2.3. Hierarchical collection structures and subcollections (INFORMATION)

dhobern #1 March 27, 2020, 10:20am

> This is topic 2.3. in the *Information* section of the ***Advancing the Catalogue of the World's Natural History Collections*** consultation. Use this topic to discuss the questions listed below.

Background
Within IH, each herbarium record usually corresponds to an institution with its own unique collection code, street address, etc. Within zoology, museums are often structured as a set of collections with differing and possibly hierarchical taxonomic scope. Specimens collected on famous expeditions or by significant researchers may have their own identity and appear as special collections. As a result, curators and researchers may wish to refer to different (potentially overlapping) sets of specimens as separate collections with their own names, identifiers and descriptions.

Other materials

The following contributed materials are particularly relevant to this topic:

- **Presentation: Index Herbariorum - Perspectives of a Long-Standing Collection Catalog**
- **Presentation: TDWG Collection Descriptions Data Standard Task Group**

Questions

- Should the catalogue support hierarchical relationships between collections (and collection records)?
- If so, how do parent-child relationships work, and do we infer information from parent to child or vice versa?

Presentation: Index Herbariorum - Perspectives of a Long-Standing Collection Catalog

Presentation: TDWG Collection Descriptions Data Standard Task Group

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》

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》

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》

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

Collections Catalogue - Daily Summaries

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

---

ekrimmel #4 April 21, 2020, 7:09pm

Situations where one collection was orphaned and subsequently adopted by another collection might be represented nicely by a hierarchical relationship. Or perhaps not; perhaps this would be better represented by a pointer from the inactive collection to its adoptive parent (which I think is how IH does it, e.g. **http://sweetgum.nybg.org/science/ih/herbarium-details/?irn=126715**)…

3 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

Rich87 #5 April 22, 2020, 4:02am

I agree - knowing where a collection is now located is important.

After reading the intro to this topic, another thought comes to mind - the user needs to be able to find both all of the collections at an insititution as well as individual collections easily. Asking a variety of potential users about their intended use of the catalog (what are they trying to find, etc.) might be a useful avenue to pursue.

1 Like

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

Integrated summary from 17 to 30 April 2020

---

abentley #6 April 22, 2020, 9:33pm

I think it would be useful to have three levels of hierarchy - institution, collection and dataset. This would allow for situations where collections are split into various pieces - be they orphan collections, research collections or simply different parts of a single collection. This level could be optional.

---

cboelling #7 April 27, 2020, 10:47am

I believe that institution, collection and dataset should be modeled as complementary elements rather than levels of a hierarchy.

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

---

abentley #8 April 27, 2020, 2:25pm

Can you explain what you mean by complementary elements?

---

dhobern #9 April 27, 2020, 11:30pm

Hi @abentley . I interpreted @cboelling to mean that we should treat these as three different types of digital object, each with their own representation and data standards (metadata, etc.), although with some well-defined relationships. This is in contrast to the idea (implied in your previous message) that these all belong to the same class of object. It may be that it was your earlier message that was misleading. Perhaps you simply meant that we needed to support hierarchies of sets of specimens all as nested collections, including a “super-collection” that represents all the materials held by an institution, a “collection” that represents a management unit that typically has a recognisable CollectionCode, and an “infra-collection” that represents a potentially unnamed set of specimens of interest in some context.

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》

Integrated summary from 17 to 30 April 2020

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

abentley #10 April 27, 2020, 11:59pm

@Dhobern and @cboelling . I agree that each of these hierarchical classes requires its own set of metadata and representation but some overlap will occur in fields of information and a true hierarchy can still be established between them. Not all three would have to be required as there may be a 1:1 relationship between institution and collection in some cases and also between collection and dataset. I think GRBio had a data model for the first two but not for datasets that may be prescribed by a project, person or orphan collection with its own identifiers.

cboelling #11 April 28, 2020, 9:28am

I find it helpful to separate thinking about the real-world entities from thinking about the information structures used to represent them in an information system (for a given set of use-cases - like the planned catalogue). The former informs the latter.

In addition, I understand a hierarchical relationship to hold when entities are related to one another by a relation that puts them in a nested sequence of increased/decreased generality:

a cat *is_a* mammal — a mammal *is_a* vertebrate.
collection-A *is_part_of* collection-B — collection B *is_part_of* collection-C

In this sense I think that institutions, collections and datasets stand in various relations to each other, but they don't form a hierarchy. I tried to sketch the basic relations between them in **this post in topic 2.1**. (I am actually unsure if the term "dataset" is used here in its general sense or in some special sense with regard to collections that I have not understood.)

> abentley:
>
> each of these hierarchical classes requires its own set of metadata and representation but some overlap will occur in fields of information and a true hierarchy can still be established between them. Not all three would have to be required as there may be a 1:1 relationship between institution and collection in some cases and also between collection and dataset.

I agree that one can and should exploit commonalities of the properties of the real-world entities for designing the information structures used to represent them. In this vein, it is certainly an option to arrange and aggregate data fields so that on the representation level (in the data structures of the information system) there is a hierarchy of representations. However, in my experience a design that mirrors the distinctions of the real-world entities we want to capture is often easier to maintain and easier to extend in the future if new functionalities (e.g., representation and query of more complex relations between and among institutions and collections or other entities which we don't consider now) are desired. I therefore would argue that also on the representation level a collection should indeed always be treated separately from its holding institution - even if, in the real world, all the holdings of an institution are subsumed in a single collection.

> dhobern:
>
> we needed to support hierarchies of sets of specimens all as nested collections, including a "super-collection" that represents all the materials held by an institution, a "collection" that represents a management unit that typically has a recognisable CollectionCode, and an "infra-collection" that represents a potentially unnamed set of specimens of interest in some context.

I would agree that this was a desirable feature and this approach keeps the clear distinction between the institution and the (possibly nested) set of collections it administers.

As you, @dhobern , note, the relations between two collections may be more complex: when they aren't nested but there is a partial overlap with regard to the specimens each collection subsumes. Given adequate data from the relevant sources this could be represented in the catalogue as a separate symmetric relation between collection (A *overlaps_with* B). Parthood (non-symmetric) discussed in this topic could then be treated as a special case of overlap (A *part_of* B ==> A *overlaps_with* B). (Also the relations between historical and contemporary collections and merge / split situations **discussed in the context of identifiers** could be modeled as sub-relations of such an overlap-relation.)

Regarding the second question of this topic, I think that if collection B is identified as a sub-collection of collection A (in the sense that B is part of A) then it is reasonable to conclude that what applies to the whole collection also applies to the sub-collection, so there is inference from parent to child but not vice versa. However, I think this very much depends on the kinds of attributes available for a collection entry. There might be edge cases where based on the parthood no other information can be legitimately transferred from the parent to the child. So may be, until there is a better understanding *what* information can be safely copied to sub-collections on the side of the catalogue indicating parthood on the basis of the provided information is what the catalogue should aim for.

neilcobb #12 April 28, 2020, 3:40pm

You also need to account for non-hierarchical structures. There are enough institutes with collections that cover the same broad theme (e.g., insects) but they are housed in different colleges or departments. Even if they are not hierarchical in practice it would be nice to see the institution take responsibility and provide an organizational chart for reference.

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》

Integrated summary from 17 to 30 April 2020

---

neilcobb #13 April 28, 2020, 7:20pm

Finally, I would suggest a different term than subcollection if there is a need for a category below collection (I agree there is a need). But I do not know anyone that would enjoy listing themselves as a curator of a subcollection, a term like “special collection” would be more palatable and I would guess more universally accepted. I am sure there are other terms, but please not subcollections.

---

dhobern #14 April 28, 2020, 11:49pm

Thanks **@neilcobb** - I was definitely not suggesting different terms than “collection” for different examples in the hierarchy. My “super-collection” and “infra-collection” were terms I selected because I thought they would avoid any suggestion I took them seriously as names. My view is that we should call all of these sets of materials “collections” without any rank-based qualifiers. Instead we should try to recommend standard qualifiers that will be useful to users and curators. I don’t know what these would be and how many categories we would think are worth distinguishing but terms that look like “institutional collection”, “special collection”, “research collection”, etc. would be better than “subcollection”, “megacollection”, “nanocollection”, etc.

1 Like

---

neilcobb #15 April 29, 2020, 1:34am

@dhobern Donald, I have to admit it I would love to have the title of “Megacollection Curator”.

1 Like

---

pzermoglio #16 April 29, 2020, 6:35pm

From **@maperalta** in this **Spanish thread**

The hierarchical relationship is useful, as it allows to keep grouped accessory collections (for example, a palinoteque) or collections with a distinctive characteristic (for example, the Archbold Expedition collection in the AMNH mammals collection). This facilitates their management.

---

pzermoglio #17 April 30, 2020, 7:52am

From **@WUlate** in this **Spanish thread**, probably related to:

> Rich87:
>
> I agree - knowing where a collection is now located is important.
>
> After reading the intro to this topic, another thought comes to mind - the user needs to be able to find both all of the collections at an insititution as well as individual collections easily. Asking a variety of potential users about their intended use of the catalog (what are they trying to find, etc.) might be a useful avenue to pursue.

The hierarchy of collections is dynamic, and sometimes they intersect. What for someone is a single institutional collection, for others are multiple collections (including a Dry Collection, a Seeds Collections, a Logs Collection, a Wet Collection). These can even be physically distributed in various locations and even in distinct institutions.

---

cboelling #18 April 30, 2020, 9:07am

> pzermoglio:
>
> The hierarchy of collections is dynamic, and sometimes they intersect. What for someone is a single institutional collection, for others are multiple collections (including a Dry Collection, a Seeds Collections, a Logs Collection, a Wet Collection). These can even be physically distributed in various locations and even in distinct institutions.

Such relations could be represented by a carefully drafted set of relations, as argued **here**.