1.8. Improvements to citation and visibility for collections (USE)

dhobern #1 March 27, 2020, 10:06am

> This is topic 1.8. in the *Uses* section of the ***Advancing the Catalogue of the World's Natural History Collections*** consultation. Use this topic to discuss the questions listed below.

Background
Research value is primarily measured in terms of visibility and impacts from published literature. Natural history collections are poorly recognised by such measures and their importance as foundational research tools is almost hidden. Users of collections are regularly urged to cite specimens examined and reference the collection. However, citation is often **https://fistfulofcinctans.wordpress.com/2016/06/23/how-and-why-to-cite-museum-specimens-in-research/**[lacking, incomplete or ambiguous]. Research infrastructures such as **https://explore.openaire.eu/search/find[OpenAIRE]** in Europe increasingly map not only linkages between researchers and publications but also datasets, projects, content providers and organisations. A catalogue could help to standardise citation of collections, making their impact visible through such knowledge graphs. Journals and editorial boards could be encouraged to require standard collection identifiers wherever collections are referenced.
Other materials

The following contributed document is particularly relevant to this topic:

- **Document: GBIF Services and Support for the Collections Catalogue**

Questions

- How might a comprehensive catalogue promote citation and attribution for collections?
- What can be done to encourage wide standardised use of identifiers from the catalogue?

Document: GBIF Services and Support for the Collections Catalogue

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

Collections Catalogue - Daily Summaries

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

建立《全球自然历史馆藏名录》

Faire Avancer le Catalogue des Collections d'Histoire Naturelle du Monde

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

建立《全球自然历史馆藏名录》

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

Advancing the Catalogue of the World's Natural History Collections

建立《全球自然历史馆藏名录》

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

建立《全球自然历史馆藏名录》

1.1. Directory to support the collections community (USE)

Integrated summary from 17 to 30 April 2020

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

建立《全球自然历史馆藏名录》

Rich87 #4 April 22, 2020, 4:16am

I think this is a very worthy goal although one that may not be as attainable as one might think.

In the herbarium community, citation of specimens is commonplace and, unless I am mistaken, one of the reasons that Index Herbariorum came into existence was so that authors would have a shorthand representation to use when citing particular collections. While it is very widely used, I believe that some journals neither

require authors to acknowledge the source of such abbreviations or to even consistently use them. It's now even a bit more important with the advent of online supplementary material that, by the admission of at least two journal policies that I have seen, such appendices are not carefully checked during the editorial process. I would hope in this age of wanting to more clearly being able to derive collection usage statistics that the attitude towards such citations would change.

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

---

Integrated summary from 17 to 30 April 2020

---

abentley #5 April 22, 2020, 9:57pm

I think this varies by discipline. The ichthyology community routinely cites the ASIH symbolic code list in publication as their source of acronyms. If we had a central source for this information that was even more informative I think it may become more common practice - in the vein of "Build it and they will come"

2 Likes

Rich87 #6 April 22, 2020, 10:25pm

I agree with both points made by abentley. Yes, this certainly varies by discipline. A common informative catalog might indeed make a stronger argument for getting editors, etc. to require collection citations.

3 Likes

---

waddink #7 April 23, 2020, 8:42pm

In addition, citation of collection Identifiers might be promoted if they are adopted in other places, like added/linked to the institute metadata in ROR, which will then also be linkable to datacite and crossref, etc.

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

---

建立《全球自然历史馆藏名录》

Integrated summary from 17 to 30 April 2020

---

sharif.islam #8 April 23, 2020, 9:10pm

Specimen/collection citation should have similar currency to publications.

We need best practices that can engage all the actors and entities involved in the lifecycle of the data (for example from data collection during expeditions to different levels of data aggregation). These engagements and best practices, along with a s sustainable technical infrastructure and standards, should entail clear and consistent guidance. However in order to sustain these best practices we need tracking and reporting mechanism as it provides incentive for the citation.

Incentive can also be provided by the funding agencies. Several funding agencies already require that data need to be open and FAIR but the requirements are less specific in terms of citations and identifiers.

1 Like

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

---

建立《全球自然历史馆藏名录》

Integrated summary from 17 to 30 April 2020

---

elyw #9 April 24, 2020, 8:13am

Citation of individual specimens, where this can be done (e.g. taxonomic treatments) is certainly important. Other use cases in this area include citation of the individual specimen from which tissue has been extracted and genetic sequencing work has been performed. In this case the issue is convincing other researchers to include a pointer back to the original specimen (as is often required as a condition of obtaining the tissue) in the repository for the results (e.g. GenBank).
More complex is the issue of citing specimens when they've been downloaded from an aggregator (such as ALA or GBIF) and used in an analysis in aggregate.

ALA and GBIF already mint DOIs to identify the download, and institutions can work back from there to the specimens. In this case, the challenge is to encourage researchers to cite these DOIs in their papers (rather than individual specimen identifiers / registration or accession numbers).
Ultimately, there are many benefits but it comes down to users of collections doing the right thing.

2 Likes

Integrated summary from 17 to 30 April 2020

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

---

Lyubomir #10 April 25, 2020, 3:44pm

@elyw you are right that citing specimens using their specimen codes (triplets) (InstitutionCode+CollectionCode+Catalog No) is tricky; the main problem is how to link the specimen codes to their resolvable GUIDs? And where is that place where the specimen GUIDs would resolve, or in other words where are the "stable http specimen IDs"?

Here is a recent blog on the topic including a proposal for introducing semantically enriched specimen data tables (appendices) in publications. Comments welcome!

- Blog: **How to get data from research articles back into the research cycle at no additional costs?**
- Exemplar paper: Patterson et al. (2020): **https://doi.org/10.3897/zookeys.929.50240**
- Proposal for a template for ontology-linked and semantically enriched specimen data in publications: **https://docs.google.com/spreadsheets/d/1h7IubZT25yh6kECM35iS4XH1TIVlZp_gv2m7lFD6SCU/edit#gid=0**

1 Like

3.3. Integrated catalogue (TECHNOLOGY)

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

---

建立《全球自然历史馆藏名录》

Integrated summary from 17 to 30 April 2020

Lyubomir #11 April 25, 2020, 4:13pm

As a publisher, I am asking myself what I would like to have as a service/infrastructure so that to be able to ensure proper citation of specimens or collections? Under "citation" I understand not just an "in-text mention" of a collection or specimen code as a text string (even it would be a well-known and widely used text string). Rather, I think that "citations" nowadays should mean "in-text pointers (mentions) to stable resolvable Specimen IDs" which could easily be linked to other digital information about a specimen (sequences, tissues, samples, literature, etc.).

Such unversal citation and linking infrastructure is still in its infancy. It is also not clear how long it would take to publishers to adopt it (I assume very long, if ever for some). Nevertheless, there is already some experience and some positive movements. So, to summarise, as a publisher I would like to have the following:

- Web service to tag Institutional and/or Collection codes in publications through GRSciCol (we have some exprience in that using the former GrBio vocabulary, see example: **https://zookeys.pensoft.net/article/51142/list/15/**)
- Asking the same for specimens would be too much at the present stage of development of collection infrastructures. Instead, we would request the authors to hyperlink their specimen codes to their either GBIF, iDigBio or DiSSCo records (the latter still to come).
- Through these links we could link a record in a specimen table to specimen's GUID, and through it to other information about the specimen, that is to its OpenDS (digital specimen) representation in the terms of ICEDID and DiSSCo.

---

dhobern #12 April 26, 2020, 12:58am

Thanks @Lyubomir . I am very interested in whether OpenRefine could give us the framework we need to handle situations such as mapping (InstitutionCode+CollectionCode+CatalogNumber+**other contextual information**) to locate CollectionID and the current and accurate version of the collection record and then determine whether the Specimen is also digitised. In the ELODINS proposal we all submitted a few years ago to seek funds to increase interlinkages between European biodiversity data sources, I called this "linking the linkable" and I still see it that way. In very many cases, a combination like (InstitutionCode+CollectionCode+CatalogNumber) is fully adequate for a trained human reader to know what collection and what specimen is referenced. We can do this even when there are some typos or glitches in the code strings, becauase we have a strong probabilistic understanding in context.

There was some initial discussion of OpenRefine options in the **Topic 3.5** thread.

Lyubomir #13 April 26, 2020, 9:43am

Thanks for commenting @dhobern . We haven't try OpenRefine for that prupose yet. For us as publishers is important to find a scalable method for linking of the triplets to the digital representation of the specimen, in most cases in GBIF. The task is not that simple as it might look like, but perhaps OpenRefine indeed could help us in that!

At present, our main goal is to set up an appropriate format for inclusion of "linked linkable" data in the articles in the form of a semantically-enriched, ontology-linked "**Appendix for primary biodiversity data**" to be implemented as an amendment to the Author guidelines across all Pensoft journal fairly soon. Comments are welcome and will be much appreciated!

- Draft amendment: **https://docs.google.com/document/d/1AXlKozlb9J70sXY-dc7O8WHEFU8OC_oJBwUTWpCC4KE/edit**
- More detail in the related blog: **How to get data from research articles back into the research cycle at no additional costs?**

---

dhobern #14 April 26, 2020, 9:49am

Thanks @Lyubomir . I believe that one of the major uses for the **integrated catalogue** should be for it to offer precisely this kind of service, combining structured data, machine learning and human curation to offer APIs that take sets of informal identifiers such as InstitutionCode, CollectionCode and taxonomic group to return the identifier for the matching collection (or a ranked set of likely matches).

Lyubomir #15 April 26, 2020, 10:18am

Fine! We would be happy to contribute as the publishing side is concerned. The implementation of the semantic Appendix in question would facilitate tracking and preserving some missing links between peer-reviewed, published, hence vetted data. Links by themeselves are vetted too by the authors by putting these together in a formally published linked table.

We could perhaps contribute to the Catalogue **under Topic 1.8** through open APIs to some of our resources, e.g. **https://refindit.org**, **https://openbiodiv.net** and others.

Lyubomir #16 April 26, 2020, 10:26am

Fully agreed and happy to participate! We would treat the links between triplets and Occurrence IDs, for example, as curated, high-quality data, if published in a peer-reviewed article following a pre-defined format. Such links will be RDF-ized and can be discovered and harvested by aggregators and other users along with their provenance record (=incentive for the authors to put a little (indeed a little!) more effort to properly publish their data).

---

agosti #17 April 26, 2020, 8:48pm

Collections codes are routinely tagged in the Plazi workflow that leads to 29K datasets, that is publications, in GBIF. They are also available in the materials citations aka ocurrences eg **https://www.gbif.org/occurrence/2608702029** extracted from a publications: **https://doi.org/10.11646/zootaxa.4407.1.2** and the respective taxonomic treatment. **http://tb.plazi.org/GgServer/html/676A87E0FFA3B14D2392FD6C47540BD1** , and the collection code is annotated with the respective identifier **http://biocol.org/urn:lsid:biocol.org:col:34871**

This workflow existed before the the GRSciol broke down, and now that GBIF resumes this service, the identifiers even resolve.

Together with CETAF EJT a guideline exists on how to publish materials citation so that this extraction is more efficient. **https://doi.org/10.5852/ejt.2019.586**

Unfortunately the collection codes have not been included because the publishers, editors, authors, scientists are very far away from realizing this potential, and the focus has been on suggesting, that materials citations should be published in a standardized way to make this data type digitally accessible knowledge. Also a potential to find out how often specimen are being used - what has been mentioned above.

In the case of GBIF, it is now possible to go from a collection code to an "occurrence" aka materialsciation, to the taxonomic treatment, to the respective publication. It also allows increasingly, based on scholarly published data, to know who collected which specimen from a collection and many other thing, all in the respective data in the submitted record.

1 Like

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

---

建立《全球自然历史馆藏名录》

Integrated summary from 17 to 30 April 2020

agosti #18 April 26, 2020, 8:52pm

There is another aspect. Publications are a very rich source of collection codes and extensiions that do not exist in GRSciol, but are obviously used by some scientists.

In taxonomic publications, generally there is a section in Materials and Methods that lists the collection codes and the full extensions that are later used in the materials examined section within the respective treatments. Very regularly, collections codes occur that we cannot match with records in GRScicol, and we would like to add. This would rapidly add to an extension of collections codes, for which there are materials citations ("occurrences") attached.

2 Likes

agosti #19 April 26, 2020, 8:57pm

Within the CETAF publishing group, also in collaboration with Pensoft, we currently are looking into pulling together an overview of identifiers used, and could be used within the publishing of scientific articles. It would probably a good thing for this group to contact here to collaborate and provide suggestions.
Since these publishers are aware of what is happening at GBIF, it would be productive if this group here could help to provide a mechanism to submit by machine collections codes that are not yet available.

agosti #20 April 26, 2020, 9:04pm

Finally, taxonomic treatments are now a sub data type within the Biodiversity Literature Repository at Zenodo of the DOI text type. There we can add the collection codes as custom metadata, using the respective DWC term, which allows to discover citation of usage of materials from collections. See eg **https://doi.org/10.5281/zenodo.3730231**. Since in many ways these treatments are the end of an unfortunate data liberation workflow (often based on closed access PDF articles), the representation of collection codes will have problems, such as a string of collection codes instead of each on individual.
For that reason, I would suggest to work with the CETAF publishing group/Pensoft to develop ways so we can generate data for immediate use by starting with recommendation, how collection codes should be published in the future.

benichou #21 April 27, 2020, 7:33am

We all agree that linking publication to Institutionscode/ collectionCode / Catalog number is clearly something that would definetely benefit to all the community :

- It would enable institutions to track all publications citing their collection and think of more relevant metrics
- Could increase citability for authors
- Would enable publisher to increase their visibility, and comply with FAIR data

However, as pointed out by Lyubo that is not so easy to achieve and he is right to think that it might take a long time for publishers to adopt such best practices. It is still very difficult so far to implement standards to properly publish (so it can be linked) the data as the authors themself are not always aware of the importance of it. We'll have to work on several front forward this goal, not only in terms of technical improvement. The set up and promotion of best practices in terms of linking data and properly published is one way on which EJT is working alongside Pensoft.

The guideline on how material citation should be structured so as to be accurately extracted was the first step (**https://doi.org/10.5852/ejt.2019.586**).

The second one is done within CETAF Publishing group to encourage and promote the use of relevant identifiers to link data in the articles on ongoing publcation.

Being able to offer APIs that take sets of identifiers such as InstitutionCode, CollectionCode and taxonomic group to disambiguate the information before publication, when editing the article would be a great improvement for publishers and/authors.

To quote Donat « For that reason, I would suggest to work with the CETAF publishing group/Pensoft to develop ways so we can generate data for immediate use by starting with recommendation, how collection codes should be published in the future. »

Finally, what seems to be missing in the discussion is the link with librarian and legacy publication.

1 Like

qgroom #22 April 27, 2020, 9:15am

> dhobern:
>
> What can be done to encourage wide standardised use of identifiers from the catalogue?

I thought it would be interesting to see what the International Code of Nomenclature for algae, fungi, and plants had to say about citing collections. I've posted the relevant part of article 40 below.
Apparently, it is very lax on the subject with only a suggestion to use Index Herbariorum codes, but institution and collection names in any language are allowable. In general I think the IAPT would be reluctant to make its rules more concrete on this issue, even though it would help interoperability. For political reasons it is not always easy in some countries to engage with international organizations and I think the IAPT would not want to block someone from engaging in taxonomy, because they can't register their institution/collection.

40.7. For the name of a new species or infraspecific taxon published on or after 1 January 1990 of which the type is **a specimen or unpublished illustration, the single herbarium, collection, or institution in which the type is conserved must be specified** (see also **Rec. 40A.5** and **40A.6**).

Ex. 8. In the protologue of *Setaria excurrens* var. *leviflora* Keng ex S. L. Chen (in Bull. Nanjing Bot. Gard. 1988–1989: 3. 1990) the gathering *Guangxi Team 4088* was indicated as "模式" [type] and the herbarium where the type is conserved was specified as "中国科学院植物研究所标本室" [Herbarium, Institute of Botany, The Chinese Academy of Sciences], i.e. PE.

Note 4. **Specification of the herbarium, collection, or institution may be made in an abbreviated form**, e.g. as given in Index Herbariorum (**http://sweetgum.nybg.org/science/ih/**) or in the *World directory of collections of cultures of microorganisms* .

Ex. 9. When 't Hart described *"Sedum eriocarpum* subsp. *spathulifolium"* (in Ot Sist. Bot. Dergisi 2(2): 7. 1995) the name was not validly published because no herbarium, collection, or institution in which the holotype specimen was conserved was specified. Valid publication was effected when 't Hart (in Strid & Tan, Fl. Hellen. 2: 325. 2002) wrote "Type … 't Hart HRT-27104 … (U)" while providing a full and direct reference to his previously published Latin diagnosis (**Art. 33.1**).

From Turland et al. (eds.) 2018: *International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017* . Regnum Vegetabile 159. Glashütten: Koeltz Botanical Books. DOI **https://doi.org/10.12705/Code.2018strong text**

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

---

建立《全球自然历史馆藏名录》

Integrated summary from 17 to 30 April 2020

---

sabaj #23 April 28, 2020, 1:36am

Hello everyone. Since 2010, I have maintained a reasonably complete list of annotated codes for historical and modern natural history collections associated with lost and extant specimens of fossil and Recent fishes, amphibians and reptiles.

The list currently includes 3784 codes anchored to about 2033 collections or institutions in 150 countries. A slightly dated version of the list is available here: **https://asih.org/standard-symbolic-codes/about-symbolic-codes**. An updated list was submitted for print publication in Copeia.

This list was generated by combing the fish and herp literature (1769 publications) for citations of codes and with help from fish/herp taxonomists and collections staff around the world.

Anyway, I can attest that authors cite specimens using a lot of different codes, some stable, many not. Given the way taxonomists work…it might be easier (at least at the beginning) to tie together codes on the "back end" rather than have taxonomists employ standardized codes on the front end (e.g., published works).

2 Likes

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

---

建立《全球自然历史馆藏名录》

Integrated summary from 17 to 30 April 2020

---

dhobern #24 April 28, 2020, 1:57am

Thanks @sabaj . You are correct that much of the integration will have to be after-the-event cleanup of a muddle of collection codes. Of course, these codes will in many cases not be unique, so we will need to experiment with frameworks like OpenRefine to find ways to simulate the contextual interpretation that allows human readers usually to understand the reference even in the most muddled cases.

2 Likes

---

agosti #25 April 29, 2020, 10:41am

I agree with @sabaj that a way to go is to collect what users of collection codes do and then create some sort of lookup table. Here is a glimpse into the collectionCodes we extract from publications. Be aware, it is dirty data, not least because of the botanists that use single letter as collection codes which makes it very difficult to mine. But this might be another starting point.
For each of the collection code we have the treatment, and the publication from where we extracted the data

Here is a CSV http://tb.plazi.org/GgServer/srsStats/stats?outputFields=colls.code+colls.name&groupingFields=colls.code+colls.name&format=CSV&separator=%2C output, and you can get more through using the Plazi stats at http://tb.plazi.org/GgServer/srsStats

---

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

建立《全球自然历史馆藏名录》

Collections Catalogue - Daily Summaries

Integrated summary from 17 to 30 April 2020

---

gdadade #26 April 29, 2020, 12:47pm

GGBN already uses stable identifiers of specimens (plus the traditional triplet) to reference tissue and DNA material back to specimens and some of their members also provide stable identifiers for the DNA and tissue samples already. Adding those resources to publications would improve visibility of (biobank) collections since not in all cases it will be possible to deposit a voucher specimen. Very often the physical tissue and DNA sample is simply forgotten when citing used material, so are the biobank collections in general. A centralised catalogue with identifiers for these important collections plus identifiers for the DNA and tissue samples will help to make them more visible and enable traceability and transparency.

---

sabaj #27 April 30, 2020, 1:13am

> gdadade:
>
> GGBN

A note about documenting tissues in publications…it is often done poorly for fishes. At least 4 identifiers might be associated with a fish tissue:

1. Collection code: ANSP
2. Specimen catalog number: 123
3. Specimen tag number: 456 (fishes generally kept in lots, so one lot might contain multiple tissue vouchers)
4. Tissue catalog number: 789
   It is not uncommon for authors to only report 3 or 4 without specifying which one.

Also…when tissue subsamples are sent from one collection to another, the receiving collection might assign its own Tissue Catalog Number…which adds two more identifiers "Code + catalog number for new collection with tissue". One might ask, why doesn't the receiving collection simply retain the original identifiers assigned by the source collection? Most of the time they do…and that is how unused and leftover tissues often get lost in freezers. Better for the receiving collection to newly catalog the tissue (alongside data on its source) so that the receiving collection may keep track of it.

Finally…whereas most fish tissue/DNA collections employ the same **code** for specimens and tissues (e.g., ANSP)…a few collections use separate codes, such as:
Museu Nacional, Universidade Federal do Rio de Janeiro uses MNRJ (fishes, herps), MNTI (fish tissue collection), MNLM (vertebrate DNA extract collection)
Museu de Zoologia da Universidade de São Paulo uses MZUSP (fishes, herps), CTMZ (tissue collection) or MZict (fish tissues)
Instituto de Investigación de Recursos Biológicos Alexander von Humboldt uses IAvH-CB (Colecciones Biológicas), IAvH-CT (tissue collection), IAvH-P (fishes), IAvH-Am (amphibians), IAvH-R (reptiles) - **I prefer this b/c the full code preserves the same institutional for each collection**
**Fish Genetics and Biotechnology laboratory, Indian Council of Agricultural Research uses NKGMF (fish specimens), VIZNK (tissue samples)
University of Kansas Biodiversity Institute uses KUI (fishes), KUIT (fish tissues), KUH (herps), KUVP (fossils)

All is to say…tracking tissues (and DNA) for fish collections adds a significant amount of complexity. Best to develop a stable list of institutions first…then the collections contained within them.

1 Like

---

gdadade #28 April 30, 2020, 6:42am

> sabaj:
>
> Also…when tissue subsamples are sent from one collection to another, the receiving collection might assign its own Tissue Catalog Number…which adds two more identifiers "Code + catalog number for new collection with tissue".

That is very true and mainly because the sending institution has no proper tissue catalog numbers. Mostly they use the same number for both specimen and tissue which is a big problem we are working on in GGBN. So for the data portal they must make the tissue catalog numbers unique by adding a prefix or suffix for example. Also often the same DNA number is used if extracted again, so noone can ever track back where the sequences are really coming from.

Within the EU funded SYNTHESYS+ project we are currently working on solution to enable a network of trusted biobanks with standardized exchange of material, like the botanical gardens community does with IPEN. Their IPEN numbers work perfectly fine for their purposes. We hope to establish something similar for GGBN and hence for all biobanks.

It is the same problem in any organism group and becomes even worse when we talk about parasites or environmental samples. That the fish is also a fish and not only an environment or host is often forgotten in such cases ☹️ 🐟

Don't know if you ever tried to search fish samples in GGBN. Vertebrates in general is the group currently best covered taxonomocially in GGBN. So might be worth a try. We are always looking for constructive feedback.

1 Like

---

pzermoglio #29 April 30, 2020, 8:18am

From **@WUlate** in this **Spanish thread**

In the same way that researchers deposit sequences in GENBank (or other) to publish, the Catalogue could be part of the best practices of any scientific publication if:

- we start with those already here (early adopters),
- we prepare directives and agreements (SYNTHESIS+, GBIF, Nodes, etc.) that recommend the adoption of the Catalogue,
- we announce the initiative to the world and gain advocates in the organizations that manage the collections,
- but we offer useful products soon (first maybe small and focused) that take advantage of the value of the Catalogue y contribute to real situations (after, some larger and more complex).

1 Like