Annex 8
The Botanical and living Collections integration

1. INTRODUCTION

To make data from their herbarium collection available and easily harvestable by the Natural Heritage (NH) portal, APM published all their herbarium data to the Global Biodiversity Information Facility (GBIF). Publication was enabled by constructing a workflow which maps herbarium specimen data to the Darwin Core standard and makes this data web-accessible in the form of versioned Darwin Core archives (DWC-A) on APM's Integrated Publishing Toolkit (IPT) server. Registration of this IPT server with GBIF also allows automated ingestion by GBIF and the subsequent benefits of strongly improved findability and data validation, in particular taxonomic validation of names to the GBIF Taxonomic Backbone. This way, a lot of double-work and constructing local taxonomy backbones that quickly fall out of date is avoided. As digitization of APM's herbarium collection is still ongoing, this dataset is regularly updated.

2. METHODOLOGY

A similar approach was undertaken for APM's collection of living plant accessions, which are now available on both the IPT and GBIF as well. Another useful collection at APM is the Flora of Central Africa series, which includes a vast number of local vernacular plant names from Central Africa. These names were curated to annotate them with ISO language codes, as much as possible, and they were made similarly available on the IPT and GBIF, allowing the NH portal to make use of them.

3. INFRASTRUCTURE

No specific infrastructure was required.

4. RESULTS AND RECOMMENDATIONS

4.1 Making botanical collections available to the NH

A key requirement for opening up collections is that the individual specimens are not only findable, but also re-usable when needed. This means that digital publications need to maintain a link with the local, physical specimens. The CETAF identifier concept was conceived to help maintain this link and constitutes a digital identifier for the physical specimen, making use of semantic web referencing methods. Data made available on the NH portal uses these identifiers as keys to find back the physical specimens, still curated by the institutions themselves. For APM, these identifiers point to collection records on the Botanical Collections web portal. These identifiers need to be maintained and for this purpose, data of APM's living collection were migrated from the old, deprecated LIVCOL website to the Botanical Collections portal. As the Botanical Collections portal is not intended to only hold collections from APM, migration of living collection data from other Belgian living plant collection holders from the PlantCol website to the Botanical Collections portal is planned as soon as the workflows for this goal can be implemented.

If the NH portal shows images, these images need to be hosted still at the local institution. For this purpose, APM herbarium images were made available on a web server so that they can easily be crawled through the IPT or GBIF DWC-A's. The same was done for the living collection images. The directory structure was brought in line with the policy already in use for the other image archives in use within APM. Images were also hashed during transfer and validated for errors afterwards. To make these images easily findable and usable, an International Image Interoperability Framework (IIIF) server was set up at APM so that images and their metadata could easily be retrieved in the format needed.

To improve findability and interoperability of specimen data, semantic annotation protocols were set up, resulting in over 1M specimens being annotated with a persistent identifier for the person(s) who collected them. This allows unambiguous identification of these collectors and easy harmonization with data from other collection-holding institutions. These annotations were made web-available on the Botanical Collections portal in the form of machine-readable XML/RDF following semantic web practices, easily retrievable through the CETAF identifiers. As soon as the standards and technical infrastructure allowed it, they were also made available on the IPT server and therefore to GBIF.

4.2 Migrating from LivCol to another database

LivCol was a bespoke living collections database created by a staff member of APM, Thierry Vanderborght. Living collections databases are integral to the management of botanic gardens, Dr. Vanderborght's database was acutely tailored to the running of the Botanic Garden. The LivCol software used is PROGRESS, a fourth generation relational database management system. It ran on a Linux platform where internal users could access it using a SSH service such as putty.exe. LivCol has been used at APM for 25 years and includes some 500 000 records distributed in 34 files and 338 unique fields. In order to facilitate the migration to the new data management system, LivCol was combined with another independent database at APM PHASEO, the Wild Bean Collection. A Web version of LivCol was also present, being extracted from the central LivCol database, using PHP and PostgreSQL technologies. As Dr. Vanderborght was due to retire by the end of 2018 and he was the sole developer of LivCol, the migration of LivCol to a new Collection Management System was imperative. The ideal database would be able to handle both preserved and living collections.

After extensive testing of different collection management systems, including DaRWIN, Specify, IRISBG, PlutoF and Botalista, APM chose Botalista (http://www.ville-ge.ch/cjb/modules_en.php) developed by the Conservatoire et Jardin botaniques de la ville de Genève (CJBG). This database had a functional Living Collection Module, however it needed to be expanded to include features that were present in LivCol. Also, as the CJBG is responsible for the development of the taxonomic backbone for the World Flora Online (http://www.worldfloraonline.org/), it was deemed a good choice. The staff members at APM did a complete analysis of all fields present in LivCol and mapped this to corresponding fields in Botalista. The reverse exercise was also done to check that all fields in Botalista were present in LivCol. Priorities were defined to enable the importation of LivCol into Botalista. A 5 day meeting was organized at APM with IT staff from Geneva. IT staff from RMCA / RBINS attended a joint meeting with CJBG 1 February 2018 to discuss interoperability issues between their database and the common portal. A schedule for the data transfer to Botalista and further development of their system were also set out. After months of analysis, discussions and planning a schedule and costing was established for minimum requirements to make Botalista LivCol ready. It was estimated that the necessary changes would take approximately 800 programming hours. This meant the earliest Botalista would be ready for import of LivCol would be mid-2019 – assuming all went according to schedule. As Dr. Vanderborght would already be long on pension by this stage, and there was no staff member that could take over the responsibilities for the new

database – the management at APM decided to migrate LivCol to BG-BASE. The latter program is already in use at APM for the preserved collections. Significant improvements in BG-BASE had been made, making it a suitable candidate for the Botanic Garden. Including LivCol in BG-BASE also meant that the Living and Preserved Collections would both be housed in the same database. Dr Engledow, database manager of BG-BASE at APM, was also involved in the planned migration of LivCol to Botalista, so was up-to-date with what was required. He would be responsible for the transfer of data from LivCol to BG-BASE, management of both Collections and the training of staff.

BG-BASE is a commercial software program for managing Living and Preserved Collections of plant material (for more info see http://www.BG-BASE.com/). APM uses BG-BASE version 9, includes 364 data tables (195 of these are currently used at APM) and over 10 000 data fields. As BG-BASE has so many data fields – most fields in LivCol could be mapped in BG-BASE. As a result no additional fields needed to be created in BG-BASE. Dr. Kerry Walter of BG-BASE was initially responsible for doing the import of the 34 LivCol tables into BG-BASE. But, it soon became clear that was progressing too slowly due to back and forth communication between APM and BG-BASE resulting from data misunderstandings. It was at this point that Dr. Engledow went to Edinburgh to work with Dr. Walter and Mr. Akbalik to create a generic import tool that would allow us to import data into any table in BG-BASE (this was not previously available). Dr. Engledow then took over the importation of all the LivCol tables into BG-BASE. All LivCol tables have been imported into BG-BASE.

However, the migration from LivCol to BG-BASE was accompanied by its challenges. These findings were presented at the BiodiversityNext Conference held in Leiden (Netherlands) in October 2019. Shared tables between the Living and Preserved collections needed to be correctly linked to associated records. The table NAMES in BG-BASE, dealing with taxonomy, was particularly challenging. As taxon authorities have not been entered in a standardized manner, the scientific names had to be matched without authorities. This worked for about 63% of the taxa, however the problem of homonyms, orthographic variants, spelling errors and manuscript names, had to be resolved before importation to avoid the problems of duplication. Similar issues arose in all shared tables. An intermediated database was created to transform the data provided from LivCol into a BG-BASE format.

Differences in database structure, degree of atomisation and field definition made transfer of data difficult. Accession information is central to Living collections, but the way one does this and the philosophy behind this may differ. The LivCol approach differed from BG-BASE (and Botalista) by affecting the structure of the data model in each. The LivCol approach was less rigorous than the BG-BASE e.g. new generations derived from existing accessions in LivCol retained the same accession number despite being not genetically identical (of seed origin), whereas in BG-BASE a new accession number would be generated with reference to the parent accession. In the data transfer LivCol accession numbers were grouped by accession number and garden location, and the inter-generation information combined in a single record in BG-BASE. The use of data ‘standards’ are key in any database, and many standards are used by both databases. However, it soon became evident that there are multiple ‘standards’ or versions for a single topic e.g. for information concerning conservation status: NatureServe Global Conservation Status Ranks; Fish & Wildlife conservation category; International Union

for Conservation of Nature (IUCN) - old and new codes (plus version); Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES); etc. This lack of scientific or other standards makes transfer or interoperability between databases very difficult. Organisations like TDWG (https://www.tdwg.org/) attempt to deal with many of these issues concerning Biodiversity data, but there is often an unwillingness of many scientists to follow. Many create their own bespoke ‘standard’ in the hope that others follow their lead, instead of being proactive in creating standards within their community / domain. At a time of data openness and exchange, the latter approach is antediluvian and should be avoided in the future.

Some data is region specific and does not map with global standards e.g. there are 3 principal regions in Belgium controlling conservation status (Brussels, Flanders, Wallonia) and each have their own approach and definitions, as this has legal implications they all need to be taken into account. The latter was done by finding close matches in IUCN (New) codes and combining them with 'non-standard' World Geographical Scheme for Recording Plant Distributions (WGSRPD). The latter TDWG standard is out of date and in many circumstances not sufficiently atomised to be of practical use. There were also certain fields that would benefit from having standards, but are at present absent e.g. invasiveness - BG-BASE uses Cronk and Fuller (1995) whereas LivCol uses AlterIAS (http://www.alterias.be/), Belgian Forum on Invasive Species (http://ias.biodiversity.be) and Lambinon et al. (1992). In some instances a ‘best fit’ solution was adopted. These are just some of the issues addressed during the importation process.

All the LivCol data is presently in BG-BASE and many of the staff are now working in the database. The number of staff working in the database continues to grow as less senior staff members are given permission to enter data for specific domains (this was not possible with the previous database). The living accessions data is regularly exported from BG-BASE and uploaded to our website (http://www.botanicalcollections.be/#/en/search/living-collection) where they can be consulted. At present this is only done for APM, but we are presently looking into expanding this functionality to the PLANTCOL network (http://www.plantcol.be/).

Bibliographic references

Biodiversity Next conference (22 – 25 October 2019, Leiden (Netherlands)): “Data Migration from One Database to Another: Nervous breakdown of a database manager!” https://doi.org/10.3897/biss.3.37302

Cronk QC, Fuller JL (1995) Plant Invaders. The threat to natural ecosystems. Chapman & Hall, London, xiv, 241 pp. [ISBN 0-412-48380-7]

Lambinon J, Langhe J-D, Delvosalle L, D'Hose R, Geerinck DJ, Lebeau J, Schumacker R, Vannerom H, Rammeloo J, Duvigneaud J (1992) Nouvelle flore de la Belgique, du Grand-Duché de Luxembourg, du nord de la France et des régions voisines (Ptéridophytes et Spermatophytes). 4. Jardin Botanique National de Belgique, Meise, cxx, 1092 pp. [ISBN 9072619072]