Identifiers and Sharing of Digital Information about Scientific Specimens

Museum based natural history collections (NHCs) are important for research on evolution, biogeography, habitat interaction, climate change, and many future research projects. "As primary archives of biogeographical data, NHCs permit rigorous analysis of changing mammalian distributions through both space and time in response to climatic and other permutations" according to McLean, et al. (2016, p. 289). The museum (through the usage of a database system) allows for the access and findability of data concerning their specimens. Biodiversity research has become dependent on digitization of specimens, "growth of multi-institutional specimen databases, and increased web connectivity" (McLean et al., 2016, p. 289). Linkage through persistent ids and normalized metadata is a way to support data accessibility (McLean, et al., 2016, p. 293).

Increased accessibility and transparency of data and research about scientific specimens will allow for more analysis, even enabling examination of the process of science itself. Societal prejudices have been embedded into scientific research as well as into protocols for recording data. For example, it is often difficult to identify historical contributions of women and minorities in the established literature. Contributions of many workers were denigrated or never recorded as the researchers only attributed research to the principal investigators. Communities in the areas of scientific study and specimen collection may not have been consulted about the researcher's goals, and many community contributions are not obvious when reading the scientific record. Current broader sharing of data about localities and species names has stimulated activity around renaming of offensive terms and acknowledgement of compromised legacies (see e.g., Chamber, 2021). Creation of methods for digital sharing of more information about scientific specimens allows for better attribution and deeper analysis of current scientific

practice. In some cases, the historical records may be improved to be inclusive and more broadly shared, allowing for more accurate study of the history of science.

Correlation of museum specimen data with observational data from scientists and citizen science initiatives (such as iNaturalist) broaden the scope of possible research for scientific and policy planning goals. In one example, the planning for wind turbine farms, scientists used data from citizen scientists to plot prevalence of eagles, whooping cranes, and other animals in the relevant areas and correlated it with other known information to make decisions about where to place the farms (Lee, 2021). Broad (2021) described how researchers can compare environmental DNA sampling data to comb jelly DNA sequences "to get a much better idea about numbers and species for these extremely delicate ocean creatures" (para 5). When "snippets of DNA that all creatures shed in their environment" (para 8) are collected and analyzed, the results give "much needed precision for biologists seeking to learn the true dimensions of ocean life" (para 5). In another study, after identifying a new phylum of archaebacteria, investigators put together the genomes and searched for DNA sequences in public databases, finding these "previously unknown organisms" in places all over the world (Gramling, 2021). Studies using these sequences depend on proper record keeping and open sharing of research data containing metadata about collection methods and event time and place.

Analysis of specimens stored in museum collections allow research to be done on evolution and extinction over time (Shaffer, et. al., 1998). Recently, researchers were able to analyze genomic and mitochondrial DNA from a Field Museum specimen collected before its population disappeared in the 1940s and showed that it was a distinct species, illustrating that a "[s]pecimen's true utility may not be clear for many years" (Buehler, 2021), when techniques change. The deposition of this specimen voucher in museum collections has enabled future

research. Existing open data requirements "rarely address specimens as primary data …leading to inconsistency in building critical scientific infrastructure" (Colella et. al., 2021, p. 408). Although saving of derivative data is also important, "irrevocable loss of physical genomic resources …poses a greater risk to national security than the potential loss of secondarily derived genomic sequences that, unlike physical specimens, can be regenerated" (Colella et. al., 2021, p.408). Requiring planning for specimen deposition and funding to support it in DMPs (Data Management Plans) and permit applications, as well as including it in annual updates, is important to encourage retention of scientific specimens (Colella et. al., 2021, p. 409).

Because it is not possible to anticipate which specimens, specimen derivatives, or research data associated with the specimens or their associated digital representations will be usable with technology developed in the future, data curators need to store as much of this information as possible (Morrison, et. al., 2017). Biological research conducted by interdisciplinary teams studying complex interactions can be facilitated by attention to the research infrastructure (Schindel & Cook, 2018). Researchers doing 'holistic sampling' of all data in an ecosystem including genetics, isotopic content, interactions, and behavior need appropriate technology and standards to store all the sampled data. Future use of these samples and research "will require durable informatics linkages among all the data derived from that original collecting event" (Schindel & Cook, 2018, p.3), including images and trait data.

The numbers and types of databases storing scientific data continue to increase. Examples include NCBI, which stores genetic sequences, genomic data (GenBank) and BioSample information (NCBI, n.d.), and Neon, which stores "comprehensive datasets of geographical regions of the U.S." (Neon, n.d.). There are databases that are used to store taxon information to be used as authoritative taxonomic backbones for collection managers and

researchers (see e.g., World Register of Marine Species, https://www.marinespecies.org). FuTRES stores functional traits in a database that provides workflow for storing functional trait data at the specimen level to allow data sharing using "a semantic model …powered by extensible parsers, a backend database, and an API" (FuTREs, n.d.). Geome is a database for storing "field and sampling event metadata associated with genetic samples (GEOME, n.d.), while Canada's Centre for Biodiversity Genomics' "BOLD is a cloud-based data storage and analysis platform" (BoldSystems, n.d.) to allow for identification of the species of unknown DNA sequences. The Zoological Information Management System (ZIMS, Species 360, n.d.) is one of several systems used for identification and management of zoological specimens. 3D Scans and other digital data are stored in Dragonfly (Bii, n.d.) or Morphosource (Morphosource Beta, n.d.). Museums also use Collection Management systems and museum servers to store some of their data.

Currently, there is an ongoing 'conversation' taking place on multiple levels concerning the best ways to implement the storage of digital extended specimens; the information about scientific specimens and all of their associated specimens and data. The optimal ways of storing the data depends on the needs of multiple stakeholders. Collection managers of science museum collections must digitize and store data about their specimens so that it is discoverable and interoperable. This is a burden given the many other important tasks needed for managing scientific specimens, documentation, and research around the specimens. Defining the important minimal requirements, along with implementing tools for storing many different kinds of related data is critical to getting more of these collections online and accessible.

The goal of institutions is to put the data online to fit the FAIR standards, which stands for findable, accessible, interoperable, and reusable (GOFAIR, n.d.). For many in medical and

scientific fields, reproducible should be added. Also important is the issue of limiting access to sensitive information. Using resolvable IDs is key to making data findable. The ID must be maintained through systems that will be persistent for at least 100 years. Analysts predict that as new technologies evolve, older IDs will be transformed to fit. Maintaining the links to the IDs will be the responsibility of designated services, institutions, or aggregators. The creation of Darwin Core (https://dwc.tdwg.org/) and other standards that specify how to standardize and store data is also important for helping with interpretability of stored information. Museum collection managers are primary stakeholders whose requirements, based on history and needs of their collections, need to be incorporated into the procedures for creation of digital extended specimens.

The topic of persistent identifiers and digital data concerning scientific specimens includes:

A) What are current ideas about types and implementations around persistent identifiers as well as requirements to allow access to Digital Extended Specimens in the FAIR data model?

B) What minimal information is needed for defining the Digital Object?

C) How do issues such as attribution, annotation, reproducibility, restricted data, and fraud prevention influence digital specimen architecture?

D) How do factors such as the history of collection management, community involvement, and the needs of providers affect the process of data sharing?

Literature Review

Open Science

The Open Science Concept directs that all data generated by scientists should be shared, although often after a defined period to allow for publication. Shared data must be findable,

shared in a way that normalizes metadata so that similar information is available in a comparable format. This reduces the amount of time redoing the same work and allows for further research.

Standardized Data (Darwin Core)

Biodiversity Data storage depends on normalization of data. Darwin Core is a biodiversity standard that has worked well because it is fairly flexible. Terms in this standard are explained with examples that have been and are continually updated by the community (Wieczorek, et. al, 2012). Various extensions are added for different communities and requirements (e.g., Shorthouse, 2017). For data that has need for fields that diverge significantly, new 'core' standards are used, including Audubon Core for audio/visual data from sources such as iNaturalist, and Humbolt Core, which is currently being developed to help the community digitizing ecosystem wide observational data (Neon, n.d., TDWG, n.d.-b.). Documentation of cross walks also facilitates usage of more structured standards, such as ABCD (TDWG, n.d.-a.), and ABCD-EFG extension for geosciences. Development of other standards in the future can be cross walked so that they will incorporate the vast amounts of data already accessible through use of the Darwin Core standard.

Mechanisms for storage

The technical architecture for providing storage and access is being developed in the computer science community. Currently, data is stored using Darwin Core standards enabled by various collection management systems, repositories, and upstream aggregators. Diverse communities and groups have adopted different solutions, which are all fed into the data aggregators for various countries and fields, to interact ultimately with the global cyberinfrastructure. It is important for people building software to talk to people using software; "programmers are good at building stuff, but not good at knowing what is needed – how are

people actually going to enter all this data in. The system gets more complicated for including more users, and standards get more complex. The community needs better discussion to make sure that data can be entered in a non-painful process" (Kwan, 2021, 54:24).

What types of PID Schemas Are Used?

There are various schemas for creating and maintaining PIDS (unique persistent identifiers), including UUIDs (Universal Unique Identifiers), ARKs (Archival Resource Keys, University of California, n.d.) and a few DOI (Digital Object Identifier) schemas. One important aspect is that these IDs should be maintained, so the organization using them should be able and willing to plan for their continued support. IDs that include an institutional prefix may qualify as being unique but may not be safe enough. This is better than a catalog number, which is unlikely to be unique. Using an organization to maintain these IDs (which is the case for DOIs) may be expensive for many collection systems unless the cost is shared. Damerow et. al. (2021) explored multiple possibilities for using PID schemas with required metadata. They called for unifying the approach to allow for collaboration across disciples because having different standards for different communities complicates the ability to reuse data, especially with the low compliance of some researchers for supplying necessary metadata (Damerow et al., 2021, p. 2). PIDs are "arguably essential for supporting data synthesis" (Damerow et al., 2021, p. 6) as assigning identifiers allows linkage of records to metadata. Since IGSNs (geoscience) and ARKs are commonly used for many scientific sample types, the authors compared them (Damerow et al., 2021, see also other comparisons in Appendix A) before choosing to use IGSNs.

Using IDs developed to community standards enables linking to atomized data so that less metadata needs to be added to each record. When a sample is subsampled or further processed, it requires the addition of persistent IDs to subsamples, so that links can be

maintained to the parent id and all related data. The 'internet of samples' project (Davies, 2021) is developing a user-friendly infrastructure to mint PIDs using the IGSN scheme, along with designing other essential functions for data providers (Damerow, 2021, Results, para 2). This iSamples system "will provide the cyberinfrastructure to facilitate such connections within and across scientific domains" (Davies, 2021, p. 2). Eventually, this approach to minting IDs could be added to functions provided by collection management systems (Hardisty, 2021, p. 28).

Guarlnick et. al (2015) mentioned that community practices often do not preserve linkages between data when specimens are divided between institutions and subsequent studies are done. Assigning a GUID (globally unique identifier) as soon as digital information is generated is an important step for maintaining these linkages. Changing the practices for handling new specimens still does not solve the problem of legacy data that have heterogeneous identifier types and metadata associated with them. Guralnick et al. argued that these legacy IDs need to be saved and linked to a new central identifier, maintained by a registration service that "enters identifiers into a database so that the resolver host can look it up and forward requests to the object's current location; for example, user interfaces and APIs exist for EZID ARKs, DataCite DOIs, Handles, and PURLs" (2015, p. 137).

The authors advised that identifier "solutions must support scientists' current practices and create minimal burden during the collecting process" (Guarlnick, 2015, p. 138). Given the complexity of the community, there is no way to get "the entire biodiversity community to adopt a single implementation for identifiers" (Guarlnick, 2015, p. 138). So, to allow for interpretation of collection data, local collection managers need to register their collection data management procedures and specify the identifier scheme.

Publishers can use GUIDs (Globally Unique Identifiers) for all data used in papers to link it to the associated collections. This includes "GUIDs for formally cited or potentially relevant data (e.g., authors, books, articles, taxon names, taxonomic treatments, gene sequences, specimens, etc.) maintained in well- established and widely used external registries" (Guarlnick, 2015, p. 146). Resolution services for persistent IDs should try to standardize their responses so that they are predictable and interoperable. Having human friendly identifiers such as catalog numbers but also using "computer-friendly identifiers (LOD, UUID, DOI, ARK, etc.) for electronic cross-linking" (Guarlnick, 2015, p. 151) is a way to support both management and access. The California Digital Library supports "linking research data to their associated publications via PIDs" (mariapraetzellis, 2021). Their DMP (Data Management Plan) tool will supply machine readable data management plans to generate connections, consistency, and the ability to update plans and use information extracted from the plan elsewhere (Dina Palto in mariapraetzellis, 2021).

DiSSCo is a European wide initiative for creating interoperable infrastructure for digital specimens. As part of its architecture design, Hardisty et al. recently (2021) published a paper that examined persistent identifier schemes. According to them, persistent identifiers should "transcend changes in the underlying technologies of their implementation" (Hardisty et al., 2021, p. 1). PID schemes considered in their paper use the Handle System, where the structure of the ID reflects "assigning responsibilities for administering portions of the entire Handle namespace" (p. 11). Two parts of this structure include the naming authority, and "a unique local name under a specific naming authority prefix" (Hardisty, 2021, p. 11). Naming authorities can follow a hierarchy, but "delegation beyond two or three levels becomes unwieldy" (p. 11).

Three segment prefixes can be used to reflect organizational divisions and responsibilities (Hardisty et al., 2021, p. 13). See Appendix A for examples of PID structures.

After exploration of many schemas and organizational implementations, DiSSCo chose to "adopt a 'driven by DOI' persistent identifier (PID) scheme customized with natural sciences community characteristics" (Hardisty et al., 2021, p. 2). The potential for trust in and adoption of the scheme by the global community was a primary consideration for choosing between systems that could conform to DiSSCo Requirements of Scalability, Trust, Persistence, Governance, Use of appropriate identifiers, and Global Suitability.

In the 'Digital Object Architecture' model, PIDs are used to identify people, organizations, and the "things they work with" (Hardisty et al., 2021, p.3). "Critically, PIDs act not only as identifiers but also as connectors -of one identified thing to another" (p. 4). Using them allows for "machine actionable data packages unambiguously identified with persistent identifiers" (Hardisty et al., 2021, p. 4), in conformance with FAIR digital principles. Thus, the original voucher specimen records can be linked to all data and digital data, publications, grants, and other associations important to stakeholders.

DiSSCo designers decided that forming their own Registration Agency under the umbrella of an established MPA (Multi-Primary Administrator) is the best organizational model for their system, at least for identifying Digital Specimens. Other identifiers are being explored for use with other data types, shown in Appendix C, Table 1 from the DiSSCo paper.

Important points about Persistent Identifiers

IDs are needed for specimens and PIDs for the digital representations of those specimens. To allow for linkage to related data, PIDs should be minted for associated and derived items

including DNA sequences, isotope data, scanning, and other research data. IDs are also needed for subdivided parts of the original specimen that may be used or loaned separately.

At the other end of the pipeline, assigning persistent ids for datasets and additions generated by users and for the results of data manipulation will allow for reuse of these products. For example, to support efficient tracking of errors in the Encyclopedia of Life (EOL), validation failures "result in the creation of new refuted interaction records that challenge the original claim…so other … users have the option to identify data records that are incorrect or controversial" (Schultz, 2020, Abstract).

Without unique persistent ids, unambiguous linkages cannot be made, resulting in confusion about what data or products were actually used, and which research results are related to a specimen. This is especially crucial for manipulation of large datasets and other current research. Virtually every scientific collection has examples of cryptic notes that cannot be deciphered because of unclear references to people, places, or procedures. Assigning persistent ids to each of these entities will allow current and future researchers to be able to reuse the data using their own procedures or new techniques.

Findability

FAIR data infrastructure implementation enabled by persistent ids is the findability and accessibility of the data. If the data is associated with a persistent id that is unique, it will be resolvable in a search. Islam et al. (2020, Kernel, para 4) mentions that the "first step towards FAIR data services" concerning digital objects is assigning a PID and providing kernel information, which is a small number of essential attributes at the record level, specifically "registered for the Digital Specimen object type." Digital Specimen Architecture also includes MIDS (minimal information for a digital specimen), (Addink, 2021, Appendix D).

FAIR and CARE principles

The community is striving to digitize information about biodiversity collections in a format that is Findable, Accessible, Interoperable, and Reusable (FAIR), which defines a 'crucial role of the information system' for specimen collections (Miller, n.d.). For FAIR implementation, persistent identifiers are required for all linked data, including for datasets, queries, and records of a download. The structure of a digital object includes the persistent id, kernel metadata, and linkages to various other data and metadata stores. A schematic of the structure of an extended digital specimen taken from Islam et al., 2020 is shown in Appendix D, along with other helpful visualizations of this digital object.

Collaborations between workers in various fields are helping to determine the required metadata. Minimum requirements include location of the digital 'specimen', date and time of creation, type of object, and ID of the associated physical specimen. There may also be more required metadata dependent on the definition of the object type. It could also be important to include a checksum such as Sha-256 (one way to save information about the stored data that allows for detection of changes) to use as a test for data integrity. Islam et al. states that "Digital Specimens now can become part of a FAIR infrastructure implementation because with kernel information and other metadata, they are findable and accessible" (2020, p. 8).

Other important considerations about data curation are embodied in the CARE principles. (Carroll et al., 2020). These allow for sensitive data to be handled appropriately by being aware of and collaborating with people involved in the communities that are affected most by the storage of the data. This means that indigenous communities will be involved in decisions involving accessibility of data. Control from these communities is more ethical than sharing data without their input as if it is the property of a researcher. Allowing their input could develop

better usage cases and 'fruitful linkages'. It also allows for data restrictions based on copyright, fear of poaching, and other important considerations, such as privacy considerations.

Adjustments and iterations are allowing people to slowly develop convergent infrastructure while allowing 'rights and wellbeing' to be the focus. The addition of CARE to FAIR principles will allow for collaboration across disciplines by forming respectful relationships while using data preeminently for a collective benefit (Carroll et. al, 2020, p. 8).

Linking and many diverse databases

One of the major challenges in developing the global infrastructure is that the many types of data that need to be stored and linked will have different requirements. Thus, sequence data may be in one database, imaging in another, and publications in another. In addition, the same data may be packaged and stored in multiple databases- for example, as part of an investigator's grant, for an institutional goal, and in discipline specific databases. Since the original sample may have been subdivided with parts sent to multiple museum collections, each collection may have digitized and stored the derived information differently in different databases. The way to disambiguate these various incarnations is by assigning a persistent id to the specimen (and to the occurrence).

In Darwin Core, the occurrence (and occurrence id) is supported by the basisOfRecord, which "started out as a way to differentiate specimen data from observation data" (Jegelewicz, 2021). Vocabulary for this field includes MaterialSample (https://dwc.tdwg.org/list/#dwc_basisOfRecord), the evidence for the species occurrence. Developers are discussing whether to change the primary linkage from the occurrence id to the specimen id, or other ways of organizing the information. The current 'flat' relationship of data

used in Darwin Core Archives (to share records with aggregators) could be expanded to include ways to associate tables of information in a relational database type of model. However,

> A radical departure of this sort couldn't happen overnight. It would need to live in parallel with the existing flattened record star schema approach to allow content providers to transition over, and folks like GBIF/iDigBio would need to "flatten" the new system to the old method to aggregate content. But eventually (as measured by usage statistics), the star schema approach would attenuate. (Jegelewicz, 2021, Deepref, Sep 12)

Influence of the History of the Digitization of Scientific Specimens

Originally, many of the databases were developed for collection management purposes, and support was provided for that activity. As the benefits of data interoperability for Open Science became more apparent, various aggregators and collection management services developed tools to help with sharing data. Because the pipeline has formed in this way, local and field specific norms have become part of the system. Currently, some data providers must upload their data directly through various tools (IPTs, Integrated Publishing Toolkits) to the aggregators, some have that service incorporated in their data management system, and some have those services provided by the state or by their institution.

Mechanisms vary greatly by field and collection. In some management systems, there is support for other id linkages. Arctos is a collection Management system that allows for storage of a large variety of alternate ids in the specimen record. Their other id code table (Arctos: Collaborative Management Solution, n.d.-a.) gives a long list of possible id types to add to the record. Collaboration between collection managers who use this solution allow for modification of this and other tables provided for input of specimen information

(https://github.com/ArctosDB/arctos/issues/4101). This is the type of discussion that allows for iterative improvement of data provided.

Cyberinfrastructure

The decisions made at higher levels will have repercussions everywhere as other suppliers of data for the system attempt to increase interoperability and FAIR compliance (Hardisty et al., 2021, p. 11). Thus, the influence of the European effort known as DiSSCo on choices around persistent id, DO structure, and cyberinfrastructure should be important in that geographic area, and may extend worldwide. GBIF and other US communities are discussing the issue of data storage and are currently debating the use of storage of aspects of the Digital Extended Specimen in parallel databases such as relational databases.

Digital Extended Specimen Concept

The Digital Extended Specimen Concept further extends the concept of extended specimen (Lendemer, et. al, 2020), which links a specimen to primary data associated with it (Webster, 2021) and secondary data derived from research using the specimen. The digital 'twin' of the specimen can also be used for research independently of the physical specimen, creating a need for more persistent ids, allowing for proliferation of more linkages. The other digital object types may have "differing circumstances of use and type specific metadata needs ..[meaning that there are different] possible PID schemes that could be adopted for each" (Hardisty et al., 2021, p. 10). Appendix C includes a table of some of these potential object types. Having IDs for collection events allows storage of and linking to metadata about these events. It is also very important is to have an ID for the institution or the collection where the physical specimens are managed. Knowledge of the collection of origin gives additional information about context of collection and provides access to specimens that aren't digitized.

More linking can be attained using location ids, collector ids, and other IDs specific to a digital object type. For example, the medical research field has RRIDs (Research Resource Identifiers) to allow for identification of crucial reagents used in a study. An important test of the scientific method is to be able to reproduce another researcher's results. Having persistent ids for specimens, protocols, products, and reagents allows for this test. The use of RRIDs allow the biological reagents to be "identified unambiguously" (Bandrowski & Martone, 2016, p. 434). Obtaining a unique and stable RRID is "fairly simple" (Bandrowski & Martone, 2016, p. 435), involving depositing the resource and registering information about it. Biological collections need Taxon ids to describe the identification of their specimens, and paleontological collections also require linkage to the appropriate era and stratigraphic data.

Attribution

Linking the specimen to persistent ids of people who worked on the specimen shares more information about time, place, and culture. Attribution also encourages deposition of specimens and digitization work. Bionomia (Bionomia, n.d.) is a resource that allows collection managers to find the correct person associated with their specimen. It uses the ORCiD (ORCID, n.d.), which allows registration of living people, and Wikidata (Waagmeester et. al, 2020), which has Q numbers as identifiers for living and deceased persons. Input from community members helps resolve ambiguities caused by multiple ways of referring to a person.

Annotation

Annotation is one of the most complicated functions enabled by persistent id and digital object structure. Annotations allow for researchers in the field and members of the public to add information about mistakes or relevant facts. To be useful, annotations need to be added to the original record. The data producers should have some input on whether the annotations have

merit, but this becomes a huge burden if they need to deal with large numbers of annotations. The history of annotations and their evaluation (why they were correct or not) needs to be stored.

One way that this process could be managed "may involve changing the biodiversity data publication paradigm to one based on the atomized transactions relevant to each individual data record" (Rios et al., 2021, para 3). In this approach, each addition of information changes the record to a new item, either through addition of data or by adding links to data stored elsewhere. This would allow for unambiguous access to the specific data that was used in a study (at whatever stage in the chain), and for tracking to identify potential addition of mistakes. Development of two-way annotations by using digital signatures linked to annotations made will help in review of changes. This model may require more types of persistent ids, and a roadmap for that is being developed by the authors and other collaborators at GBIF, to be presented at upcoming collection management meetings.

Transparency- Fraud Detection

Sharing datasets, images, metadata, and even code in open databases/repositories will allow others to evaluate data and determine whether conclusions were reached appropriately. It also allows for detection of fraud since datasets can be tested to make sure that they are likely to have been generated using the methods described. For example, fraud can be discovered when the data that was supposed to have been discovered was 'too good' (e.g., Reardon & Jain, 2020).

Methodology

The methods used in this report include examination of related topics in the above Literature Review and in interviews conducted with three members of the community involved in sharing of digital data about scientific specimens.

Results and Analysis

Results, Part 2: Review of Interviews

Mayfield-Meyer, Teresa- Project Coordinator, Arctos Collection Management Solution

Mayfield-Meyer discussed persistent ids and the concept of the digital extended specimen on September 17, 2021. Mayfield-Meyer has been working with Arctos since 2016 when she was a collection manager digitizing collections at the University of Texas at El Paso, and the Arctos community really helped her in her work then. Following that, she was hired to migrate data for the Terrestrial Parasite Tracker grant. Currently, she is a Project Manager bringing new collections into Arctos.

She emphasized that persistent ids need to be unique as well as persistent and that different structures for persistent ids are difficult to compare. Arctos does have access to some DOIs, but, so far, they are under another schema, and getting more would require more money. According to Mayfield-Meyer, the discussion about how many items need to be assigned persistent ids "is kind of like 'how to build a brain', and types and numbers of persistent ids explode out of control." For example, assigning a stable part id will help to identify which sample (associated with a specimen) is being referenced. This is complicated by the structure of the data record which can be organized around occurrence (the specimen is evidence for an occurrence) even though collection managers think of it as a specimen record. This confusion may be alleviated by assigning multiple material sample ids to each occurrence record.

Eventually, a different data structure will need to be used globally, changing the flat Darwin Core Archive to one that is more like Arctos, or a relational database. "The reality is- the aggregators do not use relational databases and I don't know if they ever will. The Darwin Core Archive is a flat file attempting some ways to convey relational data. I don't know if a relational aggregator (like Arctos) is the answer, but what happens at the aggregator level (GBIF, iDigBio)

will probably need to be more COMPLEX if we are going to aggregate 'digital extended specimens.'" (T. Mayfield-Meyer, personal communication, November 12, 2021).

According to Mayfield-Meyer, the number one concern is that there is a 'huge spectrum of experience' in data providers which causes difficulty for participation in digitization and discussion. The different experience levels make it challenging to communicate to all people what is happening. Each collection should have a data coordinator to help identify best practice. Otherwise, there is a huge hold up in improving records, and we are constantly creating a "backlog of less than good records." There also needs to be better data management downstream so that the correct record and version of that record can be identified to allow collection managers to figure out how useful their collection has been, as well as help other users to evaluate the research. Annotations to improve data need to be made more accessible, so that community members can evaluate the information using the history of annotation data. The collection manager is often not an authority and is only "taking care of the information that we have." Biology is so complicated, and this complexity needs to be conveyed by the records. This is not easy. "We are involved in PROCESS; Science is reaching toward, trying to get the correct answer." It is interesting to add new resources and a challenge to get to where half of us are using them correctly.

Krimmel, Erica- Digitization Resources Coordinator, iDigBio

Krimmel discussed PIDs on October 8, 2021. She started out as an information sciences person/geologist at the Chicago Academy of Sciences, which migrated digital information about their collections into Arctos (CMS). She is now a digitization resources coordinator at iDigBio, a national coordinating center for the Advancing Digitization for Biological Collections initiative funded by the National Science Foundation (iDigBio, n.d., para 1).

Krimmel's evaluation of the health of the field is that it is a mess but moving in a good direction. One problem is a communication gap between computer scientists and collection personnel. Computer scientists can solve problems, but they need to know what problems to solve; collection managers must talk in a way that computer scientists "need it to be phrased."

"The role of persistent ids is to track specimens and to link physical specimens with other information, both digital and physical." A key question is 'who is managing this system of ids?' and how are the links being maintained? The field needs to move to future numbers that "allow the tech systems to move". This requires an 'interchange of sorts- an aggregator of identifiers that knows all of the types' of data and interactions needed. This brings up the subject of who pays for people to maintain the PIDs. Krimmel does not think that paying for DOIs is sustainable, but she also thinks that using free services is not sustainable. This is supported by the example of the development of LSIDs, which was originally funded by a grant. Now that the funding ran out, "[s]ince they are not maintained, you can generate one yourself, just make one up, and there is no one to say that it already exists." Natural History collections already have a tradition of assigning an ID to identify specimens to facilitate organization and research. There is a problem of figuring out how to make these IDs globally unique.

"The conceptual format is not as important as functionality." Someone must maintain the IDs. "There is no agreement in the community, so they leave it up to the collections, where decisions are too big of a burden to expect consistency." Maintaining IDs is sometimes accomplished by narrowing the focus, as in ORCiDs. DiSSCo is a European initiative discussing how to provide persistent ids for specimens. Wikidata is also used for generating and storing ids. Wikidata is a place for identifiers (Waagmeester et. al, 2020), but it is not trusted, and people feel anxiety because it is not controlled, leading to misassociated identifiers.

> Another issue involves metadata and how to track annotations.
>
> Who is the authority? For physical specimens, it is often the collection manager. But Extended Specimens are distributed amongst institutions. The authority there would be whoever can be counted on to monitor it to be unique and resolvable. You need to control communications to make assertions open [transparent]. Then, users decide who they trust. If you want to be really certain, some authorities are more trustworthy, so noting this needs to systematically be part of the annotation.

Although collection managers are typically people who strive for perfection, they need to be able to share data without expecting perfection. Current discussion about changing aspects of record organization, controlled vocabulary, and type of database may result in changes in methods for sharing of extended specimen data. This may be accomplished by using Organism id as the nucleus, and some form of relational database. The transition would have to take place over time, allowing for efforts to enable inclusion of all of the past data shared through Darwin Core Archive Occurrence Records.

There also needs to be a way for humans to monitor some of the digital data management assisted by tools. If persistent ids are not 'really human readable,' some mistakes can take years to catch and be difficult to correct. Each collection really needs "data managers [who] can think critically to do things more efficiently, get data in standardized format so anyone can interpret it. This is NOT TRUE NOW."

Collection Managers need to be open to getting the conceptual details and use tools that are in use now. They also need to put in place long term plans and be welcoming and interested in new technology.

Neu-Yagle, Nicole- Assistant Collections Manager, Earth Sciences, Denver Museum of Nature and Science

On October 8, 2021, Neu-Yagle discussed ideas about persistent ids and digital extended specimens using EMu (another collection management system), in the DMNS Earth Sciences department. Collections that use EMu have to upgrade in big steps as new versions are available, and they need to pay more for any improvements or additional modules for this modular system. The Earth Sciences department has been hampered by the decision not to upgrade from version 5 to version 6 to save money during the pandemic. In her department, the only data that is assigned persistent ids is multimedia, and that is because that function is turned on in their version of EMu. IRN numbers are Internal Record Numbers generated by the database. "They are unique within our DMNS EMu database, and help us identify precise records: every party, taxonomy, catalog record, multimedia (such as an image), etc. has its own unique IRN. They become meaningless outside of our institution" (N. Neu-Yagle, personal communication, November 9, 2021). Field numbers are another number assigned by the museum researchers using their own protocol to identify localities visited. Sometimes, when new data is added about field numbers, this can be associated with past sites, giving more information about them. They can also be aligned with sites visited by other researchers at other institutions or with the same site as it is stored in other databases such as BLM (Bureau of Land Management).

When Neu-Yagle needed to assign GUIDs to records to share with a TCN, she used a program to generate UUIDs, which were then shared with the aggregator. In EMu, the collectors and other people associated with a collection are called parties. Since it is complicated to assign more than one role to a party, there are multiple records for each party to align with each of their roles. Many types of IDs are stored in the notes of an appropriate module since there is no

designated place for them currently. EMu 6 may have more places; for example, there is a space for ORCiD in EMu 6. Another issue is that fossils or cores often have multiple samples in the same cataloged item, so multiple taxon ids are associated with it. DMNS catalogs the item under the vertebrate id, since there are usually more restrictions associated with vertebrates, but adds the other IDs to the record.

To allow for sharing of data to TCNs or to aggregators, Neu-Yagle must prepare the data for the IPT (Integrated Publishing Toolkit) herself. "In order to upload data to aggregators, you Do a Report. So, you run a report, then email the CSVs, 1000-2000 records at a time, not the whole collection. This Report is not generated in a perfect format, so you have to tweek it." Neu-Yagle searches for duplicates using a script that she wrote, then fixes them. Images are transferred separately utilizing a program from Cyberduck (n.d.). She fills in fields that differ between the report and the required Darwin Core format and fills in missing data that might not have been put into the correct field. So, the sharing of records from DMNS Earth Sciences is difficult and requires tweaking and manipulation to generate the CSV files that are sent manually by the collection staff. Institutional decisions affect what is stored and how. This is dependent on history of how numbers and metadata were saved in the past and the capability of the collection management system. Lack of support in the past has made this into a complicated process. "Currently, the collections division's top priority goal is digitization of collections, so maybe the administration can't deny resources now."

Discussion

There is a contrast between the collection manager perspective and ideas about data sharing at the aggregator level. This is illustrated by the fact that the collection manager has to spend a lot of time generating and linking the data by manipulating the collection management

system and using work arounds to make the data fit Darwin Core for upload using the IPT. The use of collection management systems may help the collection manager, but different systems require different amounts of effort. Ultimately, both the choice of collection management system and of the effort needed to make the data fit for sharing is dependent on the history of collection management practices and the amount of money and administrative support available for digital data management.

Sharing data that conforms to FAIR data standards requires the assignment of persistent ids. "The role of persistent ids is to track specimens and to link physical specimens with other information, both digital and physical" (E. Krimmel, personal communication, October 8, 2021). To make data findable and accessible, "they are unique, and resolvable and maintained" (E. Krimmel, personal communication, October 8, 2021). This means that some institution needs to be able to maintain the IDs for an extended period. A key question emphasized by Krimmel (personal communication, October 8, 2021), is "who is managing this system of ids," and how do they facilitate adding the interactions that need to occur between aggregators as well as local databases? The discussion about how many items need to be assigned persistent ids "is kind of like 'how to build a brain', and types and numbers of persistent ids explode out of control." (T. Mayfield-Meyer, personal communication, September 17, 2021). Storing data in specialized packets allows it to be used for more than one digital object. For example, multiple specimens may have been collected at a common collection event, by a common collector, at a common locality, or have other information in common. The ID associated with each of these information packages needs to be persistent so that including it in the specimen record will make data entry and linking easier without having to worry about broken links and 404 messages.

The structure of the digital object also includes minimal information needed for that type of digital object, which would be defined based on the type of specimen or data described. For example, different required information would be needed for a biological specimen, for a geological core sample, for a photograph, for a species list, or for other types of data. Also included in the digital object are various linkages to related data such as scans, localities, collectors, and to other items that are stored in other places. Collections often store images on external hard drives or in Morphosource (Neu-Yagle, personal communication, October 8, 2021). If there is a Morphosource (Morphosource Beta, n.d.) file, then its ID needs to be added to the digital extended specimen (the specific type of digital object). Many of these additional databases and repositories have their own standards for assigning ids, so the choice of a repository will include evaluating the support for continued access.

The Darwin Core standard is maintained by the Biodiversity Information Standards Group (TDWG). In this group, digital data managers discuss the types of information needed for metadata fields describing scientific specimens and define required vocabulary. Task groups are formed to work on constantly changing aspects of the standard to help accommodate different types of collections and data. In this way, the standards can be made widely applicable and more interoperable. Teresa Mayfield-Meyer recently convened a task group that is discussing a type of controlled vocabulary concerning MaterialSample (T. Mayfield-Meyer, personal communication, September 17, 2021). This is one of the Darwin Core terms for the field basisOfRecord, which establishes the evidence for an Occurrence, the primary ID for records that are stored in the GBIF aggregator. The discussion about this has become quite complicated since specimens may be sampled and manipulated; thus, they may need more than one Material Sample id for each Occurrence id (Jegelewicz, 2021). In the collection management world, the

specimen is usually the object being managed, so this structure is a bit difficult to align with tasks. In fact, having an Organism id as the primary ID for the digital object could make a lot more sense, as it would be able to be linked to all of the other ids more intuitively (Arctos: Collaborative Collection Management Solution. (n.d.-c.). The organism is the entity to be sampled, labeled as to how it is related to other specimens, and linked to data that is derived from its parts. Yet, since Organism is often defined as a Living Specimen, it may not be usable in all contexts, and other organizational structures are being discussed (R. Burkhalter, personal communication, November 18 & 19, 2021).

As the Extended Specimen concept (Lendemer et. al, 2020) is implemented, it is clear that the digital object structure needs to include fields for multiple IDs. Although the central ID is not more important than some of the other IDs for object structure, it may make it easier for humans to understand organizing by an Organism (or similar) id to multiple specimen or part IDs that may be managed by different organizations when specimens are subdivided and sampled. The Arctos community is currently discussing the creation of a new Entity collection that will include Organisms and other Entities (for different types of collections). This service may be provided for Entities that have multiple parts, especially if they are in different collections. If it is useful, this Arctos collection may be provided to aggregators via its own IPT. Using this structure, the Entity id will populate the Organism id field, or whatever the aggregators allow (https://github.com/ArctosDB/arctos/issues/3765). This eventually may make data easier to manage, but it is difficult to figure out how the whole community would make the transition to a new structure (E. Krimmel, personal communication, October 8, 2021).

Another potential innovation is the transformation of the flat Darwin Core Archive records to a relational structure to accommodate the Digital Extended Specimen Concept. “At

the Aggregator Level, [it is possible that they will] use relational databases, yet Darwin Core Archive is a flat file relationally, [currently] used by GBIF. We need Arctos on a global scale to follow all the stuff needed for Extended Specimens" (T. Mayfield-Meyer, personal communication, September 17, 2021). Since this is a big change for data exporters, it may exist in parallel with current practices. As predicted by David Shorthouse, "I expect we'll have a transition period where some providers will continue to publish data via eg [sic] traditional Integrated Publishing Toolkits whereas others will use whatever infrastructure will support DES" (n.d.).

At this point, data is not this well-organized. For example, there is no agreement on persistent id structure or maintenance; "This requires an 'interchange of sorts- an aggregator of identifiers that knows all of the types' of data and interactions needed. An important question is 'Who pays?'" (E. Krimmel, personal communication, October 8, 2021). Persistent ids need to be maintained at the local level since paying DOI for services is not sustainable and using services that stop working when the funding runs out is also not sustainable, according to Krimmel (personal communication, October 8, 2021). In a post exploring this issue, Waddink asserted

> the costs of having PIDs are directly related to FAIR data, as PIDs are fundamental to FAIR data. There are many studies available about the economic cost of not having FAIR data… The costs [of having PIDs] are mainly in additional services offered, like services that check there is only one PIDs (sic) for a resource, that guard against broken links, have suitable metadata schemas, make the PIDs discoverable, link them with other PIDs etc… If these costs would be shared globally that would be very affordable and minor in comparison with the digitisation costs or the economic benefits (2021).

Compounding the difficulty of modifying collection management practices to allow for inclusion in cyberinfrastructure are the barriers for collection managers who are underfunded or in poorer countries who do not have access to many meetings and journals that give information and assistance with developing best practices (Valenzuela-Toro & Viglino, 2021). Providing resources in the collection management systems is one way to assist managers with data curation tasks. Arctos as a community has recently explored the possibility of supporting some of the collections affected by these inequities, making Equity and Redress one of their Core Values (Arctos: Collaborative Collection Management Solution, n.d.-b.).

Sometimes, the focus of an organization maintaining persistent ids needs to be narrow to allow for sustainability. This is true of ORCiDs; "ORCID is a non-profit organization supported by a global community of member organizations, including research institutions, publishers, funders, professional associations, service providers, and other stakeholders in the research ecosystem" (ORCID, n.d., para 5). This type of ID only works for living people involved in the research 'ecosystem,' and allows for research effort to be attributed to the person who did the work. Attribution is very important (Thessen, et. al, 2019) because it allows researchers, managers, and others to get credit for work done, encouraging engagement in this process. It also helps with linking of data. The Arctos community currently uses a code table (Arctos: Collaborative Management Solution, n.d.-a.) that defines types of identifiers that may be linked to a specimen record. By using the defined type, collection managers may choose to enter 'other id' numbers into the *Identifier* section of an Arctos record. There is also linking via Relationships in the *Relationship* section that uses a pick list for entry including 'same individual as.' Currently under discussion is a modification to the other id table so that more information could be added, including who assigned the other id and when, as well as, interestingly, the

relationship inferred by the other id (jegelewicz, November 10, 2021, #issue-1049962133). This would allow for more nuance as the relationship (collected by, in collection of, part sequenced) could be better defined. In Arctos, the individual (or organization) doing the assigning is referred to as an Agent. Other systems have different designations, such as the EMu designation of 'party' (N. Neu-Yagle, personal communication, October 8, 2021). All of these should link back to the ORCiD or Wikidata identifier in the CMS record.

Attribution can also be used to help end users evaluate the reliability of evidence; they can know if an identification was asserted by a collection manager with some experience, an acknowledged expert, or a student in the field who has just started their career, for example. David Shorthouse has developed Bionomia, a resource for finding people both living (derived from ORCiD) and deceased (from Wikidata) that can be updated by community members who might have information about specific workers.

The ability to create annotations also needs to be supported by the digital object architecture. Using persistent ids allows the annotation to be linked back to the original record so that the data provider (or other authority) can evaluate the additional information and accept or reject it. Thereafter, the result of this evaluation also needs to be linked to the data record to inform future users. One way to do this is to develop a transactional model for supporting the Digital Extended Specimen. According to Rios et. al (2021), such a model would "involve changing the biodiversity data publication paradigm to one based on the atomized transactions relevant to each individual data record" (para 3). There is a need for a network to push annotations such as in transactional publishing. The output would be a new entity, which only allows users to add data (not subtract). That way, each object in the 'chain of objects' will be

able to be identified independently so that researchers can figure out which set of data was used in each study. (Rios, 2021, 1:34).

"The reality is that a lot of ID [assignment]s aren't all that great, so there needs to be a tool that can double check their quality. Annotations by other people can help a ton" (T. Mayfield-Meyer, personal communication, September 17, 2021). This is an issue for all metadata associated with specimen information. If the identification of these errors in records is noted for the use of managers and end users, even if the individual collection managers cannot evaluate the records, community feedback will give information about the reliability of the information. In this context, it is important to create digital signatures for annotators so that future users can evaluate the level of trust that they should place in these annotations (Schulman, et al., 2021, Data quality management).

Collection managers try to fit their data into a normalized data system that includes mandatory data fields to make it interoperable for sharing. However, the use of persistent identifiers often depends on historical practice since the data providers need to manage the IDs for the long term. According to Krimmel, "The catalog number is conceptually a persistent id, but there is a problem with implementation. There is no agreement in the community, so they leave it up to the collections, where decisions are too big of a burden to expect consistency" (personal communication, October 8, 2021). Because of this, there are different persistent id schemas used by collections, and aggregators need to accommodate them in the sharing of data. Collection managers depend on their collection management systems to help with normalizing their data, assigning ids, creating links, and sharing to aggregators. This process is important as many collections are mainly accessible through aggregators such as GBIF. Increasing usage of tools by the CMSs should allow for availability of more standardized data. Arctos allows

collection managers to upload data to aggregators using the IPT process. "What if Specify [and other CMSs] also had an integrated IPT? That would be ideal" (E. Krimmel, personal communication, October 8, 2021). Another initiative, known as iSample in a Box, would most likely work to make data discoverable in iSamples central thorough 'installations' used in Arctos, Specify, k-EMu, etc. (RDA: Research Data Alliance, 2021, 57:48- 58:35).

In this increasingly complicated digital world, the optimal strategy for collections is for each collection to have a digital data manager. "If we want to share more [data] well, every collection should have a collection manager, but they also need to have a data person that ensures data quality, and evaluate what system is 'best quality'" (T. Mayfield-Meyer, personal communication, September 17, 2021). "Maybe many collections will have a person like Teresa [Mayfield-Meyer], whose role is shared amongst institutions. People need to be open to getting the conceptual details and use tools in use now", according to Krimmel; they also "need to put in place long term plans and be welcoming and interested in new technology" (E. Krimmel, personal communication, October 8, 2021). However, a common situation is that collections are not funded for digitization and lack necessary personnel (N. Neu-Yagle, personal communication, October 8, 2021). There is currently a huge range of experience in collection management, which is a major challenge for communication and for participation of professionals in development of standards (T. Mayfield-Meyer, personal communication, September 17, 2021). Even though collections staff may need to provide FAIR compliant data in stages, it is still important for them to make long term plans and to provide access to their practice protocols so that the data can be maintained and be interpreted by others. Work from digital data researchers discussing appropriate advances in data infrastructure will always need to

accommodate the needs of individual collection personnel as they struggle to share huge amounts of collection data and to improve the quality of digital data records.

Conclusions

The creation of a digital architecture suitable for sharing in the community and, ultimately, to other potential end users is an ongoing process, and best practices may change soon. New technologies are being developed and collaborations with other communities attempted. There is constant updating of data standards, including required vocabulary. The implementation of the Digital Extended Specimen concept may be facilitated by changes in organization of the flat Darwin Core Archive data files that are supplied to aggregators to fit a relational database type structure. As data management practices change, conversations between managers of the data and developers of new data architecture will be needed to incorporate older data into these proposed new data schemas. Collection management professionals working on the side of providing data face a difficult task of deciding how to manage historical data workflows to fit into a required structure for sharing. It would be optimal for collections to have a data manager to manage this work. In most cases, this is an iterative process, and managers can aim to provide data in increments that increase compliance with FAIR data practices. For many professionals faced with time and budgetary constraints, there is a limit on how much can be done. Efforts in this area will provide great benefits for current and future research as this mobilization of data allows for new types of analysis to be done, including ecosystem wide correlations and integration with other types of data sets, even in other fields. These future integrations will require even more types of metadata to allow for interpretation by diverse audiences.

Although collection data managers are anxious to provide perfect results, it is more fruitful to concentrate on improving data so that it can increasingly be used by researchers. Trying new methods and improving data is similar to the process of science; “We are involved in PROCESS. Science is reaching toward, trying to get the correct answer. It is interesting to add new resources and a challenge” (T. Mayfield-Meyer, personal communication, September 17, 2021) to implement them correctly.

While designers of digital data infrastructure have an important role in debating creation of aspirational digital object architecture, the needs and practices of data providers will have to be addressed. Thus, the creation of the Darwin Core biodiversity standard emphasized community involvement in the definition of terms. It continues to evolve based on input from community members as they propose new fields, map data to current fields, and develop additional ‘cores.’

The use of unique persistent identifiers for scientific specimens also depends on historic practices and needs of specific communities. Most collections started with locally defined identifiers as science requires the ability to reference specific specimens. Different regions and initiatives have addressed the need for persistence and resolvability independently, resulting the adoption of multiple schemas (some of which were developed to satisfy technical requirements for semantic web). Local needs might depend on national priorities, limited financial resources, or fit with current practices. The importance of community involvement will continue to be primary as old and new models exist side by side during any transition. A discussion of the most appropriate identifier schema is helpful, but it is more important to develop ways for the data to be harmonized. Individual collections need to standardize their practices and provide information about them, so that data provided to aggregators can be matched and used to create

an interconnected system. Continuing educational efforts can help collection data managers. On the individual collection level, digitization is sometimes frustrating and iterative, yet it allows for increasing usage of specimens and the data derived from them. The updating of the infrastructure 'backbone' on all levels is already leading to an astounding increase in the availability of data. With communication, the work of sharing improved digital records from the backlogs of undigitized and 'less than complete' specimen data can continue in an increasingly efficient iterative process.

References

Addink, W. (2021, March 5). Structure and Responsibilities of a #digextspecimen. *GBIF*. https://discourse.gbif.org/t/structure-and-responsibilities-of-a-digextspecimen/2533/22

Arctos: Collaborative Collection Management Solution. (n.d.-a.). Documentation for code table coll_other_id_type. Retrieved October 26, 2021, from https://arctos.database.museum/info/ctDocumentation.cfm?table=ctcoll_other_id_type#organism_id

Arctos: Collaborative Collection Management Solution. (n.d.-b.). Draft Mission and Vision. Retrieved November 11, 2021, from https://arctosdb.org/draft-statements/

Arctos: Collaborative Collection Management Solution. (n.d.-c.). Organism ID. Retrieved December 5, 2021, from https://arctos.database.museum/info/ctDocumentation.cfm?table=ctcoll_other_id_type#organism_id

Bandrowski A. E., & Martone, M.E. (2016). RRIDs: A Simple Step toward Improving Reproducibility through Rigor and Transparency of Experimental Methods. Neuron *90*, (May 4, 2016), 434-436. http://dx.doi.org/10.1016/j.neuron.2016.04.030

Bentley, A. (n.d.). Structure and Responsibilities of a #digextspecimen. *GBIF*. https://discourse.gbif.org/t/structure-and-responsibilities-of-a-digextspecimen/2533/11

Bii. (n.d.). Bioimage Informatics Index. *Dragonfly*. Retrieved November 13, 2021, from https://biii.eu/dragonfly

Biodiversity Collections Network. (n.d.). *Resources.* https://bcon.aibs.org/resources/

Bionomia. (n.d.). Retrieved November 10, 2021, from https://bionomia.net

BoldSystems. (n.d.). Barcode of Life Data System. Advancing biodiversity science through DNA-based species identification. http://www.boldsystems.org

Broad, W.J. (2021, October 16). Taking the pulse of the ocean's comb jellies. *The Denver Post.* 8C.

Buehler, J. (2021, August 14). Lost blue butterfly was its own species. Urban development drove California insect to extinction. *Science News.* 200(3), 14. https://www.sciencenews.org/article/xerxes-blue-butterfly-first-human-caused-us-insect-extinction

Bushbom, J. (n.d.). Structure and Responsibilities of a #digextspecimen. *GBIF.* https://discourse.gbif.org/t/structure-and-responsibilities-of-a-digextspecimen/2533/30

Carroll, S.R., Garba, I., Figueroa-Rodríguez, O.L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J.D.,

Anderson, J. and Hudson, M., (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal*, 19(1), 43. http://doi.org/10.5334/dsj-2020-043

Chamber, J. (2021, September 25). Racist legacies lurk in common names. In the scientific community, support grows for renaming animals. *Science News,* 200 (6), 12.

Colella, J. P., Stephens, R. B., Campbell, M. L., Kohli, B. A., Parsons, D. J., & Mclean, B. S. (2021). The open-specimen movement. *BioScience*, *71*(4), 405-414.

Cyberduck. (n.d.). Download Mountain Duck. Retrieved November 14, 2021, from https://cyberduck.io

Damerow, J.E., Varadharajan, C., Boye, K., Brodie, E.L., Burrus, M., Chadwick, K.D., Crystal-Ornelas, R., Elbashandy, H., Alves, R.J.E., Ely, K.S., Goldman, A.E., Haberman, T., Hendrix, V., Kakalia, Z., Kemner, K.M., Kersting, A.B., Merino, N., O'Brien, F., Perzan, Z., Robles, E., Sorensen, P., Stegen, J.C., Walls, R.L., Weisenhorn, P., Zavarin, M. and Agarwal, D. (2021). Sample Identifiers and Metadata to Support Data Management and Reuse in Multidisciplinary Ecosystem Sciences. *Data Science Journal*, 20(1), 11. http://doi.org/10.5334/dsj-2021-011

Davies, N., Deck, J., Kansa, E. C., Kansa, S. W., Kunze, J., Meyer, C., Orrell, T., Ramdeen, S., Snyder, R., Vieglais, D, Walls, R.L. & Lehnert, K. (2021). Internet of Samples

(iSamples): Toward an interdisciplinary cyberinfrastructure for material samples. *GigaScience*, *10*(5), giab028. https://doi.org/10.1093/gigascience/giab028

Deck J, Gaither MR, Ewing R, Bird CE, Davies N, Meyer C, Riginos, C., Toonen, RJ, Crandall, E.D. (2017). The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples. *PLoS Biol*, 15(8), e2002925. https://doi.org/10.1371/journal.pbio.2002925

Denver Museum of Nature and Science. (n.d.). Browse Selected Collections. *LUNA Imaging.* Retrieved November 14, 2021, from https://dmns.lunaimaging.com/luna/servlet/

FuTRES. (n.d.). https://futres.org

GBIF. (n.d.). Extending, Enriching and Integrating Data. Digital/Extended Specimen. https://discourse.gbif.org/t/extending-enriching-and-integrating-data/2421/33 https://discourse.gbif.org/c/digital-extended-specimen/21

GEOME. (n.d.). Genomic Observatories MetaDatabase (GEOME). https://geome-db.org

GOFAIR. (n.d.). FAIR Principles. Retrieved November 5, 2021, from https://www.go-fair.org/fair-principles/

Gramling, C. (2021, June 5). Climate-friendly archea don't pass gas. *Science News.* 199 (10), 12. https://www.sciencenews.org/article/climate-microbes-recycle-carbon-methane-archaea-microbiology

Guralnick, R. P., Cellinese, N., Deck, J., Pyle, R. L., Kunze, J., Penev, L., Walls, R., Hagedorn, G., Agosti, D., Wieczorek, J., Catapano, T. & Page, R. (2015). Community Next Steps for Making Globally Unique Identifiers Work for Biocollections Data. *ZooKeys* (494), 133–154. https://doi.org/10.3897/zookeys.494.9352

hardistyar. (2020, November 19). DiSSCo/openDS. Introduction to the openDS data model. *Github*. https://github.com/DiSSCo/openDS/blob/master/data-model/data-model-intro.md

Hardisty, A.R., Addink, W, Glöckler, F, Güntsch, A, Islam, S, & Weiland, C. (2021). A choice of persistent identifier schemes for the Distributed System of Scientific Collections (DiSSCo). *Research Ideas and Outcomes* 7, e67379. https://doi.org/10.3897/rio.7.e67379

iDiGBio. (n.d.) About iDigBio. https://www.idigbio.org/about-idigbio

Islam, S., Hardisty, A., Addink, W., Weiland, C. and Glöckler, F. (2020). Incorporating RDA Outputs in the Design of a European Research Infrastructure for Natural Science Collections. *Data Science Journal*, 19(1), 50. http://doi.org/10.5334/dsj-2020-050

Jegelewicz. (2021, August 19). Other Deliverable- Basis of Record review # 11. *tdwg/material-sample.* Retrieved November 10, 2021, from https://github.com/tdwg/material-sample/issues/11

Jegelewicz. (2021, November 10). Other Identifiers - add metadata #410. *Arctos DB/arctos.* Retrieved November 10, 2021, from https://github.com/ArctosDB/arctos/issues/4101

Kwan, Wai-Yin. (2021, November 4). *Identifying connections between earlier Happy Hour topics and action items. 20211104_paleo-happy-hour* [Recording of Zoom meeting]. paleo-data. Retrieved November 30, 2021, from https://drive.google.com/file/d/12hiQ07uly7D6NUg96eXtE8PhbwdtbNah/view

Lee, J.J. (May 2021). How to keep birds safe as U.S. wind farms expand. *Science News.* 199 (9), 4-5. https://www.sciencenews.org/article/wind-turbine-farms-expand-bird-safety

Lendemer, J., Thiers, B., Monfils, A.K., Zaspel, J., Ellwood, E.R., Bentley, A., LeVan, K., Bates, J., Jennings, D., Contreras, D., Lagomarsino, L, Mabee, P., Ford, L.S., Guralnick, R., Gropp, R.E., Revelez, M., Cobb, N., Seltmann, K., & Aime, M.C. (2020). The Extended Specimen Network: A Strategy to Enhance US Biodiversity Collections, Promote Research and Education, *BioScience*, 70(1) January 2020, 23–30. https://doi.org/10.1093/biosci/biz140

mariapraetzellis. (2021, September 16). FAIR Island Project Receives NSF Funding. *TAG ARCHIVES: MACHINE-ACTIONABLE DMPS: DMP Tool Blog.* https://blog.dmptool.org/tag/machine-actionable-dmps/

McLean, B.S., Bell, K.C., Dunnum, J.L., Abrahamson, B., Colella, J.P., Deardorff, E.R., Weber, J.A., Jones, A.K., Salazar-Miralles, F. and Cook, J.A. (2016). Natural history collections-based research: progress, promise, and best practices. *Journal of mammalogy*, *97*(1), 287-297.

Miller, J. (2021). Making FAIR data for specimens accessible. *GBIF.* https://discourse.gbif.org/t/making-fair-data-for-specimens-accessible/2420

mjbuys. (June 25, 2021). Persistent Identifiers, PID Schemes. *GBIF.* https://discourse.gbif.org/t/7-persistent-identifier-pid-schemes/2664/8

Morphosource Beta (n.d.). Retrieved November 13, 2021, from https://www.morphosource.org/projects/00000C427?locale=en

Morrison, S.A., Scott A., Sillett, T.S., Funk, W.C., Ghalambor, C.K., & Rick, T.C.. (2017). Equipping the 22nd-Century Historical Ecologist, *Trends in Ecology & Evolution*, 32(8, August 2017), 578-588. ISSN 0169-5347, https://doi.org/10.1016/j.tree.2017.05.006

National Science Foundation. (2021, May 21). RCN: Sampling Nature: A Network to Enhance the Natural History Value Chain for Sustainability Science, Award Abstract # 2129268.

https://www.nsf.gov/awardsearch/showAward?AWD_ID=2129268&HistoricalAwards=false

NCBI (n.d.). BioSample. https://www.ncbi.nlm.nih.gov/biosample/

Neon. (n.d.). Good Science is Built on Good Data. The Future of Science is Open. *NSF*. https://www.neonscience.org

ORCID. (n.d.). Connecting Research and Researchers. Retrieved November 10, 2021, from https://orcid.org

Reardon, S. & Jain, S. (2020, June 23). How A Scientific Paper About a Promising COVID-19 Treatment Was Debunked. *FiveThirtyEight*. https://fivethirtyeight.com/videos/how-a-scientific-paper-about-a-promising-covid-19-treatment-was-debunked/

Rios, N. (2021, October 20). SYM07. *TDWG Conference Talk*. https://ufl.zoom.us/rec/play/OCcJQXvt-Jbs9Se8RzUPfv9ZS87yqyGantbwLO42VNycp1ynSdJlPwx8NO1sVDVCqvTgt9rGyUZZGZnB.f6roH4f6265mZjI-?startTime=1634763481000&_x_zm_rtaid=JOp-5RZyRK-rbUFwtZZwLA.1638248797670.4917c837f710d00cfb3ddf1b4e1a929a&_x_zm_rhtaid=888

Rios N., Islam S., Macklin J., Bentley A. (2021). Technical Considerations for a Transactional Model to Realize the Digital Extended Specimen. *Biodiversity Information Science and Standards.* 5, e73812. https://doi.org/10.3897/biss.5.73812

RDA: Research Data Alliance. (2021, October 7). Supporting Interdisciplinary Sample Data Discovery, Integration, and Reuse [Webinar]. https://www.rd-alliance.org/ps-interdisciplinarysampledata-October-webinar

Schindel, D. E., & Cook, J. A. (2018). The next generation of natural history collections. *PLoS Biology*, *16*(7), e2006125.

Schulman, L., Lahti, K., Piiarainen, E., Heikkinen, M., Raitio, O., & Juslen, A. (2021). The Finnish Biodiversity Information Facility as a best-practice model for biodiversity data infrastructures. *Sci Data* 8, 137. https://doi.org/10.1038/s41597-021-00919-6

Schulz, K., Hammock, J., Poelen, J. H., & Agbayani, E. (2020). Management of Biotic Interaction Data in the Encyclopedia of Life (0.1). Digital Data in Biodiversity Research, Indiana University, Bloomington, Indiana. Zenodo. https://doi.org/10.5281/zenodo.4015329

Shaffer, H.B., Fisher RN, & Davidson C. (1998). The role of natural history collections in documenting species declines. *Trends Ecol Evol*. 13(1), 27-30. doi: 10.1016/s0169-5347(97)01177-4 PMID: 21238186

Shorthouse, D. (n.d.). Transactional Mechanisms and Provenance. Digital Extended Specimen. *GBIF*. https://discourse.gbif.org/t/10-transactional-mechanisms-and-provenance/2667/41

Shorthouse, D. (2017). Proposed Extension to Darwin Core for People and their Roles in the Curation of Physical and Digital Objects. Biodiversity Information Science and Standards; Sofia (Jul 27, 2017). http://doi.org/10.3897/tdwgproceedings.1.19829

Smith, J.E. (2021, September 25). Searching for life in old "ocean of forest." *The Denver Post*. Saturday, September 25, 2021, 8C.

SPNHC Biodiversity Crisis Response Committee Webinar. (2020, October 7) [Webinar]. https://www.youtube.com/watch?v=y2CIYI13SPI

SPECIES 360. (n.d.). ZIMS for Husbandry. https://www.species360.org/products-services/zims-for-husbandry/

TDWG. (n.d.-a.). Access to Biological Collection Data (ABCD) Schema. Retrieved December 3, 202l, from https://www.tdwg.org/standards/abcd/

TDWG. (n.d.-b.). Humboldt Core. Retrieved October 25, 2021, from https://www.tdwg.org/community/osr/humboldt-core/

tdwg/material-sample. (n.d.). *Github.* Retrieved November 14, 2021, from https://github.com/tdwg/material-sample/issues

Thessen, A.E., Woodburn, M., Koureas, D., Paul, D., Conlon, M., Shorthouse, D.P. & Ramdeen, S. (2019). Proper Attribution for Curation and Maintenance of Research Collections: Metadata Recommendations of the RDA/TDWG Working Group. *Data Science Journal*, 18(1), 54. http://doi.org/10.5334/dsj-2019-054

University of California. (n.d.). EZID Identifiers made easy. https://ezid.cdlib.org

Valenzuela-Toro, A.M. & Viglino, M. (2021, September 23). How Latin American researchers suffer in science : It's time to tackle the cumulative barriers and biases faced by scientists who aren't from wealthy countries. *Nature (Nature) ISSN 1476-4687 (online)* https://doi.org/10.1038/d41586-021-02601-8

Waagmeester, A., Stupp, G., Burgstaller-Muehlbacher, S., Good, B. M., Griffith, M., Griffith, O. L., Hanspers, K., Hermjakob, H., Hudson, T.S., Hybiske, K. and Keating, S.M., Manske, M., Mayers, M., Mietchen, D., Mitraka, E ,Pico, A.R., Putman, T., Riutta, A., Queralt-Rosinach, N., Shriml, L. M., Shafee, T., Slenter, D., Stephan, R., Thronton, K., Tsu, G., Tu, R., Ul-Hasan, S., Willighagen, E., Wu, C., & Su, A. I. (2020). Science Forum: Wikidata as a knowledge graph for the life sciences. *Elife*, *9*, e52614.

Waddink. (2021, June 22). Persistent Identifiers, PID Schemes, *GBIF,* https://discourse.gbif.org/t/7-persistent-identifier-pid-schemes/2664/6

Webster, M. (2021, February 25). Extending, Enriching and Integrating Data. Digital/Extended Specimen. *GBIF*. https://discourse.gbif.org/t/extending-enriching-and-integrating-data/2421/36

Wieczorek, J., Bloom , D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., & Vieglais, D. (2012). Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7(1), e29715. https://doi.org/10.1371/journal.pone.0029715

Wyborn, L., Ramdeen, S., Lehnert, K., & Klump, J. (2020, November 9). Targeting the Bullseye of Metadata for Material Samples: Can We Define a Minimum Kernel for Transdisciplinary Interoperability? (Version 1). Zenodo. https://doi.org/10.5281/zenodo.4694740

Annotated Bibliography

Broad, W.J. (2021, October 16). Taking the pulse of the ocean's comb jellies. *The Denver Post.* 8C.

Using a new technique for comparing environmental DNA sampling to comb jelly sequences, scientists are able to get a much better idea about numbers and species for these extremely delicate ocean creatures. Steven H.D. Haddock, monteray bay aquarium research institute "said the advance will give 'much needed precision' for biologists seeking to learn the true dimensions of ocean life" (para 5). Environmental DNA sampling- "collects and analyzes snippets of DNA that all creatures shed in their environment" (para 8). These same samples can be used for multiple studies.

Buehler, J. (2021, August 14). Lost blue butterfly was its own species. Urban development drove California insect to extinction. *Science News.* 200(3), 14. https://www.sciencenews.org/article/xerxes-blue-butterfly-first-human-caused-us-insect-extinction

Analysis of genomic and mitochondrial DNA from Field Museum specimen verified that this butterfly that was only found on the San Francisco Peninsula and lost by the early 1940's was its own species instead of just an isolated population. A "Specimen's true utility may not be clear for many years" when techniques change.

Carroll, S.R., Garba, I., Figueroa-Rodríguez, O.L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J.D.,

Anderson, J. and Hudson, M., 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal*, 19(1), 43. http://doi.org/10.5334/dsj-2020-043

Adding CARE to FAIR principles in saving of data. CARE stands for Collective Benefit, Authority to Control, Responsibility, and Ethics, and emphasizes innovation, self-governance, and self-determination. Keeping these principles in mind will prevent exploitation and allow for usage based on needs and desires of those most affected by the data sharing. This is a very important paper as the FAIR concepts are concerned with increasing access and interoperability important to open science, but do not address concerns about sensitive data.

Colella, J. P., Stephens, R. B., Campbell, M. L., Kohli, B. A., Parsons, D. J., & Mclean, B. S. (2021). The open-specimen movement. *BioScience*, *71*(4), 405-414.

Specimen Archival requirements do not usually include requirements for deposition of the specimen as a voucher. This article discusses that specimen derived data can often be reproduced by going back to the specimen, but loss of the specimen cuts off many lines of research as well as reproduction of results and checking information about the specimen. They advocate for requirements on the part of editors and data management plans for proper specimen deposition and money for continued management. Museums may be caretakers of the specimens and the management of data records according to FAIR data practices and community standards, allowing for continuing usage of specimens and data.

Damerow, J.E., Varadharajan, C., Boye, K., Brodie, E.L., Burrus, M., Chadwick, K.D., Crystal-Ornelas, R., Elbashandy, H., Alves, R.J.E., Ely, K.S., Goldman, A.E., Haberman, T., Hendrix, V., Kakalia, Z., Kemner, K.M., Kersting, A.B., Merino, N., O'Brien, F., Perzan, Z., Robles, E., Sorensen, P., Stegen, J.C., Walls, R.L., Weisenhorn, P., Zavarin, M. and Agarwal, D. (2021). Sample Identifiers and Metadata to Support Data Management and Reuse in Multidisciplinary Ecosystem Sciences. *Data Science Journal*, 20(1), 11. http://doi.org/10.5334/dsj-2021-011

Integration of data from large studies using physical samples is complicated by separation of the samples and related/derived data that can often result in loss of metadata needed for drawing appropriate conclusions. The authors discuss how unifying the process for datasharing will allow for collaboration across disciplines. They compared different solutions to allow for mapping and identification of common core metadata. They advocate continuing to use community feedback during development. PIDs are "arguably essential for supporting data synthesis" (Results, para 1). So, they want to provide tool to make it easy to assign PIDs and input other core metadata in a machine readable structure, as well as definition of common vocabulary and terms.

GBIF. (n.d.). Extending, Enriching and Integrating Data. Digital/Extended Specimen. https://discourse.gbif.org/t/extending-enriching-and-integrating-data/2421/33 https://discourse.gbif.org/c/digital-extended-specimen/21

This broad ranging discourse sponsored by GBIF includes subtopics about Persistent identifier (PID) schemes and Structure and responsibilities of a #digextspecimen as well

as an interesting section on meeting legal, ethical, sensitive data and one on Transactional mechanisms and provenance, which is rather fascinating, but a bit too technical for this paper.

Guralnick, R. P., Cellinese, N., Deck, J., Pyle, R. L., Kunze, J., Penev, L., Walls, R., Hagedorn, G., Agosti, D., Wieczorek, J., Catapano, T. & Page, R. (2015). Community Next Steps for Making Globally Unique Identifiers Work for Biocollections Data. *ZooKeys* (494), 133–154. https://doi.org/10.3897/zookeys.494.9352

The authors explore the issues around storage of data associated with biocollections. The authors advocate assigning a GUID as soon as data is generated and saving legacy ids by linking to an identifier that can be saved so that the current location can be found. However, "solutions must support scientists' current practices and create minimal burden during the collecting process." (Roadblocks, para 1). The paper points out the importance of making data about the collection and protocols available. "Curators should register their collections in GRBio and specify the adopted identifier scheme for the collection" (p. 143). At least, local collections need to consolidate and clarify their practices. The community needs to discuss schemas, try to maintain human and computer friendly identifiers for management and cross linking.

Hardisty, A.R., Addink, W., Glöckler, F., Güntsch, A., Islam, S., & Weiland, C. (2021). A choice of persistent identifier schemes for the Distributed System of Scientific Collections (DiSSCo). *Research Ideas and Outcomes* 7, e67379. https://doi.org/10.3897/rio.7.e67379

Persistent identifiers are 'one of the mechanisms for digitally transforming collections-based science" (p. 1). In this paper, the authors discussed many different approaches to

assigning identifiers and supported the decision for DiSSCo to "adopt a 'driven by DOI' persistent identifier (PID) scheme customized with natural sciences community characteristics" (p. 2). This will enable normalization and linking of data, since 'if everything has a resolvable id, everything is findable'.

Islam, S., Hardisty, A., Addink, W., Weiland, C. and Glöckler, F., 2020. Incorporating RDA Outputs in the Design of a European Research Infrastructure for Natural Science Collections. *Data Science Journal*, 19(1), 50. http://doi.org/10.5334/dsj-2020-050

European Research Infrastructure for Scientific Collections uses Digital Object Architecture and RDA supporting documents for data lifecycle. RDA allows for the data to be shared to multiple repositories when "different operations are performed in multiple contexts" (Metadata Attribution and Use of PROV entities, para 5) of APIs, systems, standards, etc..

Jegelewicz. (Aug 19, 2021). Other Deliverable- Basis of Record review # 11. https://github.com/tdwg/material-sample/issues/11

Basis of Record started out as a way to differentiate specimen data from observation data and is a subtype of Dublin Core dcterms:type using controlled vocabulary dwc:type namespace for classes. This review is a discussion about the term, ways to change this standard and how it is used to shared Linked Open Data. Looking at the history of the term led to ways to change it for RDF. Contributors discussed ways to structure the data, leading to a detailed but difficult discussion, including about how to envision relational databases.

"A radical departure of this sort couldn't happen overnight. It would need to live in parallel with the existing flattened record star schema approach to allow content providers to transition over, and folks like GBIF/iDigBio would need to "flatten" the new system to the old method to aggregate content. But eventually (as measured by usage statistics), the star schema approach would attenuate." Deepref sep 12

"My solution was/is that a Digital Specimen fundamentally is a "bag of links", very much like the Organism entity discussed here." Jbsatgen Sep 30

https://github.com/tdwg/material-sample/issues/11#issuecomment-94197480

Lendemer, J., Thiers, B., Monfils, A.K., Zaspel, J., Ellwood, E.R., Bentley, A., LeVan, K., Bates, J., Jennings, D., Contreras, D., Lagomarsino, L, Mabee, P., Ford, L.S., Guralnick, R., Gropp, R.E., Revelez, M., Cobb, N., Seltmann, K., & Aime, M.C. (2020). The Extended Specimen Network: A Strategy to Enhance US Biodiversity Collections, Promote Research and Education, *BioScience*, 70(1) January 2020: 23–30, https://doi.org/10.1093/biosci/biz140

This paper explores the concept of the extended specimen and how to share data to create a global network. This will maximize the benefit of information about specimens and research related to them. It requires a huge effort to digitize specimen data to allow for linkages between types of data and allow for user interfaces that promote discovery.

"Biological collections comprise the most comprehensive record of life on Earth; their potential will only be fully realized when the data contained within them are revealed and made more accessible for computational analyses" (Index US biodiversity collections and

their holdings, para 1). The paper continues to explore various challenges for tackling the development of this network.

McLean, B.S., Bell, K.C., Dunnum, J.L., Abrahamson, B., Colella, J.P., Deardorff, E.R., Weber, J.A., Jones, A.K., Salazar-Miralles, F. and Cook, J.A. (2016). Natural history collections-based research: progress, promise, and best practices. *Journal of mammalogy*, *97*(1), 287-297.

Museum based Natural history collections are vital to research on evolution, biogeography, habitat interaction, climate change, and many future research projects that may not be anticipated yet. Depositing 'voucher' specimens into museum collections enables the storage of current information about the specimen and a lot of future uses based on analysis of DNA, morphology, isotopes, etc.. Museum collection managers maintain records of loans and research associated with the loans which can help document, link, differentiate, and replicate research. Supporting quotes: "As primary archives of biogeographical data, NHCs permit rigorous analysis of changing mammalian distributions through both space and time in response to climatic and other permutations" (p. 289).

This paper points to the importance of using specimens to link together data from many different online databases. The museum's usage of a database system allows for the access and findability of data concerning the specimen.

Miller, J. (2021). Making FAIR data for specimens accessible. *GBIF.* https://discourse.gbif.org/t/making-fair-data-for-specimens-accessible/2420

Data needs to be more widely available, conforming to FAIR data principles. The concept of extended specimens leads to a theory of an "extended specimen network." The data must be findable and reusable not just by humans, but also by machines, as the digital surrogate of a physical specimen may be manipulated and reused in its own right. Sharing using FAIR data principles allows for increased opportunities for use by different stakeholders.

Morrison, S.A., Scott A., Sillett, T.S., Funk, W.C., Ghalambor, C.K., & Rick, T.C.. (2017). Equipping the 22nd-Century Historical Ecologist, *Trends in Ecology & Evolution*, 32, (8, August 2017), 578-588. ISSN 0169-5347 https://doi.org/10.1016/j.tree.2017.05.006

When anticipating what types of data will be useful in the future, the researcher needs to be broad and holistic. Basically, everything will need to be saved as there are already applications to do ecosystem wide studies. The analytic tools will only get better in the future. Underinvestment in data curation creates 'gaps in data,' which will limit the ability to understand current conditions and interpret ecosystems effects of climate change and other factors. Crowd sourcing is one possible way to expand efforts.

National Science Foundation. (2020, October 15). Collaborative Research: Frameworks: Internet of Samples: Toward an Interdisciplinary Cyberinfrastructure for Material Samples, Award Abstract # 2004642. https://nsf.gov/awardsearch/showAward?AWD_ID=2004642

Development of a system of iSamples for normalization of metadata, registration and other tasks.

Neon. (n.d.). Good Science is Built on Good Data. The Future of Science is Open. *NSF*. https://www.neonscience.org

This is an open science ecological initiative in which comprehensive datasets of geographic regions of the U.S. are obtained using multiple approaches to get exhaustive datasets of all interactions in the ecosystem, including pictures, soil samples, environmental scans. The data is then saved and stored so that data relationships are built and maintained.

Rios N., Islam S., Macklin J., Bentley A. (2021). Technical Considerations for a Transactional Model to Realize the Digital Extended Specimen. *Biodiversity Information Science and Standards.* 5, e73812. https://doi.org/10.3897/biss.5.73812

The current system has resulted in unprecedented sharing, but it is not sufficient for the Digital Extended Specimen concept. "One solution may involve changing the biodiversity data publication paradigm to one based on the atomized transactions relevant to each individual data record." (para 3). This would allow for community management of data, including annotations and attribution for work done and annotations added. Each data item then becomes a new item with each annotation.

Actual talk: Publishing dataset through IPT is easy to do, yet every time updated, there is an additional copy of the dataset. Citations point to the raw dataset, not the filtered data that was actually used. There needs to be a network to push annotations, sucked back into the data providers. Rios discusses various technical implementations, including a transaction 'tree' which tracks the chain of objects. He questions what unique ids are

needed and how to match with existing ones. The team is developing an exemplar roadmap and PID roadmap this year.

Schulz, K., Hammock, J., Poelen, J. H., & Agbayani, E. (2020). Management of Biotic Interaction Data in the Encyclopedia of Life (0.1). Digital Data in Biodiversity Research, Indiana University, Bloomington, Indiana. Zenodo. https://doi.org/10.5281/zenodo.4015329

"To support efficient feedback about potential data problems, validation failures result in the creation of new refuted interaction records that challenge the original claim. This EOL curated dataset of refuted interaction records is then indexed by GloBI, so other GloBI users have the option to identify data records that are incorrect or controversial." GloBI gets its data from diverse sources, scientific literature, databases, citizen science projects, text mining, museum specimen data, and checks for redundancies or erroneous data using validation rules. It furnishes the information using a system based on a Darwin Core Extension.

Shorthouse, D. (n.d.). Transactional Mechanisms and Provenance. Digital Extended Specimen. *GBIF*. https://discourse.gbif.org/t/10-transactional-mechanisms-and-provenance/2667/41

"I expect we'll have a transition period where some providers will continue to publish data via eg traditional Integrated Publishing Toolkits whereas others will use whatever infrastructure will support DES."

Valenzuela-Toro, A.M. & Viglino, M. (2021, September 23). How Latin American researchers suffer in science : It's time to tackle the cumulative barriers and biases faced by scientists who aren't from wealthy countries. *Nature (Nature)* ISSN 1476-4687 (online) https://doi.org/10.1038/d41586-021-02601-8

Researchers in poorer countries face many barriers, including language, gender, access to meetings and access to journals guarded by paywalls.

Waddink. (2021, June 22). Persistent Identifiers, PID Schemes, *GBIF,* https://discourse.gbif.org/t/7-persistent-identifier-pid-schemes/2664/6

Costs of Persistent identifiers are directly related to cost of FAIR data as they are integral. Not having FAIR data may have a much higher cost. Sharing costs related to implementation and maintenance of persistent ids could make them minor compared to the 'economic benefits.'

Wieczorek, J., Bloom , D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., & Vieglais, D. (2012). Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE* 7(1), e29715. https://doi.org/10.1371/journal.pone.0029715

Darwin Core is a low barrier method of creating interoperable data. Community groups are continually evaluating fields (using consensus discussions) and creating extensions for different fields. It uses a variety of encoding schemes. Once Darwin Core fields are populated, they can be transformed into Darwin Core Archives, a "combination of CSV files and a simple XML document describing the semantics of the data file columns and their relationships to each other" (Methods, Implementation Guidelines, para 2). "A normalized database structure was the primary technique for enforcing data integrity"

(Results, Development and Ratification as a Standard, para 2). Originally, collection management system emphasis was on data management rather than exchange. DC provides mappings to outdated terms and the ABCD (Access to Biological Collections Data) model, which is more highly structured, rather than flexible and adaptable, goals of Darwin Core. Different extensions and cores are added with community input. The semantic web is difficult to traverse as it is not a well defined ontology or defined relationships between terms.

Additional Sources

Addink, W. (2021, March 5). Structure and Responsibilities of a #digextspecimen. *GBIF*. https://discourse.gbif.org/t/structure-and-responsibilities-of-a-digextspecimen/2533/22

Visualization of a converged Digital extended specimen as interpreted by Addink

Arctos: Collaborative Collection Management Solution. (n.d.). Documentation for code table coll_other_id_type. Retrieved October 26, 2021, from https://arctos.database.museum/info/ctDocumentation.cfm?table=ctcoll_other_id_type#organism_id

Code table for other id numbers that are allowed for linking in the other id field of Arctos.

Bandrowski A. E., & Martone, M.E. (2016). RRIDs: A Simple Step toward Improving Reproducibility through Rigor and Transparency of Experimental Methods. Neuron *90*, (May 4, 2016) 434-436. http://dx.doi.org/10.1016/j.neuron.2016.04.030

Enabling transparency and authentication of scientific results requires key findings as well as reagents need to be findable and resolvable using identifiers. Research Resource Identifiers (RRIDs) help researchers to understand and verify results, as well as allowing for reproduction of studies, if needed, since the important reagents used in the study are identifiable and findable. Using the original name or lab designation usually will not allow resolvability without extensive querying of authors and other providers of reagents,

which often does not resolve adequately. RRIDs require assignment of a RRID prefix, unique number, and sufficient metadata. The strings are assigned by community database aggregators. Obtaining an RRID is “fairly simple” (p. 435), and creators of new resources can obtain an RRID by depositing it and registering information about it. RRIDs are unique and stable, allowing for resolvability in spite of company changes, mergers, shifts in names of products, etc.. RRIDs also allow for linkage between studies using the same reagents.

Baskauf, S. (2021). *Having your cake and eating it too: JSON-LD as an RDF serialization format* (Version 1). figshare. https://doi.org/10.6084/m9.figshare.16823473.v1 (['https://doi.org/10.3897/biss.5.74266'])

JSON- LD bypasses developers’ unfamiliarity with RDF. Used in IIIF, similar one used in Audubon Core. “serialize data in a manner that is both easily consumed by conventional applications, but which also can be seamlessly loaded as RDF into triplestores or other linked data applications” (para 1).

Bentley, A. (n.d.). Structure and Responsibilities of a #digextspecimen. *GBIF.* https://discourse.gbif.org/t/structure-and-responsibilities-of-a-digextspecimen/2533/11

How does this fit into the existing systems of data publishing and use? Visualization starting with data providers as roots, trunk as the data stores, aggregators branches, and end users as leaves.

Biodiversity Collections Network. (n.d.). *Resources.* https://bcon.aibs.org/resources/

Archive of materials developed though the Network. A resource for development of a "sustainable, networked community of practice (para 1)."

Bionomia. (n.d.). https://bionomia.net

Tool for use for linking specimens to workers in the field. Uses Orc Id for living researchers and wikidata ids for dead ones. Allows for community review and annotation.

Black, R. (2020, October 16). *A T. Rex Sold for $31.8 Million, and Paleontologists Are Worried.* Smithsonian Magazine. https://www.smithsonianmag.com/science-nature/t-rex-sold-318-million-and-paleontologists-are-worried-180976071/

Specimens selling for so much is a threat to research as scientists may not be able to gain access. Also, this includes poaching of sites, and exploitation by private landowners.

BoldSystems. (n.d.). Barcode of Life Data System. Advancing biodiversity science through DNA-based species identification. http://www.boldsystems.org

Canadian storage and analysis platform includes storage and an analysis workbench.

Bushbom, J. (n.d.). Structure and Responsibilities of a #digextspecimen. *GBIF.* https://discourse.gbif.org/t/structure-and-responsibilities-of-a-digextspecimen/2533/30

Visualization of data flow from provider, Registration Agency, Aggregators, to users, and the backbone of development of a digital extended specimen.

Buys, M. & Lehnert, K. (2021). Bringing together communities: IGSN and DataCite. *DataCite Blog*. https://doi.org/10.5438/thhf-kx17

Roadmap towards a partnership to use physical sample identifiers. DataCite has a vision to connect research and IGSN a central registration system for globally unique persistent physical sample identifiers.

Chamber, J. (2021, September 25). Racist legacies lurk in common names. In the scientific community, support grows for renaming animals. *Science News,* 200 (6), 12.

Names of animals can create barriers for communities and "enshrine harmful legacies." Birder and registered citizen of the Cherokee Nation Steve Hampton talks about seeing the Scott's oriole often when he lived in California. The name "holds a violent history" since Scott was a commander who drove native Americans from their land in the Trail of Tears. "Now that [Hampton] lives in Washington State, which is outside the bird's range, 'I'm kind of relieved,' he says.

Davies, N., Deck, J., Kansa, E. C., Kansa, S. W., Kunze, J., Meyer, C., Orrell, T., Ramdeen, S., Snyder, R., Vieglais, D, Walls, R.L. & Lehnert, K. (2021). Internet of Samples (iSamples): Toward an interdisciplinary cyberinfrastructure for material samples. *GigaScience*, *10*(5), giab028. https://doi.org/10.1093/gigascience/giab028

Defining common metadata requirements in search of a common core.

Deck J, Gaither MR, Ewing R, Bird CE, Davies N, Meyer C, Riginos, C., Toonen, RJ, Crandall, E.D. (2017). The Genomic Observatories Metadatabase (GeOMe): A new repository for

field and sampling event metadata associated with genetic samples. *PLoS Biol*, 15(8), e2002925. https://doi.org/10.1371/journal.pbio.2002925

How do natural history museum collections fit into the iSamples model?

Engelbrecht, I. & Steyn, H. (2021). Does TDWG Need an API Design Guideline? Biodiversity Information Science and Standards (5, e75372). https://doi.org/10.3897/biss.5.75372

"APIs are therefore a valuable mechanism for making biodiversity data FAIR (findable, accessible, interoperable, reusable)." Para 1. Using Http commands, search a URL and return data as JSON, which can be converted to program analysis software or common data analysis tools like Excel and Open Refine. There is currently a variety of URL and data formats, and inconsistency of application of Darwin Core standards.

European Commission. (n.d.). Research and Innovation. Open Research Europe. https://open-research-europe.ec.europa.eu/

Publication platform that includes citations to all supporting data and materials, enabling reanalyses, replication and reuse. So, this allows for rapid publication and open peer review, requires open access and FAIR data automatically.

FuTRES. (n.d.). https://futres.org

Functional Trait database that provides workflow for storing functional trait data at the specimen level. The database serves the data, the goal is to allow sharing of data at multiple levels without data loss using "a semantic model and is powered by extensible parsers, a backend database, and an API".

GEOME. (n.d.). Genomic Observatories MetaDatabase (GEOME). https://geome-db.org

Helps users to ensure that metadata is FAIR, improve quality and comply with standards.

Groom, Q., Dillen, M., Hardy, H., Phillips, S., Willemse, L., & Wu, Z. (2019). Improved standardization of transcribed digital specimen data. *Database* Vol. 2019: article ID baz129; doi:10.1093/database/baz129 https://academic.oup.com/database/article/doi/10.1093/database/baz129/5670756?login=true

Most data in paper form needs to be transcribed and managed digitally. Standards organizations need to normalize information used for metadata such as identification, geography, collector numbers, etc. to allow for interoperability between datasets.

hardistyar. (2020, November 19). DiSSCo/openDS. Introduction to the openDS data model. *Github*. https://github.com/DiSSCo/openDS/blob/master/data-model/data-model-intro.md

Schematic of RDF structure for digital specimen attribution data.

iDiGBio. (n.d.). About iDigBio. https://www.idigbio.org/about-idigbio

National resource for advancing digitization of biodiversity collections (ADBC). Data and images are being made available in electronic form.

ICPSR. (n.d.). Sharing Data to Advance Science. https://www.icpsr.umich.edu/web/pages/

International leader in data stewardship. Maintains national collections.

Krimmel, E., Little, H., Karim, T., Levitt-Bussian, C., Paul, D. (n.d.). Data Standards and how they are used for Georeferencing. Retrieved March 31, 2021, from https://drive.google.com/file/d/1hR7mvlG28qGdGdgu_zAj_F3eu51n50h_/view?usp=sharing

Data Standards allow for sharing of data so that it can be understood by multiple stakeholders. This slideshow explores data standards used for georeferencing, especially in Paleo and Earth Sciences.

Lee, J.J. (May 2021). How to keep birds safe as U.S. wind farms expand. *Science News.* 199 (9), 4-5. https://www.sciencenews.org/article/wind-turbine-farms-expand-bird-safety

Using citizen science, plotting prevalence of eagles, whooping cranes, and other animals in the relevant areas and correlating with other known information.

Levinson, M.A, Niestroy, J., Al Manir, S., Fairchild, K, Lake, D.E., Moorman, J.R., Clark, T. (2021). FAIRSCAPE: a Framework for FAIR and Reproducible Biomedical Analytics. *Neuroinformatics*. bioRxiv preprint. Retrieved Sept 2, 2021, from https://doi.org/10.1101/2020.08.10.244947

Current research involves increasingly complex and enormous datasets and computational methods, making a user friendly 'framework' helpful for storage of data about methods and results so that they can be easily accessed and understood. FAIRSCAPE is a complex but openly documented approach to providing a 'digital commons environment' that generates URI associated Evidence Graphs for result metadata that points to all of the needed information stored with appropriate persistent ids and standardization. This supports understanding and scientific verification of

reproducibility of results. Future enabling of persistent id registration services such as DOI will allow for longer term maintenance of access.

Lowndes, J.S., Best, B.D., Scarborough, C., Afflerbach, J.C., Frazier, M.R., O'Hara, C.C., Jiang, N., and Halpern, B.S. (2017). Our Path to Better Science in Less Time Using Open Data Science Tools. *Nature Ecology and Evolution.* 1, 0160. DOI: 10.1038/s41559-017-0160.

Standardizing methods and protocols allows for science and data manipulation to be more easily reproduced. It also allows the original research to be evaluated more easily once the challenge of teaching methods to the researchers is undertaken. Thus, this group was able to implement various tools such as Open Refine, RStudio, Git, etc.. to allow scientists in large ecological studies to communicate more easily and share their findings in a way that allows for versioning, collaboration, reproduction of the results and easy comparison over time.

mariapraetzellis. (2021, September 16). FAIR Island Project Receives NSF Funding. *TAG ARCHIVES: MACHINE-ACTIONABLE DMPS: DMP Tool Blog.* https://blog.dmptool.org/tag/machine-actionable-dmps/

"CDL, BIDS, and the University of California Natural Reserve System will work together to build an integrated system for linking research data to their associated publications via PIDs." One goal is to incorporate these policies into a templated data management plan. The DMP tool will supply Machine readable data management plans, foster connections, consistency (Dina Palto). The goal is connecting research assets /

data throughout lifecycle. The comment mentioned that use of the DMP tool is not just at beginning but should help to manage throughout data lifecycle. Researchers can be provided with training, so Data is Reusable and approaching FAIR (Margaret Levenstein).

Mayfield T., Campbell M., Hildebrandt K., Cicero C., McDonald D., Cook J., Demboski J. (2018). Establishment of the ARCTOS-GGBN Data Pipeline. *Biodiversity Information Science and Standards 2,* e25525. https://doi.org/10.3897/biss.2.25525

The Arctos collection management system has established a collaboration with GGBN to share data from tissue sample and sequencing data. This required standardizing data structures and controlled vocabulary in Arctos.

Miller, J., Agosti, D., Guidoti, M., Quiroz, F. A. R. (2021). Linking and the Role of the Material Citation. Biodiversity Information Science and Standards 5, e75543. https://doi.org/10.3897/biss.5.75543

Citing specimens used to describe new species requires a return to source material. Records of some occurrences are only referenced as Material Citations. (para 1). It is optimal to "track all material citations across the taxonomic history of a species." In addition, this needs to link bidirectionally to sequences and databases of collections.

mjbuys. (2021, June 25). Persistent Identifiers, PID Schemes. *GBIF.* https://discourse.gbif.org/t/7-persistent-identifier-pid-schemes/2664/8

DataCite is working on supporting research by "connecting knowledge," and is interested in collaboration with IGSN to mint new identifiers, providing "infrastructure services to

support increasing registration, resolution, and discovery" of identifiers. It is important that they anticipate being able to scale up to support large volumes.

Mortensen, H.M., Senn, J., Levey, T., Langley, P. & Williams, A.J. (2021). The 2021 update of the EPA's adverse outcome pathway database. *Sci Data* 8**,** 169. https://doi.org/10.1038/s41597-021-00962-3

Automated data pulls, integration of data, better web-user interface will allow for better usage of this database for evaluating adverse outcomes of chemical exposure.

National Academies of Sciences, Engineering, and Medicine 2020. *Biological Collections: Ensuring Critical Research and Education for the 21st Century*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25592

Designing Infrastructure for storage of specimens as well as cyberinfrastructure for storage and manipulation of data.

National Science Foundation. (2021, May 21). RCN: Sampling Nature: A Network to Enhance the Natural History Value Chain for Sustainability Science, Award Abstract # 2129268. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2129268&HistoricalAwards=false

Material Samples are basic form of data about the natural world, but not yet FAIR. Overcoming barriers requires community standards, robust infrastructure, training and education involving multiple groups of stakeholders.

NCBI (n.d.). BioSample. https://www.ncbi.nlm.nih.gov/biosample/

The BioSample database contains descriptions of biological source materials used in experimental assays. This also includes links to GenBank and other sources.

Norris, C., & Butts, S. (2014). Opinion: Let Your Data Run Free? The Challenge of Data Redaction in Paleontological Collections. *Collection Forum.* 28 (1-2), 113-118.

Norris argues that more data should be open since it is difficult to determine who should be granted access. This is a particularly difficult burden for the collection manager.

Penev, L., Koureas, D., Groom, Q., Lanfear, J., Agosti, D., Casino, A., Miller, J., Arvanitidis, C., Cochrane, G., Barov, B., Hobern, D., Banki, O., Addink, W., Kõljalg, U., Ruch, P., Copas, K., Mergen, P., Güntsch, A., Benichou, L. & Benito Gonzalez Lopez, J. (2021). Towards Interlinked FAIR Biodiversity Knowledge: The BiCIKL perspective. *Biodiversity Information Science and Standards*. (5), e74233. https://doi.org/10.3897/biss.5.74233

BiCIKL is a European initiative to build a community of stakeholders in biodiversity and life sciences. Research infrastructures "will solidify open science practices by providing access to data, tools and services at each stage of, and along the entire biodiversity research and data life cycle (specimens, sequences, taxon names, analytics, publications, biodiversity knowledge graph)." Shared, common or interoperable domain standards. Semantic publishing and access by liberating sub article data. "Data linkages may be realised with different technologies (e.g., data warehousing, linking between FAIR Data Objects, Linked Open Data)" (para 2). The FAIR data place will be a tool for searching

across different domains. Following will be a biodiversity knowledge hub for access to tools services and workflows.

Reardon, S. & Jain, S. (2020, June 23). How A Scientific Paper About a Promising COVID-19 Treatment Was Debunked. *FiveThirtyEight*. https://fivethirtyeight.com/videos/how-a-scientific-paper-about-a-promising-covid-19-treatment-was-debunked/

A study based on a dataset was publicized immediately because of its critical relevance to important societal issues (the need for Covid treatments). When concerns came forward, the dataset got reviewed. This turned out to be fraudulent. Peer review was lax due to not reviewing the database (can't share the database due to privacy issues, but still should share statistics). More caution needs to be used in the future, and people need to pay more attention to errors identified early by critics.

RDA: Research Data Alliance. (2021, October 7). Supporting Interdisciplinary Sample Data Discovery, Integration, and Reuse. [Webinar] https://www.rd-alliance.org/ps-interdisciplinarysampledata-October-webinar

Use existing community practices, build on them and integrate so that they will be FAIR. There is a need to communicate with services currently being used by collection managers.

When iSamples infrastructure is tested and live, natural history collections can adopt infrastructure (primarily isamples in a box) to help make their samples discoverable through iSamples central. This would most likely be installations using Arctos, specify, k-emu, etc..

Schindel, D. E., & Cook, J. A. (2018). The next generation of natural history collections. *PLoS Biology*, *16*(7), e2006125.

In order to facilitate "interdisciplinary research into complex" biological questions, the large amounts of data being digitized currently will be more beneficial if an investment in "research collection infrastructure" is undertaken.

Rapid changes in how things are collected, preserved, and analyzed and documented leads to "unprecedented success in sharing images, data", and research results.

Participation of all major stakeholders is needed in planning this enterprise. The authors proposed Holistic Sampling, an expanded view of collecting, improved collection management, increased visibility of results, expanded scope of materials collected to enable new technologies, genetics, isotopic content, interactions and behavior.

Facilitating this requires data standards and connectivity to repositories to evolve into "integrated global enterprise" linking connected information in different repositories.

"Using and building on these research finding will require durable informatics linkages among all the data derived from that original collecting event" p.3. including images and trait data usable by multiple disciplines.

Schindel, D. E., & (IWGSC), E. S. G. of. the I. W. G. on S. C. (2020). *Economic Analyses of Federal Scientific Collections: Methods for Documenting Costs and Benefits* (Version 1). SISP Commons. https://doi.org/10.5479/si.13241612.v1

It is difficult to document benefits generated by federal collections, and this paper describes five methods for documenting this. Returns on investment depend on decisions

involving what is accessioned and services provided. They will provide more benefit if collections and collection data is findable and usable by more stakeholders. Providing more information helps usage by future generations as well as a more diverse audience for broader applications.

Schulman, L., Lahti, K., Piiarainen, E., Heikkinen, M., Raitio, O., & Juslen, A. (2021). The Finnish Biodiversity Information Facility as a best-practice model for biodiversity data infrastructures. *Sci Data* 8: 137. https://doi.org/10.1038/s41597-021-00919-6

This paper provides a lot of information about a national aggregator and the types of services provided. In addition to other services provided by CMSs and aggregators, this one has an interesting idea of introducing three data flags for identifying how reliable the data is based on the expertise of the authority. There is also a linkage to the taxonomy used by school children and the chance for education based on their ability to create their own 'collection' based on a search. The database is set up in two instances- one is less mutable and only accessible by the collection providers while the other one is accessible to many more and allows annotations.

Sequeira, A. M., O'Toole, M., Keates, T. R., McDonnell, L. H., Braun, C. D., Hoenner, X., Jaine, F.R., Johnson, I.D., Newman, P., Pye, J., Bograd, S.J., Hays, G.C., Hazen, E.L., Holland, M., Tsontos, V.M., Blight, C., Cagnacci, F., Davidson, S.C., Dettki, H., Duarte, C.M., Dunn, D.C., Eguíluz, V.M., Fedak, M., Gleiss, A.C., Hammerschlag, N., Hindell, M.Al, Holland, K., Janekovic, I., McKinzie, M.K., Muelbert, M.M., Pattiratchi, C., Rutz, C., Sims, D.W., Simmons, S.E., Townsend, B., Whoriskey, F., Woodward, B., Costa, D.P.,

Heupel, M.R., McMahon, C.R., Harcourt, R., & Weise, M. (2021). A standardisation framework for bio-logging data to advance ecological research and conservation. *Methods in Ecology and Evolution*, *12*(6), 996-1007.

Thousands of biologging datasets are not easily available. Biologging is used to give evidence of particular species at a particular place and time. So, these comprehensive datasets will help with studies of species presence and interactions for behavioral and conservation studies. Currently different sensors have data formatted according to non-standardized manufacturer provided solutions. Standardization will allow for integration of data from multiple studies. This paper suggests 3 templates for manufacturers and researchers and repositories to evaluate the original decoded data and how it was achieved ("for tag data acquisition and decoding" p. 1000), allowing for evaluation of heterogenous data, while researchers will also develop important metadata needed for further data integration. Standardizing as early as possible in the process of data production and manipulation will allow for better integration into research studies.

Shaffer, H.B., Fisher RN, & Davidson C. (1998). The role of natural history collections in documenting species declines. *Trends Ecol Evol*. 13(1), 27-30. doi: 10.1016/s0169-5347(97)01177-4 PMID: 21238186

Documentation of species declines requires knowledge of past prevalence. This can be found on a gross scale using data stored in natural history museums. Although this is often area specific and inconsistent, using this information has been helpful in many studies. Scientists can deposit specimen vouchers in museums in order to save data about resampling and other studies.

Shorthouse, D. (2017). Proposed Extension to Darwin Core for People and their Roles in the Curation of Physical and Digital Objects. Biodiversity Information Science and Standards; Sofia (2017, July 27). http://doi.org/10.3897/tdwgproceedings.1.19829

"Here, I propose a lightweight extension to Darwin Core to accommodate new terms for agent identifiers and their roles in the curation of physical and digital objects" (para 1). This involves disambiguating terms for various activities in current fields. Dealing with legacy data will allow for participation of more collections and aggregators. Dealing with this issue also increases excitement about engagement with these activities.

Smith, J.E. (2021, September 25). Searching for life in old "ocean of forest." *The Denver Post*. Saturday, September 25, 2021: 8C.

Researchers resurvey area in Columbia near where Leo Miller, in 1912, collected more than 800 specimens for Frank Chapman of the American Museum of Natural History. There have been six expeditions organized across Columbia by a team lead by Andres Cuervo (Alas, Cantos y Colores government financing). These birds will be deposited in a Columbian Museum. Dry Ice, Liquid nitrogen to flash freeze for genetic studies. The 1912 specimens only have skeletons and skins collection location, collector, and altitude. Descriptive data about the specimens being developed. There is also census data of birds and bird songs to complement data from specimens. This gives information about composition of bird life changes over 100 years. Comparison of the analysis of these two museum collections will give important data about species diversity changes over time. They will also see "how the genetic variation has shifted" (Seeholzer, And no antbirds

called, para 9) in the species collected. This will give idea about what has changed within the species.

SPNHC Biodiversity Crisis Response Committee Webinar. (2020, October 7). [Webinar] https://www.youtube.com/watch?v=y2CIYI13SPI

How can collections community contribute to a solution for protecting Biodiversity? Escalating crisis and biodiversity loss around the world. The society's goal is to support preservation of natural history collections around the world. Need 'transformative changes to allow for mitigation of extinction factors. This committee has provided "resources of use to collections to showcase how they can be used to address conservation and biodiversity crisis themes as a tool to increase advocacy for collections in general." The collections can contribute to evaluation of past, assessment of biodiversity, and effects of various factors on distribution, including climate change and human factors. It is also helpful for disease tracking and pandemic research.

TDWG. (n.d.). Humboldt Core. Retrieved October 25, 2021, from https://www.tdwg.org/community/osr/humboldt-core/

Darwin Core extension for species inventory observational data. Creating new fields or integrate into existing Darwin Core terms, to allow for integration into existing metadata schemas (or creation of new ones). "The outcome in either case will be to provide a framework and clear semantics for sharing and integrating biodiversity inventory data."

Thessen, A.E., Woodburn, M., Koureas, D., Paul, D., Conlon, M., Shorthouse, D.P. and Ramdeen, S. (2019). Proper Attribution for Curation and Maintenance of Research Collections: Metadata Recommendations of the RDA/TDWG Working Group. *Data Science Journal*, 18(1), 54. http://doi.org/10.5334/dsj-2019-054

Many collections are not maintained or curated as well as they should be. One way to improve this is to provide better attributions metadata to encourage participation. "After 18 months, this Working Group recommended the use of PROV entities and properties to link people (Agent), the curatorial actions they perform (Activity), and the digital or physical objects they are curating (Entity)" (Abstract). Using RDA, providers can also include a Role for an Agent. This effort needs a collaboration with other initiatives.

Thomer, A.K., Twidale, M.B, & Weber, N.M. (2018). Supporting the Long-term Curation and Migration of Natural History Museum Collections Databases**.** *ASIS&T Annual Meeting 2018.* 504-513.

Data base migration is a common task for data curators but is often difficult due to complicated data storage in relational databases and parsing with schemas and database software. This paper is about practices that are commonly used and gives ideas about ways to support these activities and curation of 'complex digital objects.'

Wake, D.B. & Vrendenburg, V.T. (2008). Are we in the midst of the sixth mass extinction? A view from the world of amphibians. *PNAS* August 12, 2008, 105 (Supplement 1) 11466-11473; https://doi.org/10.1073/pnas.0801921105

Worldwide assessment shows that up to one third of amphibians are having a major threat of extinction. Global warming and pathogens are having a major effect. Amphibians are in a biodiversity crisis and have been "studied intensively since scientists have become aware of their global decline."

Webster, M. (2021, February 25). Extending, Enriching and Integrating Data. Digital/Extended Specimen. *GBIF*. https://discourse.gbif.org/t/extending-enriching-and-integrating-data/2421/36

Specimen records need to be integrated with observational and other types of data. (This is an important observation that has been the impetus for the ideas around extended specimens and digital extended specimens.) Each type of data may go into appropriate repositories but needs to be linked, perhaps through a common collecting event, date and time.

Whitaker, Anna F. and Kimmig, Julien. 2020. Anthropologically introduced biases in natural history collections, with a case study on the invertebrate paleontology collections from the middle Cambrian Spence Shale Lagerstätte. *Palaeontologia Electronica*, 23(3), a58. https://doi.org/10.26879/1106 palaeo-electronica.org/content/2020/3238-collections-biases

"Duplicate records (i.e., records of the same specimen wrongfully included multiple times, often as a result of a taxonomic change creating records bearing both the old and new assignment) can artificially swell abundances of taxa or localities."

Wyborn, L., Ramdeen, S., Lehnert, K., & Klump, J. (2020, November 9). Targeting the Bullseye of Metadata for Material Samples: Can We Define a Minimum Kernel for Transdisciplinary Interoperability? (Version 1). *Zenodo*. https://doi.org/10.5281/zenodo.4694740

Physical samples need to be "globally uniquely identified, well described, and findable in online catalogues" (para 1). IGSN is expanding from the geoscience community, and it is not possible to use the same metadata for diverse fields. A "minimum set of attributes common to all samples" (para 2) would be defined as a core kernel- in order to obtain this, establish a clearinghouse for schemas so that the community can share preferred schemas and work together to define the core kernel.

Appendix A
Tables Showing Characteristics of Persistent Identifier Schemes and Structures

This appendix includes tables showing information and examples about multiple types of identifiers. Many authors have compared the various types, and it is interesting to look at similarities and differences between these types. Note that not all of the identifiers are unique or persistent. Some of the examples can no longer be resolved and some were never resolvable. Some of them might be found by going to the collection management system or database for the institution that minted them but cannot be found otherwise. Some can be redirected by going to the aggregator, such as GBIF, which has maintained the link. Others have not been maintained at all. Although there are probably reasons for choosing all of these types of identifiers, the most important factor is that there is some organization maintaining the link for a reasonable period of time and that when/if a new identifier system is chosen, a redirect for the older link is automatically made to the new location of the record.

Table A1

Supplemenal Table 1. Comparison of ARK and IGSN sample identifiers characteristics.

PID characteristics	ARK	IGSN
Community for metadata and standards around samples	no	yes
Option to use existing infrastructure for PID minting, sample data management, and metadata sharing through sample-oriented web landing pages	no	yes
Infrastructure available for minting IDs (e.g. EZID, NOID) when setting up as a naming authority	yes, only for minting	no
Use of shoulders, namespace splitting	yes	no
Use of containment and variant qualifiers	yes	yes, not currently standard practice but can be easily done
Existing resolution services	yes	yes
Cost	free to mint IDs, but there are signfiicant costs to maintain identifiers and infrastructure. If using EZID infrastructure may cost 1500 per year, if not through UC.	free to obtain IGSNs from SESAR currently; to become allocating agent there is annual membership fee, cost of building infrastructure and maintaining if allocating agent
Sustainability	support of CDL, ARKs in the Open project	actively working to improve and sustainability and business plans
Central catalog to search for samples	Not really. If using EZID- can search existing ARKs there, but no sample specific metadata.	Currently, can only search by allocating agent. SESAR is the largest allocating agent.

(Damerow et al., 2021)

Table A2

Table 1
Examples of PIDs that have been used for samples, modified from Guralnick et al, (**2015**).

IDENTIFIER TYPE	IDENTIFIER EXAMPLE	SCOPE
ARK	ark:/12148/btv1b8449691v	Flexible
URN	urn:catalog:UMMZ:Mammals:171041	Flexible
HTTP URI	*http://data.rbge.org.uk/herb/E00115694*	Flexible
DOI	10.7299/X7VQ32SJ	Flexible, mostly papers and datasets
UUID	EF0A4D3E-702F-4882-81B8-CA737AEB7B28	Flexible
IGSN	*IGSN: IECUR0002*	Geoscience, working to become general physical sample identifier
CETAF URI, based on HTTP URI	*http://data.rbge.org.uk/herb/E00421503*	Species Occurrence, Specimens from CETAF institutions
RRID	*RRID:MGI:5630441*	Biomedical Research Resources
BioSample accession number	*SAMN03983893*	Biological source materials used in experimental assays

Excel | CSV
Acronyms: ARK = Archival Resource Keys, URN = Uniform Resource Name, URI = Uniform Resource Identifier, DOI = Digital Object Identifier, UUID = Universally Unique Identifier, IGSN = International GeoSample Number, CETAF = Consortium of the European Taxonomic Facilities, RRID = Research Resource Identifier.
(Damerow, et. al., 2021)

Table A3

Table 1.

Examples of identifiers in use for biological samples in the GBIF database.

GBIF occurrence	Identifier type	Identifier	Catalog number	Collection
872747863	LSID	urn:lsid:biosci.ohio-state.edu:osuc_occurrences:OSUC__169968	OSUC 169968	C.A. Triplehorn Insect Collection
896421698	URN	urn:occurrence:Arctos:MVZ:Bird:157675:1526959	MVZ 157675	MVZ Bird Collection
784060956	URN	urn:catalog:UMMZ:Mammals:171041	UMMZ 71041	UMMZ Mammal Collection
575336458	HTTP URI	http://data.rbge.org.uk/herb/E00115694	E00115694	Royal Botanic Garden Edinburgh Herbarium
1050474791	HTTP URI	http://arctos.database.museum/guid/UAM:Ento:230092	UAM 230092	UAM Entomology Collection
1050474791	DOI	10.7299/X7VQ32SJ	UAM 230092	UAM Entomology Collection
624211191	UUID	EF0A4D3E-702F-4882-81B8-CA737AEB7B28	UF 161444	UF FLMNH Ichthyology
476850316	Darwin Core Triplet	MCZ:Mamm:8831	MCZ 8831	Museum of Comparative Zoology, Harvard University

(Guralnick et al. 2016, p. 135)

Table A4

Table 2.

Identifiers schemes according to key characteristics noted in part in Box 2.

Identifier characteristics	DataCite DOI	EZID ARK	OCLC PURL	Self-minted HTTP URI*	LSID	DwC Triplet	UUID
Globally Unique	yes	yes	yes	yes	yes	no	yes
Service Metadata Required for global uniqueness	yes	yes	yes	yes	yes	no	no
Per-identifier Cost	per id or subscription fee	yearly subscription fee	free	free	free	free	free
Identifier Issuance	registration	registration **	registration	local	local	local	local
Human-Friendly	provider dependent	provider dependent	provider dependent	provider dependent	provider dependent	high	low
Opacity	partial	partial	partial	provider dependent	provider dependent	low	high

Adoption by biodiversity informatics community	biodiversity publishing	low	low	high	low	collections community	variable
Adoption by broader informatics infrastructures	variable	low	variable	high	low	low	high
Dereferencing Service Integration	yes	yes	yes	yes	yes	no	no
Dereferencing Characteristics							
Dereferencing Type	central	central	central	distributed	distributed	N/A	N/A
Structured Identifier Responses directly from resolver ***	HTML, RDF/XML	HTML	HTML	provider dependent	yes	N/A	N/A
Redirection	yes	yes	yes	possible	possible	N/A	N/A
Clear Namespace policy and contract	yes	yes	no	no	no	N/A	N/A
Resolution service backed by institutions	yes	yes	no	provider dependent	no	****	****

^*

Self-minted HTTP URIs may include ARKs or PURLs as well

^**

ARKs have special mechanisms to extend scalability

^***

Structured metadata responses may be available after redirection, depending on the provider (e.g. dublincore.org returns RDF/XML for PURLs)

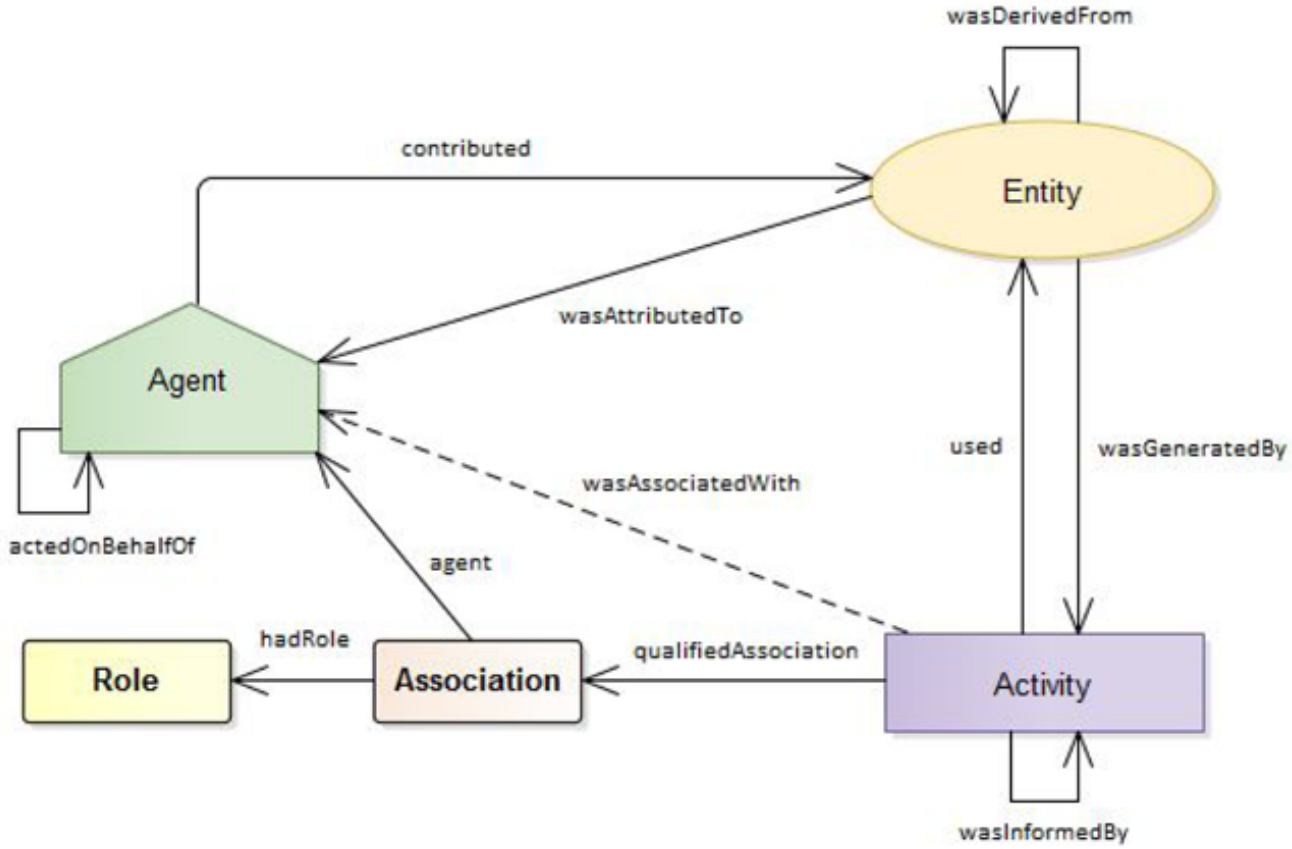
^****

Perhaps, if hosted by a general service (e.g. GrBio for Biocollections, GBIF for occurrence records, etc.)

(Guralnick et al. 2016, p. 141)

Appendix B
RDF Model for Attribution Metadata

For the information to be machine readable, it needs to be stored in a structure. In the case of RDF triplets, the information can be organized to include a relation linking subjects, such as ‘parent of,’ ‘generated by,’ and other possible linkages, which can be saved in triple stores or other types of storage. This schematic shows examples of possibilities for attribution metadata in the Digital Extended Specimen.



Simple model for attribution metadata from
harditya r, 2020

Appendix C

A Short List of Examples of Digital Objects needing PIDs

Table 1.

Categories of digital object needing identifiers.

Kind of object	Scenario of use*	PID scheme
Digital Specimen	Internal, external	Topic of the present article
Digital Collection	Internal, external	DOI
Collection Description	Internal, external	DOI
Institution / facility	Internal, external	GRID/ROR#
Loans / visits transactions	Internal only?	t.b.d.†
Annotations / interpretations	Internal, external?	t.b.d.†
Provenance events	Internal, external?	t.b.d.†
Documents	Internal, external	DOI
Persons	Internal, external	ISNI/VIAF/ORCID

* Internal to DiSSCo means PID needs to be resolvable within DiSSCo infrastructure. External to DiSSCo means PID needs to be globally and publicly resolvable.

Sometimes we may need to (internally) reference institutions that do not have a GRID/ROR, e.g., institutions that no longer exist (but their codes are still found in literature and collections), or service providers that are not research organisations.

? Exact scenarios of use need to be studied further to determine whether internal only or both internal and external resolution are necessary.

† The PID type is still to be determined (t.b.d.). Whilst still likely to be selected from one of the Handle System variants, requirements are more 'internal' than 'external' and with lower profile/importance than for Digital Specimens and Digital Collections.

(Hardisty, et al., 2021)

Appendix D
Digital Specimen Architecture Visualizations

This appendix includes various visualizations of the digital extended specimen to facilitate thought about the structure and required elements.

Figure D1

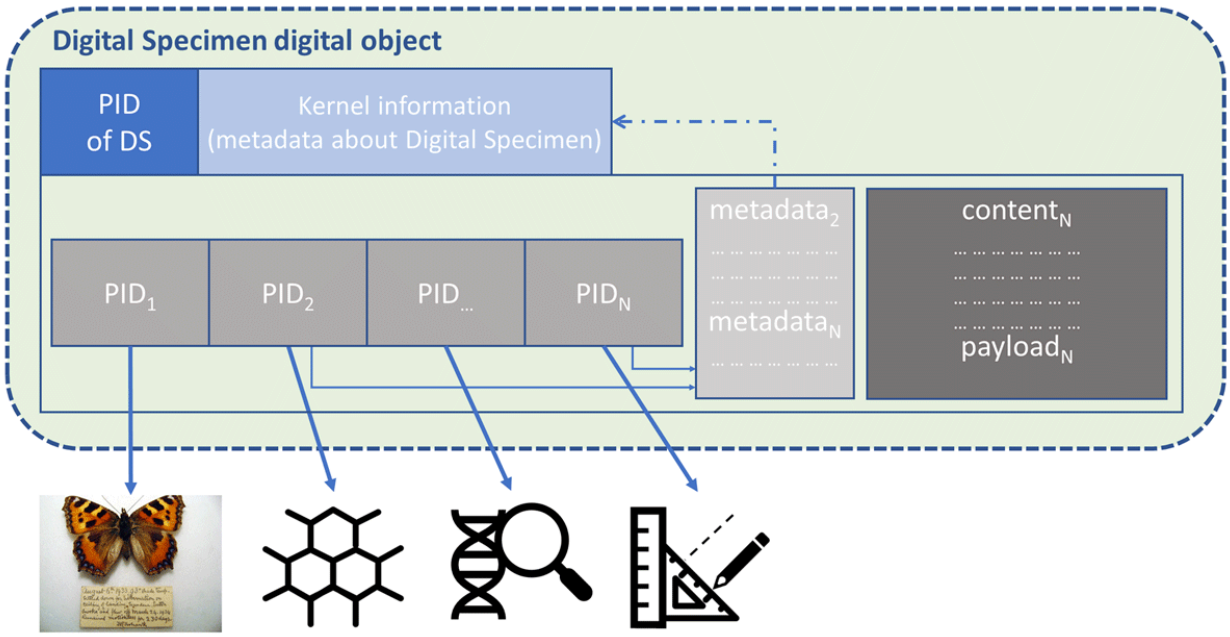
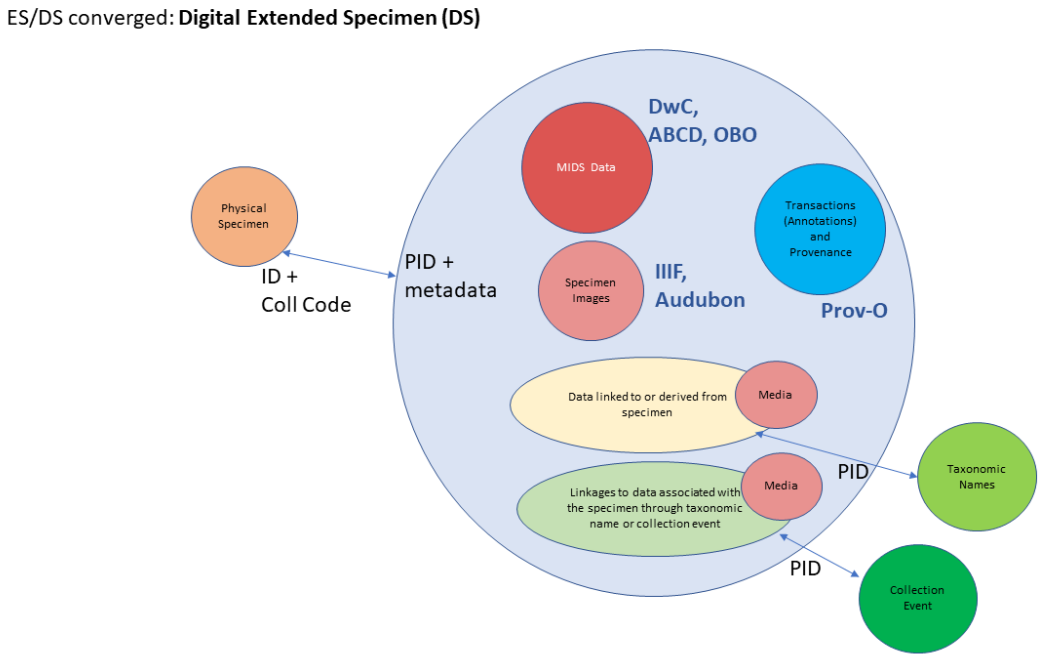


Figure 3b

Basic structure of a Digital Specimen (DS). A DS acts as a container for pointers, metadata and embedded content, i.e., information about and derived from the corresponding physical specimen including but not limited to, for example, necessary information about the specimen, image(s), molecular data, genetic sequence data, and morphological measurements.

(Islam et al., 2020, p. 6)

Figure D2



"And this is a picture how a converged Digital Extended Specimen would look like, based on my interpretation of the earlier comments in this consultation. It would look like a digital specimen with the extensions worked out in ES, with data not only linked to but also derived from specimen in the 'secondary' part, and with a different section for specimen images. All the component in the object can have their own PIDs to be linked to the DS PID. Taxonomic names and collection events would also need PIDs to be linked" (Addink, 2021)

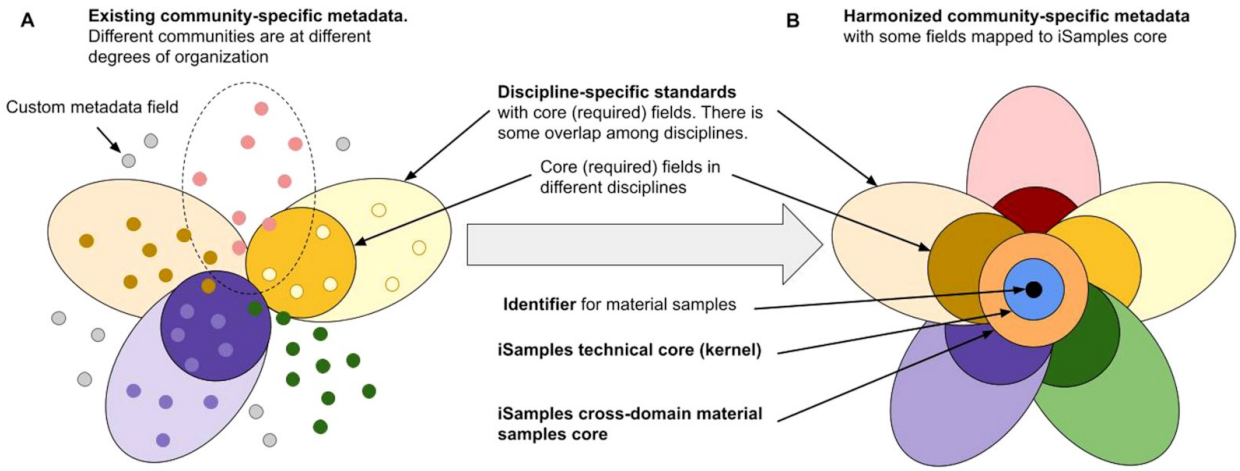
Table D3

Table 2: Simple example of PID Kernel Information for a Digital Specimen. Example PID: 123prefix/uuid-27a9edf63.

Attribute	Value Type	Example Value
Location	url	http://example-dissco-repo/uuid-27a9edf63
Created	date and time	2019-04-24T11:07:11.771Z
Type	type definition	typedef123/DigitalSpecimen
PhysicalSpecimenId	string	BMNH:1905.5.30.352

(Islam et al., 2020, p. 7)

Figure D4



Organization of metadata fields in the iSample cyberinfrastructure to facilitate interoperability of digital objects via assignment of persistent identifiers and definition of cross-domain core fields (Davies et al., 2021, p. 3)

Appendix E

Interviews

Research questions of my paper:

Persistent identifiers and digital data
Concerning scientific specimens

a) How used? FAIR data model
b) Material info for DO
c) Attribution, annotation, reproducibility, restricted data, fraud prevention

Prepared Questions:

1) What is your position? Title- Responsibilities
2) From that point of view- what issues interest you around digital asset management?
3) Digital Voucher = Digital specimen twin?
   -links in data
   -types of persistent ids
   -DOIs from datacite
   ARKS for?
   Media IDs
   ID for a specimen
   Q: who maintain persistent id for 100 years?
   Q: minimal information (metadata) for a DO?
   Darwin Core- how does that fit in?
   Q: Lately, I have noticed comments about persistent ids for PARTS?
   Why/how assign? How about deassign when mistakes are made? OR parts/specimens lost?
   OR used up/ in that case, keep persistent ids?
   Can you discuss this from your point of view?
   What is the goal of Collections personnel?
   -Sharing with researchers and the public?
   -can they implement Darwin Core?
   - In this context, what is Minimal information needed for a digital specimen for preservation and ability to use?
   -How implemented? XML, RDF
   -Containers, linkages
   Do we need more persistent ids?
   Why? Not linkable (or links disappear)
   What do you find interesting, exciting, important?
   What are uses of persistent id- attribution, locality, occurrence id, material sample?
   Preserved specimen, fossil- stratigraphy, geology
   Is it important to make it easier for primary data producer, and what is secondary ???
   Can IPTs be used for DES?
   Is the Authority for the record the originator of the record?

1) Mayfield-Meyer, Teresa- Project Coordinator, Arctos Collection Management Solution. "like you, we all have way too many jobs to do" (to Arctos Working Group, 14 Oct 21) Interview September 17, 2021

Mayfield-Meyer has been working with Arctos since 2016 when she was a collection manager digitizing collections at the University of Texas at El Paso, and during that time, the Arctos community really helped her with her work. Following that, she was hired to migrate data for a terrestrial parasite tracker grant, working for the Museum of Southwestern Biology and the New Mexico Museum of Natural History and Science. Currently, she is a Project Manager bringing new collections in to Arctos in five planning phases. 5-6 collections joined more than a year ago and did not have enough help with their data.

Persistent ids

Collections are struggling because PIDs are not easily obtained for purposes for which we want to use them. The issue is how to be unique and persistent; DataCite, a provider of registration with a DOI and metadata for research data (https://datacite.org) allows authors to obtain DOIs for publications. It is difficult to understand the difference between different DOI formats. DOIs can be reserved in groups. In order to get their own DOI for Arctos' needs, they could form a group with others to negotiate for some, starting with Arctos. It is a normal DOI, and there would have to be money to pay for this annually. One question is 'what kind of organization?' of the DOI structure. Also, you need to understand what is a digital object? One exploration of this is taking place in the TDWG (Biodiversity Standards) task group on material sample that is being convened by Teresa (tdwg/material-sample, n.d.). They are discussing controlled vocabulary around basisOfRecord, specifically, the definition of Material Sample.

[This discussion is quite complex, involving separating definitions of Organism and Material Sample as well as types of specimens and objects!].

The discussion about how many items need to be assigned persistent ids "is kind of like 'how to build a brain', and types and numbers of persistent ids explode out of control."

Some of the most important things concerning persistent ids include:

If an id is not supported anymore, there needs to be a redirect, same as for other deprecated links.

People can change GUID prefixes, for catalog numbers; they still exist, so you need to take responsibility to monitor the table somewhere.

Unique persistent ids want to be able to find one thing, and not disambiguate two things.

At the time of creating the physical catalog, the collection manager also needs to make a digital representation of what is known.

So, for GBIF ecological work, occurrence id was used for identifying what species were where and when. For example, tigers are all occurrences, and occurrence id is assigned to all parts. In the part table, each sample is appended with a part number, which makes it unique. To clearly do that for occurrence, you need to edit number of specimens for the event. Sometimes, it is difficult to understand which sample is from which time, and the way to convey this information is to assign stable part ids.

Occurrence id is defined by TDWG, its usage depends on many factors. When GBIF was set up, they had not set up occurrence id yet. In order to trust the occurrence id, you need to appreciate the new relevance of material sample id. This may require a different format of a Darwin Core Archive. Currently, the Occurrence id is overstated, and the Occurrence is primary. Material Sample is also important for the archive, as is catalog number, and using a stable part id. Archives could be packaged up differently if they have stable part ids.

At the Aggregator Level, they will use relational databases, yet Darwin Core Archive is a flat file relationally, used by GBIF. We need Arctos on a global scale to follow all the stuff needed for Extended Specimens. [The reality is- the aggregators do not used relational databases and I don't know if they ever will. The Darwin Core Archive is a flat file attempting some ways to convey relational data. I don't know if a relational aggregator (like Arctos) is the answer, but what happens at the aggregator level (GBIF, iDigBio) will probably need to be more COMPLEX if we are going to aggregate "digital extended specimens." Mayfield-Meyer, personal communication, November 12, 2021]

Superchallengers.

The number one concern is that there is a 'huge spectrum of experience' in data providers which causes difficulty of participation in digitization and discussion. The levels challenge how to communicate to all people what is happening. Therefore, the social parts are the difficult part of the upgrades. This effects efforts to discuss the usage of persistent ids. To communicate that it allows everyone to talk to one another, in different systems using the model with Darwin Core. If we want to share more well, every collection should have a collection manager, but they also need to have a data person that insures data quality, and evaluate what system is 'best quality.' In order to let users get a dataset out of GBIF, IT collection managers need to check that it is "clean and be careful." The reality is that a lot of ids aren't all that great, so there needs to be a tool that can double check their quality. Annotations by other people can help a ton. From Arctos, you may have 212 new acquisitions in UTEP records, if no one looks at them, they may not detect that nothing is there, and issues may not be addressed. Many collections can't take annotations into consideration. So, along with the explosion of persistent ids, there is a need for

collection managers to review, and decide how to respond to annotations. So, there is a Huge hold up in improving records, and we are constantly creating a backlog of less than good records. Complicating this, when reviewing papers, it is often not clear- "which data did they use?" Did they use the later "Normalized" data or the "original" data as first entered?

Also, the researchers may have done a download from 50 different collections, and my collection is cited, but "did they use any data from the collection?" Collection managers want to know how useful their data has been, but unless the research cites every record, I can't know how useful each record was. More detailed reports are needed, another set of tools, and who pays for that? All datasets downloaded from GBIF have a unique id, activity can be related to the collection, but can't go record by record.

To get into even more 'WEEDS,' there are discussions about nature of id. For example, is the collecting source wild caught, or wild? Two things are meant by that term, and you need to separate it out.

A sequence attributed to a species of mouse belongs to that species of mouse, but also to the specimen and to the part… Arctos has multiple ids, managers can say a record is linked to an id in another resource, but it is not clear how the link is made. At least you can know the history of the record by saving past links.

But, this is stuff to examine and test and analyze; it is not the job of the collection manager to say what it is- instead, you should leave it to the user to decide. The user needs to be able to bring their previous experience and to use information about the determination to make their own decision. So, the collection manager is not the authority, they are only "taking care of the information that we have." Biology is so complicated, we don't even have a good definition of species, so we take the 'huge mess and organize' so that people can find and use and make

sure that it is organized correctly. This is hard to do, but everyone wants it to be easy. It is never easy, you just always try to do your best. Some collections which are in public museums are thinking about how the public can view their information. They are making more collections more open to searching and are changing standard practice to make it usable for diverse users. When we are doing this, we need to make it known that there is always new information that may show that "we were wrong." We can never know everything, and we keep changing the answers.

"We are involved in PROCESS." Science is reaching toward, trying to get the correct answer. It is interesting to add new resources and a challenge to get to where half of us are doing it correctly.

Email 10/7/21

Other IDs are the main way to link catalog records to other stuff, especially external stuff (when that is possible), but links can also be made in roundabout ways through media, publications, projects, taxonomy, geography, locality, agents, and probably some other stuff I am forgetting! All of these "links" in Arctos provide some sort of context or relationship about the link so that there isn't just some link without "same individual as", "collected by" or "cited in" and so on. Also in almost every case we have the ability to record WHO asserted the link, when they asserted it and why.

You can see all of the Other Ids in Arctos along with definitions and base url (if there is one) in the code table (Arctos: Collaborative Collection Management Solution, n.d.a.).

[This includes DRYAD Other id for datasets stored there and Morphosource Other ID for images stored in Morphosource, with potential for editing or adding other ids as different outside resources update their policies.]

2) Krimmel, Erica- Digitization Resources Coordinator, iDigBio (national coordinating center for natural history collection digitization information, including TCNs in its umbrella). (personal communication, October 8, 2021).

Krimmel started out as an information sciences person/geologist at the Chicago Academy of Sciences, which migrated information into Arctos. It was a relatively small collection, but not 'small' in many respects. Arctos works in the community to tackle big problems. So, Arctos has a solution for many things that are being tackled in the field. Although it is a nice part of the community, the Arctos user interface was terrible, so Erica used to have her students put data into Excel.

Her evaluation of the health of the field is that it is a mess but moving in a good direction. Computer scientists can solve the problems, but not those related to collection history- they need an idea of the problem that has to be solved. The role of persistent ids is to track specimens and to link physical specimens with other information, both digital and physical. Collection managers do not talk in a way that computer scientists need it to be phrased. A key question is 'who is managing this system of ids?' How do you add interactions that need to occur between aggregators, and also local databases? How to make data resolvable from a journal article, where the catalog number may be different? The field needs to move to future numbers that allow the tech systems to move. This requires an 'interchange of sorts- an aggregator of identifiers that knows all of the types' of data and interactions needed. An important question is 'Who pays? Who maintains?, and the solution is at the local level, not in paying DOI to maintain DOIs, which is not sustainable. DiSSCo is planning to contribute to maintenance of

DOIs. It is also not sustainable to get everything for free. This was illustrated in the example of Life Science ids. Originally, they were grant funded, but when the grant ended, it was no longer maintained, so they are less reliable. Since they are not maintained, you can generate one yourself, just make one up, and there is no one to say that it already exists.

The catalog number is conceptually a persistent id, but there is a problem with implementation. There is no agreement in the community, so they leave it up to the collections, where decisions are too big of a burden to expect consistency. What you can do with technology is one thing, the social problem is still there. Persistent identifiers have two important aspects, they are unique, and resolvable and maintained. Natural History Museums are doing really well as catalog numbers are unique and resolvable. A URI is a form of identifier which is a concept of a term with a text value. The conceptual format is not as important as functionality. The main problem with identifier schemes is a social reason, 'who is monitoring the id schema?'. This is also true of ORCiD. ORCiD is only used to identify living people. It needs a narrow scope to be sustainable. Wikidata is broader and used for dead people. Wikidata is a place for identifiers, but it is not trusted [Waagmeester et. al, 2020]; people feel anxiety because it is not controlled, so it has misassociated identifiers.

Who is the authority? For physical specimens, it is often the collection manager. But Extended Specimens are distributed amongst institutions. The authority there would be whoever can be counted on to monitor it to be unique and resolvable. So, you need to control communications to make assertions open [transparent]. Then, users decide who they trust. If you want to be really certain, some authorities are more trustworthy, so noting this needs to systematically be part of the annotation.

There is value in collections where no one has a specimen other than that one. Other people could have an opinion. For example, georeferencing requires more knowledge of localities than a collection manager has; sometimes there is success in finding random people in the area who have more knowledge. There is no final answer. Collection manager personality types like to have things organized. But they have to work on increasing perfection on professional (and personal) level without needing perfection.

iDigBio accepts multiple ids, including ARKs, GUIDs, Darwin Core Triplet, Organism ID. In discussion recently- sharing currently uses Darwin Core Occurrence records as the nucleus. In the future, the Extended Specimen concept may use Organism as the nucleus. How is the community going to make the transition? This is not sure. Organizations like Arctos provide specimen management. This is confusing researcher needs for specimen occurrence, Organism ID, collection sites. There are different entry points for data access. For an example, 'How to manage data locality'? GBIF supports IPT software deployed in several instances. Some data is published through Vert Net. iDigBio provides IPTs for small collections. Arctos has more collection managers to update data through IPTs. What if Specify also had an integrated IPT? That would be ideal.

One example of a problem to do with persistent id is the data for the mineral collection at Kansas. The persistent id associated by Specify to the IPT was misaligned, and the wrong persistent id was linked to specimen records. This took three years to catch since the id was not really human readable. A human did not notice the mistake, so for three years, users got data for persistent ids that were meaningless. It was difficult to do the redirects- they used Resource Related id, but it is complicated to redirect groups of ids.

Sometimes the occurrence record is applied to the incorrect object. The collections manager needs to be a data manager. For TCNs, they get funding from NSF for multiple positions including a data managers' network. Data managers can think critically to do things more efficiently, get data in standardized format so anyone can interpret it. This is NOT TRUE NOW. Maybe many collections will have a person like Teresa [Mayfield-Meyer], whose role is shared amongst institutions. People need to be open to getting the conceptual details and use tools in use now. They also need to put in place long term plans and be welcoming and interested in new technology.

3) Neu-Yagle, Nicole- Assistant Collections Manager, Earth Sciences, Denver Museum of Nature and Science. (personal communication, October 8, 2021)

Neu-Yagle discussed persistent ids and datasharing using EMu (another collection management system).

Her department is using version 5 of EMu- they were planning to move to version 6 before the pandemic, but all spending was put on hold at that time. In version 6, many modules were changed to upgrade looks dramatically. Unlike Arctos, which has constant slow changes, EMu has dramatic changes. Digital data managers at DMNS use the multimedia tab [in EMu] to assign an id (GUIDs). EMu is navigated using the Task bar. Different modules hold the data, then you can get a list, and once you click to specimen records, there are detailed new tabs. The Admin tab shows who used EMu last and what updates they made. Security determines who can see it and the Audit tool shows all edits as they are made. GUIDs in Admin show the Field number, which is assigned by the collector, the Morphotype number, and other stuff numbers. Some of the numbers are in other modules. To explore this, you have to open the module, then search for any that can be named by other numbers, (you can program how searches work).

Some of the DMNS specimens were catalogued in two different collections, so the records are messy. You can also add multiple taxon numbers to an object. The staff have set it up so that the vertebrate taxon overrides the invertebrate, but it can have two numbers. Specimens are organized by organism, and most restrictions are on vertebrates, so, to be safe, they cataloged as a vertebrate, and then include both numbers.

Overall, for persistent ids, DMNS is not a wonderful example- only images have persistent ids, and new field numbers have persistent ids. As an example of this, Stu Holgarth has a trilobite collection at the Smithsonian, so many studies double check data with Stu Hogarth's numbers. Many times, localities numbers are with records that have skeletal information, and staff are filling in the data later.

For location, managers often can update old locality records when there is a new collection event. In the Transactions tab, they use Accession to document field collections. They then create a dataset agreement [correlation] to previously saved field collections.
For the Fern project [pteridophyte fossils], Nicole had to assign a GUID to share with the TCN.
Copyright- how to enforce copyright that restricts selling things?- the Business staff look for infringement. They track citations in order to determine how much is used and how much of the collection is producing new research. When Nicole loans a specimen, she often checks to see what research is produced. Geology checks researchers, also the registrar does more, and also the collection manager. Of course, they don't care about non-profit or educational use, only care about if it is making money. Nicole will pull requests for high resolution data, and see if the researchers publish a book, check for agreements, rights, and make sure that the citations are correct. The business office has a program that automatically alerts staff- it is set up to get trending usage. This program searches based on name (not on specimen id). Reputable journals

require usage of specimen numbers and won't publish numbers with restrictions or legal issues like private land. Loans for paleobotany require notification of usage.

One researcher had a private collection that had about 40 specimens. In order to write a paper, the researcher was forced to donate the collection to a museum for keeping. Sometimes the collection still belongs to the forest service, or BLM (Bureau of Land Management), or the state of Colorado, or a government agency. They create their own databases, so the museum asks for their localities in order to group specimens together. Another issue is if it is a model or the specimen that is used for the research. Other things that are stored are Localities, Collectors, Land ownership, coordinates, and links to data bases for all localities and anthropological sites. The BLM people want to track land ownerships which may have changed so that the specimens may not have been found on BLM land in the past. Earth Sciences has a lot of weird hold backs to sharing. Private land is not divulged to anyone as well and other redactions that are due to fear of poaching.

Sharing with GBIF and PBDB.

At Yale, EMu expert Larry wrote a script for automatically sharing data. At DMNS we are manually exporting. We are also waiting for them to hire a new archivist to include our digital management, as well as for the Anthropology Collection. An institutional DAMS needs to be established. [LUNA is still up and searchable, but I highly doubt anyone is currently updating or managing it (Denver Museum of Nature and Science, n.d.), Neu-Yagle, personal communication, 10/14/21].

We'll need a new Archives staff before we can figure out what the future holds with it. The director has emphasized Digital Asset Management; the need to evaluate and get a plan set up. Currently, the collections division top priority goal is digitization of collections, so maybe

the administration can't deny resources now. There have been morale issues here, and the archivist was in a big hall all alone with no socializing.

3D images and SCANS are too big to store in EMu. Neu-Yagle's export for datasharing [with the pteridophyte TCN] took days. With a remote server, the special request was overloading the computers at DMNS. The scans were put on External Hard Drives. DMNS also uses Morphosource for some storage and links to Morphosource by putting Morphosource id into the notes of the specimen record. You can also attach a paper directly to the specimen record. The instance of EMu at DMNS does not have a minimum number of fields to be filled. It is sometimes easier to create separate party records for each role so that one person has a record as a curator and another as a collector. The department records still have some bad data from the old Argus database. We don't merge functions but keep them separate. Institutional decisions affect what is stored and how. When work is done at some other place, we attach specimens to it. Denver Museum is one separate instance. When loans are made to some other entity there is no obvious other id place to have it. When EMu was built at DMNS, they selected what fields they want; every institution has a different build and looks different. Institutions have to pay Axiell to build modules. Mostly, things are kept the same, if there are multiple requests for a function, all benefit from the new build. For example, the Integrated Pest Management module was modified and built to track management.

In order to upload data to aggregators, you Do a Report. So, you run a report, then email the CSVs, 1000-2000 records at a time, not the whole collection. This Report is not generated in a perfect format, so you have to tweek it. Neu-Yagle uses Excel, deletes duplicates (multiple anything in the collection will create two rows, so she runs scripts to detect this). She creates groups matching a search parameter to produce a special report. If the images are stored

separately, she uses Cyberduck. (Cyberduck, n.d.). [Cyberduck is a free server and cloud storage browser with an FTP for transferring files over the internet.] She transfers images with a little program that she wrote, so she does what she can. The report obtained from EMu does not match Darwin Core format locality information. For an example, it just has period, epoch, age, and you need to add eon for Darwin Core. There is no tech support to get the report to the IPT. Also, some files have meaningless IRN numbers (upload as just an IRN with no data), so you need to connect them, give new data, and set up the IPT. [An IRN is an "Internal Record Number," a number generated by our database. They are unique within our DMNS EMu database, and help us identify precise records: every party, taxonomy, catalog record, multimedia (such as an image), etc. has its own unique IRN. They become meaningless outside of our institution. Neu-Yagle, personal communication, November 9, 2021]. Refresh of the IPT is done by sending email. In order to submit the data to the [pteridophyte] TCN, you have to figure out how to do an attachment, and you have to plan how often to do updates as the TCN wants a plan for long term maintenance.

Multimedia.

A few pictures have UUID, or a column form of GUID. To see it Linked to original object in the catalog module, open the specimen record. For our GUID, we have EMu automatically generate GUIDs. We have numbers that correspond to EMu fields, locality number, use for picture tags; there is a separate module for locality. One field cleanup centrally in EMu is really nice. When collection items are moved, we upload the new location as csv to change the location. For this, we use notes written out when doing the movements. We don’t have room where new specimens should go, so some are out of temporal order, not in order of Period and State. Geology arranged specimen number within the cabinet, which results in

wasted space, minerals from the county are organized by that first, then smallest number to the largest.

DMNS has asked for adaptations; apparently EMu 6 has a spot for the ORCiD. All IMu [public face of EMu] records have static web addresses which are google created.

Email, 10/14/2021

Sam [the archivist] turned on a function in EMu to assign GUIDs (Globally Unique IDentifiers) to all Multimedia records (which included the PCC [pteridophyte TCN] photos). I assigned GUIDs to all the PCC relevant Catalog records. EMu generates UUID4 GUIDs if you ask it to. But we're not ready, mentally I think, to turn it on for all other collections. When we are up with an IPT I'm sure we'll do so. But I needed them for the ferns. I attached my upload template [to the email] which had the instructions notes to myself and how I generated the GUIDs for the catalog records [instructions for a program for generating UUIDs was attached].

Our catalog numbers, at least for Earth Sciences, are in the format (Museum Acronym) (Collection Acronym).(Number)

We use DMNH instead of DMNS for historical purposes even though DMNS is our recognized museum ID. So when you look up our specimens on the PCC TCN their catalog number is a DMNH EPI.#### for each, so it's kinda like DMNS DMNH EPI.#### on that list, and I think it's the same in morphosource.

The catalog numbers look like

DMNH EPI.5200 for a plant or invertebrate fossil (**D**enver **M**useum of **N**atural **H**istory, **E**arth Sciences, **P**aleontology, **I**nvertebrates & Paleobotany)

DMNH EGM.600 for a mineral (**DMNH**, **E**arth Sciences, **G**eology, **M**inerals)

For field numbers, the curator makes those up in the field. Each curator has their own system. It is up to the collections managers and the curators to accurately transfer the data from the curator's field notebook/brain to being assigned a DMNS locality number, and then getting uploaded to the database. Each locality has a DMNH #####, a field number, and then all the associated coordinates, stratigraphy, landownership, etc.

Kirk Johnson, Ian Miller, and Gussie Maccracken all use a similar field number system (as they've passed it down in mentorship): Initials, year, locality visitation sequence. So the 14th site Kirk Johnson collected in 1998 would look like: KJ9814

And the third site Gussie collected in 2021 would look like: GM2103 (or maybe GM21003)