

Natural Collections Descriptions (NCD)

**2006 Workshop
13 – 15 July 2006**

Smithsonian Institution, Washington DC

1. Summary

NCD is based on the collection-level data standard used in the BioCASE project, which has now been extended to cater for library and archive collections, in addition to collections of specimens and observations. This extension work has been undertaken by a group of librarians and archivists known as the RAVNS (Resources Available in Natural Sciences) under the auspices of RLG (formerly the Research Libraries Group).

This workshop is the first time that representatives of the three communities (science research, libraries and archives) have met face-to-face to discuss how we can work together so that NCD can fulfil our collective needs. In future, the standard will be developed through a single group melded from the RAVNS and the TDWG NCD Subgroup, following the new rules for standards development being implemented by the TDWG Infrastructure Project (TIP).

The NCD standard is primarily intended for resource discovery, particularly of collections that have no item-level database. It is lightweight, pitched between very general resource discovery standards such as Dublin Core (DC) and rich collection description standards such as the Encoded Archival Description (EAD), but it is possible to extract a Dublin Core record from an NCD record or, conversely, to use an NCD record as the basis of an EAD record as and when resources allow.

This will be the second meeting devoted to NCD, but the first at which opinions outside the RAVNS have been solicited. The schema has reached a point where any further modifications necessary should be informed by any problems revealed by the testing of the standard.

Day 1 was primarily a background day, with presentations and discussion about audience identification and their needs.

Day 2 was more technical with demonstrations of tools that might be helpful for collection owners in various communities and discussion on how best to test NCD and its longer term prospects.

Day 3 was a half-day session during which the requirements of TIP and TAG were reviewed and their impact on the work of developing NCD were assessed.

Grateful thanks are expressed to the organisations that contributed funds towards the cost of holding this workshop, especially the Gordon & Betty Moore Foundation, GBIF and SYNTHESYS. Several individuals were supported by their own institutions.

A list of participants is provided as Appendix A and further details about the meeting can be requested from Neil Thomson n.thomson@nhm.ac.uk

2. Goals

- To reach agreement on who are the users of collections descriptions and to identify some specific needs.
- To agree a way of testing the NCD standard which will lead to a stable version by March 2007.
- To clarify the long-term prospects for products and services based on NCD.
- To determine the future development of NCD through a single group and to ensure that the work of the group fits with the new TDWG / GBIF expectations for standards development.

3. Out Come

- There is a conflict between records generated as “slices” of an item-level database (e.g. all the items from China) and true CLD records, which would result in an inaccurate item count. It may be necessary to include a flag field to indicate record type (e.g. these records make up the totality of the Smithsonian and they have been sliced by e.g. collector or by taxonomy) or could use <CollectionClass> = “virtual”
- We need to consider an element indicating how collections are arranged or ordered within an institution
- It should be possible to combine national and thematic surveys in the same format, to build the “phone directory” of collections
- An organisation (as a collection of collections) should be identified by coden, which will need expansion and disambiguation through a GUID. A subset of NCD could be used for this
- It was noted that the Library of Congress are now working with CLDs since they have too much material to allow cataloguing at item level
- We need to develop concatenation rules for extracting Dublin Core records. Mapping to DC “Unqualified” is sufficient. DC Type is a standard = “collection”
- A mapping to and from MARC should be added
- A scoping note could be used in EAD to handle the NCD “strength” attribute, marked as an encoding analogue
- It should be determined whether NCD could be used by the long-term monitoring community
- Consider changing the <CollectionClass> element to be mandatory and validated by pick-list
- If someone cites e.g. Library of Congress Names as a source, then the guidance notes should encourage them to do so in a consistent way
- **Archivists toolkit – Lee Mandell**
 - A Beta release is expected at the end of July
 - The toolkit is output neutral – it can output in most standards, but is based around the archival standard ISAD(G)

- It can't import the whole of EAD, but nothing is thrown away – overflow goes into “ingest problem” area
 - The toolkit can be used as a METS authoring tool – except for technical metadata. Future versions may look to work with JHOVE and Dspace
 - It will load screens specific to the type of collection e.g. archival, NCD or visual etc.
 - It may be possible to adapt to NCD in about 6 months – at least for the data model and export
 - MARC ingest is supported
 - Ingest converts characters to Unicode and gets rid of entities
 - There are useful “Babelfish” functions – import in whatever, output whatever
 - It works with MySQL (or other SQL) backend – 25 dialects of SQL are understood
- **National Nodes toolkit – Larry Speers**
 - If NCD could be used as a tool by the big organisations, we could get 80% of the way to estimating the global resource
 - Metadata sharing could be achieved through the GBIF Directory Tool. This would also manage codens, TAPIR interface and NCD import/export
 - The toolkit is not yet in use and needs testing. It has an install wizard and would be particularly good for smaller institutions
 - TDWG is thinking about using OAI for metadata harvesting as it is indexed by Google
 - For RDF it is possible to create a UML class diagram then can export either RDF or XML schemas
 - The toolkit / online editor should produce RDF and ensure that records are linked to each other so users can use Web crawling techniques
 - **GBIF technical architecture – Markus Döring**

Markus spoke from his BioCase/TDWG experience. BioCase has 31 national nodes with collection data from National surveys. Markus sees four models, each with benefits and costs:

- Model 1: Decentralized “social network”
 - direct from institutions
 - low coverage
 - heterogeneous data
 - Model 2: Fully centralized authority
 - low coverage. Many languages
 - many organizational structures.
 - difficult to keep updated
 - Model 3: National node centres (hybrid)
 - national surveys
 - each needs funding and responsible person
 - aware of national language and structures
 - potentially up-to-date and high coverage
 - Model 4: Nodes + individuals
 - duplication more likely
 - more up-to-date and high coverage
- RDF has been developed for metadata. It is hierarchy-based and expandable, but

- maybe not best for large datasets?
- In light of recent TDWG GUID meeting's proposed recommendations to move toward RDF for metadata (and small datasets < 1 million records?), some NCD decisions can not be settled yet.
 - Discussion took place on the objectives of GBIF, the relationship between GBIF and TDWG and the impact of the TIP and TAG projects on the work of NCD
 - Discussion also took place about the integration of the RAVNS and the TDWG NCD Subgroup into a single task force
 - The TIP project should not affect the work of the NCD subgroup substantially since the work is already proceeding in the recommended way
 - The TAG project will substantially affect the group and a decision must be made on when and how to move to RDF

4. Conclusions

- The first priority is to update the standard before the tool can be developed
- The second priority is to write wrappers for BioCASE and Index Herbariorum data – it will be necessary to map existing data to NCD
- The third priority is to implement the data collection tool. If we wish to extend the Archivist Toolkit to NCD, this will need money. The Nodes Toolkit would also need funding.
- A timeline is needed for future actions, including the move to RDF
- For the test phase, BioCASE, RAVNS and Index Herbariorum records will be transferred to NCD and some new records will be created. A variety of records is more important than a large quantity, e.g. MARC records, organisations, networks
- A telephone interview approach to help others to fill in the online NCD form will be developed
- We should look at NBII (ex-FGDC) collection standard to see if this is still active
- For test need
 - Data collection
 - Aggregation (GBIF)
 - Query / report / diagrams
- We need to implement the correct XML Namespace. Current recommendation is `res.tdwg.org/ncd/<version>`
- The group will await guidance from TDWG on moving to RDF
- The RAVNS and TDWG NCD Subgroup will merge and make use of single facilities, such as mailing list, wiki and conference calling

Appendix A: Attendees

| Name | Email | Institution |
|--------------------|--|--|
| Carol Butler | butlercr@si.edu | Smithsonian Institution, USA. |
| Jo DeVeer | | Museum of Comparative Zoology, Harvard University, USA |
| Markus Döring | m.doering@bgbm.org | Berlin Botanic Gardens and Museum, Germany. |
| Scott Federhen | | Genbank, USA. |
| Michael Fox | michael.fox@mnhs.org | Minnesota Historical Society, USA |
| Doug Holland | doug.Holland@mobot.org | Missouri Botanical Gardens, USA. |
| Tom Hollowell | hollowellt@si.edu | Smithsonian Institution, USA. |
| Jackie Kallunki | | Index Herbariorum, NYBG, USA |
| Tony Kirchgressner | | Index Herbariorum, NYBG, USA |
| Barbara Mathé | mathe@amnh.org | American Museum of Natural History, USA. |
| Lee Mandell | lee@nyu.edu | New York University, USA |
| David Schindell | | Consortium for the Barcode of Life (CBOL), USA |
| Larry Speers | lspeers@gbif.org | Global Biodiversity Information Facility (GBIF), Denmark |
| Neil Thomson | n.Thomson@nhm.ac.uk | Natural History Museum, UK |
| Günter Waibel | Guenter_Waibel@notes.rlg.org | RLG, USA |