Towards a Roadmap for Advancing the Catalogue of the World's Natural History Collections

Donald Hobern[‡], Laurence Livermore[§], Sarah Vincent[§], Tim Robertson[|], Joseph T Miller[|], Quentin Groom[¶], Marie Grosjean[|]

‡ Atlas of Living Australia, Canberra, Australia
§ The Natural History Museum, London, United Kingdom
| Global Biodiversity Information Facility, Copenhagen, Denmark
¶ Meise Botanic Garden, Meise, Belgium

Corresponding author: Donald Hobern (dhobern@gbif.org)

Abstract

Natural history collections are the foundations upon which all knowledge of natural history is constructed. Biological specimens are the best documentation of variation within each species, increasingly serve as curated sources for reference DNA, and are frequently our only evidence for historical species distribution. Collections represent an enormous multigenerational investment in research infrastructure for the biological sciences, but despite this importance most of the holdings of these institutions remain invisible on the Internet, inaccessible to taxonomists from other countries and hidden from computational biodiversity research.

Although comprehensive digitisation of the complete holdings of each natural history collection is the long-term goal, this is an expensive and labor-intensive task and will not be completed in the near future for all collections. However, many benefits could quickly be achieved by publishing high-quality metadata on each collection to increase its visibility, provide the foundations for further digitisation and enable researchers to discover and communicate with collections of interest.

This paper summarises the results from a consultation activity carried out in 2020 as part of the SYNTHESYS+ (Synthesys of Systematic Resources), *"Developing implementation roadmaps for priority infrastructure areas as part of cooperative RI for biodiversity"* project. This consultation was primed through an ideas paper, and introductory webinars and conducted as a facilitated two-week online multilingual discussion around 26 topics grouped under four broad headings (Users, Content, Technology and Governance). The results of these discussions are summarised here, along with the wider context of existing and planned initiatives.

Keywords

biodiversity, natural history collections, data standards, data linking, taxonomy

1. Introduction

The creation of a catalogue of the world's natural history collections is central to the shared vision and goals of a large number of institutions, projects and other stakeholders and initiatives within the natural history and wider science collections landscape. However, the number and diversity of interested parties brings with it key challenges around unification of approach, interoperability of already developed and widely used systems, and the differing requirements of such a wide range of user groups.

Even in the absence of data on the specimens held in these collections, information about each collection contributes to a map of the resources supporting taxonomy and biodiversity research and assists researchers in locating and contacting the holders of specimens. Collection records contribute to the development of a fully interlinked biodiversity knowledge graph, showcasing the existence and importance of museums and herbaria and supplying context to available data on specimens (Page 2016). These records also open avenues for novel use of these collections and for accelerating the full online digitisation of their holdings.

There is currently no definitive figure for the number of specimens held by collections globally, but estimates range between 1.2 and 2.1 billion (Ariño 2010). A catalogue would go some way towards refining these estimates, and this in turn would provide an opportunity to gauge the economic value of collections and collection-based services (Hobern et al. 2020), opening up new funding opportunities based on both quantitative and qualitative measures of importance. Although some collection types are not amenable to simple valuation, alternative metrics may showcase societal as well as economic importance.

1.1 Community consultation process

The basis for this report is the ideas paper by Hobern et al. (2020) and the subsequent community consultation on the topic *Advancing the Catalogue of the World's Natural History Collections*, held virtually by the *alliance for biodiversity knowledge* in March and April 2020. The discussion topics of the virtual consultation are summarised in Table 1. The consultation and resulting roadmap form part of Task 5.1 under SYNTHESYS+ (Synthesys of Systematic Resources), *"Developing implementation roadmaps for priority infrastructure areas as part of cooperative RI for biodiversity"* (*Smith et al. 2019*). The ideas paper was first issued for review on the 25th February 2020 and was subject to initial review and revision by task participants. Stakeholders were gathered through existing project and community networks*[1], and the online consultation format was adopted to facilitate wider

participation by minimising the need for travel and allowing contributions to be collected over an extended period of time, removing time zones, travel costs and other commitments as a barrier. All materials were uploaded to the Global Biodiversiity Information Facility's (GBIF) Community Forum site on the 6th April 2020, and discussion threads were opened to the public from the 17th – 29th April. Stakeholders were asked to consider questions relating to the 26 topic areas outlined in the ideas paper, grouped under four key categories. To remove language barriers, dedicated threads with translated summaries were available in Spanish, French and Chinese. In addition to the ideas paper, presentations were contributed by organisations and interest groups to give participants a clear idea of the current collections information landscape, and demonstrate current tools and activities that may inform or form part of the development of a future catalogue. The timeline of the consultation process is described in GBIF's Discourse site (Copas 2020), archived here as Suppl. material 1. and archived on the Internet Archive.

This paper uses the outcomes of this consultation to identify common themes, priorities, areas of consensus, and areas of dispute. These will be used to propose a vision for how a global collections catalogue may be developed, covering use cases, information, maintenance, resourcing and sustainability.

1.2 Articulating the need

The ideas paper outlines a range of potential use cases based on those collected by the TDWG Collection Description Interest Group*[2], as well as work done by ICEDIG in preparation for the Distributed System of Scientific Collections (DiSSCo, a European research infrastructure programme for natural science collections) (Hobern et al. 2020, van Egmond et al. 2019). These collectively illustrate the potentially extensive value and the benefits that could be offered by a global collections catalogue, whilst also highlighting the difficulties in adequately scoping the catalogue to fit the needs of a large and varied user community.

Four broad headings are described by the ideas paper:

1. Uses for the catalogue
2. Information in the catalogue
3. Technology for the catalogue
4. Governance of the catalogue

2. Current landscape

The sections in this landscape overview are based on the contributed materials for the community consultation supplemented with additional research to give an overview of the key platforms and databases, collections management systems, data standards and other

community activity. The aim has been to provide background information for readers but not comprehensively cover the current landscape.

2.1 Platforms and databases

A number of existing catalogues for institution, collection and specimen-level information are already in use or development, driven by several community-driven initiatives and projects. There are other broader sources of information that could be integrated or used in a future platform. To prevent record duplication and minimise the level of resource required to create collection catalogue records, the scope, controlled vocabularies and preferred identification schema of the most relevant systems should be investigated and incorporated during development of the collection catalogue data architecture.

Atlas of Living Australia Natural History Collections - the ALA Natural History Collections page (formerly known as the "Collectory") is an example of a national information resource on natural history collections. ALA has a high calibre informatics and software development team and receives strong institutional support and engagement on the national level. ALA collection records do not currently use a standard vocabulary and the repository is struggling to de-duplicate collection-level records contributed for different views of the same collection (Atlas of Living Australia 2020,Belbin et al. 2021).

CETAF Collections Registry/CETAF passports - the Consortium of European Taxonomic Facilities (CETAF) provide a central source for information about its 63 European member organisations. 'CETAF passports' are contributed as a condition of membership and include high-level categorisation of collections including non-mandatory collection size metrics. CETAF is currently building on the functionality of CETAF passports with the development of the CETAF Collections Registry and has proposed assigning unique institutional acronyms to each member, which may cause some overlap/conflict with existing identifiers (Semal et al. 2019).

The Global Registry of Scientific Collections (GRSciColl, including GRBio) was initially developed as a global 'clearing house' of information for institutions and collections before being incorporated by GBIF in 2019. GRBio held information on biodiversity collections and was a subset of GRSciColl which is open to all categories of scientific collection. Although its content is currently incomplete, GRSciColl is considered a viable framework for expansion and is currently in a new phase of development. So far synchronisation has been established with Index Herbariorum (see below) and content from the iDigBio collection database has been integrated, with GRSciColl now powering the iDigBio collection portal. GBIF are now actively developing the codebase, where a role-based authentication model enables wider contributions. During the consultation, key priorities for 2021 were identified and these have been implemented. The draft GBIF work programme for 2023 includes a goal to "enrich GRSciColl through the integration of collection description information, compatible with the Latimer core, to support use cases such as priorities for data mobilization" (GBIF 2022b).

iDigBio web portal - iDigBio is the United States national resource for digitised information about natural history collections. The iDigBio specimen portal makes available millions of records from neontological and paleontological specimens curated at museums and other institutions in the US. The data held in their repository follows the Darwin Core and Audubon Core data standards and iDigBio has contributed upwards of 1.5k collection-level records to GRSciColl to date (iDigBio 2021).

Index Herbariorum – This is the most successful and established collections catalogue and covers the world's botanical collections. Indeed, its use is recommended in the International Code of Nomenclature for algae, fungi, and plants (Turland 2018). Herbaria can provide/edit their records and updates can be submitted through email or other channels. Existing biodiversity data tools such as the Integrated Publishing Toolkit (Robertson et al. 2014) which currently facilitates the creation of EML metadata could be used to deliver collection records to Index Herbariorum. Collection records from Index Herbariorum have already been integrated into GRSciColl.

Wikidata is already recognised as an identifier broker with potential to advance biodiversity knowledge graph development (Sachs et al. 2019) and is being used by successful community initiatives like Bionomia (Shorthouse 2020). It could be used to semantically link people, taxa, places, collections, institutions and more (Groom et al. 2020).

The Global Research Identifier Database (GRID) and the **Research Organisation Registry** (ROR) are existing databases of globally unique persistent identifiers and associated metadata for education and research-related organisations across all disciplines. Each service holds data on more than 100,000 organisations, and their identifiers are interoperable. GRID is a commercial product managed and owned by Digital Science. GRID provided the seed data for ROR, which is a community-led initiative.

These databases could potentially be used as a starting point for institutional identifiers.

2.2 Collections management systems

Collections Management Systems (CMS) are databasing tools that are used to organise, control and manage information on behalf of natural history collections. They support many tasks that are important to operation within each collection, including inventory management, creation and publication of descriptive specimen and collection metadata, risk management, collection conservation and assessment, exhibition management, loans and research requests, and as stores of legal information regarding the acquisition and use of collections.

Collections management systems are likely to be one of the fundamental sources of natural history collections data but there are several challenges using them as to contribute and maintain entries in a catalogue of collections. Many different systems are in use. A survey of European collections conducted by DiSSCo (Casino et al. 2019) identified over 37 different systems ranging from general-purpose database management systems (e.g., Microsoft Access, Filemaker) through in-house (e.g., Kotka, PlutoF) and open-source

solutions (e.g., DINA, Koha, Specify) to commercial products (e.g., Adlib, ActiMuseo, EMu). Some respondents did not use any sort of database and stored their collections data in spreadsheets or text documents. The number and variety of systems in use around the world will be even larger although some are more frequently used in particular countries (e.g. Specify in North America) and for particular taxonomic groups (e.g., BRAHMS for plants).

There are no studies evaluating these various systems as a source of standardised collections metadata. CMS interoperability has been studied at a limited scale with a focus on specimen/observation data. Dillen et al. (2019) concluded that we are far from being able to seamlessly import and export data between different CMS platforms.

2.3 Data standards and interoperability

The standards summarised in Table 2 are highly relevant to development of a collections catalogue. Consistent adoption of compatible data standards will ease aggegation and integration of collections metadata and avoid duplication of effort.

2.4 Community activity and stewardship

GRSciColl

The GBIF Secretariat coordinates updates to GRSciColl. Edits are performed by data managers from the GBIF Secretariat and iDigBio. Other changes are imported from Index Herbariorum or through contributions by staff from institutions and national nodes within the GBIF network. Any user can suggest changes for inclusion following review by the GBIF Secretariat and community. Data from some national surveys have been uploaded directly into GRSciColl.

3. Community Priorities

This section presents the community's priorities for a collection-level catalogue as a summary of notable areas of consensus and concerns that emerged during the consultation process. We have followed the four high level categories (Use, Information, Technology, Governance) and 25 subcategories used in the community consultation. In a few instances, we have referenced comments from other subcategories if these more naturally relate to the topic under discussion. Where applicable, we have provided links "(ref)" to the original discussions in the GBIF Community Forum which are also archived in Suppl. material 1. Integrated summaries of all forum threads can be found here.

3.1 Use

The Use category included nine topics.

3.1.1 Directory to support the collections community

By establishing natural history collections as a global scientific infrastructure we make it easier to foster new collaborations, resource research, fund opportunities and support sustainable data infrastructure. By standardising our institutional acronyms and the collections held within them, we improve collection discoverability and citability, making it easier to demonstrate impact and importance (ref). We can make use of existing persistent identifiers (PIDs) in GRID or ROR, so we are not establishing a set of new PIDs and benefit from integration and re-use (Addink 2020) (ref).

There are many collections that are mostly invisible due to the predominantly specimen-based approach to digitisation. Specimen-level digitisation is often costly. Publication of collection-level data should be recognised as an important and cost-effective starting-point (ref). We recognise that understanding and serving the needs of different users will be important and that keeping the collections data up-to-date will be a challenge (ref).

3.1.2 Locating specimens and genetic materials

A catalogue that provides summary information on the holdings of each collection would be a highly useful resource, if the summary information was relevant, reliable and could be kept up-to-date.

Previous initiatives relating to the creation and aggregation of collection-level catalogue records have increased use of and interest in items in the collections (ref). Summary collection information acts as a 'signpost' for end-users to help them narrow down which of the world's collections may hold items of interest and facilitates further investigation and communication with collection managers. Collection-level records would also help to document key networks and linkages between specimen data and existing related data platforms such as the International Nucleotide Sequence Database Collaboration (INSDC) databases (ref). This can be expected to increase the points of discovery and entry for underserved or non-traditional users.

The minimum level of information required for collection records to be a useful resource is likely to vary across disciplines, user groups and geopolitical contexts (ref). There is a general consensus that details on the institution holding the collection, taxonomic scope, and metrics on the size of the collection should all be mandatory fields (ref). These could be augmented with optional fields that allow additional data to be shared where available (ref).

3.1.3 First step towards databasing collections

We need to provide guidance and support to the community, particularly to collections staff. This includes the need for good tools and tutorials for curating, updating and disambiguating collections records (ref). The community will need region-specific roadmaps and strategies as levels of support and motivations vary (ref). Current emphasis

on publishing specimen records lessens potential data sharing of less well-resourced collections that are effectively excluded. The GBIF dataset classes (Resource, Checklist, Occurrence, Sampling-event)*[3] offer a hierarchy of complexity and can serve as a stepwise path towards a goal of specimen digitization of a collection from simple collection-level metadata (Resource) through species lists (Checklist, which could serve as a simple tool to list species held in a collection) and specimen-level data (Occurrence) to opportunities even to model field-collection activities (Sampling-event).

Publishing a metadata-only dataset (Resource) could be sufficient to advertise a collection and information about its holdings. The collection would become Findable, even if not digitally Accessible, Interoperable or Reusable. The Integrated Publishing Toolkit supports metadata-only datasets.

If a collection is then in a position to add a checklist dataset summarising species held - this was quite a common category of web page 15 years ago - the collection could be listed in simple ways on GBIF species pages, again further raising its profile for wider access and use. This adds some Interoperability. Databasing as DwC specimen data then takes things forward and allows for full "FAIRness" (Wilkinson et al. 2016).

3.1.4 Assessing the scale and value of collections

Estimates of collection size are already widely held and used by collection-holding institutions, but these metrics are decentralised and typically provide little information on the assessment methodology used.

High-level estimates of collection size would be useful to external stakeholders such as government agencies (ref). Collection size estimates can be used to represent the ‘value’ of collections on the national and global scale and would be invaluable in helping the community to ‘build funding cases, show current (often national) capacity, and highlight gaps‘ (ref) (Leggatt 2019,Council of Australasian Museum Directors 2018).

To be useful, such estimates would need to be either developed under a shared methodology (e.g., the One World Collection project) (ref), or contain sufficient methodological information to allow users to assess the applicability of the record for comparing collections or aggregating metrics (ref). The former approach would make the catalogue easier to use, but the latter would facilitate data collection and re-use of existing information.

Standardised methodologies for valuing collections based on scale and scope are already in active use (ref), but there is risk attached to following a single model in this respect: the value of collections will ultimately depend on the requirements of those seeking to use them (ref).

3.1.5 Increased value for data on specimens, taxonomic publications, etc.

We recognise that better linkage of collections metadata, including information on the main collectors who helped to build the collection, with other external identifiers and authorities like ORCID, Wikidata, and VIAF will improve discoverability both inside and outside our community (ref). If we are able to combine collector information with understanding of the taxa present in collections (e.g. at a checklist as opposed to a specimen/occurrence level) we would have a better understanding of what makes a collection unique (ref). Detailed information on specimen preservation methods is important for collections users (ref).

3.1.6 Reducing duplication of effort

A large amount of information about collections is already available on institutional websites, but effort is required to pull this together and maintain it over time. It would be helpful to provide a template or other pro-forma data collection mechanism to let collection managers update summary data quickly and easily (ref). Some institutions already record curatorial assessments for their collections; it would be benefiical to support these assessments, along with all supporting information, as part of a world collections catalogue (ref).

Providing reusable collections data and standardised institution and collection names would reduce the overhead on other specialised collection catalogues such as the Global Genome Biodiversity Network (GGBN) which currently maintains its own general collections registry and could instead focus more time and community effort on collections biobank metadata (ref).

There are recent discipline-based examples of assessing the state of collections (Cobb et al. 2019,Sierwald et al. 2018) including use of software to harvest data for a U.S. fish collection catalogue, although this still requires relatively high effort and may miss significant collections (Singer et al. 2018) (ref). This informatics-based approach could be applied to other disciplines but would need community support to generalise the software (ref).

3.1.7 Foundation for new and enriched services

A collections catalogue would make collections more findable and accessible (in the sense of the FAIR Data Principles, Wilkinson et al. 2016) for new audiences and users (ref). Without an awareness of what resources are available in our collections or clear channels to contact the collection managers, potential funders will overlook our holdings (ref). We should be making our biodiversity information more available to environmental managers, policy makers and other government agencies (ref). Our collections have a role as part of cultural heritage within the wider arts and humanities research community (ref). Some of the necessary linkages with identifiers for people are described in section 3.1.5.

Cooperation with other initiatives like the International Nucleotide Sequence Database Collaboration (INSDC) is crucial to allow linking sequences that lack references to collections to corresponding voucher specimens and samples. Building tools to help researchers submit better metadata is important (ref).

When considering new and enriched services, we should also be mindful of focus, delivery and utility. While new downstream use is important, we should consider focusing narrowly on what queries the catalogue can support best in the short to medium term and that correspond to a sufficiently important audience (e.g. large, high impact, well-resourced, etc.) (ref). We can look at other adjacent sectors for analogue data infrastructures and what makes their core services successful (Leonelli 2013).

3.1.8 Improvements to citation and visibility for collections

Research value is primarily measured in terms of visibility and impacts from published literature. To be recognised by such measures, the citation and attribution of natural history collections needs to be agreed and standardised across the community and made visible and useful to stakeholder groups such as publishers, funding bodies and data aggregators (Rouhan et al. 2017).

Understanding the community's existing practices and data quality issues in this area is key to successfully developing the collection catalogue so that citation of collection-level records is sustainable, measurable and more fit-for-purpose than current practices (ref). Outcomes from this analysis, such as comprehensive lookup tables of identifiers used for particular collections or institutions (even when these are not unique within the broader collections community) (ref), could improve discoverability of collections from an end-user perspective, feed into current initiatives to unlock the historical scholarly record (ref) and aid in the discovery and embedding of linkages between related outputs (ref).

Previous initiatives around standardising citation and attribution have stalled due to lack of uptake (ref); a critical mass of adopters is required before stakeholders outside of the core community (e.g., publishers and aggregators/content banks) will change their working practices to incorporate a particular standard. Additional barriers to user uptake include a lack of guidance around attribution practices both for collection users and for collection-holding institutions and uncertainty around proper citing procedure for collection data from aggregators and other secondary sources (ref). It may be difficult to get authors to consistently use a standard abbreviation. It might be easier to simply link multiple abbreviations to a single, stable PID (ref).

Engagement may be encouraged via links with other data repositories, especially those with established infrastructure and dataflows related to the identification and resolution of research citations. ROR, GBIF and Wikipedia, for example, already integrate with Datacite and Crossref (ref), both of which provide impact metrics that would incentivize both contribution to the catalogue and adherence to related standard citation practices (ref).

3.1.9 Support for national and regional needs and applications

One of the biggest issues we face is demonstrating the role and value of collections (value is covered in more detail in section 3.1.4). This is often a national challenge because this funding primarily operates at this scale , but on occasion becomes a continental or global challenge (ref). A more integrated model of the natural world, founded on observations and collections, would provide evidence of where we are deficient in data, and to identify which organisations might coordinate to fill these gaps at a national or regional level (Meineke et al. 2019).

Uniqueness of collections can help focus prioritisation for digitisation (and other activities) at a national and regional level. It can act as a starting point for understanding how to effectively collaborate and pool resources (ref).

In other sections it was noted that some countries have minimal online catalogues, or resources shared in languages that may make them less internationally discoverable (ref). National legislation can play an important role in motivating data sharing and coordinating national activities (e.g. the *Registro nacional de Collecciones in Colombia*) (ref). This may be an example other countries or research councils could adopt.

3.1 Recommendations

- A collection catalogue should mandate a minimum number of standard fields such as: taxonomy, holding institution and collection scale metrics which could be augmented with additional fields where available.
- Strong guidance and support materials must be available to the community to support the catalogue.
- There is a need for ongoing methodological standardization while maintaining flexibility for institutions and for national, regional and taxonomic networks.
- Collection records should maintain linkages with other external identifiers and authorities.
- The collections catalogue should be built in such a way that it can also support use as a national resource.

3.2 Information

The Information category included six topics.

3.2.1 Scope for the catalogue and definition of “collection”

The definition of a natural history collection is broad and reflects the goals and uses of the collection as well as its contents. At its core, a natural history collection represents

evidence of biological and geological diversity on Earth, but collections may include related objects such as extraterrestrial geological specimens and anthropological artifacts. Living collections, whether in an active or dormant state, can be included. Furthermore, the collection objects themselves are not necessarily items of biological and geological diversity but may include associated materials such as field notebooks, photographs and ethnobotanical objects. Collections can be eclectic or have a specific focus and *raison d'être*, such as a xylarium.

There is also a wide range of different usage-based goals for a collection. Some are purely used for taxonomic research, but there are others that focus on education, history, material science etc. The group is not even managed under the same set of regulations, management and ethical considerations. Living collections, human remains and objects of cultural significance have specific requirements that determine how the collection operate. One cannot even state that each collection must be under the control and hosted by a single institution, since we need to be able to refer to collections that no longer exist in this way, having been destroyed or divided up.

In some cases collections are defined as the complete set of materials held at the level of a single institution, as is true for most herbaria listed in Index Herbariorum. However, in many other collections the material forms seveal separately identified collections, divided perhaps by curatorial practices, by taxonomy or by the collection's origins.

Collections often map to organisational structure and to curatorial approaches rather than adhering to consistent definitions. This conflation of institutional structure with institutional collection(s) is too frequent to have occurred by chance; it seems reasonable to assume that operational concerns and priorities (e.g., naming/defining a collection to reflect acquisition or provenance events) play a key role in shaping the community notion of a ‘collection’.

Ultimately, the easiest way to define a collection in the Catalogue may be purely in terms of usage: collections are the entities to which we need to refer when organising information about the materials they hold. If we need to be able to refer to these collections in a reliable way, they each need an entry within the Catalogue. Consideration must also be given to the advantages that the collections and the collections-holding institutions may gain from being listed.

Differentiating “natural history” collections from associated collections is important, but we need the ability to reference and link to holdings that are regularly treated as adjunct collections (archives, field notebooks, registers, photographic collections) and born-digital collections (e.g. sound records, camera trap images). These may be considered and identified as discrete collections in their own right (see section 3.2.5).

The broad consensus was that the scope for the Catalogue should be broad and inclusive, including all collections that are useful for natural science, natural history or natural heritage. This includes xylaria, ethnobotanical, paleontological collections and

anthropological collections. Some of these collections will have sensitivity and legal restrictions that need to be managed when sharing their descriptions.

3.2.2 Identifiers for collections

Multiple collection identification schemes exist and are in actively use. Collections are often identified in parallel in multiple schemes, a situation which reflects the flexible definition of a collection as discussed earlier (ref). A number of identifier schemes are provided by or derive from data platforms and services: GRSciColl, ROR, ALA Natural History Collections and the GBIF Registry (ref). Identifiers for an organisation or unit within an organisation have also been widely adopted as a shorthand to refer to the collections they hold, even if the original organisational entity no longer exists in an operational sense (ref).

It may be the case that only these more traditional collection identifiers (e.g. the identifier for a specific herbarium) need to be human-readable because of their historical use in previous and current registries (ref). We need to avoid conflating the purpose of and requirements for human and machine-readable identifiers: machine-readable identifiers need to be globally unique, persistent and resolvable. They should provide unambiguous identification of a collection — even if the contents or environment of the collection changes over time — and facilitate wider data linkages. Human-readable identifiers need to be succinct, descriptive, memorable and, if not unique and persistent, contextually flagged clearly enough to enable software systems to distinguish and accommodate this (ref).

One approach to prioritising existing identification schemes within the Catalogue would be to select those that most closely map to a discrete class of collections within the Catalogue (ref). It would also be prudent to prioritise identification schemes on their technical capacity, accessibility, underlying infrastructure and accompanying data services.

Usage of preferred identifiers could be promoted by the development of resources and activities focused on community engagement and increasing the wider awareness of the benefits and availability of the selected schemes (ref).

3.2.3 Hierarchical collection structures and subcollections

An important consideration is whether the Catalogue should represent the complex historical and contemporary relationships between collections and subcollections that may be important to different communities. The alternative would be a simple flat catalogue which treats all included entities as equivalent.

Hierarchical relationship structures would be useful for collections that have changed ownership or location over the course of their lifespan. For example, a subcollection record could be linked to a ‘parent’ collection record to reflect provenance and facilitate discovery (ref). Flexible parent-child relationships of this kind could go beyond fixed hierarchies and also represent alternative classifications of subcollections. Hierarchies are less suitable for use in scenarios where a single collection object falls under the scope of several different

collections (ref). Such nested scenarios are common and, unless carefully handled, could lead to double-counting and inflation of collection size metrics.

A system that is not fully hierarchical could be a workable model capable of handling most use cases so long as a few primary classes of entity (e.g, institution, collection, subcollection, dataset) are properly defined, standardized and incorporated during its design (ref). Standardised relationships between instances of these few classes could maintain the simplicity of the Catalogue while allowing most situations to be represented appropriately through judicious mapping of real-world entities against the classes. The development of well-defined classes would also enable aggregators and other platforms to validate the integrity of the Catalogue, reason over relationships and logically constrain the operations that can be applied to different classes of catalogue record (ref). The nature and scope of each class of collection record needs to be communicable to end-users to allow for different search strategies based on their data requirements (ref).

3.2.4 Description of a collection

The community broadly supports use of the TDWG CD standard for collection descriptions (Collections Descriptions interest group 2019) for the collection catalogue. The only addition recommended was for a field specifying how the collection should be cited (ref).

The TDWG CD model centres on a small number of mandatory fields and a larger range of optional fields. This approach allows different classes of collection description records to be described using dimensions most appropriate for the discipline, while still controlling the quality and integrity of the data for core fields and allowing some level of class interoperability (ref). The flexibility to describe different collections using optional, discipline-specific fields is widely seen as essential to successful uptake and use of a collection-level data standard and accompanying discovery systems and catalogues (ref).

Controlled vocabularies should be identified or developed for as many fields as is feasible (ref). Fields most urgently in need of a controlled vocabulary could be identified via analysis of existing specimen-level records containing equivalent DwC fields.

Any consensus/community level collection data standard should not be considered complete until it has undergone adoption or testing in institutional data workflows and projects to ensure that it is fit for purpose (ref). Real-life testing and early adoption of the standard for a small set of use-cases and collection description classes would facilitate the identification and subsequent development of those fields most suited for machine access.

3.2.5 Wider data linkages

Fields that support an plurality of identifiers and links between the catalogue and external services will enable discovery and use by non-traditional users, e.g. visitors to a Wikipedia page following a citation link to the collection catalogue (ref). It will also improve the usability of the collection catalogue by allowing users to easily navigate to external, authoritative sources of information on topics associated with the specified collection (ref).

Fields selected for use in this manner need to be carefully evaluated and prioritised: creating and maintaining linkages between data silos is a non-trivial undertaking and the benefits to contributors, system providers and external data sources must be clearly defined (ref). There is general consensus that the following core fields should be explored: collector, species/taxa, specimen-level information, notable and/or primary collectors and associated publications (ref). Linkages should be bidirectional wherever feasible, taking into account each external data source's sustainability and technical capacity in areas such as link resolution, identifier integrity and reporting (ref).

Fieldwork notes and images, type specimens, and taxonomic treatments were also mentioned as possible candidates for linkage (ref), but these fields may be more appropriately and usefully associated with specimen-level records (ref). External linkages with sources that provide usage and impact metrics could be valuable mechanisms for boosting engagement. Without support and clearly defined benefits for catalogue contributors, this may lag in existing areas of poor data-density such as south-west Asia (ref).

3.2.6 Information services relating to collections

All of the information services proposed in the ideas paper were recognised as components that would enhance the value of the Catalogue:

- Assess the growth, scale and value of the world's collections
- Provide a collection digitisation dashboard to monitor and highlight progress
- Discover the location of biological materials or the likely presence of biological materials for any taxon
- Develop discovery services for accessing information on type specimens or communicating with the relevant collection where the specimen is not digitised
- Identify sections of collections that should be digitised to answer specific questions
- Match gap analysis of published specimen data against the collection catalogue to prioritise digitisation for filling taxonomic, geographic, or other gaps
- Discover holdings that make a particular collection unique, and therefore of even higher value
- Develop and fund collaborative digitisation programmes focused on understanding of the holdings of the network as a whole
- Develop cross-institutional loan systems and taxonomic workbenches
- Develop citation models for collections and track their impact
- Perform risk assessment of the sustainability of a collection

Partnerships with existing digital repositories (e.g., CoL, GBIF, BHL) to deliver shared or complementary services would be beneficial for encouraging both development progress and collaboration within the existing ecosystem of research infrastructure services, tools and platforms (ref).

3.2 Recommendations

- A collection catalogue would be broad and inclusive to be used across many disciplines that maintain collections.
- Collection identifiers initiation must be accompanied by community engagement.
- Controlled vocabularies should be identified or developed for TDWG CD standard for collection descriptions (Collections Descriptions interest group 2019).
- Core fields should be used for linking to external data.

3.3 Technology

The Technology category included five topics.

3.3.1. Pathways and tools for publishing collection records

Good software and infrastructure will be critical to building a global collections catalogue, and creating and maintaining these is likely to be one of the more significant costs associated with the Catalogue (ref). The proposed approach would be to maintain a single master record for each collection in GrSciColl and to use existing publishing mechanisms to keep them up-to-date (ref). Wikidata might serve as a broker between other identifier systems, although it should not itself be considered an authoritative source (ref). Wikidata could also allow many more members of the community to make enhancements to data about collections and would make the collections data more discoverable.

There are national platforms that could be integrated with a global collections catalogue (e.g. Colombia's *Registro Nacional de Colecciones* and Argentina's *Sistema Nacional de Datos Biológicos*) but a review is required of the update frequency and data richness of each such source when compared with direct information feeds from each individual collection (ref, ref).

3.3.2. Community catalogue

There are several community catalogues that are established and widely used and that should retain their own identity. These catalogues (including Index Herbariorum and GGBN) could maintain the primary version of the collections data for their focus communities and then synchronise data with GRSciColl (ref). In some cases institutes themselves will maintain their own information on local systems, or get support for publishing these data at a national level (e.g. iDigBio, Atlas of Living Australia) (ref). This will require careful consideration of how to model and manage role-based access permissions for editing collection information and nominating which source(s) should be used as the primary copy. The data standards used across the community catalogues and

the global catalogue should normally be the same, but where there are differences mapping will be required to ensure they are discoverable and interoperable (ref).

Where there are other community initiatives that are also building discipline-specific catalogues there should be discussions between these communities and GBIF to understand how they can contribute to or use GRSciColl functionality (ref).

3.3.3. Integrated catalogue

A successful integrated catalogue needs tools to easily customise, create, and update collections records. This will depend on a combination of manual and automated approaches, including tools to support the community resolve and map informal collection identifiers (ref, ref).

3.3.4. Collection management systems

While collection management systems hold the potential to be efficient data sources for a collection catalogue, maintaining a CMS should not be a requirement for participation: a significant proportion of organisations manage their collections data solely through spreadsheet tools (ref). The GBIF IPT goes some way to reducing participation barriers for spreadsheet data at the specimen level (ref) (Robertson et al. 2014), but even the GBIF IPT requires a degree of infrastructure and technical resources that may not be available everywhere (ref). In addition, the GBIF IPT does not facilitate round tripping of data. Collection catalogue records are likely to be simpler to create and less numerous than specimen-level records, so a simple web-form could be a suitable mechanism for collections without a CMS to add data to the Catalogue (ref).

For organisations where the CMS plays a central role in all aspects of the collection data lifecycle, the ability to manage collection-level records in the same system would have significant benefits. Inclusion of collection-record management functionality would reduce double-entry of data, enable links between specimen and collection records, simplify high-level reporting, enable better tracking of digitisation progress, promote consistency between common fields and potentially drive workflows around automated enhancement of specimen level records (ref).

CMS systems could automate the creation and updating of collection-level records: both descriptive and quantitative collection metadata could be produced by aggregating specimen-level records over a limited set of dimensions (ref). Specify and Symbiota both already hold some capacity for interoperability with the IPT and EML: a similar approach incorporating fields from the TDWG CD standard may be a suitable mechanism for data exchange between a CMS and the collection catalogue (ref).

Elements of this architecture are already operating in GRSciColl, including metrics derived from aggregated GBIF specimen records (ref). The MIDS (minimum information about a digital specimen) metadata standard (Haston and Hardisty 2020) may be an appropriate

digitisation progress metric, but further thought is required on how this could be best adapted to reflect digitisation status at the collection level (ref)

3.3.5. Interfaces, APIs and client modules

A “one-size-fits-all” approach rarely works when attempting to integrate data from a variety of systems. Flexibility and agility will be important when designing the interfaces and underlying APIs (ref). The users of a global collections catalogue will have varying technical capabilities and we need to ensure participation for all, so we need to support spreadsheet uploads and web form editing. In terms of APIs and harvesting data we need to take a gradual approach at connecting, partnering and building on established infrastructures wherever possible.

Interpreting and validating data will be critical when building the global collections catalogue. Lessons from Bionomia’s implementation of an OpenRefine reconciliation endpoint would be useful in designing services. Careful consideration and potentially editing the collection model in Wikidata would allow us to more easily use Wikidata in our own reconciliation efforts and share our data more effectively (ref). The content of collection records should be interpreted and validated as much as possible so its utility and value as data can be maximized. Implementations must be designed to support and display both human- and machine-readable data and to underpin high-quality metadata management, standards compliance, reliable update mechanisms and clear provenance reporting (ref).

3.3 Recommendations

- A single master record for each collection is required and existing publishing mechanisms should be used to keep them up-to-date.
- The existing community catalogues should retain their own identity and synchronized with the global system.
- Link data from existing CMS to reflect digitisation status at the collection level.
- System should be accessible to both human users and machines.

3.4 Governance

The Governance category included six topics.

3.4.1. Ownership of information for each collection

The starting assumption is that each institution should have responsibility and control for information on its own collections. Under some conditions, responsibility and access control may be delegated to a third party where local informatics resources are limited or non-existent (ref, ref).

Indigenous labels and worldviews should be included in collections descriptions where possible (ref).

Even when there are local resources we will need to encourage active maintenance through mixed approaches, such as training and educational outreach, how data are presented to users, and how editors are recognised and credited (ref, ref, ref). Formally incorporating the maintenance of collections information into organisational roles would be ideal, but this has been challenging in the past (ref).

Although it is assumed that institutions and, by implication, curators will provide and maintain collection information, there is an obvious concern that they may not engage with this international initiative to take ownership of their information. Training and incentives may help to change this. Without appropriate incentives, curators may not necessarily benefit directly from improvements in publicly accessible data for their collections.

3.4.2. Communities of practice

For some communities, metadata on collections (or parts of collections) are already included in multiple collection catalogues owing to overlaps in scope (ref). We need to avoid duplication of effort wherever possible through integration and interoperability.

There are several examples of national organisations which may act as intermediaries, or already curate national collections data (e.g. NatSCA's FENSCORE, iDigBio and Atlas of Living Australia). These could champion the global catalogue at a national level and help to broker data using established networks and infrastructure (ref, ref).

Publishers of scientific literature have a significant role in existing communities of practice: they are among the largest users of collection codes and could effectively promote their use and encourage linkage. They may also serve as a source for data on collections that may not be recorded elsewhere (e.g. private collections) (ref).

Further discussion is needed to identify the best ways to encourage, support and engage existing communities since these will be critical in encouraging and facilitating voluntary additions and updates to the catalogue. At some level, a federated architecture will be required to allow the global catalogue to be constructed as a mosaic of contributions from different communities and services, each with their own focus and strengths (ref).

3.4.3. Technical infrastructures

There was limited discussion and was covered in more detail in section 3.3 Technology.

3.4.4. Governance arrangements

This discussion was merged into section 3.4.2. Communities of Practice.

3.4.5. Incentives for contributors

The catalogue can raise awareness of collections and act as a free advertisement by displaying branding and use of rolling highlights on the home page, etc. It would also be possible to develop functionality that generates metrics that may be of use when reporting to stakeholders, preparing funding requests, prioritising internal curatorial efforts, seeking to understand the value of collections or seeking potential collaborators. It should however be noted that metrics and metadata on collection activities are not universally considered to have a positive effect. Some stakeholders may have concerns that such information could be used to impose changes in performance management approaches or could lead to undesirable public recognition (ref). More consideration is necessary to understand how to establish metrics while mitigating these perceived risks.

While not an incentive, lowering the technical barriers for editors and contributors makes participation more likely (ref). This could be achieved through financial support for training courses or for projects to improve collections data. Free collection management software and technical support may also significantly increase engagement.

A sense of ownership is important for long-term engagement, and it is more sustainable to equip contributors to take control than to provide ongoing data support (ref).

3.4.6. Funding and sustainability

Governance and technical infrastructure both require funding and support. This could be achieved by formally including responsibility for the catalogue in the mission of GBIF or another trusted infrastructure partner. National and regional consortia (e.g CETAF) that would benefit from a collections catalogue have a vested interest in ensuring the long-term sustainability of the solution (ref). Even with such support, long-term funding will be challenging. Government agencies, including research councils, and large collections are also potential sources of funding and support (ref).

Some regions will be able to contribute staff time and potentially funding, but there are areas where economic or legal constraints will make contributions difficult. To ensure global participation and sustainability we must consider how we can support less-resourced regions (ref).

Stakeholders will require metrics and performance indicators justify long-term support. Sustained growth, data quality and fitness-for-use are some of the potential metrics that will need to be monitored (ref).

3.4 Recommendations

- Mechanisms for outreach and training are critical for success.
- Governance should build on existing communities of practice.

- Formal acknowledgment of the work of collections through metrics and metadata is critical, but this will not by itself be sufficient to secure success.
- Metrics and performance indicators will be required not just for the individual collections but for the catalogue itself.

4. Development of GRSciColl

As recognised during the consultation, GRSciColl can serve as the central linkage point at the global scale for dispersed activity towards developing and maintaining the catalogue of the world's natural history collections. GBIF continues to work to develop services and improve the value of GRSciColl so it can serve as this foundational resource.

4.1 GBIF's GRSciColl Catalogue priority roadmap 2021

The GBIF Secretariat prepared a priority roadmap in 2021 (Robertson 2021) for ongoing development of GRSciColl, building on previous work to connect Index Herbariorum, import metadata from iDigBio and link GBIF occurrence records to the entities in GRSciColl, and focusing on the development necessary to allow wider external contribution, and to mature the processes around editing.

The roadmap identified six key priorities to progress.

Reduce the amount of duplicate records

Linking to Index Herbariorum and iDigBio enriched the catalogue, but also increased the number of duplicate entities requiring manual intervention. Future duplication of records to be addressed by:

- Documenting guidelines on how a data manager can resolve duplicate issues [REG-316]. The guidelines will provide example scenarios, explain the recommended approach to defining codes and explain the implications on external systems (see master data management below).
- Develop tools that help identify potential duplicate records and alert data managers [REG-191 *-now implemented*]

Allow any user to propose changes

At present, the process for feedback and corrections is weak and does not allow proposed changes to be supplied in a structured form, to be addressed by:

- Developing an interface allowing any user to propose a change to any/all fields and state whether they have authority to approve these changes. Changes are then to be reviewed and applied by the editorial team [REG-CONSOLE-376 - *now implemented*].

Improve documentation

Support materials wll be improved in the following ways:

- Documenting the technical aspects of the system, focusing on the data model [REG-317], authorization rules [REG-310] and the details around master data management (see below).
- Documenting the guidelines for data editors including the decision process of merging entities and assigning IDs and codes [DP-3] [REG-316].

Grow the pool of editors

Curation tasks have in the past been handled by a small editorial team. Resources to be increased by:

- Presenting the system at the GBIF global nodes meeting and inviting GBIF node managers to assign staff to assist with specific identified tasks (arranged as a to-do list and allowing contributions and community involvement to be measured).
- Reviewing the authorization rules so that editors can be granted access to work on only those areas they are responsible for [REG-310 - *now implemented*]

Define and implement the master data management solution

Multiple metadata sources may exist for the same collection and require resolution. For example, information may be available from a metadata description associated with a specimen dataset, an existing GRSciColl entry and an Index Herbariorum record. This is a problem known as master data management (Wikipedia 2022), to be addressed by:

- Defining, implementing and documenting the approach taken by the catalogue for handling differing views of metadata [REG-319, now implemented]

Develop a richer user interface

Many improvements are possible to support users of GRSciColl, including:

- Implementing a new user interface based on visual concepts including:
    - Institution and collection search and detail pages
    - Integration of specimen-related occurrences (search, maps, gallery, detail, clustering)
    - Capability for any user to “suggest a change” [now implemented]
- Exploring citation tracking based on data mediated through GBIF for GRSciColl institutions and collections [REG-323]

- Attending to branding with a call for institutions to review their data and clear instructions on how to suggest edits.

4.2 Progress and next steps

By August 2022, progress had been made against most of the priorities in the GRSciColl roadmap:

- Editors may now be given scoped responsibility at institutional or national level. Induction webinars have been held with several nodes, resulting in a global team of 45 editors and 12 mediators actively curating content in the registry since July 2021. Training videos are being developed.
- The iDigBio collection catalogue is now powered by GRSciColl, through its open APIs. iDigBio data managers edit directly through the online editing interface.
- Documentation for editors has been developed. All GRSciColl fields are associated with an English-language description available in the online forms. Capabilities for anyone to suggest a change were deployed in May 2022. Proposed data changes are reviewed by the pool of editors and mediators before being applied.
- Capability for user interface translations is set up to support multilingual content. Editors and external communities that support the catalogue are invited to propose translations to support their work. This option has already been taken up by the Society for the Preservation of Natural History Collections (SPNHC) Biodiversity Crisis Response Committee.
- A service has been deployed allowing the linking of collections in GRSciColl to specimen records in GBIF. This has resulted in 134 million records being linked to GRSciColl entries. A basic data dashboard is now available for institutions and collections.
- Options for a richer user interface for GRSciColl are being considered within the hosted portal framework, but these depend on improvements to the data model and how data clustering can facilitate linkages between related data items.
- European nodes (e.g. through DiSSCo) are exploring adoption of persistent identifiers, such as Research Organization Registry (ROR) identifiers, with some nodes piloting use of ROR IDs on their entries.
- Integration with the CETAF registry remains an objective, but has not yet started due to the effort required to enable external editors and focus on content issues. Piloting a profile of the TDWG Collection Descriptions to capture collection-level metadata has also been delayed and will be considered for 2023.

During 2022/23, GBIF will continue to work in the following areas:

- Complete outstanding tasks to deploy an enriched GRSciColl providing search and access of collections, specimens and people.

- Focus on content of GRSciColl: cleanup of existing entries and registration of new ones by promoting use and giving training and support to editors, and promoting consistent use of codes within data shared.

- Seek to identify links between journal articles and collections based on the collection codes, within the framework of the EU-funded BiCIKL project.
- Support user interface translations for GRSciColl.
- Explore synchronization of content with the Consortium of European Taxonomic Facilities (CETAF) Registry (under development).

Funding program

H2020-EU.1.4.1.2. - Integrating and opening existing national and regional research infrastructures of European interest

Grant title

SYNTHESYS+ (submitted as SYNTHESYS PLUS), Grant agreement ID: 823827

Author contributions

Authors:

Donald Hobern: Conceptualization, Investigation, Writing – Original Draft, Writing – Review & Editing

Sarah Vincent: Investigation, Writing – Original Draft

Laurence Livermore: Investigation, Writing – Original Draft, Writing – Review & Editing

Tim Robertson: Conceptualization, Investigation, Writing – Original Draft, Writing – Review & Editing

Joseph T.Miller: Conceptualization, Investigation, Writing – Original Draft, Writing – Review & Editing

Quentin Groom: Writing – Original Draft, Writing – Review & Editing

Marie Grosjean: Writing – Review & Editing

Contribution types are drawn from CRediT - Contributor Roles Taxonomy.

Conflicts of interest

References

- Access to Biological Collections Data Task Group (2005) Access to Biological Collection Data (ABCD). Biodiversity Information Standards (TDWG). http://www.tdwg.org/standards/115. Accessed on: 2022-4-19.
- Addink W (2020) Identifiers for our institutes – GRID and ROR. https://dissco.tech/2020/04/11/identifiers-for-our-institutes-grid-and-ror/. Accessed on: 2022-4-19.
- Ariño A (2010) Approaches to estimating the universe of natural history collections data. Biodiversity Informatics 7 (2). https://doi.org/10.17161/bi.v7i2.3991
- Atlas of Living Australia (2020) Atlas of Living Australia Strategy 2020-2025. Atlas of Living Australia, Publication Series No 2, Canberra, Australia. URL: https://www.ala.org.au/app/uploads/2020/07/ALA-Strategy-2020-25-Public-June-6-2020.pdf
- Belbin L, Wallis E, Hobern D, Zerger A (2021) The Atlas of Living Australia: History, current state and future directions. Biodiversity Data Journal 9 https://doi.org/10.3897/bdj.9.e65023
- Casino A, Gödderz K, Raes N, Addink W, Koureas D, Hutson A (2019) DiSSCo Partner Capabilities Survey 2017. Zenodo https://doi.org/10.5281/zenodo.2653707
- Cobb N, Gall L, Zaspel J, Dowdy N, McCabe L, Kawahara A (2019) Assessment of North American arthropod collections: prospects and challenges for addressing biodiversity research. PeerJ 7 https://doi.org/10.7717/peerj.8086
- Collections Descriptions interest group (Ed.) (2019) Collection Descriptions (CD), in development. Biodiversity Information Standards (TDWG). https://github.com/tdwg/cd. Accessed on: 2022-11-08.
- Copas K (2020) The consultation process. https://discourse.gbif.org/t/the-consultation-process/1716
- Council of Australasian Museum Directors (2018) Australian framework for the valuation of public sector collections for general purpose financial reporting. https://camd.org.au/files/2018/11/CAMD-Collections-Valuation-Framework-1-Nov-2018.pdf. Accessed on: 2022-4-19.
- Dillen M, Groom Q, Hardisty A (2019) Interoperability of Collection Management Systems. Zenodo. https://doi.org/10.5281/zenodo.3361598
- Fichtmueller D, Berendsohn W, Droege G, Glöckler F, Güntsch A, Hoffmann J, Holetschek J, Petersen M, Reimeier F (2019) ABCD 3.0 Ready to Use. Biodiversity Information Science and Standards 3 https://doi.org/10.3897/biss.3.37214
- GBIF (2022a) Dataset classes. https://www.gbif.org/dataset-classes. Accessed on: 2022-5-15.
- GBIF (2022b) GBIF Work Programme 2023: Annual Update to Implementation Plan 2023–2027. URL: https://docs.gbif-uat.org/2023-work-programme/en/gbif-work-programme-2023.en.pdf

- Groom Q, Güntsch A, Huybrechts P, Kearney N, Leachman S, Nicolson N, Page RDM, Shorthouse DP, Thessen AE, Haston E (2020) People are essential to linking biodiversity data. Database 2020 https://doi.org/10.1093/database/baaa072
- Haston E, Hardisty A (2020) An Introduction to the Minimum Information about a Digital Specimen (MIDS) Digitisation Standard. Biodiversity Information Science and Standards 4 (e59214): 1-2.
- Hobern D, Asase A, Groom Q, Luo M, Paul D, Robertson T, Semal P, Thiers B, Woodburn M, Zschuschen E (2020) Advancing the Catalogue of the World's Natural History Collections. v2.0. GBIF Secretariat, Copenhagen. https://doi.org/10.35035/p93g-te47
- iDigBio (2021) iDigBio Portal. https://www.idigbio.org/portal. Accessed on: 2021-3-02.
- Leggatt J (2019) Valuing priceless objects: can you put a price tag on heritage assets? https://www.intheblack.com/articles/2019/12/01/can-you-put-a-price-tag-on-heritage-assets
- Leonelli S (2013) Global data for local science: Assessing the scale of data infrastructures in biological and biomedical research. BioSocieties 8 (4): 449-465. https://doi.org/10.1057/biosoc.2013.23
- Meineke EK, Davies TJ, Daru BH, Davis CC (2019) Biological collections for understanding biodiversity in the Anthropocene. Philosophical Transactions of the Royal Society B 374 (1763): 20170386. https://doi.org/10.1098/rstb.2017.0386
- Morris RA, Barve V, Carausu M, Chavan V, Cuadra J, Freeland C, Hagedorn G, Leary P, Mozzherin D, Olson A, Riccardi G, Teage I, Whitbread G (2013) Discovery and publishing of primary biodiversity data associated with multimedia resources: The Audubon Core strategies and approaches. Biodiversity Informatics 8 (2). https://doi.org/10.17161/bi.v8i2.4117
- Natural Collections Descriptions interest group (2008) (2008_ Natural Collections Descriptions (NCD), version 2008-08-12. Biodiversity Information Standards (TDWG). http://www.tdwg.org/standards/312
- Page R (2016) Towards a biodiversity knowledge graph. Research Ideas and Outcomes 2 https://doi.org/10.3897/rio.2.e8767
- Petersen M, Glöckler F, Kiessling W, Döring M, Fichtmüller D, Laphakorn L, Baltruschat B, Hoffmann J (2018) History and development of ABCDEFG: a data standard for geosciences. Fossil Record 21 (1): 47-53. https://doi.org/10.5194/fr-21-47-2018
- Robertson T, Döring M, Guralnick R, Bloom D, Wieczorek J, Braak K, Otegui J, Russell L, Desmet P (2014) The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. PLoS ONE 9(8) (e102623): 1-7. https://doi.org/10.1371/journal.pone.0102623
- Robertson T (2021) GRSciColl Roadmap. GitHub. Release date: 2021-3-05. URL: https://github.com/gbif/registry/blob/1b574a3b40e66781808624da6a7145fbd60d3355/roadmap-grscicoll.md
- Rouhan G, Dorr LJ, Gautier L, Clerc P, Muller S, Gaudeul M (2017) The time has come for Natural History Collections to claim co-authorship of research articles. Taxon 66: 1014-1016. https://doi.org/10.12705/665.2
- Sachs J, Page R, Baskauf SJ, Pender J, Lujan-Toro B, Macklin J, Comspon Z (2019) Training and hackathon on building biodiversity knowledge graphs. Research Ideas and Outcomes 5 https://doi.org/10.3897/rio.5.e36152

- Semal P, Tilley L, Theeten F, Casino A (2019) The new CETAF Registry of collections and integration of the current CETAF passport: A collection information hub for the European Natural Science Community. https://collections.naturalsciences.be/cpb/cetaf-passport-and-collections-registry-manual
- Shorthouse D (2020) How it works. https://bionomia.net/how-it-works. Accessed on: 2021-4-02.
- Sierwald P, Bieler R, Shea E, Rosenberg G (2018) Mobilizing Mollusks: Status Update on Mollusk Collections in the U.S.A. and Canada. American Malacological Bulletin 36 (2). https://doi.org/10.4003/006.036.0202
- Singer R, Love K, Page L (2018) A survey of digitized data from U.S. fish collections in the iDigBio data aggregator. PLOS ONE 13 (12). https://doi.org/10.1371/journal.pone.0207636
- Smith V, Gorman K, Addink W, Arvanitidis C, Casino A, Dixey K, Dröge G, Groom Q, Haston E, Hobern D, Knapp S, Koureas D, Livermore L, Seberg O (2019) SYNTHESYS+ Abridged Grant Proposal. Research Ideas and Outcomes 5 https://doi.org/10.3897/rio.5.e46404
- Thessen A, Woodburn M, Koureas D, Paul D, Conlon M, Shorthouse D, 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). https://doi.org/10.5334/dsj-2019-054
- Turland NJ, et al. (2018) International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, 2017. In: Turland NJ, Wiersema JH, Barrie FR, Greuter W, Hawksworth DL, Herendeen PS, Knapp S, Kusber W-, Li D-, Marhold K, May TW, McNeill J, Monro AM, Prado J, Price MJ, Smith G (Eds) Regnum Vegetabile. 159. Koeltz Botanical Books, Glashütten. https://doi.org/10.12705/Code.2018
- van Egmond E, Willemse L, Paul D, Woodburn M, Casino A, Gödderz K, Vermeersch X, Bloothoofd J, Wijers A, Raes N (2019) Design of a Collection Digitisation Dashboard. Zenodo https://doi.org/10.5281/zenodo.2621055
- 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). https://doi.org/10.1371/journal.pone.0029715
- Wikipedia (Ed.) (2022) Master data management. https://en.wikipedia.org/wiki/Master_data_management. Accessed on: 2022-11-10.
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18
- Woodburn M, Paul DL, Addink W, Baskauf SJ, Blum S, Chapman C, Grant S, Groom Q, Jones J, Petersen M, Raes N, Smith D, Tilley L, Trekels M, Trizna M, Ulate W, Vincent S, Walls R, Webbink K, Zermoglio P (2020) Unity in Variety: Developing a collection

description standard by consensus. Biodiversity Information Science and Standards 4 https://doi.org/10.3897/biss.4.59233

- Woodburn M, Buschbom J, Vincent S, Webbink K, Trekels M, Jones J, Grant S (2022) Latimer Core Guidance Documentation. https://github.com/tdwg/cd/wiki. Accessed on: 2022-9-16.

Endnotes

*1 GBIF website and associated communication channels (social media, mailing lists to all node managers, newsletter etc), the Alliance For Biodiversity mailing lists, SYNTHESYS+ mailing list and TDWG communication channels.

*2 https://github.com/tdwg/cd/tree/master/reference/use_cases

*3 GBIF 2022a - https://www.gbif.org/dataset-classes

Table 1.

Discussion topics from the virtual consultation.

Category	Topic
Use	1.1. Directory to support the collections community
Use	1.2. Locating specimens and genetic materials
Use	1.3. First step towards databasing collections
Use	1.4. Assessing the scale and value of collections
Use	1.5. Increased value for data on specimens, taxonomic publications, etc.
Use	1.6. Reducing duplication of effort
Use	1.7. Foundation for new and enriched services
Use	1.8. Improvements to citation and visibility for collections
Use	1.9. Support for national and regional needs and applications
Information	2.1. Scope for the catalogue and definition of “collection”
Information	2.2. Identifiers for collections
Information	2.3. Hierarchical collection structures and subcollections
Information	2.4. Description of a collection
Information	2.5. Wider data linkages
Information	2.6. Information services relating to collections
Technology	3.1. Pathways and tools for publishing collection records
Technology	3.2. Community catalogue
Technology	3.3. Integrated catalogue
Technology	3.4. Collection management systems
Technology	3.5. Interfaces, APIs and client modules
Governance	4.1. Ownership of information for each collection
Governance	4.2. Communities of practice
Governance	4.3. Technical infrastructures
Governance	4.4. Governance arrangements
Governance	4.5. Incentives for contributors
Governance	4.6. Funding and sustainability
Language	Adelantando el Catálogo de Colecciones de Historia Natural del Mundo
Language	Progressant le Catalogue des Collections d’Histoire Naturelle du Monde
Language	建立《全球自然历史馆藏名录》
Process	Comments on this virtual consultation process

Table 2.

Significant data standards relevant to cataloguing collections.

Standard	Description	More Information
Darwin Core (DwC)	Darwin Core is the most widely used standard for sharing data on natural history specimens and biodiversity observations. It builds on existing metadata standards (like Dublin Core) and is supported by the majority of specimen-level data repositories and community tools/platforms.	Wieczorek et al. 2012
ABCD	The Access to Biological Collections Data (ABCD) Schema is an alternative standard for specimen data. ABCD is a comprehensive, complex, structured standard for biodiversity data.	Access to Biological Collections Data Task Group 2005, Fichtmueller et al. 2019
ABCDEFG	ABCDEFG (Access to Biological Collection Databases Extended for Geosciences) is an extension to ABCD developed to support palaeontological, mineralogical and geological digitized collection data.	Petersen et al. 2018
TDWG Attribution project	A collaboration between TDWG and the Research Data Alliance to enhance existing and create new standards for giving attribution for the maintenance, curation, and digitization of physical and digital objects with a special emphasis on biodiversity collections.	Thessen et al. 2019
Audubon Core	Audubon Core (AC) is a set of vocabularies designed to represent metadata for biodiversity multimedia resources and collections of such resources. The vocabularies address such concerns as management of media, descriptions of content, taxonomic, geographic, and temporal coverage, and appropriate ways to retrieve, attribute and reproduce them.	Morris et al. 2013
Natural Collections Descriptions (NCDs)	The NCD standard arose from an earlier TDWG attempt to define a collection-level data standard. NCDs are actively used by several platforms outlined in 2.1., but subsequent development efforts stalled and as a result this standard has not been more widely taken up. The TDWG CD model (see below) is acknowledged as the natural successor/ continuation of the NCD standard.	Natural Collections Descriptions interest group 2008
TDWG Collection Descriptions (Latimer Core)	Building on earlier work in the NCD standard, the TDWG Latimer Core collection descriptions data standard will define a set of classes and properties that can be used to represent groups of collection objects and their associated information. These incorporate common characteristics used to describe, group and break down collections, metrics for quantifying those collections, and properties such as persistent identifiers for tracking collections and managing their digital counterparts. Coupled with flexible underlying data models, the CD standard is intended to support use cases from simple, high-level collections summaries to detailed quantitative collection breakdowns and assessments.	Woodburn et al. 2020, Woodburn et al. 2022

Supplementary material

Suppl. material 1: Advancing the Catalogue of the World's Natural History Collections - Consultation Materials

Authors: Hobern, D., Asase, A., Groom, Q., Luo, M., Paul, D., Robertson, T., Semal, P., Thiers, B., Woodburn, M. & Zschuschen, E.

Data type: Discussion documents

Brief description: PDF archive of materials shared as part of the consultation process on the GBIF Discourse site. All pages from this discussion are included, with separate threads for each discussion topic and additional comments in Spanish and Chinese. Daily summaries and information on the process itself and presentation materials shared.

Download file (11.96 MB)