Minutes of the joint CETAF Digitisation Group and ISTC Meeting, Vienna, 13-14 February 2019

Executive Summary

The meeting originally planned for London took place in the form of a shorter teleconference on 20 April 2020 due to the corona situation. The agenda is publicly available at https://cetafdigitization.biowikifarm.net/cdig/ISTC_DWG_Meeting_Spring_2020_London. Presentations as well as these minutes will be made available on the same website. The meeting was attended by 45 participants from 18 CETAF member institutions, the CETAF secretariat, and the University of Cardiff.

Action Items

- To assess the possibility of a Georeferencing Hackathon in 2021.
- To integrate the currently developed data archiving manual (Mobilise WG4) into the DiSSCo Data Management Plan (DMP).
- To publish results of the CETAF “Botany Pilot” more widely.
- To revise Strategic Plan table.

Participants

Anton Güntsch (BGBM), Elspeth Haston (RBGE), Dominik Röpert (BGBM), Wouter Addink (Naturalis), Quentin Groom (Meise), Mathias Dillen (Meise), Maarten Trekels (Meise), Pieter Huybrechts (Meise), Stefan Seifert (SNSB), Dagmar Triebel (SNSB), Wiebke Walbaum (SMNS), Sharif Islam (Naturalis), Roger Hyam (RBGE), Heimo Rainer (NHMW), Ari Taponen (Luomus), Patricia Mergen (Meise and RMCA), Agnes Wijers (Cultural Connections), Laurence Livermore (NHM London), Falko Glöckler (MfN Berlin), Matt Woodburn (NHM London), Sarah Phillips (RBGK), Karol Marhold (SAV), Ana Casino (CETAF), Laura Tilley (CETAF), Luc Willense (Naturalis), Jiri Frank (NMP), Patrik Mráz (NMP), Peter Grobe (ZFMK), Björn Quast (ZFMK), Jean-Marc Herpers (RMCA), Franck Theeten (RMCA / RBINS), Henry Engledow (Meise), Rob Turner (Kew), Paul Braun (MnhnL), Anne Koivunen (Luomus), Celia Santos (MNCN-CSIC), Pierre-Yves Gagnier (MNHN Paris), Fredrik Berger (MfN), Rob Cubey (RBGE), Robyn Drinkwater (RBGE), Sally King (RBGE), Sofie De Smedt (Meise), Alex Hardisty (Cardiff University and DiSSCo Technical Team), Marie-Hélène Weech (Kew), Josh Humphries (NHM London)

Digitisation Working Group

Adoption of Agenda

The agenda was approved.

CETAF Strategy and Strategic Development Plan revisited (Elspeth Haston)

The Actions and Targets for the Digitisation Working Group (https://cetafdigitization.biowikifarm.net/cdig/CETAF_Strategy_ISTC_DWG#Focus_Area_.233_-_Natural_history_collection_management_and_access_to_collections) were reviewed. The development of a standard for the Minimal Information for a Digital Specimen (MIDS) is a critical part of being able to measure this target. The various tasks involved in institutes meeting the target formed the basis of the DWG meeting.

MIDS update (Alex Hardisty & Elspeth Haston)

Version 0.10 of Minimal Information for a Digital Specimen (MIDS) is coming soon. Entering TDWG standardisation process is planned. Mapping to ABCD/DwC terms under way. Institutes have expressed interest to implement the standard. It was noted that additional guidance will be helpful for institutes, particularly if they plan to adapt digitisation processes to align with MIDS levels. Standardisation with 3 levels of data: Basic (imaging and persistent identification, enabling data extraction); Regular (meaningful scientific fields present); Extended (enriched records). NSID has been taken out of the MIDS specification so that we don't have to wait for an agreed ID scheme.

A new version of the MIDS Specification will be made available by the end of this week (by Alex Hardisty). If anyone has questions on MIDS contact Alex Hardisty or Elspeth Haston. An ad hoc workshop to be organized (date tbd) to work on the next version of MIDS.

Transcription options and standards update (Mathias Dillen & Quentin Groom)

This presentation was based on the publication of the ICEDIG Deliverable (https://academic.oup.com/database/article/doi/10.1093/database/baz129/5670756: https://doi.org/10.1093/database/baz129). It includes guidance for missing data (not yet transcribed, withheld or not on label) and proposes a defined set of terms for missing data. Decisions are required about the handling of verbatim data and this will depend on the use being made of the data. Additional issues highlighted included: Partial dates, date ranges (some systems can’t handle these), Excel date corruption (if Excel is part of the workflow). With geoference data, the transformation of grid reference squares into point radius circles was . The abiity to correct data, as well as annotate the correction was requested, to ensure that

corrections were not changed back to the original. The use of Excel is discouraged in some institues, but if it is being used, versioning can and should be implemented.

It was agreed that versioning needs to be included as a function in CMS software and implemented by data managers. The TDWG conference was suggested as a forum for this.

Imaging best practice update from ICEDIG (Agnes Wijers)

Agnes Wijers presented the results from several ICEDIG project deliverables focussing on image protocols and best practice. This work covered a wide range of material and equipment, processes etc, and included difficult collections such as liquid. The question of prioritising specimens or labels was raised. The use of robots in for digitising collections has been explored. The status of imaging protocols was summarised for the different collection types:

- Herbaria - conveyor belts, quality control - system working. Information still needs to be shared - particularly on issues resolved
- Pinned insects - novel techniques (6 research pilots) - combinations of robotics, machine learning etc opportunities, imaging on conveyor belts with multiple webcams at different angles which were stitched to enable label text to be visible for transcription (not yet at mass level)
- Liquid - 3D images may be needed, but at least labels would be useful. Small pilot for curved labels on a rotary table with multi-images to produce readable labels
- Microscope slides - 2 pilots with workflows to get to MIDS 0-1. Issues of slide variability for mass digitisation
- 3D - many different options for technology. Techniques are not yet at mass digitisation level and may not be needed for all specimens. Selection may need to be made.

Big decisions need to be made about the level of digitisation required for all 1.5 billion objects. Not same level may be required for all objects. We can use lessons learned and techniques developed to bring together in a mix and match system. We can also involve industries although it is not always easy to attract interest from SMEs. Bringing institutes together to create larger client groups might attract larger players and good business cases with precise questions and requirements are needed.

More information can be found in several deliverables within the ICEDIG project here: https://icedig.eu/content/deliverables

DiSSCo Synchronisation Group 4: Digitisation update (Laurence Livermore)

Laurence Livermore presented an update on the DiSSCo Synchronistation Group 4. The work of this group includes looking at gaps in DiSSCo, costs of filling the gaps and possible funding sources, along with bringing experts together. The SG4 currently comprises five people, but membership is open all interested.

Synergies with DiSSCo-linked projects include: a) ICEDIG (with the Blueprint resulting from WP8); SYNTHESYS+ including Digitization of collections on Demand, Virtual Access and Specimen Data Refinery; and also DiSSCo Prepare, with Capacity Building and enhancement (WP3), business frameworks and cost book (WP4) and digitization protocols (under WP5)

Some of the issues within the scope of the SG4 are: Costs of setup and the concept of distributed digitisation teams; Digitisation warehouses as facilities which digitise for multiple institutes/countries; More sharing of digitisation protocols – there is no single place to look for standards, guidelines, recommendation or who to contact about new ideas and developments; Collection Management Systems – a good collection management system would go a long way to solve many of the data issues we are facing at the moment.

Concerns were raised about possible overlaps with other SWGs, such as SG3, in reference to standards, or SG5 on Training and Capacity Building. In terms of training programmes for digitisation, collaboration should include the COST Mobilise training school and CETAF DEST, which is being transformed into an e-training version (https://www.mobilise-action.eu/the-second-training-school-on-digitization-and-data-management-of-collections/; http://taxonomytraining.eu/about). PDF presentations of the MOBILISE Second Training school are available here (https://drive.google.com/drive/u/0/folders/1ozrTyfaT7u9qD4sXto74wzcLj41iNcdU). If you would like powerpoints contact : voreadou@nhmc.uoc.gr.

ELViS update (Wouter Addink)

Wouter Addink presented an update on the development of the European Loans and Visits System (ELViS). The website is here: https://elvis.dissco.eu/welcome.

Version 1 will be ready early next year (2021), but an early production version will be tested for the SYNTHESYS+ Virtual Access Call. Two years of development are needed for the fully functional system. ELViS is being developed to use a range of existing infrastructures, including GBIF, EOSC, etc.

Breakout Session: Calculating % CETAF Collections Digitised

This session ran through the process of calculating the percentage of collections digitised in CETAF institutes as part of the CETAF Strategy targets.

The first step for this exercise is to decide what we mean by digitised, which links directly to the MIDS definition. We will then need to find methods of determining the percentage of collections at each level.

Using GBIF may be good case study, but will not cover all collections, since geology and petrology are not in GBIF. This led to the idea that there are different routes to count the percentage of specimens digitised. One option is the top-down approach which uses

aggregated resources to go from institute, collection down to the specimens. Another option would be to use a bottom-up method of calculation, starting with the specimens and which is closer to the collections. This could include using the IPT, BioCASE or event querying the CMS directly. so that MIDS can also be used at other levels, institutional. Scripts may have to be adapted to the local collections situation. The importance of stable identifiers for this methodology was stressed.

There is currently a consultation exercise taking place, led by GBIF on Advancing the Catalogue of the World's Natural History Collections. More information can be found here: https://discourse.gbif.org/t/advancing-the-catalogue-of-the-worlds-natural-history-collections/1710

ISTC Meeting

Adoption of Agenda

The agenda was approved.

RDA biodiversity data integration IG & Joint TDWG RDA Taskgroup (Wouter Addink)

Wouter Addink gave a short introduction to the Research Data Alliance (RDA, https://www.rd-alliance.org/), whose aim is to build social and technical bridges to enable open sharing of data. There are RDA interest groups and working groups, which produce primarily recommendations. Membership is free for individuals and work is done on a voluntary basis.
The Biodiversity Data Integration Interest Group (BGI IG, https://www.rd-alliance.org/groups/biodiversity-data-integration-ig.html) has 163 members and subgroups for Global Names Architecture, defragmentation of species data management, and vernacular names infrastructure. The following key working areas have been proposed: OpenDS, MIDS/MICS, PIDs, FAIR digital objects, collection access policies harmonisation.
The RDA is an opportunity to advance issues related to the cross-disciplinary interoperability of biodiversity data. Wouter Addink wants to encourage CETAF members to get involved in the RDA process. Interested parties can address the contacts on the website.

Brief Reports from Mobilise (Warsaw)

A Mobilise Cost Action meeting took place from 10 to 14 February 2020 in Warsaw (Poland) focusing on challenges around data mobilisation, publication and re-use derived from natural science collections (https://www.mobilise-action.eu/2019/11/14/mobilise-meeting-in-warsaw/). ISTC members gave a report on the following workshops:

Wikidata workshop (Quentin Groom)

Purpose of the Wikidata workshop was to "develop the data model for taxonomic and nomenclatural data in Wikidata"

(https://en.wikipedia.org/wiki/Wikipedia:Meetup/Cost_MOBILISE_Wikidata_Workshop). Several working groups addressed topics such as taxonomy, people, and “the big picture”. The results of the workshop can be found in the blog post https://data-blog.gbif.org/post/wikidata/.

OpenDS workshop (Alex Hardisty)

The aim of the workshop was the introduction and discussion of the OpenDS concept, which is one of the technical bases of the DiSSCo infrastructure. The concept was generally well received. Various technical and content-related aspects were compiled and discussed in smaller working groups. A summary of the results is still in work.

Georeferencing workshop

The workshop discussed quality problems related to the georeferencing of collection objects and possibilities for cooperative solutions. A summary of the results can be found at https://zenodo.org/record/3734848#.Xp7gSshKiHs.
The possibility of a georeferencing hackathon for the year 2021 should be examined. This could for example be organised in connection with one of the next physical CETAF meetings.

Data Archiving workshop (Dagmar Triebel)

(see next agenda point)

Data archiving strategies in regard to CETAF facilities and planned DiSSCo services - highlighted by COST Mobilise (Dagmar Triebel)

The COST Mobilise WG4 is mainly focused on data archiving and long-term data preservation and not on backup and storage challenges of big data. The working group has 26 active members 13 of which took part in the Warsaw meeting (see https://costmobilise.biowikifarm.net/wiki/WG4_Workshop_%22Data_storage_and_archiving_strategies:_Towards_a_documentation_and_guideline%22_in_Warsaw).
The aim was to develop a common understanding of long-term data preservation and archiving as an essential part of the FAIR data life cycle. Archiving should be based on OAIS principles (http://www.oais.info/). Also relevant are the activities of the German Federation for Biological Data (https://gfbio.biowikifarm.net/wiki/ISO_Standards_for_Digital_Archives).
The workshop participants pre-structured the publication of a guideline for data archiving. It was agreed that this work nicely complements the short DISSCo DMP chapter on storage and archiving.

Linked Open (Collection) Data: the Botany Pilot (Dominik Röpert)

Dominik Röpert presented the status of the development of the "Botany Pilot", which demonstrates the possibilities of linking collection data with Linked Open Data (LOD) technologies. The pilot brings together botanical collection data from a number of CETAF collections (e.g. Meise, RBGE, BGBM) and integrates them via links with external resources (e.g. collectors in WikiData). The BGBM has additionally started to cluster locality data using GeoNames IDs. The methods for this were integrated into OpenRefine. The next step will be to publish the results more widely.

Unique identifiers for collections (Laura Tilley)

Laura Tilley reported on the implementation of the CETAF/DiSSCo Collection Description Dashboard. The dashboard requires Unique Identifiers for Collections and the question is which system should be used for this.

It was pointed out that there are currently a number of initiatives globally dealing with identifiers for collections. GBIF is currently conducting a consultation on this. People interested should contribute to the consultation on Advancing the Catalogue of the World's Natural History Collections (17-29 April) https://discourse.gbif.org/t/advancing-the-catalogue-of-the-worlds-natural-history-collections/1710

Important questions are, for example, which standards for metadata should be used and who should assign the identifiers (centralised vs. decentralised). In any case, the system should be flexible regarding the delimitation of collections, since not all collections are formed according to taxonomic or geographical criteria (e.g. Humboldt collection, medical plants, etc.). These challenges are also being worked on by the TDWG Collection Descriptions Data Standard Task Group: https://github.com/tdwg/cd.

The recommendation would be to first use local identifiers for the development of the dashboard and then to align with the outcome of the GBIF consultation.

CETAF Strategy and Strategic Development Plan revisited (Anton Güntsch)

A table of activities relevant to ISTC and DWG has been created and updated: https://cetafdigitization.biowikifarm.net/cdig/CETAF_Strategy_ISTC_DWG.

Activities and objectives of the strategic plan should be reviewed regularly, as we are now already in the middle of the reference period (2015-2025). The most effective approach is for a smaller group (e.g. DWG/ISTC Lead, DiSSCo SG leads, Wouter Addink) to revise the table again and then send it to the whole ISTC and DWG for feedback.

AOB, next meetings

The participants hope that the next meeting can again take place as a physical meeting. This would then take place in mid-April 2021 at the NHM in London.