SYNTHESYS+
Synthesis of Systematic Resources
a DiSSCo project

[D4.2: A GUIDE OF BEST PRACTICES DOCUMENTATION]

[A. GÜNTSCH, A. PLANK]

Grant Agreement Number | 823827

Acronym | SYNTHESYS PLUS

Call | H2020-INFRAIA-2018-2020

Start date | 01/02/2019

Duration | 48 months

Work Package | [4]

Work Package Lead | [Quentin Groom]

Delivery date | [27.07.2020]

Contents

[*To update Contents, click 'Update Table' then select 'entire table'. This will update page numbers as well as Section titles. Section headings used below are examples*]

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+
Synthesis of Systematic Resources
a DiSSCo project

Summary

Within the framework of SYNTHESYS+ Task 4.2, European natural history collections cooperate in the standardisation and extension of the system of stable specimen identifiers for physical collection objects introduced by the Consortium of European Taxonomic Facilities (CETAF, https://cetaf.org/). The focus is on best practices for the syntax of such identifiers, rules for their assignment as well as the harmonisation of machine-readable specimen data in the form of RDF.
The "Guide for best practices documentation" (deliverable 4.2, https://cetafidentifiers.biowikifarm.net) bundles previously scattered information on CETAF identifiers on a single wiki page of the Biowikifarm in a comprehensible way and thus addresses both the curators of the participating natural history collections and the technical staff responsible for the implementation. It is complemented by

i) a GitLab repository with the software required for RDF publication https://git.bgbm.org/cetaf/stableidentifiernegotiation)

ii) an overview page of recommended data standards (https://cetafidentifiers.biowikifarm.net/wiki/CETAF_Specimen_Preview_Profile_(CSPP)) , and

iii) a discussion (wiki) page where best practices for specific questions can be developed by the community (https://cetafidentifiers.biowikifarm.net/wiki/Questions,_problem_solutions_and_further_discussions_(Guide_of_best_practices)).

Description of Deliverable

Background

Inspired by a system for Specimen Identifiers implemented at the Royal Botanic Garden Edinburgh (Hyam et al. 2012), collections organised in the Consortium of European Taxonomic Facilities (CETAF) have started to establish a uniform identifier system since 2013 (Groom et al. 2017). Identifiers are assigned by the collection institutions themselves in the form of HTTP URIs (Uniform Resource Identifiers), usually based on a locally established barcode system (Güntsch et al. 2017). A software script installed on the collections' web servers redirects requests for CETAF Identifiers to human-readable landing pages or machine-readable RDF representations (Fig. 1). The harmonisation of the identifier systems used for physical objects is an important contribution to international harmonisation of identifiers for digital objects, as is being promoted, for example, within the framework of the DiSSCo initiative. A key aspect and infrastructure component of DiSSCo is the use of "Natural Science Identifiers" (NSIDs) for the representation of digital surrogates of physical collection objects (Hardisty 2020). NSIDs offer a persistent access to digital collection objects and their links to relevant research objects, where the physical object (identified by the CETAF ID) is in the centre.

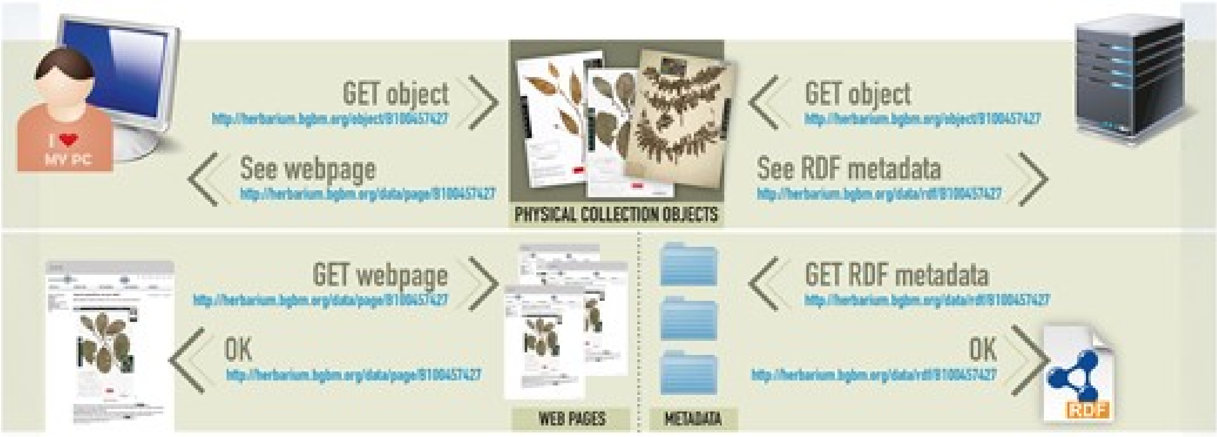
I ♥ MY PC
GET object
http://herbarium.bgbm.org/object/B100457427
See webpage
http://herbarium.bgbm.org/data/page/B100457427
PHYSICAL COLLECTION OBJECTS
GET object
http://herbarium.bgbm.org/object/B100457427
See RDF metadata
http://herbarium.bgbm.org/data/rdf/B100457427
GET webpage
http://herbarium.bgbm.org/data/page/B100457427
OK
http://herbarium.bgbm.org/data/page/B100457427
WEB PAGES
METADATA
GET RDF metadata
http://herbarium.bgbm.org/data/rdf/B100457427
OK
http://herbarium.bgbm.org/data/rdf/B100457427
RDF

Figure 1: Redirection of CETAF IDs to human- and machine-readable representations of physical specimens.

Within the framework of SYNTHESYS+ Task 4.2, the participating collections are working on broadening the implementations and harmonizing the data standards and protocols used. In this context, a number of resources have been created to support existing and new implementers and application developers. These include, among others:

- A registration of the syntax forms used in the participating collections for specimen identifiers, which can be integrated into software systems via an open Google API (https://docs.google.com/spreadsheets/d/1vHl2xDghffm6HfQhVeruHV6ZAWAnrc-2LPasq0fOyF4).

- An overview of the maturity level of the different implementations (https://docs.google.com/spreadsheets/d/1bRDbRk9eTTWX4fk0UUr0BSvUxZP1NWPBltGHnhihORQ).
- An RDF Triple-Store (https://cetafidentifiers.biowikifarm.net/wiki/CETAF_Specimen_Catalogue, fig. 2), which allows application developers to access CETAF specimens via a central (SPARQL) interface and link them to other semantic resources (e.g. Wikidata, GeoNames, etc.).

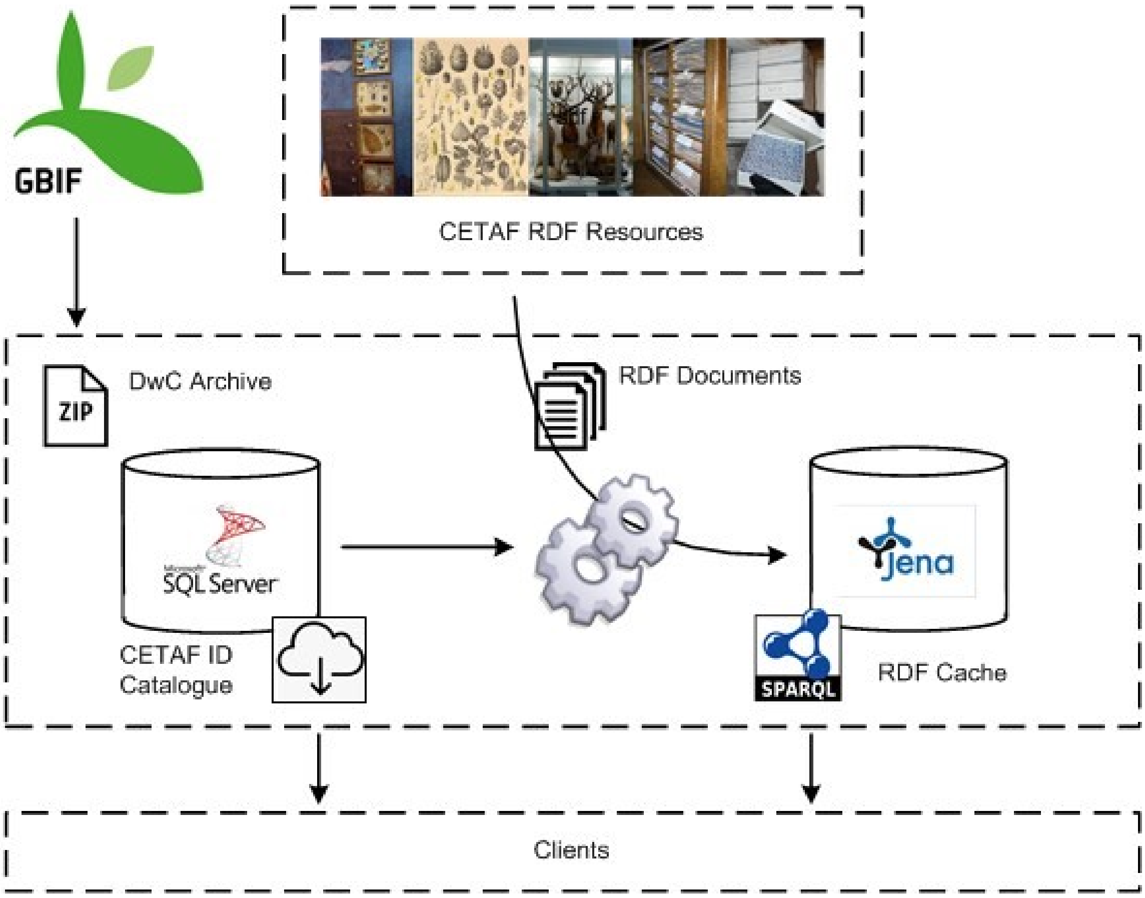

Figure 2: “CETAF Specimen Catalogue” implemented as an RDF triple-store accessible via a SPARQL endpoint.

Accompanying the various measures to harmonise and broaden the implementation of CETAF IDs, a manual should also be produced to support curators and technical staff in implementing an identifier strategy (deliverable 4.2). It was agreed that this guide should not be a static document, but rather a wiki-based website set up on the Biowikifarm maintained by the community. The different components of the manual are described in the following section.

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

The guide of best practices documentation

Main section

The central wiki page of the Guide (https://cetafidentifiers.biowikifarm.net/wiki/Main_Page) summarises the instructions and information essential for the implementation of CETAF IDs. This includes

- best practices for the choice of a suitable identifier syntax,
- information on the redirection mechanism and HTTP-specific aspects of implementation,
- examples for the use of CETAF IDs,
- the definition of reasonable implementation levels,
- technical specifications for integrating CETAF IDs into the GBIF publication process, and
- a compilation of further information resources.

Standards section

On the standards page (https://cetafidentifiers.biowikifarm.net/wiki/CETAF_Specimen_Preview_Profile_(CSPP)) a list of 13 data elements is specified, which are recommended as core elements for the RDF data belonging to CETAF IDs. The elements have all been compiled from existing standards, such as DarwinCore and DublinCore.

Data elements that come in addition to the recommendation in the continuous consensus process of the community are documented in a separate section “additional recommended data terms”.

As a blueprint for the RDF representation to be used, a sample document is provided that contains all recommended elements in their preferred representation. Also included in the sample document are examples of links to IIIF-compatible image servers (task NA 4.3) and examples of semantic enrichment of data elements with links to external resources.

Discussion section

Recurring questions concerning the use of identifiers, especially from a curatorial point of view, often require a longer coordination process. We have therefore linked a section to the main page of the guide where discussions on such issues can be documented and consensus can be reached (https://cetafidentifiers.biowikifarm.net/wiki/Questions,_problem_solutions_and_further_discussions_(Guide_of_best_practices)). Results agreed within the project should then be transferred to the main section.

Software

For institutions wishing to implement CETAF IDs, but also to ensure the highest possible level of implementation consistency, it is desirable to develop and use jointly the software components required for the redirection of IDs and the provision of RDF. Therefore, the available software was revised and cleaned up and made available as an open source project via GitLab (https://git.bgbm.org/cetaf/stableidentifiernegotiation).

Future development

The guide for best practices and the accompanying resources will be curated within the project but also beyond the project framework and will be maintained, e.g. within the framework of CETAF. A special focus will be the linkage with the concepts for Natural Science Identifiers (NSIDs) and OpenDS, which are developed within DiSSCo. The aim is to link the decentralized management of Specimen IDs in the collections themselves with the creation of a central index and the corresponding identifiers in such a way that the connection to the original information can always be maintained and kept up-to-date.
Another focus will be the streamlining of methods for semantic annotation of collection data. The methods developed and used so far are still in a developmental stage and are being calibrated in many ways. In the future, they should be stabilized and incorporated into best practices as standardized workflows.

References

- Groom, Q., Hyam, R., & Güntsch, A. 2017: Data management: Stable identifiers for collection specimens. Nature, 546(33). https://doi.org/10.1038/546033d
- Güntsch, A., Hyam, R., Hagedorn, G., Chagnoux, S., Röpert, D., Casino, A., Droege, G., Glöckler, F. Gödderz, K., Groom, Q., Hoffmann, J., Holleman, A., Kempa, M., Koivula, H., Marhold, K., Nicolson, N., Smith, V.S., Triebel, D. 2017. Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. Database (Oxford) 2017; 2017 (1): bax003. doi: 10.1093/database/bax003
- Hardisty 2020: Natural Science Identifiers & CETAF Stable Identifiers. DiSSCoTech Blog – Technical posts about the design of the DiSSCo infrastructure. https://dissco.tech/2020/05/28/natural-science-identifiers-cetaf-stable-identifiers/
- Hyam, R., Drinkwater, R. E., Harris, D. J. 2012: Stable citations for herbarium specimens on the internet: an illustration from a taxonomic revision of Duboscia (Malvaceae). Phytotaxa 73:17-30

Contributors

Alex Hardisty, Carole Goble, Andreas Plank, and Maarten Trekels

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+
Synthesis of Systematic Resources
a DiSSCo project