

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

Executive Summary

The joint CETAF ISTC (Information Science and Technology Commission) and DWG (Digitisation Working Group) meeting took place at the Natural History Museum Vienna, 13-14 February 2019. The agenda is publicly available at https://cetafdigitization.biowikifarm.net/cdig/ISTC_DWG_Meeting_Spring_2019_Vienna. Presentations as well as these minutes will be made available on the same website. The meeting was attended by 30 participants from 19 CETAF member institutions, the CETAF secretariat, and the University of Cardiff.

Action Items

- To extend the new registry for CETAF-ID implementers with an auto-generated statistics for digital images accessible via CETAF IDs (A. Güntsch).
- To extend the (Wiki-) table of ISTC-related activities and targets of the CETAF strategic plan with digitisation-related topics (E. Haston).
- To contact ICEDIG WP leaders and ask them to add relevant project results to the table (W. Addink).
- To identify deliverables of DiSSCo-related projects, which can be developed jointly with CETAF ISTC and DWG in order to avoid duplicate efforts (All).
- To continue the development of the MIDS standard with a view to reaching wider agreement and publishing the standard as a citable resource (A. Hardisty & E. Haston).
- To identify potential workshops and other activities which could be funded by MOBILISE and which would aim to deliver both the targets of CETAF as well as DiSSCo (All).
- To provide an additional user story for "system for defining annotation hot spots" (A. Güntsch).

Participants

Anton Güntsch (BGBM Berlin, chair & minutes), Elspeth Haston (RBGE, chair & minutes), Heimo Rainer (Vienna), Dominik Röpert (BGBM Berlin), Peter Grobe (ZFMK), Björn Quast (ZFMK), Martin Stein (Copenhagen), Dagmar Triebel (SNSB), Mathias Dillen (Meise), Ayco Holleman (Naturalis), Alex Hardisty (Cardiff), Roger Hyam (RBGE), Falko Glöckler (MfN Berlin), Simon Chagnoux (Paris), Wouter Addink (Naturalis), Frederik Berger (MfN Berlin), Laurence Livermore (NHM), Celia Santos (CSIC), Patrik Mráz (Charles University, Prague), Patricia Mergen (Meise/Tervuren), Ana Casino (CETAF), Carlos Monje (SMNS Stuttgart), Joachim Holstein (SMNS Stuttgart), Karol Marhold (Bratislava), Xavier Vermeersch (CETAF), Thomas

Hörnschemeyer (Senckenberg), Luc Willemse (Naturalis), Jiri Frank (NM, Prague), Anne Koivunen (UH, Luomus), Claudia Kamcke (Braunschweig)

ISTC Meeting (February 13)

Adoption of Agenda

The agenda was approved.

QoS workshop Copenhagen - report, next steps

Anton Güntsch gave a summary of the ISTC “Quality of Service” workshop which took place in June 2018 in Copenhagen

(https://cetafdigitization.biowikifarm.net/cdig/ISTC_QoS_Workshop_Copenhagen_2018). In the workshop, different implementations of CETAF identifiers were analysed and measures to improve stability of identifiers and resolution services were discussed. The participants of the workshop agreed to stop the maintenance of distributed identifier registries (CETAF, RBGE, BGBM) in favour of a central registry with an API used by different software implementations and portals. In addition, it was agreed to profile and bundle semantic annotation activities towards a “Botany Pilot” demonstrating the potential of linked open data for collection data. The following presentations are contributions to this pilot system.

Semantic enrichment of collector information (M. Dillen)

The presentation gave an overview of Mathias Dillen’s and Quentin Groom’s work on the semantic annotation of collector names of the specimen data provided by the Meise digital herbarium. 71% of 1.7M digital specimens have annotated collectors with links to various identifier systems such as HUH, ISNI, and VIAF. Outstanding issues are (not existing) semantic resources for collector teams, the verification of multiple entries for the same person, GDPR and living persons.

In the discussion, it was agreed that “GDPR for living persons” should be analysed more deeply in the coming Mobilise workshops. It was also agreed that semantic annotation activities in different CETAF collections should become more strategic. A solution could be a central index of CETAF IDs and basic metadata which can be used to identify and define priorities for annotation. The BGBM will continue to develop this idea in the framework of the different DiSSCo projects.

CETAF Specimen URI Tester - new developments (R. Hyam)

Roger Hyam gave an overview of the URI Tester developed by RBGE, which is capable tool for checking CETAF IDs and their redirection functions (<http://herbal.rbge.info>). New features include the support of both http and https as well as a monitoring service providing automated testing of registered collections. The monitoring capabilities are already used by RBGE, Meise, and BGBM. All CETAF institutions are invited to contact Roger Hyam and register their identifier implementations.

In the following, Roger Hyam presented an IIF “proof of concept place” showing the integration of CETAF IDs with IIF (<http://iif.rbge.info>). The system enables specimen images from various institutes to be viewed in the same way. IIF will be further implemented in CETAF institutions in the context of SYNTHESYS+.

ID Implementers Registry (A. Güntsch)

Anton Güntsch presented a (working) draft registry of CETAF ID implementers which is to replace the existing registries held by the CETAF secretariat, RBGE, and BGBM. In order to avoid new software implementations, the registry is implemented as a simple Google spreadsheet (<https://docs.google.com/spreadsheets/d/1vHI2xDghffm6HfQhVeruHV6ZAWAnrc-2LPasq0fOyF4>), which can be accessed via the existing Google API. As a proof of concept, Jörg Holetschek implemented a function for auto-generation of the number of available CETAF IDs in the GBIF index (presently ca. 20 Million). It was agreed that auto-generated statistics should include the number of imaged specimens.

"Botany Pilot" (D. Röpert)

Dominik Röpert gave a demonstration of the current development state of the “Botany pilot”. The BGBM runs a (BlazeGraph) triple store with RDF-data harvested from CETAF collections. With this, identical collectors can be identified across CETAF collections and integrated with specimens as well as external resources such as WikiData and BHL. The pilot will serve as a use case for related activities in DiSSCo projects such as SYNTHESYS+, Mobilise, and DiSSCo prepare.

In the discussion the question was raised whether moving all collection data to WikiData could potentially “solve all our problems”. All but one of the participants found the idea at least “interesting”. It was agreed that a future workshop should analyse the potential and risks more deeply.

DiSSCo Technical Infrastructure

Alex Hardisty (Cardiff University) gave an overview of the planning of the DiSSCo technical infrastructure, which is centered around the concept of “digital specimens” providing containers for specimen metadata and links to related object classes (e.g. names, literature, etc.). DiSSCo digital specimens will be identified by a “Natural Science Id” (NSId) representing the “box”. The specification follows the definition of Digital Objects as worked out by RDA Data Foundations and Terminology group. Specimen is central to other information as opposed to GBIF, which uses observation as their central unit. The infrastructure will provide a one-stop shop for users but control is still held by collection-holding partners. A trial implementation demonstrating digital specimens and NSIds is available at <http://nsidr.org/>. Example data has been extracted from the benchmark specimen set compiled by the ICEDIG project. The system uses NSId Handles along with CETAF stable identifiers for the physical object. A major priority for the implementation will be given on community acceptance of the digital object architecture. In the discussion, the following points were raised:

- The question of “what’s in the box?” and who is controlling the information needs more discussion. Will supplementary information need to be accepted by institutions

controlling authoritative information? The question should be addressed as a work item for MOBILISE.

- The wide range of information stored “in the box” can potentially lead to a huge harvesting and synchronisation task. Which kind of infrastructure will be able to deal with this? Will (external) data be archived by DiSSCo rather than counting on their stability for the next decades? DiSSCo will very likely not be able to take responsibility for external services.
- The DiSSCo architecture provides a clarification of the relation between NSIDs and CETAF IDs. However, the wording should be re-considered to explain the distinction between physical and digital object in a better way.
- Agreeing on data standards and processes for standards will become more important as we move into the implementation of DiSSCo. MOBILISE workshops should be used to enable discussion and collaboration at European (CETAF) and global (TDWG) levels.
- It will be important to explain the unique characteristics and profile of DiSSCo compared to GBIF. DiSSCo focusses on specimens as the center of the knowledge graph. It is also driven by scientific uses rather than acting as just a data repository.

DiSSCo related activities

COST Mobilise WG4 "Development of Standards and Guidelines for data archiving and long-term preservation" (D. Triebel)

Dagmar Triebel gave a general introduction to MOBILISE COST Action and the Working Group 4 dealing with data archiving and long-term preservation. Internal and external communication is handled with the OSF wiki. A first WG4 workshop will be held in March in Sofia with 12 participants from 7 countries. Further meetings will be organised as video conferences. Results will be documented in an open wiki system. The working group aims to cooperate with other MOBILISE WGs as well as related projects and initiatives. A symposium is planned for the 2019 Biodiversity next conference. Active partners, esp. Technical experts are welcome.

GeoCAsE – current situation and next steps (F. Glöckler)

Falko Glöckler gave a summary of the current situation of GeoCAsE (<http://www.geocase.eu/>). Currently, (apart from fossils) geological data are not include in GBIF. Europeana deals only with multimedia data. Most of the data concepts are identical in bio- and geo-collections. Differences are addressed by ABCD-EFG (Extensions for Geosciences), which has been formally published recently. Presently, there is no funding for improvement of the technical platform. Tasks for developments have been included in the DiSSCo prepare work plan. Ideas for next steps are i) content improvement from Earth Science meetings, ii) improvement of search interface and machine readability, iii) landing pages. The discussion of tools and services should take place in ISTC. Please send suggestions for improvements to Falko Glöckler!

COST Mobilise Person ID workshop (E. Haston)

Elspeth Haston informed the group of the upcoming workshop on the authority management of people names which will be held in Sofia, Bulgaria at the MOBILISE meeting. This workshop will bring together key contributors from across the relevant range of disciplines to agree a strategy for adoption and implementation for the authority management of people names in Natural History Collections and cross-discipline research and exposure. A wiki for the workshop has been created <https://osf.io/qwegk/wiki/home/> where people can find out more information and follow the outcomes of the workshop.

On the Way from Nagoya to Legal Handling of Accession Units in collections (P. Grobe)

Museum Koenig now employed someone to help implement Nagoya protocols in the collections. The CETAF code of conduct and best practices have been helpful as a basis for a range of workflows and activities allowing to trace every specimen with information about permit documentation. Workflows have been integrated into the Diversity Workbench Collection software.

Similar implementations in other collection management systems developed by CETAF organisations were discussed (e.g. DINA/Specify and JACQ). Systems should have automated flags showing the existence of relevant documents and providing links to them. It is up to the users to deal with old and “other” languages. The CETAF ABS group has two workshops lined up dealing with issues related to Nagoya implementation in collections.

CETAF strategy and development plan - targets and activities

Anton Güntsch gave a short introduction of the CETAF strategy and development plan, which defines CETAF activities and targets for the years 2015 - 2025

(https://cetaf.org/sites/default/files/final_strategy_and_strategic_development_plan.pdf). He created a wiki page with a copy of ISTC-related activities

(https://cetafdigitization.biowikifarm.net/cdig/CETAF_Strategy_ISTC_DWG). The page will be a tool for i) identifying existing activities and products which support the fulfilment of targets, ii) identifying remaining gaps, and iii) aligning works with DiSSCo-related activities.

It was agreed that DWG-related activities/targets should be added to the Wiki-page because there is a significant overlap between digitisation and informatics targets. Anton Güntsch and Elspeth Haston will continue to maintain the page on the BioWikiFarm and not migrate to Teamwork at this point. Wouter Addink will contact ICEDIG WP leaders and ask them to add relevant project results to the table. All ISTC/DWG members to review the table and fill the gaps.

AOB

The next joint ISTC/DWG meeting will be held in spring 2020 at the NHM/London (tbc).

DWG Meeting (February 14)

Adoption of Agenda

The agenda was approved.

CETAF strategy and development plan - targets and activities (E Haston & A Güntsch)

Continuing from the discussion in the ISTC meeting, it was decided to add the actions and targets of the CETAF strategy and development plan relevant to the DWG into the same wiki table as those for the ISTC. This will enable a clearer overview of the relationship between the two groups in terms of the CETAF Strategy. It was agreed that the actions and targets need to be disseminated more within other projects such as SYNTHESYS+, MOBILISE, ICEDIG and DiSSCo Prepare. The main focus area within the Strategy for the DWG is Focus Area #3: Natural History collection management and access to collections. The key target for the DWG is "10% of our 1.5 billion natural history collections are databased, digitised and digitally available and scientific collection visits increase by 10%".

Minimal Information for Digital Specimens (MIDS) (A Hardisty)

Alex Hardisty presented the current proposal for the standard for Minimal Information for Digital Specimens, work being carried out by jointly within the ICEDIG project and the CETAF Digitisation Working Group. The concept is based on the MIxS in the genomics community which specifies the mandatory and optional information elements that must be present. The aims of MIDS are: 1) to provide clarity about minimum quantity/quality of information, and 2) to offer a framework for monitoring progress and setting priorities. The proposal recommended the use of 4 levels (0=catalogue, 1=basic, 2=partial, 3=extended). Level 2 (partial) would be the minimum standard to aim for as best practice recommendation. A draft specification of object image types (resolution etc) was also included in the proposal. Future work included aiming to achieve a global consensus potentially in the form of a TDWG standard, and the publication of the MIDS. Feedback was requested and received.

- The use of the term mandatory was discussed and the alternative of 'recommended' could be considered. The mandatory property enables the assessment of the progress of the community as a whole, and for individual institutions it provides guidelines on what needs to be achieved, also helping with the planning of digitisation efforts and for reporting.
- In terms of images, the current inclusion of images as optional elements of many of the MIDS levels was questioned as was the inclusion of image recommendations in terms of format and resolution given the variation between collections and the rapid development of imaging quality. Removing image characteristics will be considered.
- Data quality was raised as a concern and a suggestion was made to add an element for "data has been checked".

- It was agreed that most current digitisation efforts would potentially qualify for Level 1 or Level 2. More detail will be needed for the standard to become a guideline for collections, including mapping to ABCD and DwC.

The final conclusion was that there was general agreement that the standard should be developed further, and that it would also be suitable for Earth Sciences. It was agreed that MIDS would be developed further independently from TDWG, involving key persons including people within iDigBio, and that it could become a TDWG standard later.

Review of digitisation within DiSSCO-related projects and work programs

SYNTHEsys+ (E Haston & General Discussion)

Elsbeth Haston presented a summary of SYNTHEsys+, focussing on the relationships between SYNTHEsys+, MOBILISE and the CETAF ISTC and DWG. Key areas include Virtual Access which is based on a novel European Loans and Visits system (ELViS) which will be designed and built within SYNTHEsys+ and which require some form of integration with collection management systems. Wouter Addink who is leading this Workpackage within SYNTHEsys+, provided some additional information on ELViS. The Virtual Access Programme will be released in summer 2020 and there will be two pilot calls. The work aims to lead into DiSSCo in a practical way, through digitisation on demand. The monitoring of users will be an important aspect and will also be interesting to the DWG. All Milestones and Deliverables will be published on Pensoft RIO. Participants at the meeting were reminded that MOBILISE can/should be used to organise and finance meetings and workshops. In terms of the ISTC, JRA3 Specimen Data Refinery, is most relevant. See <http://www.synthesys.info/home.html> for updated project information and also information on previous phases of SYNTHEsys.

ICEDIG Survey Results (X Vermeersch)

Xavier Veermersch presented on the survey carried out under ICEDIG Task 8.1 Social effects and capacity building, which was an analysis of digitisation process and a survey of existing digitisation workforces. The results, summarising the initial findings, are published as ICEDIG Milestones 48. There were ca. 200 responses, most of which were representing botanical collections. Main findings included: most institutions are short of staff, difficult to hire, difficult to keep; the importance of better training was stressed; ranging from data management to software skills, and taxonomic knowledge; digitisation has a positive influence on collaborative work; decrease in loans and visitors. A pdf with a summary is available by contacting Xavier (xavier.vermeersch@cetaf.org). The question of whether a decrease in the number of visitors potentially caused by digitisation could be an issue for institutes was raised. We should see an increase in the number of downloads and we will need to consider this in how we report in the future.

ICEDIG Gathering of user stories (A Hardisty)

Alex Hardisty briefly presented the gathering of user stories within ICEDIG. So far there are 78 stories which have resulted from a community consultation. From 1,200 people consulted there were 378 responses, resulting in 69 stories. Additional stories have been added since. It was recommended that we join this with similar national lists (e.g. Germany, Belgium). ICEDIG will consider making their list public. Anton G agreed to provide an additional user story for "system for defining annotation hot spots" mentioned in the ISTC meeting. It would be good to look at GBIF analysis of usage of GBIF data although this is probably not possible with ICEDIG resources.

MOBILISE (E Haston & General Discussion)

Elsbeth Haston presented a brief overview of the MOBILISE Cost Action, emphasising the flexibility including the option of additional partners joining. People are encouraged to consider joining the working groups. Funding is possible if you belong to one of the accepted countries.

More information can be found on the MOBILISE website (<https://www.mobilise-action.eu/>) and on the MOBILISE wiki (<https://osf.io/j6psx/wiki/home/>). In principle, there is a close connection between WG2: Development of Standards and Guidelines for Data gathering and large-scale digitisation of collection objects and the DWG, and a close connection between WG3: New Concepts and Standards for Data management in relation to content, curation, quality management, technical framework and documentation and the ISTC. Some points to consider:

- Joint workshops can be a problem.
- There is no funding for PMs. Work should therefore be carried out during the workshops as far as possible.
- Budget needs to be set for each year. Budget for Year 2 will be voted on in Sofia and submitted to the EU for ratification.
- Year 2 starts end of April, send ideas/views to the management committee as soon as possible.
- There are also short term scientific missions (comparable to Synthesys) but without funding for the host; developer meetings could use these instruments
- Mobilise training school is running a course; Digitisation and data management challenges in small collections (14-15 March 2019 in Sofia) dealing, for example, with data cleaning, dq. will use GBIF material, which can also be used after the meeting.

The difficulties of finding contact information was raised and will be looked into.

CETAF Collections Group (C Quaiser)

Christiane Quaiser presented an overview of the work of the Collections Group and the current vision for the revised group moving forwards. She emphasised the close relationship between the CCG and the DWG, particularly given that the Digitisation Working Group was formed at a meeting of the Collections Group. The potential for joint actions between the CCG, DWG and ISTC was proposed, with the recognition that we would need to find methods of exchanging information.

In both the CCG and the DWG the relationship between the physical and digital collections has been discussed and this is a clear area for collaboration between the two groups. The IT capacity within institutes was discussed with reference to collections, and the technological training required in some institutes. There are existing initiatives that can be implemented, eg DarwinCore Hour. We can also look at software demonstrations, workflow discussions, with the idea of finding low budget ideas that would have a large impact for collections and collections staff.

The idea of further joint meetings between the three working groups was discussed and this discussion will continue.

AOB, next meetings

NHM London offered to host the next joint meeting of the ISTC and the DWG in Spring 2020.