Minutes of the joint CETAF Digitisation Group and ISTC Meeting, Copenhagen, 21-22 February 2018

Executive Summary

The joint CETAF ISTC (Information Science and Technology Commission) and DWG (Digitisation Working Group) meeting took place at the Natural History Museum Copenhagen, 21-22 February 2018. The agenda is publicly available at http://cetafdigitization.biowikifarm.net/cdig/ISTC_DWG_Meeting_Spring_2018_Copenhagen. Presentations as well as these minutes will be made available on the same website.
The meeting was attended by 23 participants from 16 CETAF member institutions and the CETAF secretariat. The purpose of the meeting was to report on important activities of the ISTC and DWG, to plan objectives for the next year, and to align activities to the CETAF Strategy and Strategic Development Plan 2015-2025.

Action Items

- To submit an abstract for the identifier workshop at TDWG/SPNHC 2018 (A. Güntsch).
- To organise a “quality of services workshop” addressing harmonisation and stability of identifier services provided by CETAF member organisations (CETAF ID implementers).
- To setup and maintain a dynamic list of actions and products contributing to the fulfillment of targets defined by the CETAF strategy and development plan 2015-2025 (ISTC and DWG).
- To organise a meeting for discussing EU calls and potential CETAF participation (CETAF).
- To submit an abstract for the results of the digitisation surveys to TDWG/SPNHC 2018 (S. Phillips)
- To complete the analysis and review of the digitisation surveys (E. Haston & S. Phillips)
- To produce a list of digitisation challenges with potential “owners” who can report back to the group (DWG)
- To produce a list of digitisation case studies to represent the definition of digitisation (DWG)
- To calculate the number of digitised specimens within CETAF institutes (E. Haston & GBIF)

Participants

Quentin Groom (Meise), Anton Güntsch (BGBM), Patricia Mergen (RMCA / Meise), Dominik Röpert (BGBM), Wouter Addink (Naturalis), Ana Casino (CETAF), Karsten Gödderz (CETAF), Björn Quast (ZFMK), Falko Glöckler (MfN Berlin), Dagmar Triebel (SNSB), Ingimar Erlingsson (Swedish Museum of Natural History), Elspeth Haston (RBGE), Wilfred Gerritsen (Naturalis),

Sarah Phillips (Kew), Dare Talvitie (Finnish Museum of Natural History), Jere Kahanpää (Finnish Museum of Natural History), Jiří Frank (Prague), Martin Soucek (Prague), Jörg Lange (Stuttgart), Simon Chagnoux (Paris), Heimo Rainer (Vienna), Christian Steinwender (Vienna), Martin Stein (Natural History Museum of Denmark)

ISTC Meeting

Adoption of Agenda

The agenda was approved.

Stable IDs summary of activities and discussion of next steps

Anton Güntsch gave a summary of achievements in 2017. As planned in the last ISTC meeting, CETAF stable identifiers have been advertised in several conferences and workshops as well as in a Nature communication article. In addition, an entertaining video has been produced by RBGE in the context of SYNTHESYS+. ISTC/DWG members agreed to continue their outreach activities in particular in their own institution. It would also be useful to talk to publishers and discuss CETAF IDs as a standard format for specimen citations. A poster and powerpoint slides are available for these purposes. An abstract for the identifier workshop at the TDWG/SPNHC meeting in New Zealand will be submitted. ISTC and DWG members are invited to co-author this abstract.
Another important activity in 2017 was the virtual “LOD Sprint”, which took place in the week October 16-20. During the sprint, options for implementing a future European Identifier catalogue were discussed and documented. Furthermore, effective mechanisms for semantic annotation of collection data were documented. Results of the meeting are accessible at http://cetafidentifiers.biowikifarm.net/wiki/IDs_and_LOD_Discussion and provided the basis for activities towards the “semantic specimen catalogue” in the SYNTHESYS+ proposal.
Our assessment of the present identifier CETAF infrastructure (15 implementers, 22M records) showed that the implementations are heterogeneous and provide different levels of redirection mechanisms as well as manifold representations of specimen metadata. Although different levels and a certain heterogeneity is of implementations is part of the basic specification, it was agreed that a higher degree of standardisation and stability of services should receive more attention. To this end, a “quality of services” workshop will be organised for June 2018. The workshop will also identify useful pilot applications, which are easy to achieve a demonstrate for the potential of CETAF IDs and linked open data (LOD).

The Netherlands Biodiversity API

Wilfred Gerritsen gave an overview of the Netherlands Biodiversity API, which provides an open service-based interface to both unit-based and taxon-level data objects. The API integrates with a range of international networks and has a swagger-based list view for documenting and testing individual service endpoints. It was agreed that harmonisation of service APIs between CETAF institutions would be a useful activity and could be integrated into the best-practices

documentations, which are part of the CETAF Strategic Development Plan. The work should be aligned with the TDWG Biodiversity Services and Clients Interest Group (BSCI).

Die Herbonauten

The CETAF/ISTC collaboration on Les Herbonautes / Die Herbonauten was presented by Dominik Röpert. The German translation of the system developed by MNHN Paris is successfully running at the BGBM. So far, 6 missions have been launched, 4 of which have been completed already (http://herbonauten.de/missions). The results with regard to speed of data entry and quality are extremely positive. RGBE is working on an English instance of the platform and Copenhagen considers a Danish translation. It was agreed that citizen science approaches for transcribing specimen label data should be part of the SYNTHESYS+ project.

Other projects

IndExs - Index of Exsiccatae

Dagmar Triebel presented the DiversityExsiccatae and IndExs modules, which are part of the Diversity Workbench. The web-interface as well as further information is available at http://indexs.botanischestaatssammlung.de/. Cooperation for compiling images exist with 41 herbaria. The project is now seeking cooperation with larger networks such as CETAF, DiSSCo, as well as H2020 EU projects.

MORPHYLL

Jörg Lange presented the Morphyll database for fossil leave data, which is a PostgreSQL based information system for morphometric data with a focus on fossil leafs (http://morphyll.naturkundemuseum-bw.de). The system will now be implemented towards a web-based system with annotation and analysis functions.

The CETAF Strategy and Development Plan (2015-2025)

Biodiversity Informatics and information technologies are one of six focus areas addressed by the CETAF Strategy and Development Plan (2015-2025). It was agreed that ISTC should from now on revisit in every meeting the actions defined by the plan and align the ISTC activities on the strategic targets. Several targets have already solutions which have been developed over the last years (e.g. identifiers, interoperable standards, etc.) and need to be documented as part of a data curation guideline. However, there are several activities, which need to be interpreted and defined more clearly. It was agreed that the ISTC will work collaboratively on a Wiki table consisting of i) actions mentioned under focus area #4 of the strategic plan, ii) a more precise interpretation of actions, ii) targets each of the action is contributing to, and iv) existing or required CETAF (or other) products, which fulfill these targets and contribute to a best practices documentation. It was also agreed that a CETAF best practices documentation for data curation will be a dynamic document, which can evolve over time.

DiSSCO

Wouter Addink presented an overview of DiSSCO technical architecture. There is a great overlap between envisaged DiSSCO implementation plans and ISTC activities (e.g. identifiers, LOD, semantic enrichment, etc.). It was agreed that ISTC will fully support the DiSSCO infrastructure and take it always into consideration when planning future activities. To ensure that DiSSCo and ISTC activities will be aligned, the ISTC strategy and development plan documentation will be extended with a DiSSCo-column defining how ISTC activities can contribute to the DiSSCo development.

EU Opportunities

Patricia Mergen presented an outlook on EU FP9 planning as well as potentially relevant calls in H2020 (see presentation). It concerns calls about Integrating Activities for Advanced Communities (2019), participation to the European Open Science Cloud (EOSC), information and communication technologies and Socioeconomic and Cultural Transformations in the context of the fourth industrial revolution. It was agreed that a meeting should be organised where CETAF members can discuss participation in the coming calls. The meeting should take place as a side-meeting of CETAF43 in London.

SYNTHESYS+ Overview of ISTC-related activities

Elspeth Haston gave an overview of the SYNTHESYS+ proposal to be submitted in March. ISTC members are contributing in particular to the work package NA4, which is lead by TDWG and deals with standardisation. A strong focus is on identifiers and their application in larger networks such as JACQ and DINA as well as the implementation of a “semantic identifier catalogue”. The work package will also work on the standardisation of APIs for images and image metadata based on iiiF. In this context, ISTC and DWG should revisit the TDWG AUDUBON core to assess its role in a iiiF-based infrastructure. The work package NA2 deals with “harmonisation of processes” and will very likely also have a relevance for the ISTC/DWG process.
In the discussion it was noted that IT infrastructures, which have been erected as part of a CETAF initiative should be “brought back” to ISTC and discussed in the context of new technical developments. Examples include BioCASE, OpenUp, and PESI.

DWG Meeting

Adoption of Agenda

The agenda was approved.

Review of digitisation based on recent surveys

Digitisation Resources & Gap Analysis Survey

Elspeth Haston presented a preliminary review of the digitisation resources and gap analysis survey. There were some very clear areas which have big challenges for many institutes highlighted in the survey. Some early analyses of the survey results were able to categorise actions which would have impact into four categories: Technology, Workflows, Training and Staff. The correlation and complementarity of the institutes was also investigated for one of the sections to assess if it could be useful to identify shared needs and potential collaborations where knowledge can be shared. More work is needed to go through the survey more thoroughly and complete the analysis.

SYNTHESYS3 State of Digitisation survey

Sarah Phillips presented the results of a state of digitisation survey that was carried out as part of the SYNTHESYS3 project. This review summarises some of the main components and underlying issues with respect to digitisation workflows as evidenced from questionnaire responses from SYNTHESYS partners, as well as providing some key recommendations based on these findings.

Discussion

Following the two presentations, there was discussion relating to the overall outcome of these surveys and the next steps for the community.

The following next steps were suggested:

- Include the needs which have been identified in funded projects where possible
- Find a way of pulling together existing knowledge, expertise, processes, workflows etc, in such a way to be useful in helping other institutes
- Use the information as an evidence base when requesting resources

It was also suggested that institutes who are already working on a particular challenge could take ownership of that challenge and help find solutions through projects and collaborations.

The following challenges are being prioritised by institutes which could form a starting point for this to happen.

Challenge	Institute
Audio Visual workflows	NMP
Image QC	BGM
Digitisation in the Field	MfN

Prioritisation of Digitisation	BGBM

Projects which include work on solving some of these digitisation challenges are:

ICEDIG (2017-?)
SYNTHESYS+ (to be submitted)
Herbadrop (dates)

It was agreed that we could look at publishing more workflows etc to journals such as the European Journal of Taxonomy, Biodiversity Data Journal (workflows), Journal of Biotechnology (infrastructure, digitisation), DATABASE: The Journal of Biological Databases and Curation (database research, biocuration).

We can also produce more videos and webinars to help transfer knowledge and information.

Digitisation Definition for Collections

Elspeth Haston presented the challenge that has been set by the CETAF Executive Committee to produce a set of definitions for digitisation in the context of collections. Following discussion, the complexity of the task was clear, particularly in terms of the different terminology in place and the different collections.

It was agreed that a useful step would be to produce a list of digitisation use cases with information about the definitions and standards used in each case. This list would be created on the Digitisation Working Group wiki and openly available for more use cases to be included.

The CETAF Strategy and Development Plan (2015-2025)

Actions:
Digitising and databasing collections
i. Exploring methods and best practices in data storage and data curation
ii. Supporting the development of digitisation policies, techniques and technology
iii. Exploring the development of storage and digitisation infrastructures

Targets:
10% of our 1.5 billion natural history collections are databased, digitised and digitally available

It was clear that the Actions are being covered within several EU projects including SYNTHESYS3, ICEDIG, Herbadrop and potentially SYNTHESYS+ and BETHS in the future.

Elspeth Haston will discuss pulling the data from GBIF with reference to the target of 10%.

SYNTHESYS+ continued

This item is covered in the previous day's minutes further up.

ICEDIG Project call for test data

Sarah Phillips and Quentin Groom asked the group for submissions to create a test dataset. The specimens need to have already been imaged and transcribed. The dataset would potentially include a range of specimens in terms of date, taxon and geography. Some discussion took place on the statistical robustness of the dataset, in terms of bias and random selection.

More information on this dataset would be available at the forthcoming ICEDIG meeting in Helsinki.

TDWG World Schema for Plant Geographic Distributions

Quentin Groom informed the group that TDWG had been reviewing some of the standards in terms of ownership. One of these was the World Schema for Plant Geographic Distributions. There was some discussion about what exactly was currently included within this standard, but it was agreed that the standard should be updated to include shape files. The next step will be to find someone who would be willing to take ownership of the standard with responsibility for updating it. All members are asked to advertise this. Anyone interested should contact Quentin Groom in the first instance.

Research Agenda for CETAF

Ana Casino asked the group for contributions of ideas for the development of the research agenda for CETAF. These should be brought to the next full CETAF meeting at the NHM London in April.

Raising Awareness of Geological and Earth Science Collections

Falko Glöckler spoke to the group to raise awareness of the geological and earth science collections. The GEOCASE portal is the equivalent to the GBIF platform but lacks the public awareness of GBIF, and the funding to maintain it. The digitisation working group should therefore include these collections more in any discussions, and members are encouraged to help raise awareness more generally in their institutes and with their colleagues. If people are looking at projects, efforts should be made to include these collections.

Next Meeting

The next joint ISTC & Digitisation Working Group Meeting will be held in February 2019 in Vienna.