

NA3 Task 2.3 - Metadata on European Collections – Report and Forward Plan

Roger Hyam <r.hyam@rbge.org.uk>

The vision of NA3 Task 2.3 is to improve the curation and quality of metadata on European collections held in the Biodiversity Collections Index (BCI) as the successor to the BioCASE metadata database. This work started in November 2010 and will continue (on a one person 50% time basis) till November 2012.

The initial work plans is to:

- Analyse current European metadata holdings in BCI
- Survey of the needs of participating institutions and taxonomists for metadata on collections.
- Prioritise work to cleanup and expand on existing metadata.
- Describe changes to BCI to meet these requirements – if any.
- Increased integration with other global collections metadata registries – notably biorepositories.org

Nature of Biodiversity Collections Index

BCI was established in 2008 as a cooperative project between GBIF, TDWG (Biodiversity Information Standards) and Royal Botanic Garden Edinburgh. Its initial aim was to provide globally unique identifiers for know collections, reflect the contents of existing collections registers (such as Index Herbariorum) and allow the wiki-style community editing of data. It achieves this mix of authoritative and community sourced data by having multiple records for each collection all hanging off a single globally unique ID.

In order to allow for the incorporation of information from multiple sources there are very few required fields for each record and the data across records is heterogeneous.

BCI handles the notion of hierarchical relationships for the purposes of navigation but there are no constraints on parent-child relationships. Each ID represents a 'collection' that could be a museum containing multiple departmental collections.

Different subject domains (e.g. Botany, Zoology, Entomology) may use different acronyms for collections that reside in the same institution. Institutions change acronyms through time. The same acronym may be used by different collections.

BCI currently contains 21k collections but the majority of these are “collectors collections”. These were generated from data in Index Herbariorum and were meant to help show where major collectors materials now reside. There are 4,465 real physical collections. Of these 1,410 are in Europe – see below.

Community buy-in has been good considering the amount of publicity that BCI has had but is not sufficient to make a major difference to curation. There are 436 registered users 170 of whom have contributed data though the vast majority have only changed one collection. Two hundred collections have been created and around four hundred modified by volunteers since 2008.

BCIs has been a success but has several clear weaknesses that need addressing:

- It does not enable browsing by taxonomic or geographic scope of collection.
- It is over complex with representation of multiple records for any one collection.
- It only has a simply notion of hierarchical collections which can not prevent double counting of specimens in reporting.

European Data Within BCI

The table below contains summary statistics for European collections in BCI as of December 2010.

Region	Country	Collections ¹	Size ²	Email ³	Web ⁴	Specimens ⁵
Eastern Europe	Belarus	6	5	3	2	323,300
Eastern Europe	Bulgaria	8	4	5	3	391,800
Eastern Europe	Czech Republic	52	44	40	16	7,842,461
Eastern Europe	Hungary	13	9	7	5	8,212,675
Eastern Europe	Moldova, Republic of	2	2	1	0	221,200
Eastern Europe	Poland	38	27	21	6	4,195,833
Eastern Europe	Romania	20	14	10	1	2,722,464
Eastern Europe	Russian Federation	94	85	76	21	15,206,973
Eastern Europe	Slovakia	18	17	11	4	1,066,370
Eastern Europe	Ukraine	26	16	15	5	3,144,857
Eastern Europe		277	223	189	63	43,327,933
Northern Europe	Denmark	7	6	4	4	17,100,000
Northern Europe	Estonia	12	8	9	5	1,304,800
Northern Europe	Finland	20	14	14	10	14,125,071
Northern Europe	Guernsey	2	0	0	0	
Northern Europe	Iceland	3	2	2	0	234,000
Northern Europe	Ireland	12	6	5	3	6,850,517
Northern Europe	Isle of Man	2	1	0	0	3,000
Northern Europe	Jersey	3	0	0	0	
Northern Europe	Latvia	10	6	4	2	472,000
Northern Europe	Lithuania	5	3	4	1	455,100
Northern Europe	Norway	9	8	9	6	3,747,000
Northern Europe	Sweden	18	14	15	13	12,092,004
Northern Europe	United Kingdom	424	128	127	89	55,470,174
Northern Europe		527	196	193	133	111,853,666
Southern Europe	Albania	1	1	0	1	100,000
Southern Europe	Bosnia and Herzegovina	1	1	0	0	110,000
Southern Europe	Croatia	9	8	7	2	817,272
Southern Europe	Greece	10	8	7	2	518,050
Southern Europe	Italy	164	67	54	29	10,129,593
Southern Europe	Malta	1	1	0	0	10,000
Southern Europe	Portugal	24	18	16	6	1,969,401
Southern Europe	Serbia	4	3	3	1	700,000
Southern Europe	Slovenia	3	2	2	1	200,000
Southern Europe	Spain	66	54	49	17	10,447,383

Southern Europe		283	163	138	59	25,001,699
Western Europe	Austria	40	20	15	14	19,694,010
Western Europe	Belgium	22	12	11	10	25,830,521
Western Europe	France	77	53	45	28	64,727,069
Western Europe	Germany	122	67	66	62	20,694,425
Western Europe	Liechtenstein	1	0	1	1	
Western Europe	Luxembourg	2	1	2	2	50,000
Western Europe	Monaco	1	0	0	1	
Western Europe	Netherlands	29	13	18	18	4,921,019
Western Europe	Switzerland	29	16	16	16	18,092,500
Western Europe		323	182	174	152	154,009,544
All Europe		1,410	764	694	407	334,192,842

- 1) Number of collections within the country/region in BCI.
- 2) Number of collections for which BCI contains an estimate of size.
- 3) Number of collection for which BCI contains a contact email address.
- 4) Number of collections for which BCI contains a web address.
- 5) Sum of the number of specimens in collections for which BCI has a estimated size.

Those familiar with the collections in Europe will appreciate that these numbers are on the low side. An example of missing data is given by this 2009 report (http://www.gbif.es/InformeColecciones_in.php) for Spain which lists 119 collections and almost 20 million specimens whilst BCI only contains 66 Spanish collections and 10 million specimens. Suggesting that BCI is underestimating by 50%. The UK, on the other hand, is likely to be very over represented. BCI lists 424 collections but only has estimates of size for 128 of these. When compared with France, Germany, The Netherlands this number appears far too high. The majority of these UK collections are likely to be the results of thirty year old survey of local museums which are unlikely to include globally important material.

If we downgrade the estimate of the total number of collections in Europe by three hundred, to take into account the UK over estimate, then the BCI current estimate of collections would be around 1,100. If the Spanish study were to hold true for all countries then the total for Europe would be more like 2,200 collections. Taking these two figures as extremes it appears reasonable to estimate total significant collections in Europe to be around 1,600.

Requirements

Community requirements for the future of collections metadata were summarised at a workshop as part of the TDWG Conference Montpellier, France in 2009 (http://www2.gbif.org/GSAP-NHC_TDWG09_WorkingSession.pdf). The outcomes can be summarised by a single term “Geotaxonomic Index”. This includes three clear concepts:

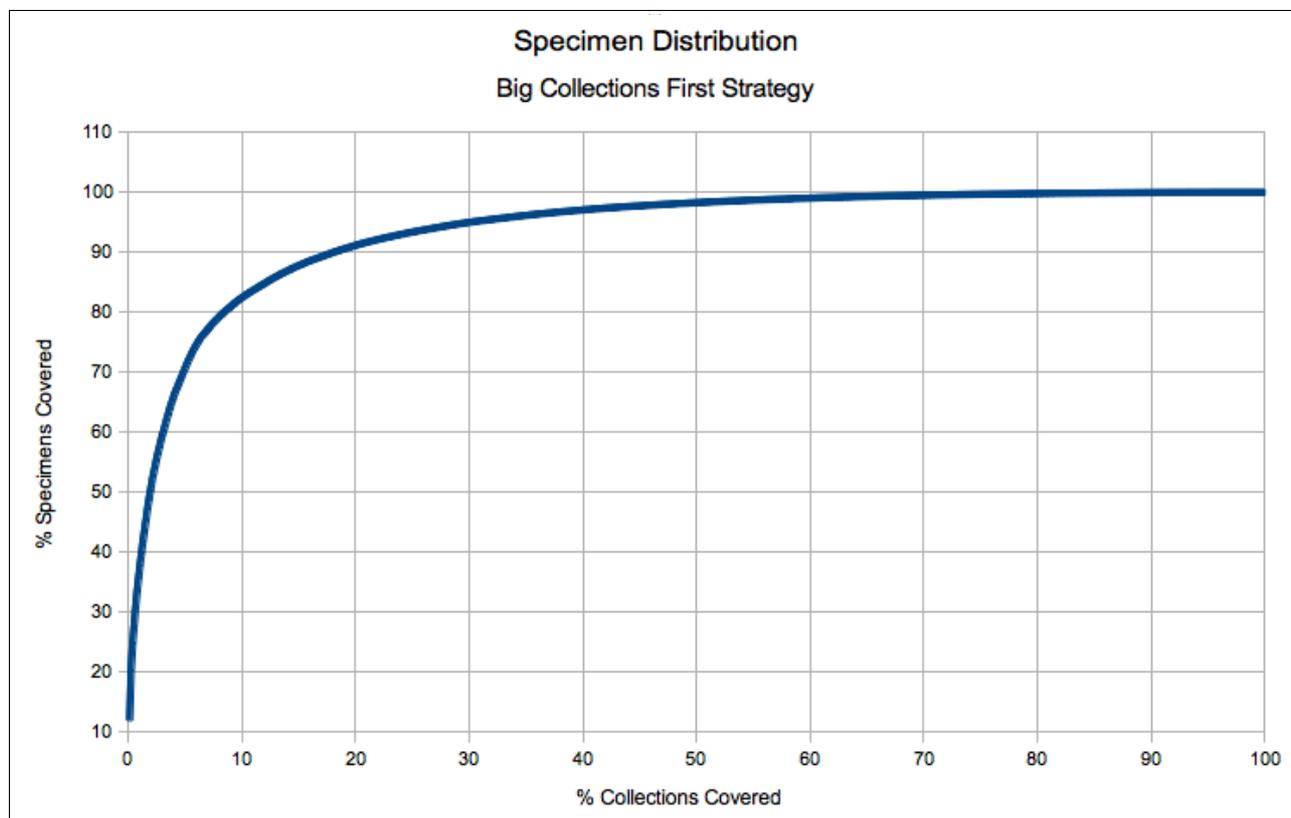
- **Geographic Scope of a collection.** Where are the specimens from?
- **Taxonomic Scope of a collection.** What taxa do the specimens belong to?
- **Location of a collection.** Where is the collection physically housed?

The indexing of collections (and sub-collections) on these three properties would allow the system to answer such questions as “Where are the frog specimens from Ecuador?”. Clearly to answer this to the species level would require each specimen to be digitised (a task for the distant future) but it would be possible to answer such questions for higher level taxa if groups of specimens were scored appropriately. No system contains such information for European collections. The data would need to be gathered afresh by surveying collections – carrying out the equivalent of a cabinet level inventory. BCI has relatively little, non-structured information on taxonomic and geographic scope and does not have mechanisms for capturing it in a query-able way.

Strategy For Improving Coverage

Specimens are not evenly distributed between collections. There are a few major collections and many smaller ones. The graph below shows this relationship. It was generated by taking the 764 collections in Europe for which BCI has an estimate of size, ranking them by size and then saying, “If we started with the largest collections and worked towards the smallest how quickly would we approach coverage of all specimens?” The answer appears to be very rapidly. 90% of specimens would be covered by 20% of collections. 50% of all specimens would be covered by a cabinet level inventory of just fifteen collections. Given the size of the collections held by the Synthesys partners doing such an inventory of their collections would cover much of what is available in Europe.

Such an approach is not without its faults. Smaller collections are often better curated than larger collections and may contain more scientifically useful material because of their niche roles and better curator:specimen ratio. A big-first strategy should not therefore be pursued at the expense of smaller collections especially if smaller collections can be incorporated without interrupting the surveying of large collections.



For the immediate future the following strategy will be adopted:

- Design and prototype an extension or replacement for BCI that is capable of storing structured Geographic and Taxonomic scoping data for collection and sub-collections in a hierarchical fashion.
- Test the prototype on the collections of RBGE
- Seek feedback from the Synthesys partners and BCI partners and users.

Synthesys Partners in BCI

All Synthesys partners now have records in BCI and associated identifiers:

Botanic Garden and Botanical Museum Berlin-Dahlem (BGBM)	http://biocol.org/urn:lsid:biocol.org:col:15534
Hungarian Natural History Museum, Budapest (HNHM)	http://biocol.org/urn:lsid:biocol.org:col:35274
Johannes Gutenberg-University (Mainz)	http://biocol.org/urn:lsid:biocol.org:col:15330
Museo Nacional de Ciencias Naturales, Madrid	http://biocol.org/urn:lsid:biocol.org:col:35275
Museum für Naturkunde, Berlin (MfN)	http://biocol.org/urn:lsid:biocol.org:col:35273
Museum National d'Histoire Naturelle, Paris(MNHN)	http://biocol.org/urn:lsid:biocol.org:col:34988
Nationaal Herbarium Nederland, Leiden (NHN)	http://biocol.org/urn:lsid:biocol.org:col:35036
National Museum of Natural History Naturalis, Leiden (NNM)	http://biocol.org/urn:lsid:biocol.org:col:34992
Natural History Museum, London (NHM)	http://biocol.org/urn:lsid:biocol.org:col:34665
Naturhistorisches Museum. Wien, (NHMW)	http://biocol.org/urn:lsid:biocol.org:col:35270
Real Jardin Botanico	http://biocol.org/urn:lsid:biocol.org:col:15682
Royal Belgian Institute of Natural Sciences. Brussels (RBINS)	http://biocol.org/urn:lsid:biocol.org:col:35271
Royal Botanic Garden Edinburgh, Edinburgh (RBGE)	http://biocol.org/urn:lsid:biocol.org:col:15670
Royal Botanic Gardens Kew, London (RBGK)	http://biocol.org/urn:lsid:biocol.org:col:15867
Royal Holloway University London, London (RHUL)	http://biocol.org/urn:lsid:biocol.org:col:35278
Royal Museum for Central Africa. Tervuren (RMCA)	http://biocol.org/urn:lsid:biocol.org:col:34210
The Swedish Museum of Natural History (NRM)	http://biocol.org/urn:lsid:biocol.org:col:35276
The University of Copenhagen, (UCPH)	http://biocol.org/urn:lsid:biocol.org:col:35272
York University, York (York)	http://biocol.org/urn:lsid:biocol.org:col:35279
Zoological Museum of the University of Amsterdam, Amsterdam (UvA)	http://biocol.org/urn:lsid:biocol.org:col:35207

