*Biodiversity Informatics*, 7, 2010, pp. 81 – 92

# APPROACHES TO ESTIMATING THE UNIVERSE OF NATURAL HISTORY COLLECTIONS DATA

ARTURO H. ARIÑO

*Department of Zoology and Ecology*
*University of Navarra, Pamplona, Spain, artarip@unav.es*

*Abstract*.— This contribution explores the problem of recognizing and measuring the universe of specimen-level data existing in natural history collections around the world, and in absence of a complete, world-wide census or register. Estimates of size seem necessary to plan for resource allocation for digitization or data capture, and may help to represent how many vouchered primary biodiversity data (in terms of collections, specimens or curatorial units) might remain to be mobilized. It further helps to set priorities, and assess certainties.

Three general approaches are proposed for further development, and initial estimates are given. Probabilistic models involve crossing data from a set of biodiversity datasets, finding commonalities and estimating the likelihood of totally obscure data from the fraction of known data missing from specific datasets in the set. Distribution models aim to find the underlying distribution of collections' compositions, estimating the hidden sector of the distributions. Finally, case studies seek to compare digitized data from collections known to the world to the amount of data known to exist in the collection but not generally available or not digitized.

Preliminary estimates of size range from 1.2 to 2.1 gigaunits ($10^9$) of which a mere 3% at most is currently web-accessible through GBIF's mobilization efforts. However, further data and analyses, along with other approaches relying more heavily on surveys, might change the picture and possibly help to narrow the estimate further. In particular, unknown collections not having emerged through literature are the major source of uncertainty.

*Key words*.— Natural history collections, size, estimates, primary biodiversity data

The Global Biodiversity Information Facility (GBIF) aims to make the world's biodiversity information freely and openly available via the Internet (GBIF, 2003), as recommended in the OECD bioinformatics report (OECD, 1999). A significant proportion of this information comes from specimens in natural history collections, generally hosted in museums whose mission includes documenting and studying life on Earth and therefore biodiversity (Krishtalka & Humphrey, 2000). GBIF recently convened a Task Group to catalyze the development of a global strategy and action plan for further mobilization of natural history collections data worldwide (GSAP-NHC), which among other objectives seeks to tackle the task of providing metadata describing the scope of natural history collections and the current status of their digitization (GBIF, 2010). In addition, GBIF encourages a national, regional and thematic implementation and enrichment of the Global Biodiversity Resources Discovery System (GBRDS) to facilitate the decentralized discovery of all biodiversity datasets worldwide (GBIF, 2010) to make them generally available for research.

Unfortunately, no complete, global repository of metadata about the contents of such collections (indeed, a single inventory of the world's collections of natural history) seems to exist at present. There are extensive lists and catalogues for selected biological groups, such as Index Herbariorum (Thiers, 2010), or initiatives seeking to index existing collections, such as the Biodiversity Collections Index (BCI, 2010), but that have yet to encompass all known collections. In addition, biodiversity collections in private

**BOX 1: DEFINITIONS**

The "size" of a collection is highly dependent on the "units" of measure. In general, this chapter uses the following concepts:

- Specimen: An individual organism (or part thereof if treated as an individual), which may be accessioned and stored either separately or together with other specimens, i.e. a herbarium sheet, a pinned moth, or each springtail in a single vial holding hundreds.
- Lot: A specimen or a set of specimens that are stored and generally handled together, i.e. individuals from a single species collected at a single site in a single sampling event.
- Unit: A specimen or set of specimens that has been registered as a single data record and treated as a whole. Generally coincidental with a Lot.
- Collection: A set of units that are held or linked together according to some criteria, i.e. the Coleoptera collected by an expedition or the Lichens from a herbarium.
- Primary Biodiversity Record (PBR): A single data record that points to a unit, and that is backed by an actual observation or vouchered specimen. Generally include at least a taxonomical identification, a location, and a time of capture or observation. A museum record for a specimen or a field observation is a PBR, but a datapoint in a map from a distribution model not backed by an actual observation is not.

hands or poorly documented public collections may not have emerged through literature or registers. At present, it is thus impossible to have a census of all biodiversity data backed by vouchered specimens, and therefore the magnitude of this mass of data cannot be known with certainty.

However, estimates of size seem necessary to plan for resource allocation for digitization or data capture, and may help represent how many vouchered primary biodiversity data (in terms of collections, specimens or curatorial units) might remain to be mobilized.

Current published estimates have been derived mostly from curatorial units (lots, specimens, and collections) and yield and aggregate 2.5 - 3 G (billion) specimens in collections worldwide (Duckworth et al., 1993; OECD, 1999, in Chapman, 2005). Between 2005 and 2007, about one third of GBIF node managers had reported an estimated 407 M (million) to 737 M sharable specimen holdings in potential data providers in their countries only (GBIF, 2008b, 2006); although these may seem low figures, quite a number of data-rich countries had not yet joined GBIF by then so the actual estimate was expected to grow as data from these countries became accessible.

These data, however, are very dependent on the composition of the curatorial units. Whereas in some fields (i.e. herbaria) accessions and specimens can generally be used interchangeably, this is not so in many zoological sections, or when dealing with collected field samples. Therefore, as it respects to the unit of interest for GBIF (the "primary biodiversity record" or PBR), it is relevant to establish the meaning of "size"; to try to translate "specimen" into PBR; and to try to refine current size estimates by figuring out what part of this "size" may have not been accounted for yet. This "size" marks the outer boundary within which useful, quality, "fit-for-use" data (Hill et al., 2010) can be found to answer specific questions in, e.g. biodiversity, sustainability, or conservation.

Recognizing that surveys are incomplete (for example, not all nodes in GBIF reported size estimates in their countries) and that a mandatory register of biodiversity collections does not exist yet, other ways to independently estimate this size (or size range) should be explored for any efficient planning of digitization efforts. In this paper, I propose several approaches and apply them with

**BOX 2: SEBER PROBABILISTIC MODEL ANALOGUE**

Seber (1982) applied probability theory to the problem of recognizing how many tagged animals had lost their marks in a recapture experiment, and therefore would have shown up in a sample as "untagged". We may analogously define a collection as "tagged" if such collection is included within a repository, which is itself actually a sample of the universe of existing collections that, if fully known, should have been all included in the repository. If two independent repositories, *A* and *B*, exist which contain a fraction of all collections, a collection may belong to both repositories ($R_{AB}$), to one ($R_A$), the other ($R_B$), or none ($R_0$). Ideally, the total number of collections would be

$$R = R_A + R_B + R_{AB} + R_0$$

but since we do not know $R_0$, we can only estimate *R*. Seber (1982) shows from probability theory that this estimate can be derived as

$$\hat{R} = \frac{1}{1-k}\left(R_A + R_B + R_{AB}\right)$$

where *k* could be interpreted, in our context, as the product of the probabilities for a collection of not belonging to one repository when belonging to the other:

$$k = \frac{R_A R_B}{(R_A + R_{AB})(R_B + R_{AB})}.$$

Seber and Felton (1981) noted that there is a certain bias in some estimates of the multinomial function used to derive *k*, and this bias, being negligible for sample sizes approaching the universe being sampled, may grow significantly for small sample sizes.

sample data to derive initial estimates of size for first overview, planning and strategic use. Further development and refinement of these approaches, and their application to a larger sample of data, could result in more precise estimates.

For digitization purposes, and towards contributing to GBIF's stated goal of mobilizing about one billion specimen primary records (GBIF, 2006, 2009), one may try to give (i) estimates of the number of accessions ("units", with one or more specimens in each), as there will generally be a one-to-one correspondence between collection unit and occurrence record in GBIF's indexes, and (ii) the number of Collections to be indexed. Whereas the first set of estimates should itself serve as proxy for the amount of person-time work ahead, the second one would probably best related to the number of digitizing teams in the digitizing effort. I will first deal with collections, as the unit numbers may depend on this.

## COLLECTIONS

Several approaches can be followed in order to estimate the number of collections to be digitized, and with different results. I outline here some of these, adapted or derived from ecological and sampling theory.

### *Data crosscheck*

The Biodiversity Collections Index (BCI, 2010) inherits BioCASE metadata (Berendsohn et al., 2000, 2002), and lists *Index Herbariorum* herbaria (Thiers, 2010), Invertebrate codens (Samuelson and Evenhuis, 1998), and other collections. As of October, 2008, BCI included data about 4,265 primary collections (plus 13,865 subcollections embedded within the primary

**BOX 3: GENERALIZING THE SEBER MODEL**

Collections in the universe may have been unlisted ($R_0$), listed in repository *A* ($R_A$), in repository *B* ($R_B$), in *C* ($R_C$), etc., of *L* repositories, and therefore may have been also simultaneously listed in more than one repository (for example, $R_{AB}$ or $R_{ABC}$). Assuming that listings are independent from each other; that there are *m(x: 0, A, B, C, ..., AB, AC, ..., ABC, ...)* possible outcomes; and that a collection has a probability $\Theta_x$ of belonging to the outcome *X*, all possible outcomes follow a multinomial distribution

$$f(R_0, R_x, ..., R_m \mid R) = \frac{R!}{R_0! R_x! ... R_m!} \theta_0^{R_0} \theta_x^{R_x} ... \theta_m^{R_m}.$$

which is the generalized version of the particular case for four outcomes proposed by Seber and Felton (1981, eq. 11). Therefore, similar substitutions can be made as in that case, resulting in a generalized analogue where the correction factor is the product of the probabilities for a collection of not belonging to one repository when belonging to any other. Thus, the case with *L*=3 has *m*=8 outcomes and is

$$k = \frac{R_B + R_C + R_{BC}}{R_T - R_A} \frac{R_A + R_C + R_{AC}}{R_T - R_B} \frac{R_A + R_B + R_{AB}}{R_T - R_C}$$

where $R_T$ is the number of collections belonging to at least one repository. The general case is

$$k = \prod_{S\{A,B,...L\}} \left[ \frac{R_T - \left( R_{Si\{A...L,\neg S\}} + R_{Si\{B...L,\neg S\}j\{C...L,\neg S\}} + ... + R_{ABC..L} \right)}{R_T - R_S} \right].$$

collections, and other data sources), of which 58.6% declared specimen holdings, or whose holdings numbers could be obtained from other sources, accounting for 640 million specimens.

BioCASE and other initiatives sought to discover how many collections (or units) existed already in digital form. For the purposes here, it was interesting to know how many collections had already reported metadata in GBIF. A simple crosscheck of tables was not possible, as the common data field did not abide to a common structure. However, by taking a sample of BCI records one could estimate the number of collections already present at least in part in GBIF indexes, and therefore having been digitized to some extent.

Thus, a random subsample of 197 BCI collections was manually searched in the entire GBIF database 19 collections of the sample contributed data to GBIF, or a 9.64% rate of findings in the sample. Conversely, about 90% of the sample consisted of collections not contributing data to GBIF. Therefore, if the sample was representative, 90% of BCI, or 3.8 K (thousand) already known collections would be either not contributing records, or not exist in a digitized form.

Conversely, one may ask how many of GBIF's datasets pertaining to collections were not in BCI. Thus, one may derive the degree of completeness of BCI: this would show collections already in digital form but unknown to BCI. About 320 GBIF datasets were examined, of which only 16.5% included actual collection data. 41.5% of GBIF collection datasets in the examined subsample corresponded to BCI records.

From this degree of incompleteness the number of completely "obscure" collections (neither in BCI nor GBIF) could in turn be estimated. Seber's (1982, in Krebs, 1999)

probabilistic model for mark loss in capture-recapture methods provided an interesting analog that could be readily adapted. Collections would belong to one or the other set (BCI or GBIF), both, or none. The three first classes, known from the sample, could be used to estimate the fourth (unknown) case, as shown in Box 2, therefore obtaining an estimate of the total number of collections (both known and unknown) in the subsample, from which the total number of collections in the sample (in our case, the universe of collections) can be extrapolated. The resulting number of *extant* collections is 8.5 K, of which all of GBIF's and at least 10% of BCI's would have digitized, contributed data. This would leave out some 6 K collections to be potentially contributed or digitized, including 1.9 K totally obscure collections known to neither BCI nor GBIF.

Seber and Felton (1981) discussed the confidence intervals and bias of their estimate, that can be large for small overlaps. A possible extension of this probabilistic model could involve further datasets, thus increasing precision. This might require trying to generalize the Seber model to more than two overlaps (box 3). As a test of concept, marginal data from the survey performed by the GSAP-NHC TG (see Macklin et al., this volume) were sampled for known collections. A 14.2% subsample of the survey's results was thus searched both in BCI and GBIF, and conversely, the BCI and GBIF samples were searched in the survey. A number of candidate collections emerging in the survey sample could be found in neither previous sample, whereas many were common. Applying the generalized model, the estimated number of extant collections remained at 8.4 K. Assuming that the survey was also random and independent from BCI and GBIF entries, this close result seems to suggest that the crossover of data from independent listings may provide an adequate estimate of the size of the universe of collections.

If these results hold through refinements of the model, the use of further independent listings, larger samples, and (possibly) a specifically-designed survey, the crossover of listings could yield more precise figures but would nonetheless point to the existence "in the wild" of a significant number of yet unknown collections.

### *Distribution model*

Declared holdings in BCI and known records in GBIF can be used to determine the distribution model of the records. From this model, one can try to derive an estimate of the total number of elements (collections) in the set.

The approach I have followed here implies looking at the collections as categories having a given frequency (the number of records). This is akin to considering collections as operative taxonomical units (OTUs), the "sample" being the full inventory of known collections having at least one unit.

One may explore the Whittaker diagram (in Krebs, 1999) of the log number of records in each collection for BCI against their rank by magnitude (fig. 1). The plot very strongly resembles a broken-stick distribution (MacArthur, 1957). However, it should be noted that the Whittaker plot drops significantly at $e^{8.5}$, which corresponds to collections having less than 5,000 specimens. Berendsohn (pers. comm.) pointed out that data in *Index Herbariorum* (a main source for BCI) has a cutoff at that size. Therefore, the actual plot could also be suggestive of a lognormal model (inverted S-curve) if the sharp drop is an artifact of the BCI selection. Data from the GSAP-NHC survey (see Appendix, Fig. 1) may support this hypothesis, as the sizes of collections were reported by their curators without limitations. In addition, the steep curve at the left of the distribution may represent the relative effort in BCI to capture the metadata of the largest collections, which would not have been sampled but censused.

If the lognormal distribution does indeed underlie the distribution of collection sizes (which could eventually be tested by having an unbiased sample if collections of all sizes were present), one could predict the number of unknown collections (i.e. missing from BCI, belonging to the veiled sector of the distribution) by estimating the corresponding lognormal distribution parameters. The best fit would be achieved with a spread of 0,2295, which in turn yields a 39% veiled sector for the distribution. This sector should now be added to the full dataset for an upper limit of the complete universe, resulting 5.9 K extant

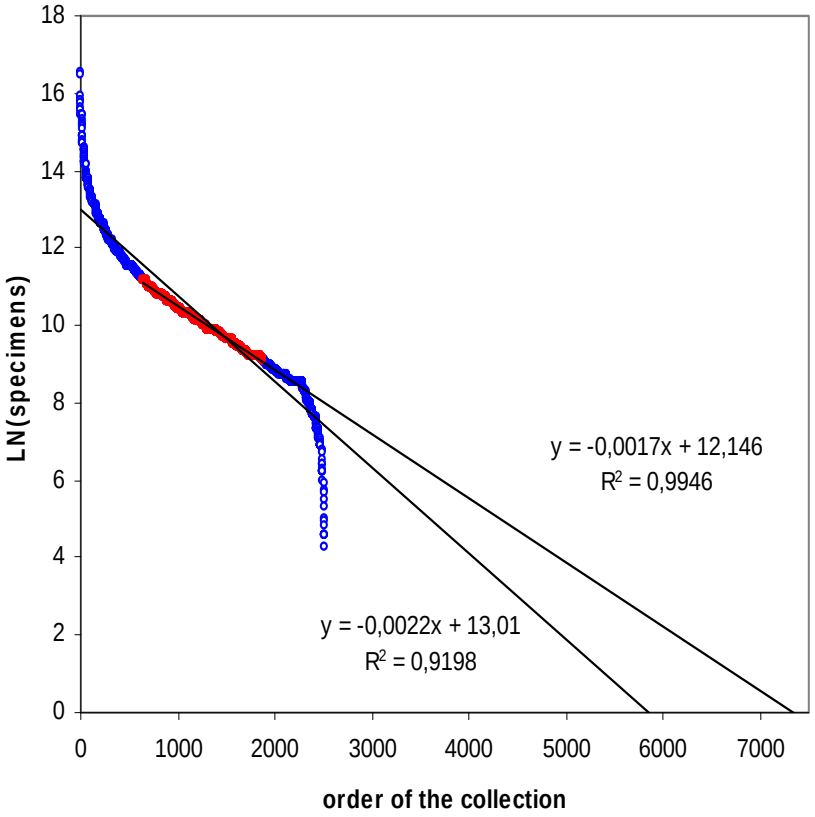



Fig. 1: Plot of log of number of specimens from BCI collections against their rank when ordered by size. Blue dots suggest a truncated lognormal distribution with a cutoff at units=5000. Red dots are the interquartile range. The upper regression line ($R^2$=0.99) corresponds to this interquartile range, and the lower ($R^2$=0.92) to the full range.

collections**.** As we know from GBIF data that digitization rate is 9.64%, according to this model 5.4 K potential collections would be not digitized.

An alternate, simpler, but perhaps more robust approach (especially when taking into account the exponential effects of small differences when using logs), could take the central interquartile range of the distribution data, fitting it a lineal regressive model (fig. 1). This should reduce the bias caused both by the size cutoff in BCI due to Index Herbariorum, and the weight from the largest collections, admittedly not sampled but censused. The intercept with the X axis, representing the order of the collection by number of specimens, should encompass all existing collections following a similar distribution. To these, one should add the zero-unit collections, but the model should not be applied to them. Thus, from the regression coefficients one can calculate 8.6 K extant collections, of which 7.8 K potential collections would not be digitized. Note the similarity of these estimates to the ones coming from the probabilistic model above.

### *Cases*

A third approach should derive from known, specialized cases where the actual number of collections is known through specialized literature, as well as the number of these collections that have emerged in general repositories or listings. A previous work (Ariño, unpublished) dealing with invertebrate collections showed that 16 collections of Collembola could be accessed through web-based searches, BioCASE, GBIF or other databases, whereas a specialized publication listed

185 collections not known elsewhere. This 1115% "specialty" rate, if applied to the then known (in general repositories, such as Samuelson & Evenhuis' codens) invertebrate collections, should yield a worst-case of 6.8 K hidden invertebrate collections. Other groups could be treated similarly, i.e. identifying specialized publications citing collections not known elsewhere. In this case, this 6.8 K figure should be complemented with similar data for entomological, plant, and other general Natural History collections, yielding a larger (albeit unknown) figure for which 6.8K would simply represent the lower bound. Paradoxically, this approach seems the least robust one given that it depends heavily on the ratio between "sampled" data in general repositories and full datasets in comprehensive repositories for specific groups. However, it should be noted that, theoretically, a complete register of all existing collections should bring this ratio to 1, and then all uncertainty should disappear, therefore having the potential to become the *most* robust possible estimate (in fact, not an estimate but a census).

## SPECIMENS

Published estimates about the number of extant specimens are in the 2 - 3 G range (Duckworth et al., 1993; OECD, 1999, in Chapman, 2005). One can independently try to refine and contrast this against actual and modeled data. Currently, BCI lists collections that, either as declared by BCI contributors or obtained elsewhere, amount to 640.5 M (million) specimens. GBIF's GSAP-NHC survey returned 833.7 M specimens reported by respondents after removing duplicated data. From these data, their distribution, GBIF distribution, number of collections, and other experimental data one may try to derive the total amount of data in collections, and what fraction has not been digitized or is not accessible through GBIF.

### *Distribution data*

Only 58.6% of BCI collections declare holdings. Assuming that the underlying distribution data of the holdings across the undeclared collections was similar, one would then have 1.09 G specimens from the known BCI collections alone. However, as I have shown before, there may also be a number of "obscure" collections. For instance, the probabilistic model returns a 23% to 29% of unknown collections.

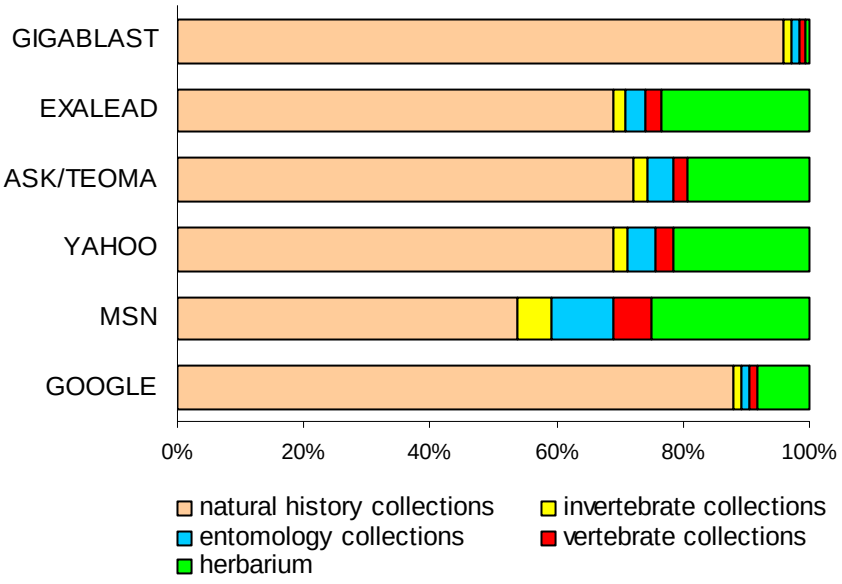



Fig. 2. Distribution of number of results returned by some search engines when queried for certain keywords (legend: keywords used) in 2006.

Taking these into account, and assuming that their size distribution would be similar to the known distribution, a bracket may result from 1.40 G for the number of extant collections calculated by the lognormal model, to 2.01 G specimens if using the probabilistic model. The number of occurrences involving specimens from actual collections already in GBIF can be estimated between 59 M and 96.1 M, depending upon how certain values given as basis of record are interpreted. If occurrences in GBIF can be equated to specimens in BCI (which is not always true) the remaining digitizable mass would bracket between 1.27 G and 1.97 G specimens.

It should be noted that the assumption of similar underlying distribution of specimens across collections, which is required in this approach, has been regarded as a weak assumption (Berendsohn, pers. comm.). Besides the fact that certain requirements apply for insertion into one main source of BCI (namely, active curation and minimal size), Berendsohn points out that **"***it is correct to assume that most of the big collections declare at least an estimate of total numbers, while small collections either hide their numbers or give very exact figures.*" Fully testing this would require knowing the actual numbers of records in every collection (as opposed to their published numbers), which does not seem practical. However, an indirect approach is possible by looking at the relative precision with which numbers were reported. The implicit precision (the relative number of significant figures when reporting a collection size) tends to veer in the way predicted by Berendsohn, although the effect seems too shallow to affect the estimates greatly (see Appendix, figure 2).

### *Rate of digitization in GBIF-mobilized data*

A similar approach but using GBIF-mobilized data can be derived from digitized specimen counts. A random sample of 132 BCI collections having declared numbers of specimens was searched in GBIF indexes, and the amount of corresponding records in GBIF tabulated. For all paired matches, 4.67% of the declared records did exist in GBIF. If this is the rate of digitization in data received and distributed by GBIF, then this rate, applied to the actual number of specimens

| | collections | | specimens | | s/l | lots |
|---|---|---|---|---|---|---|
| Arthropods | 4 | 0,08% | | | | |
| Bacteria | 1 | 0,02% | | | | |
| Data aggregator/indexer | 1 | 0,02% | | | | |
| DNA Bank | 1 | 0,02% | | | | |
| Entomology (Insects/Spiders) | 1303 | 26,10% | 257483000 | 40,20% | 1,0 | 257483000 |
| Herbarium | 3006 | 60,20% | 342725710 | 53,51% | 1,0 | 342725710 |
| Herpetology (Amphibians and Reptiles) | 53 | 1,06% | 2662300 | 0,42% | 2,3 | 1158181 |
| Ichthyology (fishes) | 10 | 0,20% | 6861238 | 1,07% | 1,9 | 3579002 |
| Images | 1 | 0,02% | 2000 | 0,00% | 1,0 | 2000 |
| Invertebrates | 1 | 0,02% | 550000 | 0,09% | 9,0 | 60956 |
| Library/Archives | 1 | 0,02% | 500000 | 0,08% | 1,0 | 500000 |
| Mammals | 12 | 0,24% | 315641 | 0,05% | 1,5 | 212899 |
| Natural History Museum (Diverse Collections) | 69 | 1,38% | 9802000 | 1,53% | 7,4 | 1331793 |
| Not specified | 486 | 9,73% | 1379481 | 0,22% | 1,0 | 1379481 |
| Ornithology (Birds) | 31 | 0,62% | 4181408 | 0,65% | 1,0 | 4043915 |
| paleontology | 1 | 0,02% | 300000 | 0,05% | 1,0 | 300000 |
| Sound/Film | 1 | 0,02% | | | | |
| Vertebrates | 3 | 0,06% | 1754000 | 0,27% | 1,0 | 1754000 |
| Zoology | 8 | 0,16% | 12000000 | 1,87% | 1 | 12000000 |
| total | | | 640516778 | | | 626530938 |

Table 1. Estimated number of curatorial units from collections in BCI assuming that the mean lot size is as in the example case (MZNA, Fig. 3).

from collections in GBIF (est. 59 – 96.1 M), should estimate the amount of unacquired specimens "in the wild", either declared in BCI or not. The resulting figures are 1.26 G – 2.06 G specimens.

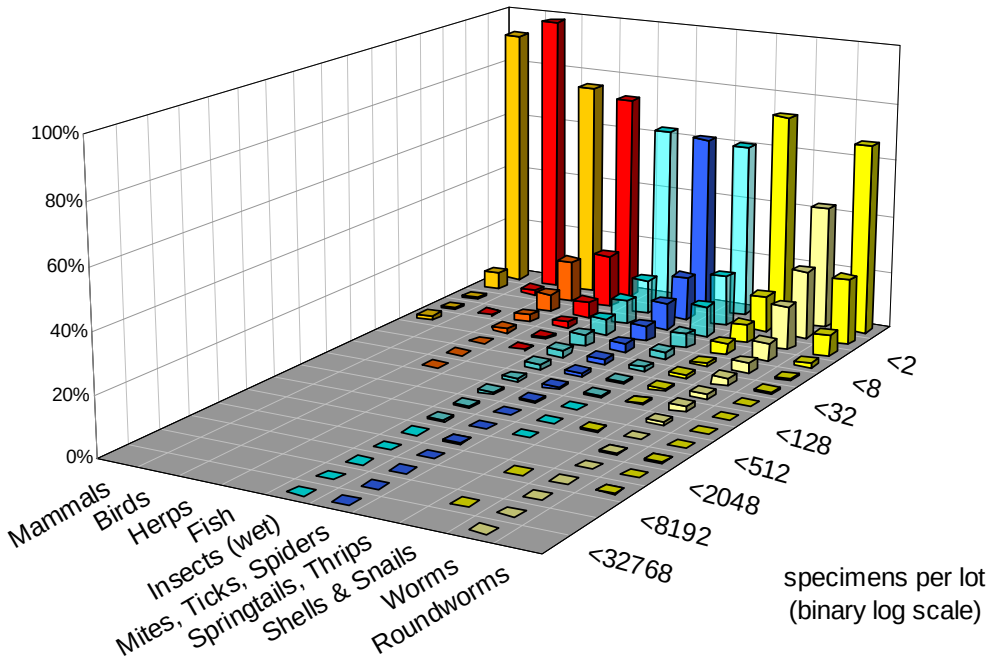
## UNITS

The most relevant figure for the digitization goal would possibly be that of the digitizing unit. Generally, this will coincide with the curatorial unit, although it is not always the case (i.e. multiple data from a single curatorial unit, such as successive exiccata or multi-taxon slides). It may also coincide with specimen data, especially for plants or pinned insects, but much less so for certain animal groups such as wet collections of small invertebrates or uncountable specimens.

The expected number of units, thus, can be either sampled from known cases, or derived from the expected number of specimens or collections based on their types thereof. For that, a distribution of collections and specimens in units may be necessary.

Data from BioCASE, BCI or GBIF can be used to estimate the approximate composition of the types of units to be expected. Also, the composition for data not yet covered in these repositories can be indirectly estimated by using search engines. 60.2% of BCI collections are

Fig. 3. Distribution of the composition of lots in a case study of a zoological collection (MZNA[1], approx. 200 K accessions, 2 M specimens). This refers to already digitized data only. Pinned entomological collections were not included.




1 MZNA is the coden for the Museum of Zoology of the University of Navarra, Pamplona, Spain. URL: http://www.unav.es/unzyec/eng/

Table 2 Summary of various size estimates of extant collections along with known data, ordered by estimate.

| Concept | Source | Estimates |
|---|---|---|
| Collections | *BCI (known)* | *4265* |
| | *GBIF (known)* | *2346* |
| | Log-normal model | 5941 |
| | Case study (invertebrate only) | 6779 |
| | Probabilistic model from metadata | 8372 - 8528 |
| | Regressive model | 8625 |
| Specimens | *BCI (known)* | *640.5 Million (M)* |
| | *GBIF (known)* | *59 - 96.1 M* |
| | *GSAP-NHC survey (known)* | *833.7 M* |
| | *GBIF NODES reports (known)* | *407 – 737 M* |
| | Distribution data | 1.61 – 2.04 billion (G) |
| | Digitisation rate | 1.26 – 2.06 G |
| Units | Composition of collections | 1.23 – 1.97 G |

herbaria and 26% insect collections, reflecting the initial population from Index Herbariorum and Bishop's Museum codens. Search engines yield similar compositions, with a dominant number of references to herbaria (Fig. 2). However, entomological collections amount for 40.2% of the specimens. As these do not distinguish pinned collections (mono-specimen) from wet collections (generally multi-specimen), it is therefore difficult to estimate the correlation for this quarter of data. We may derive an estimate from the analysis of an example of such a collection. Reporting collections may have used interchangeably specimen and unit, and thus the number of specimens per lot of this analysis may be used as a lower bound, while a 1:1 relation would represent the upper bound. Table 1 shows BCI and analytical data from a case study collection (Fig. 3, MZNA collection). Entomological collections have been assumed to be pinned (1:1), although it should be noted that a typical lot composition for entomological wet collections at MZNA is about 16 specimens per lot.

After applying both the distribution of lots and its composition, we obtain a correction factor of -1,022. Therefore, the number of units in collections can be estimated to range, depending on the type of estimate selected, from 1.23 G to 1.97 G units.

## STRATEGY

It has been argued that the rate of accrual currently exceeds the available capacity for digitizing (Beaman, unpublished), although this clearly depends on the amount of people and resources appropriated for the task. These numbers should provide a baseline for gauging the effort needed in order to both reach the 2 billion record goal, and to have most of the metadata from Natural History Collections mobilized and readily accessible. The highest estimate (2,06 G

specimens) yields more than twenty times the amount of collection records already in GBIF, and it should be noted that of the 20 largest known collections (amounting to one third of the total amount of specimens), almost one third had no data in GBIF as of October 2008, while the remaining contributed less than one-tenth of their holdings (Appendix, table 1). Also, the large proportion of herbaria in the known lists of collections may not be mirrored in reality, as zoological collections (except vertebrates) have more slowly started to share data or to be digitized at all: zoological material lends itself less easily to automated digitization procedures, due to very varied physical layouts. A careful identification of the most critically-needed metadata should translate into an adjusted estimate of the human resources to be allocated for this task.

While the above preliminary estimates decrease significantly the previously published estimated size of the digitizable universe, it should be taken into account that these figures are initial estimates, whose precision could be increased through more refined modeling and more extensive data collection. This paper suggests ways to estimate a more reliable range of size, better suited for planning and allocation of resources, and for obtaining certainties. As more records become available, and in particular as all available records are processed, narrower estimates could be obtained. For example, the following activities could all help to refine the current estimates:

- Execute a wider survey among respondents to the GSAP survey and BCI curators with the specific purpose of getting the approximate numbers of digitized data;
- Derive from that survey data to help estimating not only the universe of collections but also the depth of ongoing digitization;
- Expand the analyses to use the fullest available datasets, including subcollections within BCI, through allocation of resources for manual checking;
- Reanalyze all data independently by types of collections.

With better estimates, perhaps a way to allocate efforts for digitisation can be found, for the goal of having quality biodiversity data available for research.

Acknowledgements.- The author is indebted to Walter Berendsohn, James Macklin and two anonymous referees that provided very valuable criticism and comments, and to GBIF for setting up the GSAP-NHC Task Group that prompted part of this research.

## REFERENCES

Berendsohn, W.G., M.J. Costello, C. Emblow, A. Güntsch, A. Hahn, J. Koenemann, C. Thomas, N. Thomson and R. White, 2000. Concepts for a European Portal to Biological Collections. In: Berendsohn, W.G. (ed.): Resource Identification for a Biological Collection Information Service in Europe, Botanic Garden and Botanical Museum Berlin Dahlem, ISBN 3-921800-44-7 [2]

Berendsohn, W.G., M. Döring, M. Gebhardt and A. Güntsch, 2002. BioCASE - A Biological Collection Access Service for Europe. Trends and Developments in Biodiversity Informatics. Symposium: Key Innovations in Biodiversity Informatics, Indaiatuba, SP, Brasil 2002. [3]

[BCI] Biodiversity Collection Index [continuously updated]. http://www.biodiversitycollectionsindex.org/static/index.html (accessed July 30, 2010)

Chapman, A., 2005. Uses of Primary Species Occurrence Data, version 1.0. Copenhagen: Global Biodiversity Information Facility. 106 pp. ISBN: 87-92020-01-1[4]

[GBIF] Global Biodiversity Information Facility, 2003. Global Biodiversity Information Facility Annual Report 2001-2002. Copenhagen.[5]

[GBIF] Global Biodiversity Information Facility, 2006. GBIF Strategic and Operational Plans 2007-2011: From Prototype towards Full Operation. Copenhagen.[6][GBIF] Global Biodiversity

2 http://www.bgbm.org/biocise/Publications/Results/11.htm

3 http://www.cria.org.br/eventos/tdbi/bis/biocase

4 http://www2.gbif.org/UsesPrimaryData.pdf

5 http://www2.gbif.org/annual_report_2001_2002.pdf

6 http://www2.gbif.org/strategic_plans.pdf

Information Facility, 2008. GBIF Work Programme 2009-2010. Copenhagen.[7]

[GBIF] Global Biodiversity Information Facility, 2008b. GBIF Annual Report 2008. Copenhagen.[8]

[GBIF] Global Biodiversity Information Facility [continuously updated]. Data Discovery and Mobilisation. http://www.gbif.org/informatics/primary-data/data-discovery-and-mobilisation/ (accessed July 30, 2010).

[GBIF] Global Biodiversity Information Facility [continuously updated]. Global Plan for Natural History Collections Data. http://www.gbif.org/informatics/primary-data/task-groups/gsap-nhc/ (accessed July 30, 2010).

Güntsch, A., A. Hahn and W.G. Berendsohn, 2001: Biological Collection Databases in Europe. Pp. 9-17 in: Riede, K. (ed.), New perspectives for monitoring migratory animals - Improving knowledge for conservation. BfN, ZEF, Bonn.

Hahn, A. 2000. Information Resources - The BioCISE Survey. Resource Identification for a Biological Collection Information Service in Europe. Results of the Concerted Action Project. http://www.bgbm.org/biocise/Publications/Results/7.htm (accessed July 30, 2010).

Hill, A., J. Otegui, A.H. Ariño and R. Guralnick, 2010: GBIF Position Paper on Future Directions and Recommendations for Enhancing Fitness-for-Use Across the GBIF Network, version 1.0. Copenhagen: Global Biodiversity Information Facility, 25 pp. ISBN: 87-92020-11-9. http://www.gbif.org.

Krebs C.J., 1999: Ecological Methodology, 2nd ed. Addison-Wesley,

Krishtalka, L., and P.S. Humphrey, 2000. Can Natural History Museums Capture the Future? BioScience, 50 (7): 611-617.

Mac Arthur, R.H., 1957. On the relative abundance of bird species. Proc. Natl. Acad. Sci. 43, 293-295.

[OECD] Organisation for Economic Co-operation and Development, 1999. Final report of the OECD Megascience Forum Working Group on Biological Informatics. Paris: OECD Publications. [9]

Samuelson, A. and N. Evenhuis, 1998+: Abbreviations for Insect and Spider Collections of the World. Bishop Museum, Honolulu.

Seber G.A.F. and R. Felton, 1981. Tag loss and the Petersen mark-recapture experiment. Biometrika, 68 (1): 211-219.

Seber G.A.F., 1982: The Estimation of Animal Abundance and Related Parameters. The Blackburn Press, New York.

Thiers, B. [continuously updated]. Index Herbariorum: A global directory of public herbaria and associated staff. New York Botanical Garden's Virtual Herbarium. [10] (accessed July 30, 2010).

7 http://www2.gbif.org/WP2009-10.pdf

8 http://www2.gbif.org/annual_report_2008.pdf

9 http://www.oecd.org/dataoecd/24/32/2105199.pdf

10 http://sweetgum.nybg.org/ih/