CETAF
CONSORTIUM OF EUROPEAN TAXONOMIC FACILITIES

Exploring and documenting diversity in nature

Minutes of the joint CETAF Digitisation Group and ISTC Meeting, Stuttgart, 27-28 March 2017

Executive Summary

The joint CETAF ISTC (Information Science and Technology Commission) and DWG (Digitisation Working Group) meeting took place at the Naturkundemuseum Stuttgart, 27-28 April 2017. The agenda is publicly available at http://cetafdigitization.biowikifarm.net/cdig/ISTC_Meeting_Spring_2017_Stuttgart. Presentations as well as these minutes will be made available on the same website.
The meeting was attended by 26 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 identify fields for cooperation, and to plan concrete activities for the next year.

Action Items

- To implement a machine-readable catalogue of available CETAF stable identifiers to facilitate fast queries across CETAF collections and advanced inference and data integration functions (Berlin-BGBM, Berlin-MfN).
- To implement mechanisms for linking (CETAF) collection metadata records to external resources such as person databases, gazetteers, and scientific name services (Paris, Vienna, Bonn, Berlin-MfN, Berlin-BGBM, Meise).
- To organise a workshop addressing the International Image Interoperability Framework (iiiF, http://iiif.io/) and its applicability in the context of CETAF digital image processing (Edinburgh).
- To provide a small and nice web-enabled CETAF icon which can be used for branding CETAF stable identifiers in web-portals (Secretariat).
- To advertise the use of stable identifiers for scientists (All)
- To create and send out a digitisation resources gap analysis survey (Edinburgh)
- To request figures on the number of specimens digitised from partner institutes in order to assess progress on target to digitise 10% of collections. Figures will also be pulled from GBIF and compared (All)
- To propose to CETAF that the figures for number of visits in the the CETAF targets should include digital visits to the search landing page of institutes. If CETAF agree, institutes will be requested for these figures (Edinburgh)
- To supply examples of use cases of stable identifiers and upload them to the wiki (All)

Participants

Wouter Addink (Leiden), Ana Casino (CETAF), Simon Chagnoux (Paris), Johanna Eder (Stuttgart), Jiří Frank (Prague), Katarina Gatialova (Prague), Falko Glöckler (Berlin, MfN), Karsten Gödderz (CETAF), Peter Grobe (Bonn), Quentin Groom (Meise), Anton Güntsch (Berlin, BGBM), Elspeth Haston (Edinburgh), Thomas Hörnschemeyer (Frankfurt), Joachim Holstein (Stuttgart), Jana Hoffman (Berlin, MfN), Ayco Holleman (Leiden), Jiří Kvacek (Prague), Patricia Mergen (Patricia Mergen (RBINS/ RMCA/Meise), Juan Carlos Monje (Stuttgart), Björn Quast (Bonn), Dominik Röpert (Berlin, BGBM), Martin Stein (Copenhagen), Christian Steinwender (Vienna), Ari Taponen (Helsinki), Dagmar Triebel (Munich), Marie-Hélène Weech (London, Kew)

ISTC Meeting

Adoption of agenda

The agenda was approved.

Report from identifier initiative, discussion of next steps

Anton Güntsch gave a summary of the results of a workshop of the CETAF stable identifiers implementers group, which took place in the morning before the ISTC meeting. The initiative is well underway with two new implementations by Tervuren and Meise. The URI Tester (http://herbal.rbge.info/) implemented by RBGE is available for testing existing implementations. The Wallich Catalogue (http://wallich.rbge.info/) can be used as a demonstrator for linking specimens from distributed CETAF collections into a web-based information system. The initiative published the CETAF approach in Database Oxford journal (https://academic.oup.com/database/article/doi/10.1093/database/bax003/3053443/Actionable-long-term-stable-and-semantic-web).

For 2017, the following potential activities were discussed:

- Additional implementations of stable identifiers in CETAF institutions.
- Improve Linked Open Data capabilities by linking out to external resources.
- Improve visibility and use of CETAF IDs in external information systems.
- Create CETAF Index for fast searching, improved inference, and high visibility.
- Monitoring availability of redirection mechanisms and validity of response documents.
- Communicating the use of stable identifiers for scientists.

It was agreed that activities in 2017 should be focused on implementing a joint index (catalogue) for CETAF identifiers and to work on mechanisms to link out to external resources such as person database, gazetteers, and taxon name databases. In addition it was agreed that the use of stable identifiers in a scientific context should be promoted in the following ways:

- Presentation at SPNHC 2017
- Presentation at TDWG 2017
- Presentation at the Int. Bot. Congress 2017
- Nature Letter to the Editors
- Internal presentations in CETAF institutions.

It was further agreed that CETAF stable identifiers could be branded using a small CETAF icon. A design for this icon could be provided by the CETAF secretariat.

Report from LOD Hackathon, discussion of next steps and collaboration opportunities

Ayco Holleman gave a report of the "Experiments and documentation for the Linked Open Data mini-Hackathon" (25-26 April 2016 in Leiden) with participants from Naturalis, NHM, and BGBM. In the hackathon harmonised machine-readable access APIs to CETAF collection data and their semantic integration were discussed and implemented as prototypes (https://github.com/naturalis/lod-hack/wiki). It was decided that in future both the CETAF identifier initiative and the LOD initiative should go together under the umbrella of ISTC.

Other ISTC/CETAF initiatives

ISTC members gave presentations of biodiversity informatics initiatives related to CETAF (presentations available from):

- Herbadrop (Simon Chagnoux)
- BHL-E (Jiří Frank)
- Geo-referencing (Ayco Holleman)

The Herbadrop consortium (Paris, Berlin-BGBM, Edinburgh, Digitarium, CINES) is presently discussing the submission of a proposal for phase II. Experimenting with different OCR qualities and their impact on text-mining capabilities and other metadata analyses (e.g. duplicate detection) should be part of the second phase. Comparing results from different OCR software packages using data analysis platforms (e.g. OpenRefine) could also be an interesting activity. Additional partners from the CETAF consortium can be included. A proposal-writing workshop will take place in April or May 2017 in Montpellier.

The geo-referencing project, which aims to provide a software platform implementing optimised geo-referencing workflows, is delayed. Leiden will continue with the works and ask CETAF partners for contributions once a first demonstrable is available.

Other developments of ISTC members

Further biodiversity informatics related projects and systems where presented by ISTC members (presentations available from http://cetafdigitization.biowikifarm.net/cdig/ISTC_Meeting_Spring_2017_Stuttgart):

- Common name services (Christian Steinweder)
- GFBio pipelines for collection data (Dagmar Triebel)
- MNHN Collection website and 3d gallery (Simon Chagnoux)
- Collection portal developments at the Botanic Garden Meise (Quentin Groom)

The Common Name Service provided by Vienna still accepts additional data sources provided by CETAF partners. They can be provided in the form of data files or as service APIs, which will be linked dynamically to the central common name service.

TDWG Biodiversity Information Standards

Quentin Groom gave a report on current activities of TDWG Biodiversity Information Standards including the ratification of the GGBN data standard, the formalisation of the standardisation process itself, and the specification of important DarwinCore elements and associated controlled vocabularies. It was agreed that TDWG activities are highly important for CETAF and that ISTC members should try to attend the annual TDWG meetings. The 2017 TDWG annual meeting will take place in Ottawa, Canada.

DiSSCo

Wouter Addink gave an update on the DiSSCo initiative. An important component of the envisaged infrastructure would be a central RDF triple-store providing the basis for harmonised access to different data types provided by European collections (specimens, observations, traits, etc.). The ISTC works on stable identifiers and Linked Open Data provide an excellent basis for this architecture.

EU Funding opportunities (Patricia Mergen)

Patricia Mergen provided an overview of existing and drafted EU calls, with potential relevance for CETAF and CETAF-ISTC. There are several calls addressing big data topics, where CETAF could play a role as a use case (e.g. ICT14 - ICT17). Big data would then be interpreted as "highly complex data" rather than big volumes of data. There is another call (ICT-43, Reinforcing European presence in international ICT Standardisation), which could be used for strengthening the role of CETAF in TDWG Biodiversity Information Standards. The information about draft calls is partly still confidential. More details can be requested directly from Patricia Mergen.

Ideas for collaboration

The International Image Interoperability Framework (iiiF, http://iiif.io/) is a platform which aims to provide harmonised access to distributed image repositories and associated metadata as well as convenient mechanisms for image annotations. To learn more about the platform and assess its applicability in the CETAF context, an iiiF workshop should be organised.

Next meeting

The next full ISTC meeting will be held in spring 2018 in Copenhagen.

DWG Meeting

Adoption of agenda

The agenda was approved.

CETAF Strategy & Targets

The 25 CETAF targets were reviewed and targets relevant to the ISTC and DWG were identified. During this process it was realised that more clarification on the targets and the metrics required for them are needed.

The targets that were discussed in more detail to determine relevance include:

Research in systematics & taxonomy, integrating new tools and innovations

- 90% of the journals of our member institutions are digitized and digitally available

  This was considered by the participants of the meeting to refer to the journals published by the member institutes. As such, it was estimated that this target has been achieved, but it is not yet completely clear from the passport data if members' journals are digitally available.

Science policy and key performance indexes

- CETAF position papers and statements are published regularly

  These potentially include ISTC Stable Identifier papers and statements

Natural history collection management and collection access

▸ 10% of the 1.5 billion specimens in our natural history collections are databased, digitized and digitally available, and scientific collection visits increase by 10%

It was agreed that we could aim to measure the percentage of digitised specimens in our collections using two methods. one was the inclusion of the number of digitised specimens in the CETAF passport where the total number of collections has already been recorded. The other is to pull as much information as possible from GBIF to see if these data can be used to represent the figure - this has the potential to save work by the curators. GBIF data will not include some collections such as minerals, etc.

It was also agreed that increasing the number of scientific visits will be difficult to achieve, given the ending of SYNTHESYS3 and the general push to encourage more digital visits. It was also agreed that we should make a recommendation at the CETAF meeting that visits should include digital visits. This would be measured by the number of visits to the institutional collections main search pages.

▸ CETAF best practice and common collections policies are implemented in the majority of member institutions

This would potentially include the Recommendations for a Management Policy on Digital Collections. This was an output from the SYNTHESYS3 project and has been formally handed over to the Digitisation Working Group. The stable identifiers could be included here but these are included specifically in a later target.

Biodiversity informatics and information technologies

▸ CETAF interoperable standards for biodiversity data and natural history specimen databasing are adopted in 80% of CETAF institutions

These standards potentially need to be clarified. Essentially, these could refer to TDWG standards such as Darwin Core and ABCD in which case we have probably reached the target.

▸ CETAF standard identifiers are adopted in 80% of CETAF institutions

Standard identifiers are now being used in 13 CETAF institutions. There are 34 members in total.

▸ Digital data curation guidelines are produced and adopted

Some discussion included need to look at other initiatives including iDigBio and Arthur Chapman's work.

▸ Digital storage capacity infrastructure is established within the Consortium

Some discussion took place about the feasibility of a shared storage infrastructure and whether this was a realistic target unless it includes the concept of a system of separate infrastructures.

Communication, outreach and societal relevance

▸ Two citizen science or crowd-sourcing initiatives are undertaken

Some institutes are undertaking these at present but potentially not including the CETAF logo and branding.

Update on SYNTHESYS3 digitisation

Elspeth Haston presented an overview and highlights of the Joint Research Activity (JRA) workpackage within the SYNTHESYS3 project. A point that had been identified at the JRA meeting in Edinburgh in March 2017 was the need to make the outputs more accessible. The end date of the project is August 2017 and so there is a need to raise awareness and availability of the outputs before this date.

Digitisation at the Naturkundemuseum, Stuttgart

Joachim Holstein presented an overview of the digitisation programme at the Naturkundemuseum in Stuttgart. This included information about the development and use of the Diversity Workbench, their current web portals and the digitisation projects with which they are involved.

Implementation of Stable IDs in the physical collection. Implications for collection workflow

Falko Glöckler presented some workflows and use cases for the CETAF Stable Identifiers for specimens, including embedding them within QR codes, use cases for mass digitisation, collection management, as a generic tool for forms and as part of image analysis. A call was made for more examples of use cases and a wiki page has been created as a repository for them:
http://cetafdigitization.biowikifarm.net/cdig/Use_cases_for_stable_URIs_in_collection_worksflows

CETAF Proposal for COST Action: MOBILISE

Dagmar Triebel presented an update of the COST Action proposal, MOBILISE. The main aim of the project is to "build up a cooperative, inclusive, bottom-up and quickly responsive network in Europe to support research activity for biodiversity informatics". There are a total of 37 proposers from 17 countries. This was submitted in December 2016. The evaluation period is 6 months and a decision is expected in June 2017. If it is successful, it is projected to start in Autumn 2018.

Digitisation Resources Gap Analysis

Elspeth Haston presented a draft survey created to identify gaps in resources for mass digitisation. The draft survey had been circulated to those members of the Digitisation Working Group who had expressed interest and some feedback had already been received. There was some concern in the length of the survey and it was recommended that there should also be an option available for people to summarise their need of specific software/hardware/infrastructure/workflow requirements (eg, "if I had a system reading this or that information automatically from a scan, I could save a lot of time"). This would allow us to gather these requirements in a rather unstructured form and then try to identify the issues where certain developments can have a significant impact. It was agreed that the survey would incorporate this recommendation. The survey will therefore be updated and then circulated to CETAF and SYNTHESYS members.

Next Meeting

The next full Digitisation Working Group meeting will be held jointly with the ISTC in spring 2018 in Copenhagen.