3.5. Interfaces, APIs and client modules (TECHNOLOGY)

dhobern #1 March 27, 2020, 10:45am

> This is topic 3.1. in the *Technology* section of the ***Advancing the Catalogue of the World's Natural History Collections*** consultation. Use this topic to discuss the questions listed below.

Background
The value of a shared commons-based resource can be maximised by ensuring that interfaces and APIs support the needs of all key stakeholder groups, including addressing issues around content delivery to the fullest extent possible in multiple languages. Some needs may be addressed by offering reusable client components that can be embedded in other applications.

Other materials

The following contributed materials are particularly relevant to this topic:

- **Presentation: TDWG Collection Descriptions Data Standard Task Group**
- **Document: GBIF Services and Support for the Collections Catalogue**

Questions

- What interfaces and APIs are required to maximise access to the collection catalogue?
- How can the catalogue best support diverse user communities, including speakers of different languages?

Presentation: TDWG Collection Descriptions Data Standard Task Group

---

Document: GBIF Services and Support for the Collections Catalogue

---

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

---

Integrated summary from 17 to 30 April 2020

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》

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

---

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

建立《全球自然历史馆藏名录》

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

---

dshorthouse #4 April 16, 2020, 3:55pm

OpenRefine, perhaps also with other properties that adjust scoring of reconciled searches **https://reconciliation-api.github.io/specs/0.1/**

2 Likes

3.1. Pathways and tools for publishing collection records (TECHNOLOGY)

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

建立《全球自然历史馆藏名录》

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

Integrated summary from 17 to 30 April 2020

---

dhobern #5 April 18, 2020, 8:31am

Thanks, **@dshorthouse** - OpenRefine is certainly promising for our needs in this and in many other situations for handling data federation and linkage. If we are indeed to recognise and link references to collections based on shared attributes, it strongly reinforces the need for clarity and good alignment between 1) our **definition of what a "collection" is** and what it means for us to say we are referring to the same collection, 2) the **data standards and attributes** we use to describe a collection, and 3) how we adopt OpenRefine or similar protocols.

I see you **have just deployed** an OpenRefine reconciliation endpoint for **Bloodhound**. Are there any immediate lessons to apply to collection reconciliation?

---

dshorthouse #6 April 18, 2020, 1:41pm

There are indeed some immediate lessons and you've highlighted them here. I'll get into more specifics.

It was new to me that OpenRefine reconciliation endpoints can receive properties and values that may be used to enhance the likelihood score on a term being reconciled – this is a significant feature and it does deserve careful consideration. In particular, it will force us to think very carefully about types and properties (what you called attributes) from the perspective of a user who needs to turn strings to things and in what context they are working. Are they resolving a reference list and want a linked & citable reference to a collection? Are they wanting a linked collection code for a collection? Are they wanting a collection's parent institution? Are they wanting the collection's name in an alternate language?

Wikidata, understandably, has a robust reconciliation endpoint and if I'm not mistaken, is available in OpenRefine at install. It would be in our best interest to model what is a collection in wikidata. The immediate advantage of doing so is access to users who can verify that the reconciliation endpoint that results from this work suits their needs.

---

waddink #7 April 23, 2020, 8:02pm

The mockups for the catalog, although visually appealing, seem to distract me a little from the use cases that indicate that an api and machine actionable metadata might me more important than a human browsable listing of collections. For instance the need to use collections descriptions in a service like ELViS. Also they do not show anything about possible operations, metadata, standards, update mechanisms, provenance 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

---

dhobern #8 April 24, 2020, 12:21am

Thanks **@waddink** . I think the mockups can be seen in two ways:

- A chance to show what GBIF or another aggregator could do to make useful information about collections more accessible and integrated - all of it (and more should equally be accessible and useful via APIs and other interfaces)
- Raising the vision expectation among collection managers and the wider community about what aggregators can and should deliver - and helping to build the case for us all to make this happen (which is also the goal of this consultation)

You are correct that we need an exciting vision of how services such as ELViS can be fully integrated into the framework of a catalogue. I've just added my own views **here under the TDWG CD Presentation**. With a model like the one discussed there, I believe we could readily start plugging many standard and bespoke services into models like this mock-up.

trobertson #9 April 24, 2020, 7:00am

Thanks, **@waddink** . I can provide some background on the evolution of the visuals-

The approach evolved from over a series of discussions starting with a *very* basic sketch. Initially, GBIF was exploring what could be done to improve the representation of information feeds available and enhance GRSciColl. Since many contributors are non-technical, we found it useful as a mechanism to validate if people understood what each other were saying. It was/is an exercise in seeing if this is would be useful to design and build, rather than a technical design for how it should be built.

An example of an improvement GBIF desperately *need* to progress relates to "duplicate" occurrence records. For example **this museum record** has **a sequence record** and was cited in **this treatment**. The clustering algorithms we're exploring allow us to detect these and better represent data e.g. **the specimen illustration**.

When it comes to the technical aspects in addition to the aspects you raise, I'd comment that:

- Similar to **GBIF.org**, I assume the information would be available through open APIs in addition to user interfaces
- Many well-established infrastructures are in place and we should look to connect, partner, and build on those initiatives wherever possible. 100's millions specimens and collections from ~90 countries are already sharing data through connected infrastructures, there are now well-functioning identifier mechanisms in place to partner with etc. I favor a step-wise approach to integrate and bolster these efforts.
- We've learned that a one-size-fits-all solution rarely works. Collection metadata, citations, and specimen related data *will* be curated in a variety of systems and we should design for that flexibility and be agile.
- The technical threshold for participation needs to be low and open to all. At the simplest, we need entry points that support those using Excel and a web form.

1 Like

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

Integrated summary from 17 to 30 April 2020

waddink #10 April 24, 2020, 8:09am

Thanks **@trobertson** . For me is it already clear that this would be useful so rather than to further discuss the if and why I am more interested in discussing how this is going to be technically implemented. But I am probably going too fast here, there should be enough room still to discuss the if and why to see if others also conclude that this is useful. When it comes to the technical aspects I agree with all the additional aspects you mention.

cweiland #11 April 28, 2020, 2:50pm

> trobertson:
>
> The technical threshold for participation needs to be low and open to all. At the simplest, we need entry points that support those using Excel and a web form.

That's clear that one has to find the right balance between sth understandable, usable by humans but at the same time parsable, checkable by machines … Just thinking a bit towards enabling more rigid validation (=raising the threshold): More rigid requirements/validation helps users to understand the semantics of the data structures used. At least for me validation is much easier implemented and maintained using a web ui (or bulk upload via api) converging into a yaml, json-ld and finally rdf pipeline… I think the maintenance of the "low threshold interface" is a pain with sth like Excel, if there are significant changes in the data model upstream the validation pipeline … but maybe that's a question of available capacities/human resources to support these entry points.

建立《全球自然历史馆藏名录》

Collections Catalogue - Daily Summaries

Integrated summary from 17 to 30 April 2020

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo