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 | 31.07.2020

Contents

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))
   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.

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/1vHI2xDghffm6HfQhVeruHV6ZAWAnrc-2LPasq0fOyF4).
- An overview of the maturity level of the different implementations (https://docs.google.com/spreadsheets/d/1bRDbRk9eTTWX4fk0UUr0BSvUxZP1NWPBltGHnhihORQ).

SYNTHESYS+
Synthesis of Systematic Resources
a DiSSCo project

- 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.).
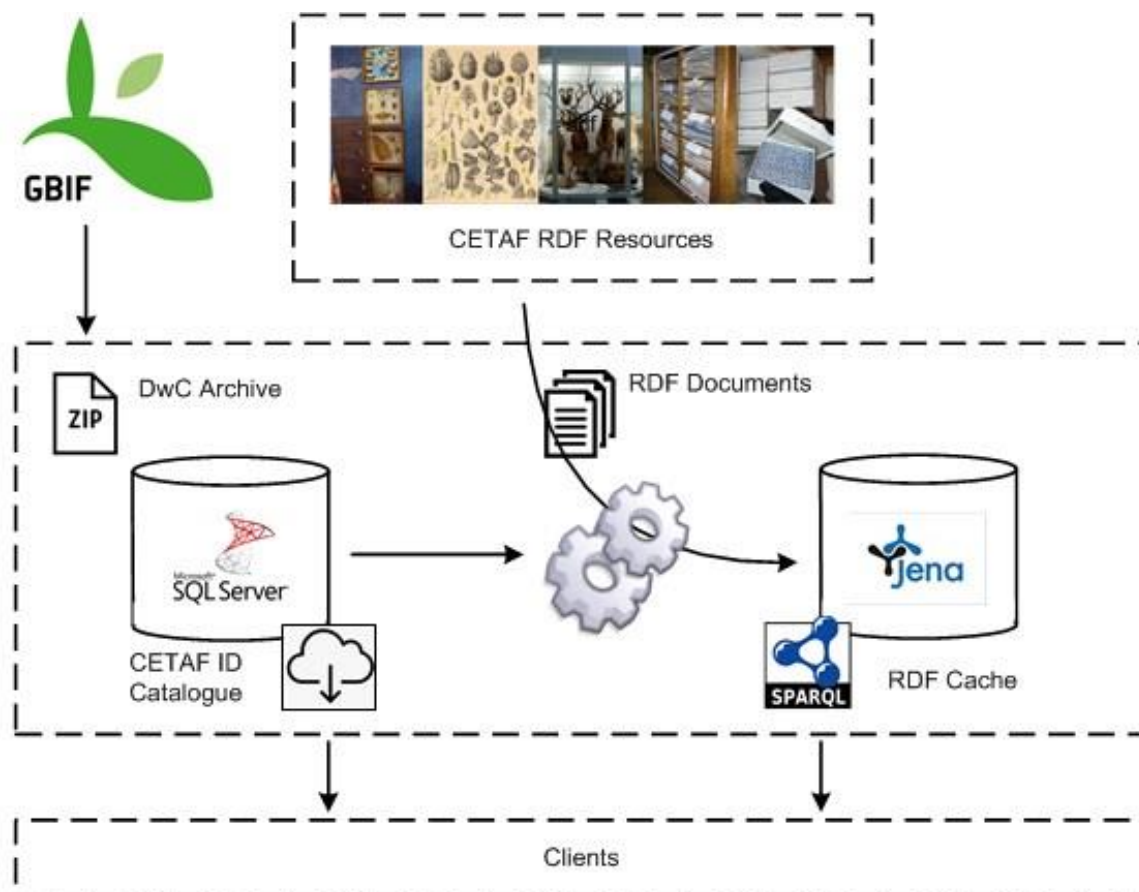

Figure 2: "CETAF Specimen Catalogue" implemented as an RDF triple-store accessible via a SPARQL endpoint.

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+
Synthesis of Systematic Resources
a DiSSCo project

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+
Synthesis of Systematic Resources
a DiSSCo project

Annex – Main Wiki Page of the CETAF Stable Identifier Guide

CETAF

Main page | Discussion | Read | View source | View history | Search [🔍]

Main Page

CETAF Stable Identifier Guide

Main page
Community portal
Current events
Recent changes
Random page
Help
Donate

▼ Tools
What links here
Related changes
Upload file
Special pages
Printable version
Permanent link
Page information
Cite this page

Contents [hide]

CETAF ISTC Stable Identifier Initiative

The Stable Identifiers of the Consortium of European Taxonomic Facilities (CETAF) are globally unique, consistent and reliable identifiers for specimens in natural and botanical collections. These identifiers are used in the world wide web to redirect users and systems to images, websites and metadata of the physical objects and to integrate them with the semantic web.

SYNTHESYS+
Synthesis of Systematic Resources
a DiSSCo project

How do CETAF Stable Identifier look like?

The CETAF identifier system is based on HTTP-URIs and Linked Data principles. It is simple and future-proof. Each collection object as well as its associated information resources (e.g. multimedia, RDF, webpages) are identified by stable HTTP-URIs that will never change. The URI Syntax for the objects is chosen and maintained by the institution owning them. This flexibility is one of the main advantages of the CETAF Stable Identifier system as it allows e.g. to include branding and local scope identifiers into the CETAF Stable Identifier URI. There are however some best practices for stable URIs. Examples are:

> http://herbarium.bgbm.org/object/B100277113
> http://www.botanicalcollections.be/specimen/BR0000005516339
> http://data.rbge.org.uk/herb/E00421509

How are CETAF Stable Identifiers resolved?

A CETAF Stable Identifier allows the access of information about the corresponding collection object in various ways. If a human user tries to access a collection object by typing it's CETAF Stable Identifier into a web-browser, he will be redirected to a human-readable representation (e.g. html web-page) of it. If a software-system tries to access the collection object via the same identifier, it will be redirected to a machine-processable RDF-encoded metadata record. The identifier is therefore integrated with the semantic web and can also be used in other RDF representations to link to the belonging collection object.

Example for the syntax of a CETAF Stable Identifier.

Resolving URI-based collection identifiers using standard HTTP-redirection mechanisms.

What can CETAF Stable Identifiers be used for?

As described above, CETAF Identifiers can first of all be used to redirect users and systems to images, websites and metadata of the physical objects they belong to. They can also be used to precisely reference specimens needed in scientific studies and serve as basis for data retrieval, integration and reproducibility of data experiments. Additionally, the stable identifiers enable new applications in the semantic web domain. An example for this is the Biology Pilot. The Botanic Garden and Botanical Museum Berlin, Meise Botanic Garden and other collections annotated thousands of specimens with the HUH and WikiData IDs of their collectors. The CETAF Stable Identifiers of the annotated specimens are available on GBIF and a server is crawling the identifiers to organize the RDF information in a Blaze Graph triple store. This graph enables us to search for specimen by their collector ID of HUH or WikiData, which is invariant to the different spelling variants the individual institutions may be using. The query will return all relevant specimens available in the joined set of specimens regardless of their origin institution. If the number of institutions using stable identifiers grows and the amount of machine readable annotations increases, this technology could be used to basically create a "google for specimens".

How can I implement CETAF Stable Identifiers for my collection?

The CETAF Stable Identifiers can be implemented in three levels. They are described in detail in herbal.rbge.info's documentation.

Following conditions have to be met to reach the corresponding implementation levels

Level 1 …	→ Level 2	→ Level 3
✓ you assigned a stable URI to each object of your collection, which will be never changed and preferably follows the best practices for stable URIs	✓ you reached *Level 1*	✓ you reached *Level 2*
✓ there exists a human-readable representation (web-page) for each of your collection objects	✓ there exists a machine-readable RDF metadata record for each of your collection objects	✓ the machine-readable RDF metadata record of each of your collection objects encodes application specific data (e.g. is compliant to the CETAF Specimen Preview Profile —CSPP)
✓ a user trying to access a collection object by typing the stable URI of it into a web-browser will be redirected to the human-readable representation (web-page) of the object (you can test this by using the CETAF URI Tester)	✓ a machine trying to access a collection object via its identifier with `application/rdf+xml` header will be redirected to the objects machine-readable RDF metadata record (you can test this by using the CETAF URI Tester)	

HTTP vs. HTTPS versions of CETAF URIs

As far as the Semantic web is concerned *http://xyz* and *https://xyz* are different things because they are different URIs. The recommendation for new implementations should be just to use HTTPS. If you have only HTTP or HTTPS versions, *or* want to change it you should take notice of the following:

HTTP	HTTPS
You	You
• have issued *only HTTP* versions of CETAF URIs and want to keep it that way	• have issued *only HTTPS* versions of CETAF URIs
• have nothing to add technically, just have the usual 303 HTTP redirect to RDF or HTML resources in place	• don't need to resolve then HTTP if you have never issued any, because they aren't out there to be resolved.

Want to change HTTP to HTTPS

You

- have issued HTTP versions of CETAF URIs but want to change to HTTPS
- have to keep resolving with a 303 redirect to HTTPS of the RDF or HTML resources. The RDF should contain an `owl:sameAs` assertion linking the HTTP and HTTPS versions of the URI, therefore only minor configure stuff for providers and transparent for users.
- could change to telling people to cite HTTPS rather than HTTP for your specimens but it shouldn't matter too much as these things are linked together. The recommendation would be to cite as HTTPS if you have it implemented as at some point in the future a client may refuse to trust even a redirect from an HTTP URI (which is a bit paranoid but may happen).

Publishing CETAF IDs to GBIF

If your institution is using CETAF IDs and yout want them (and potential Specimen RDF) to be included into the CETAF Specimen Catalogue, they need to be used as GUIDs in the specimen data fed to GBIF. As described in CETAF Specimen Catalogue, the GBIF Index is used to discover CETAF IDs.

- If DarwinCore is used, the IDs must be mapped to occurrence ID.
- For ABCD, the concept UnitGUID should be used.

Publishing CETAF IDs to GBIF

If your institution is using CETAF IDs and yout want them (and potential Specimen RDF) to be included into the CETAF Specimen Catalogue, they need to be used as GUIDs in the specimen data fed to GBIF. As described in CETAF Specimen Catalogue, the GBIF Index is used to discover CETAF IDs.

- If DarwinCore is used, the IDs must be mapped to occurrence ID 🔗
- For ABCD, the concept UnitGUID 🔗 should be used.

How can I discover specimens with CETAF IDs and corresponding Linked Open Data (LOD)?

You can discover specimens of institutions of the Stable Identifiers Implementers Group by using the CETAF Specimen Catalogue maintained at the BGBM, which offers a web service for getting a list of valid CETAF IDs. For implementers of level 2, who provide RDF representations of their specimens, a cache triple store with a SPARQL access point will be available soon.

What data fields or elements are recommended or standardized?

The CETAF Specimen Preview Profile (CSPP) is developed as a minimal set of agreed (RDF) collection metadata elements implemented consistently across CETAF organisations. Its purpose is to provide a stable resource enabling preview functions in specimen portals. The CSPP is not meant to be comprehensive, which means that Linked Open (collection) Data implementations of CETAF institutions will usually provide much richer metadata with additional RDF-elements.

Further Questions

See on Questions, problem solutions and further discussions (Guide of best practices) and in general also in Category: Discussion.

Useful Links

- Best practices for stable URIs (wiki.pro-ibiosphere.eu) 🔗
- CETAF Specimen Preview Profile (CSPP)—A set of standard data components for data exchange
- Source code and example documents (git.bgbm.org) 🔗
- CETAF URI Tester (herbal.rbge.info) 🔗
- The Standards Compliance Dashboard of collaborating institutions
- Category: Guide for CETAF Stable Identifiers—Collection of pages related to this guide or handbook

Further reading

Kuzmova, I. 'Pro-IBiosphere - Stable Identifiers for Specimens – A CETAF ISTC Initiative Supported by pro-IBiosphere'. *EUBON*. 1 July 2013. URL: http://www.pro-ibiosphere.eu/news/4296_stable_identifiers_for_specimens_-_a_cetaf_istc_initiative_supported_by_pro-ibiosphere/ 🔗.

Güntsch, A. et al., 'Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects', *Database (Oxford)*, vol. 2017, Jan. 2017. URL: https://doi.org/10.1093/database/bax003 🔗.

Groom, Q. et al., 'Stable Identifiers for Collection Specimens', *Nature (Correspondence)*, 546.7656 (2017), 33. URL: https://doi.org/10.1038/546033d 🔗

Hardisty, A. 'Natural Science Identifiers & CETAF Stable Identifiers'. DiSSCoTech (blog). 28 May 2020. URL:https://dissco.tech/2020/05/28/natural-science-identifiers-cetaf-stable-identifiers/ 🔗.

Wouter, A. 'Identifiers for Our Institutes – GRID and ROR', DiSSCoTech (blog), 11 April 2020, https://dissco.tech/2020/04/11/identifiers-for-our-institutes-grid-and-ror/ 🔗.

McMurry, J. A. et al., 'Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data', *PLOS Biology* 15(6):e2001414 June 2017. URL: https://doi.org/10.1371/journal.pbio.2001414 🔗.

Poster: CETAF stable identifiers for specimens (1.4MB, www.cetaf.org) 🔗

Meetings

- Edinburgh Meeting (June 2013) 🔗
- Joint ISTC/pro-iBiosphere workshop Berlin (October 2013) 🔗
- Geneva Meeting (October 2015)
- Joint CETAF-ISTC / CETAF-DWG meeting (May 2016) 🔗
- Joint CETAF-ISTC / CETAF-DWG meeting (March 2017) 🔗
- (Virtual) LOD Hackathon (October 2017) 🔗
- Joint CETAF-ISTC / CETAF-DWG meeting (February 2018) 🔗
- ISTC QoS Workshop Copenhagen (June 2018) 🔗
- Joint CETAF-ISTC / CETAF-DWG meeting (February 2019) 🔗

Category: Guide for CETAF Stable Identifiers