Supplementary material: Estimates of numbers of PIDs needed

This supplementary material provides an estimate of the number of PIDs likely to be needed throughout the DiSSCo lifetime and beyond. Such estimates are necessary for assessing scalability, performance and cost of the alternative PID schemes to be analysed. Estimates don’t need to be accurate, just ballpark to provide an indication of scale.

## How many specimens are there to be identified?

A conservative estimate of the total number of specimens in collections worldwide (Ariño 2010) was in the range 1-2 billion. However, that estimate does not consider the vast number of private collections that exist.

From [Index Herbariorum](http://sweetgum.nybg.org/science/ih/), lists for herbarium collections, an estimated 390 million specimens are known to have been deposited. From surveying capabilities among the DiSSCo partners (Casino et al. 2019) the share of herbarium specimens in collections overall is known to be around 15-20%. Extrapolation suggests that the total number of specimens globally is therefore at least 2 billion (and probably more) just for institutional collections. Again, this does not include private collections.

The EC-funded ICEDIG project estimated that privately held collections in Europe account for millions more specimens – anywhere in the range 9 to 33 million.

CETAF estimates the total number of specimens in Europe alone at 1.5 billion, of which around 10% have been digitized.

## Estimate based on GBIF growth rate

Currently there are about 175 million specimen records in GBIF (one third being from Europe). Over the past 20 years growth has been linear for specimen records, with about 6 million new records added each year. These are records that have been digitized up to MIDS level 2 and higher.

Specimens that have been digitized only to MIDS level 0 and MIDS level 1 are generally not shared to GBIF. The SYNTHESYS+ Collection Dashboard pilot (2019/20) with six large European institutes shows almost twice as many MIDS level 0 and MIDS level 1 records as MIDS level 2 records.

It seems to be safe to assume that currently at least 15 million digital records a year are created globally, of which about one third come from Europe. Thus, DiSSCo needs at least 5 million new PIDs per year based on current rates.

With DiSSCo enabling consolidations in mass digitization capacity and process, together with new workflow and technology innovations this number is likely to grow rapidly. DiSSCo is aiming for a ten-fold increase in digitization speed to have 50% of the collections digitized over the next 20 years. Thus, in perhaps 10-15 years’ time, DiSSCo will need 50 million new identifiers per year for specimens.

## Minimum numbers

If DiSSCo would start with giving persistent identifiers to all specimens currently digitised, this would already be 60 – 150 million new PIDs.

DiSSCo views a Digital Specimen as a container of other objects; each of which needs to be persistently identified. Annotations, interpretations and loans/visits transactions all need to be persistently identified. Especially in the case of annotations, which might be generated by automated procedures in the future, it would not be unrealistic to assume a tenfold increase in the required PIDs.

So, once digitized, how many additional identifiers does a Digital Specimen need?

## The life of a digital specimen on the Internet

Once digitized, a specimen as a usable/researchable digital object begins its life history on the Internet. Placing some ballpark figures on events in that life history: A digitized specimen could go on to accumulate fifty updates (improvements to the captured data) or enhancements (addition of links to supplementary information). Each of these updates/enhancements causes an associated provenance event to be recorded, so that a life history timeline is maintained. Each provenance event should itself be persistently identified to assist with tracing the life history and supporting queries across life histories of multiple specimens. A specimen can be annotated and/or have new interpretations placed on it. The specimen may be loaned/visited with a record of each of those transactions needing to be kept. Annotations, interpretations and loans/visits transactions all need to be persistently identified – let’s say forty to fifty per specimen as a rough estimate. Plus a few more for identifying evolving versions of a specimen. So, for a specimen during its lifetime we might say in round terms that an additional one hundred identifiers are needed in addition to the original assigned identifier. Transitioning to a curation model that is increasingly community centric might lead to higher numbers of associated identifiers for life history/provenance and attribution purposes.

## How many PIDs will be needed in total?

Given the ballpark figures above – 5 million specimens digitized annually, increasing to 50 million, with 10 – 100+ identifiers per specimen lifecycle – a rough estimate starts at 50 million+ PIDs annually in the first few years rising to 5billion+ per year after 20 years.

The rate of growth will accelerate over time with changes in working practice. Long-term it would be safe to assume that DiSSCo needs tens, even hundreds of billions of PIDs over the next 20-25 years.

## How many PIDS could the Handle system create per day?

[Performance testing](http://handle.net/performance.html) has shown the Handle system is capable today of creating a maximum of 11532 handles per second (with an average latency of 70 ms when that maximum throughput was observed). For a day this would be maximum 86400 x 11532, equalling about 1 billion per day. It can therefore be expected that the Handle system can perform at the level required to create the amount of PIDs needed in DiSSCo.

## References

Ariño A (2010) Approaches to estimating the universe of natural history collections data. Biodiversity Informatics 7 (2). <https://doi.org/10.17161/bi.v7i2.3991>

Casino A, Gödderz K, Raes N, Addink W, Koureas D, Hutson A (2019) DiSSCo Partner Capabilities Survey 2017. Zenodo <https://doi.org/10.5281/zenodo.2653707>

END.