Presentation: TDWG Collection Descriptions Data Standard Task Group

dhobern #1 April 6, 2020, 9:39am

> This presentation is particularly relevant to discussions under the following topics:
>
> - 1.3. First step towards databasing collections (USE)
> - 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)
> - 3.1. Pathways and tools for publishing collection records (TECHNOLOGY)
> - 3.5. Interfaces, APIs and client modules (TECHNOLOGY)

Presentation on the TDWG Collection Descriptions (CD) Working Group and data standard from Matt Woodburn (NHM London) and Deb Paul (iDigBio), April 2020 - available as Powerpoint or as PDF.

The TDWG CD standard offers a generalised data model for capturing and sharing information on natural history collections.

2 Likes

2.1. Scope for the catalogue and definition of "collection" (INFORMATION)

Document: GBIF Services and Support for the Collections Catalogue

---

Advancing the Catalogue of the World's Natural History Collections

---

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

---

1.3. First step towards databasing collections (USE)

2.2. Identifiers for collections (INFORMATION)

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

---

2.4. Description of a collection (INFORMATION)

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

---

建立《全球自然历史馆藏名录》

建立《全球自然历史馆藏名录》

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

---

Integrated summary from 17 to 30 April 2020

abentley #2 April 22, 2020, 9:51pm

I would like to see a field for database system used included in the CD data standard. I have brought this up before but do not see it included. This would be helpful to the collections and software communities for various reasons.

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

Integrated summary from 17 to 30 April 2020

mswoodburn #3 April 23, 2020, 4:34pm

Thanks @abentley , I'll add this suggestion to the issues in the group's Github repo. To pick your brains a bit more, do you mean this to reflect the original source database of the collection description data, rather than where the current record is stored e.g. in an aggregator? Should it be just the software (e.g. Specify, MySQL, Excel…), or include the specific implementation (e.g. Specify - MfN, MySQL - Join the Dots database, NHM…). And finally, what would you envisage the collections/software/research communities using the information for?

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

abentley #4 April 23, 2020, 5:44pm

Thanks @mswoodburn . I was just thinking of a field to articulate what software package the collection was using as its CMS. I guess there could also be a field for version # too. There are actually three communities that could use this information: 1) Collections - communities of collections using the same software would encounter similar digitization, databasing, data management and publishing issues and allowing a community of those users to self identify would allow them to assist each other with implementation questions/problems/issues; 2) The software community - it would assist them in getting a sense of who is using their software (especially open source packages that do not have traditional self-identifying user bases) for community and user engagement; 3) The research community - researchers wanting to use collections data for various purposes may gain from understanding the data model associated with the data being published - data model, relationship, field restrictions or interactions.

1 Like

mswoodburn #5 April 23, 2020, 5:59pm

That's very useful, thanks @abentley - I can definitely see the value in that. That might also open up the potential to link those records to public resources for those systems - schemas in Github, forums etc, and also be potentially useful metadata for any specimen records linked to the collection, without necessarily having to include that field on each individual record.

waddink #6 April 23, 2020, 7:01pm

Am not sure I agree on this. I see the usefulness of having an overview of CMSes in use, but this seems to me not a property that describes a collection. I think it is rather a property to describe in a catalog of institutes as part of describing their facilities and services. Of course you can link them together with collections in a service using the institute id if needed.

dhobern #7 April 24, 2020, 12:12am

Thanks @waddink - I can see the value. Here are two reasons:

Given our experience with data publishing over the last couple of decades, we know that different data holders manage and can provide the same information in different ways. So long as we understand the implications of the CMS being identified, we can infer up or down from institution to collection (just as we often do

from dataset to specimen/occurrence or from checklist to species).

More importantly, I can see ways that this could progress beyond simple identification of the CMS and version. It could be or could serve as an advertisement of the digital services available for the collection offers that may exceed what is available through a static CD document or through standard data downloads. Imagine a situation where we defined a set of standard (TDWG-developed) service types that could be implemented by a CMS (retrieve digitisation summaries, submit specimen annotation, initiate loan request, etc.). This could help GBIF, DiSSCo, iDigBio, etc. to start bridging the gap between push-based data publication and the more interactive world we all want. The CMS can offer services that enhance its digital accessibility and value. The aggregating infrastructures can supply the glue that uses these service definitions and showcases the collections' capabilities. New specialist tools could be built on the same framework. Effectively the CMS to aggregator interface can become a many-to-many client-service framework.

3 Likes

3.5. Interfaces, APIs and client modules (TECHNOLOGY)

---

trobertson #8 April 24, 2020, 7:42am

I think the question @waddink raises is whether this information is seen as a *collection descriptor* or stored elsewhere. For example, DiSSCo proposes to use Digital Object Architecture (DOA) and Digital Object Interface Protocol (DOIP), whereby the operations one could perform would be registered alongside the object while the descriptive content would likely be embedded in metadata within it.

It makes good sense to me that simple requests like @abentley offers are accommodated to help ensure many can participate. I don't expect CD covering this will restrict other ideas and may help advance them (e.g. DiSSCo infrastructure could infer technical endpoints knowing the collection is available on Specify v8.2).

1 Like

Adelantando el Catálogo de Colecciones de Historia Natural del Mundo

---

Integrated summary from 17 to 30 April 2020

---

waddink #9 April 24, 2020, 8:27am

@trobertson that is correct. @dhobern something along the line of standard service types to be implemented by a CMS is exacly what we planned to discus in a workshop under SYNTH+ JRA1, although restricted to loans and visits (to connect with ELViS). That workshop unfortunately had to be postponed to autumn because of the Covid-19 outbreak. I like the idea of doing that as part of a broader set of standard service types to develop under TDWG.

2 Likes

DagEndresen #10 February 4, 2021, 6:54am

I think the CMS is rather a property of the collection than of the institution. At the museum in Oslo different collections are managed in different CMS DBs (and all specimens in the same collection by the same CMS). Two additional points. (1) The CMS used changes through time. Maybe useful to register also CMSs used previously? (2) We preferably want to be able to report a PID (URI) resolvable to machine-readable data for the CMS used - not just a literal. (Maybe simply use Wikidata to describe a CMS? unless an actual catalog of museum CMS systems were built).

1 Like

---

waddink #11 February 4, 2021, 10:30am

Regarding PIDs for CMSes: ideally there would be a PID metadata profile for software instances (similar to PIDInst for instruments), but I think that does not exist yet. For PIDInst there are experimental implementations using DataCite, ePIC, and now also EUDAT: **https://b2inst-poc2.eoschub-surfsara.surf-hosted.nl/** (experimental, needs a lot of improvement still!)

1 Like

---

abentley #12 February 4, 2021, 3:15pm

I have tried to get a field included in the TDWG CD standard that would describe what CMS is used by the collection. That way it is tied to the individual collection description metadata. The hope is that the CD standard will be adopted by the GBIF collections catalog to provide collections-level metadata.
Surely the URI for the CMS could be the vendors landing page - **www.specifysoftware.org**; **https://arctosdb.org/**; etc.?

DagEndresen #13 February 9, 2021, 7:08pm

Maybe a given CMS software might have more than one URL. URLs have a bad tendency to change - even the main homepage? Maybe something like an ontology entry or maybe a Wikidata QID might be a more persistent URI for a CMS software?

Debbie #14 February 11, 2021, 10:43pm

Yes **@DagEndresen** because you'd likely also want to know it is, for example, Specify 5, or Specify 6, or … etc. Versions will (do) matter for the use cases being outlined here.

1 Like