ARPHA Preprints

PENSOFT

Methods

Author-formatted document posted on 03/09/2026

Published in a RIO article collection by decision of the collection editors.

DOI: https://doi.org/10.3897/arphapreprints.e214295

Functional optimised EU-Nomen processes and services up-and-running

Anton Güntsch, **Andreas Müller**

ARPHA Preprints *Author-formatted document posted on 03/09/2026.* DOI: https://doi.org/10.3897/arphapreprints.e214295

Funded by
the European Union

TETTRIs

TETTRIs

Transforming European Taxonomy through Training, Research, and Innovations

D2.1 Functional optimised EU-Nomen processes and services up-and-running

WP 2/Anton Güntsch

Authors: Andreas Müller, Anton Güntsch

Version: v1
Date: 30/05/2025

TETTRIs related product

Identification	Value
Title	Functional optimised EU Nomen processes and services up-and-running
Author(s)	Andreas Müller Anton Güntsch
Affiliation	FUB-BGBM
Contributors (Affiliation)	NHMW, NHMD
Publisher	CETAF
Identifier of the publisher	
Doc. Version	v1
Resource	
Publication year	2025
Sensitivity	PU - Public
Date	30/05/2025
Citation	

Abstract:

EU-Nomen is a centralized, expert-curated taxonomic backbone designed to standardize species names across Europe, supporting scientific research, biodiversity monitoring, and policy-making. Initially developed under the EU-funded PESI project, EU-Nomen integrates data from major taxonomic sources—Euro+Med PlantBase, Fauna Europaea, ERMS, and Index Fungorum—into a unified structure. Hosted at FUB-BGBM and maintained by VLIZ, the infrastructure harmonizes nomenclature and taxonomy using a refined merging workflow based on the EDIT Platform for Cybertaxonomy. This modernized workflow replaces an earlier SQL script-based approach that was inefficient and difficult to maintain. Key features of the updated system include modular imports, automated validation, round-trip data consistency checks, and semi-automated taxonomic merging guided by source priorities. EU-Nomen supports open data directives and provides programmatic access via web services, ensuring interoperability with European biodiversity platforms. Recent

Funded by the European Union

enhancements include the integration of new data (e.g., bryophytes, lichens, and animal common names) and increased automation that allows more frequent releases. Together, these developments position EU-Nomen as a critical infrastructure for managing and delivering authoritative European taxonomic information.

Keywords:
EU-Nomen, taxonomic backbone, taxonomic checklist, data integration

Revision:

no.	Reviewer	Status	Notes
v0.1	Anton Güntsch	In work	Introduction and objectives
v0.2	Andreas Müller	In work	The workflow and others
v0.3	Walter .G. Berendsohn; Anton Güntsch; Thomas Papp	Under review	Internal review WP2
v0.4	Vincent Kalkman; Conrad Gillett	Under review	Internal review TETTRIs
v1	Marta León	Approved	Approved

Document Control Information

Settings	Value
Document Title:	Functional optimised EU Nomen processes and services up-and-running
Project Title (Acronym):	TETTRIs
Document Authors:	Andreas Müller, Anton Güntsch
Project Coordinator	Technical Coordinator: Ana Casino, CETAF Financial Coordinator: Frederik Hendrickx , RBINS
Doc. Version:	v1
Sensitivity:	PU
Date:	30th of May 2025

Document Approver(s) and Reviewer(s):

Funded by the European Union

NOTE: All Approvers are required. Records of each approver must be maintained. All Reviewers in the list are considered required unless explicitly listed as Optional.

Name	Role	Institution	Action	Date
Vincent Kalkman	Reviewer	Naturalis	Reviewed	20/5/25
Conrad Gillet	Reviewer	LUOMUS	Reviewed	20/5/25
Marta León	Approver	CETAF	Approved	30/5/25

Document history:

The Document Author is authorised to make the following types of changes to the document without requiring that the document be re-approved:

- Editorial, formatting, and spelling
- Clarification

To request a change to this document, contact the Document Author or Owner. Changes to this document are summarised in the following table in reverse chronological order (latest version first).

Revision	Date	Created by	Short Description of Changes

Configuration Management: Document Location

The latest version of this controlled document is stored in <location>.

Funded by the European Union

INDEX

INDEX 5
ACRONYMS 6
1. Introduction 7
1.1 EU Nomen data sources 7
1.2 Basic components of the data merging pipeline 8
2. Objectives of the deliverable 9
3. The previous workflow 10
3.1 Migration to a single database instance 10
3.2 The database schema 10
3.3 Workflow 11
3.4 Obstacles of the previous approach 12
3.5 Migration of Euro+Med and Fauna Europaea 12
3.6 Migration of Euro+Med and Fauna Europaea 12
4. The new workflow 13
4.1 Euro+Med 13
4.2 Adding Fauna Europaea 13
4.3 Adding ERMS 13
4.4 Adding Index Fungorum 14
5. Merging 14
5.1 General considerations 14
5.1.1 Priorities 15
5.1.2 Name merging 15
5.1.3 Taxon merging 15
5.2 Algorithm 16
5.3 Implementation 17
6. Export to the Data Warehouse 17
7. New Data 17
7.1 Bryophytes 17
7.2 Lichens 17
7.3 Common names for Animalia 18
8. Automation 18
9. Outlook 19
10. ACKNOWLEDGEMENTS 19
11. REFERENCES 19

Funded by the European Union

ACRONYMS

CETAF	Consortium of European Taxonomic Facilities (General Secretariat)
CDM	Common Data Model (the EDIT Platform data model)
DBMS	Database Management System
ERMS	European Register of Marine Species
EDIT	European Distributed Institute of Taxonomy
EU	European Union
Euro+Med	Euro+Med PlantBase
FRDBI	The Fungal Records Database of Britain and Ireland
FUB-BGBM	Freie Universität Berlin - Botanischer Garten und Botanisches Museum Berlin
NHMD	Natural History Museum of Denmark
NHMW	Naturhistorisches Museum Wien
PC	Project Coordinator
PESI	Pan-European Species directories Infrastructure
PM	Project Manager
RBINS	Royal Belgian Institute of Natural Sciences
SMNS	State Museum of Natural History Stuttgart
TETTRIs	Transforming European Taxonomy through Training research and Innovations
UI	User Interface
UUID	Universally Unique Identifier
VLIZ	Flanders Marine Institute
WORMS	World Register of Marine Species
WP	Work Package
WPL	Work Package Leader

Funded by the European Union

1. Introduction

EU-Nomen is a centralized, authoritative taxonomic reference for European species, designed to support the consistent use of species names in scientific research, biodiversity data management, and policy-making (de Jong et al. 2015). EU-Nomen services were implemented for the most part in the context of the EU 7th Framework project PESI (Pan-European Species directories Infrastructure, Contract no. RI-223806) which established unified and streamlined taxonomic information across Europe (https://www.vliz.be/projects/pesi/).

By integrating existing taxonomic databases and coordinating the work of experts from multiple disciplines and regions, EU-Nomen ensures that species names and classifications are accurate, up-to-date, and standardized. This harmonized taxonomy enables more effective biodiversity monitoring, reporting, and conservation planning at both national and European levels. The PESI web portal, which makes EU-Nomen data accessible, allows users to explore accepted species names, identify synonyms, examine taxonomic hierarchies, determine species distributions within Europe, and trace the source of taxonomic decisions through expert-validated references.

EU-Nomen is formally recognized as part of the implementation framework for European biodiversity data infrastructures under EU Directive 2019/1024 on open data and the re-use of public sector information. Within this context, EU-Nomen serves as a core component of the European Commission's efforts to ensure consistent and open access to taxonomic information. Its role is specifically referenced in supporting the INSPIRE Directive (2007/2/EC), which establishes spatial data infrastructures across Europe, where standardized species nomenclature is essential for environmental and biodiversity reporting. These directives emphasize the need for interoperable, reliable, and openly available taxonomic resources which provide a stable anchor for linking regional or national, taxon-related data (Güntsch et al. 2024).

1.1 EU Nomen data sources

Several authoritative databases supply validated taxonomic information to EU-Nomen, each focusing on specific organism groups. For example, Euro+Med PlantBase compiles verified taxonomic data for vascular plants found in Europe and the Mediterranean region, accessible via https://europlusmed.org (Fig. 1). ERMS (European Register of Marine Species) provides an authoritative list of marine species in European waters and can be explored through https://www.marbef.org/data/erms.php. Index Fungorum contributes fungal names and taxonomic structures globally, including European entries, and is maintained at http://www.indexfungorum.org. Fauna Europaea offered a detailed checklist of all European terrestrial and freshwater animal species.

Unfortunately, Fauna Europaea currently has no web presence. However, a new host has been found in the Natural History Museum in Stuttgart, which has taken responsibility for the reestablishment of the Fauna Europaea services.

Each of the above systems offers web portals and data services that contribute to the harmonization of taxonomic information across Europe via EU-Nomen. Together, they ensure a reliable, expert-verified taxonomic infrastructure for scientific, conservation, and policy applications.

Funded by the European Union

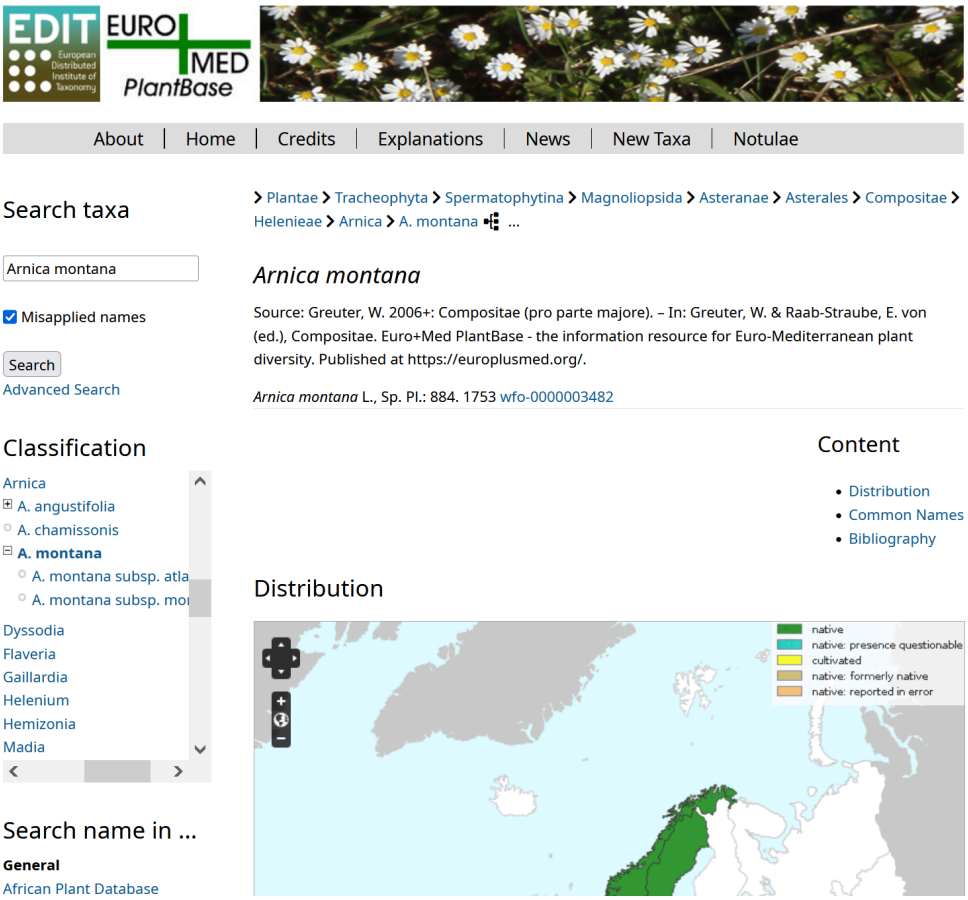

Figure 1: the Euro+Med plantbase portal provides free and open access to vascular plants data from Europe, the Mediterranean and the Caucasus and serves this information to the EU-Nomen infrastructure.

1.2 Basic components of the data merging pipeline

At the Botanic Garden and Botanical Museum Berlin (FUB-BGBM), expert-curated taxonomic EU-Nomen datasets are integrated into a unified data structure. This process involves harmonizing nomenclatural standards, resolving conflicting entries, and aligning classification systems across the contributing sources. Each dataset, originally developed for specific taxonomic groups, undergoes validation and mapping into a shared schema to ensure consistency and compatibility.

The merged database enables cross-referencing and interoperability between organism groups and geographic regions. FUB-BGBM uses specialized tools and workflows to manage data imports and updates while preserving source attributions. Once consolidated, the complete database forms the backbone of the EU-Nomen infrastructure. This unified taxonomic reference is then exported and made accessible through a central portal and web services. These online tools are developed and maintained by VLIZ (Flanders Marine Institute), ensuring robust access and integration capabilities. The data portal supports search, download, and citation of standardized taxonomic data from Europe. Web services provide real-time access to validated taxonomic content for use in biodiversity informatics platforms and environmental reporting systems.

1.3 Data access services

The EU-Nomen web services, developed and hosted by VLIZ, provide programmatic access to the taxonomic backbone created through the EU-Nomen workflows. These

Funded by the European Union

services allow users and applications to search, retrieve, and integrate standardized species information from the consolidated EU-Nomen database. The core services support functions such as querying scientific names, retrieving accepted names and synonyms, accessing taxonomic hierarchies, and checking species occurrences in Europe. Users can also validate names against the EU-Nomen checklist to ensure consistency in scientific data usage. The main entry point for these services is the PESI web portal at https://www.eu-nomen.eu/portal.

Figure 2: Overview of EU-Nomen web services documented on the PESI web site.

The actual web service interface is available via https://www.eu-nomen.eu/portal/webservices.php, where the full list of service operations is documented (Fig. 2). These include methods like “Taxon match”, which checks a species name against the database and returns the accepted version, and GetTaxonDetails, which delivers detailed information on a taxon, including its classification and synonyms. Other services provide access to expert networks, source databases, and geographic data. The portal supports both RESTful and SOAP services.

The web services are designed for integration with biodiversity platforms such as GBIF, LifeWatch, and national data infrastructures, promoting interoperability. Overall, they serve as an important component in delivering harmonised European taxonomic data to both human users and automated data workflows.

2. Objectives of the deliverable

The main objective of Deliverable 2.1 is to modernize the processes for merging taxonomic checklists from distributed and heterogeneous checklist infrastructures. To this end, the merging services based at FUB-BGBM have been completely revised and documented and can now be operated robustly, efficiently and sustainably. In addition to the technical revision, important data sources have been imported that were not previously available via

Funded by the European Union

EU-Nomen. The new implementation of the merging processes and the content extensions are described in the following.

3. The previous workflow

When PESI started in 2008, the integration and merging process was implemented in the EDIT Platform for Cybertaxonomy, which was a relatively new product at that time. The idea behind this was to create synergies by using existing functionality of the platform for PESI and integrating newly developed functionality for PESI into the platform. This worked well in an initial pilot implementation and a first merge could be carried out using the platform. However, it turned out that the implementation of the platform at that time was very unperformant when handling large amounts of data, as is the case with PESI, and therefore did not scale well. The import of the Fauna Europaea data into the common EDIT instance alone took about a week.

For this reason, it was decided to carry out the import in the meantime using SQL scripts. This was possible because all databases were available as relational SQL-based databases. They only had to be imported to a common relations database instance.

The following describes the previous SQL script based workflow for creating the EU-NOMEN target database from the individual original sources. The process of migrating and merging the data from the single data sources into the target data warehouse is elucidated.

The disadvantages of the described workflow, which made regular updates of the EU-NOMEN database time-consuming, are discussed.

3.1 Migration to a single database instance

As mentioned, at the time when EU-NOMEN was developed, the four source databases Euro+Med, Fauna Europaea, ERMS, and Index Fungorum were all available as relational databases.

Only the dialects differed. Euro+Med and ERMS were based on MS SQL Server, while Fauna Europaea used Oracle, and Index Fungorum (with integrated data from FRDBI, https://www.frdbi.org.uk/) was made available as an MS Access database.

The integration of the data into a single database system was carried out at FUB-BGBM and the target system with portal and web services was hosted at VLIZ. At the time, both institutes used MS SQL Server as the primary database management system (DBMS) for their taxonomic information systems. This made it obvious to also use MS SQL Server as DBMS for the EU-Nomen target database as well.

3.2 The database schema

As the EU-Nomen dataportal and web services were developed at VLIZ, the chosen target database schema was basically a fork of the schema used by the APHIA system (Vandepitte et al. 2015) developed by VLIZ. APHIA is the system on which the marine species register WORMS (of which ERMS is a subset) is running. This way code could be reused after small adaptations both for the portal and for the web services.

The target data schema is shown in Figure 1.

Funded by the European Union

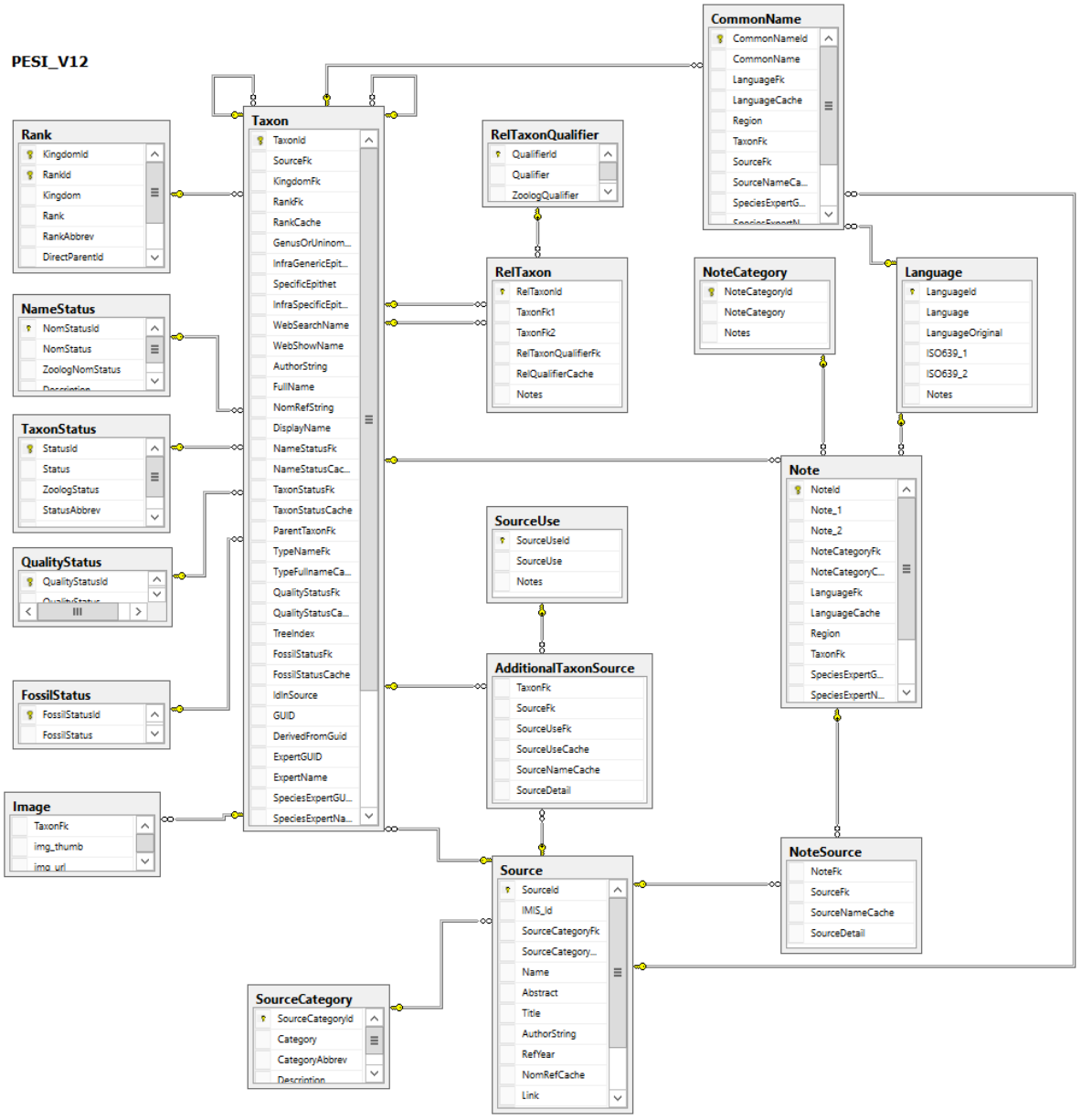

Figure 3: The EU-NOMEN target data schema.

3.3 Workflow

Based on the above considerations, the workflow used was as follows:

1. Migration of all original data into individual MS SQL Server database schemas on the same MS SQL Server instance
2. Import each source instance into the final database schema using a large number of Transact-SQL (T-SQL) Scripts. T-SQL is the extended SQL script language for MS SQL Server
3. Running scripts to unify formatting of data
4. Running scripts to validate the correctness of the import result (only for the 2014 version)

Funded by the European Union

5. Merging and deduplicating the taxonomic classifications where overlapping
6. Enriching data with additional data from other databases (e.g. common names)

3.4 Obstacles of the previous approach

Running all data manipulations in T-SQL is sufficiently performant in terms of time the scripts are running as no additional software layer is involved.

However, T-SQL scripts are unwieldy to write and difficult to read and maintain, as SQL itself is a declarative programming language and not a procedural one. This also includes the rather time consuming and cumbersome handling of errors that occur, which can, for example, lead to aborts or incorrect behavior of the scripts.

All this resulted in cutting the scripts into small pieces (Figure 3) that all had to be called up one after the other.

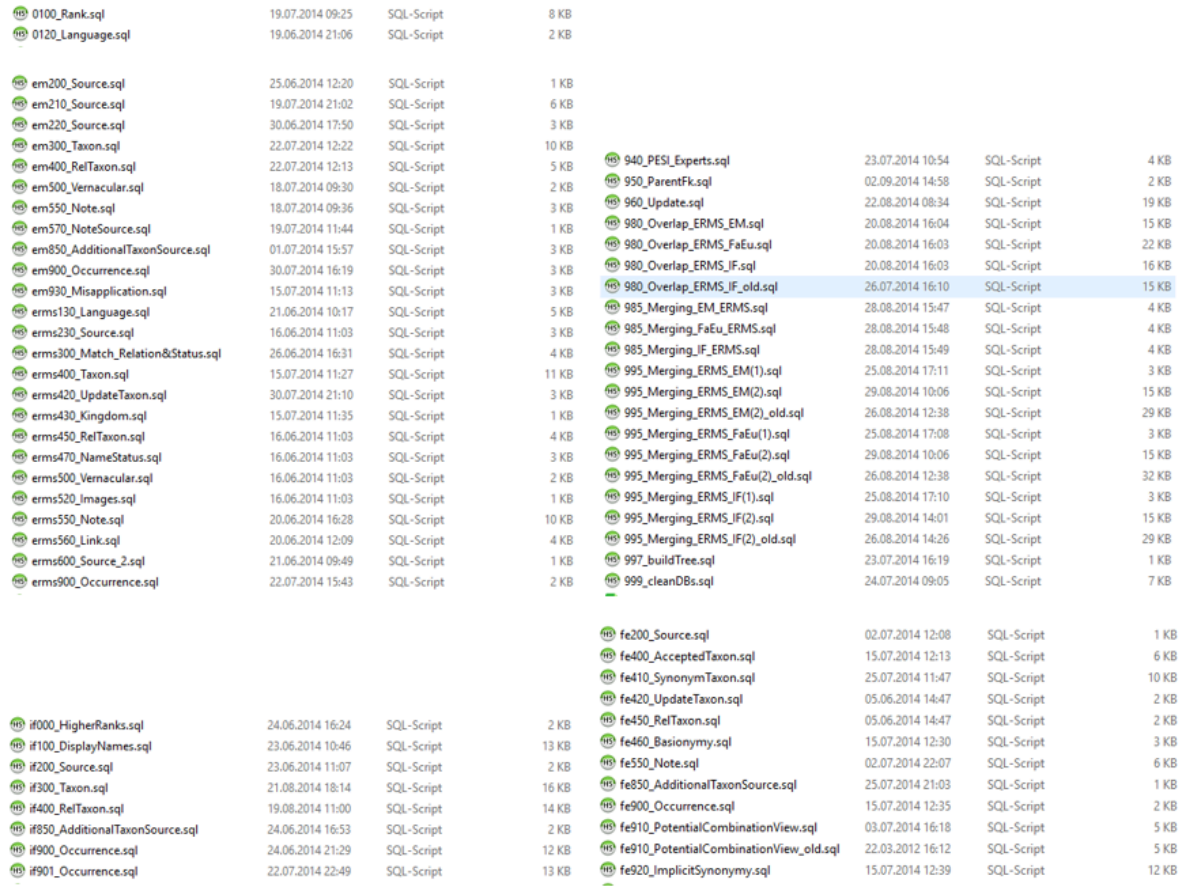

Figure 4: Scripts to run in the previous workflow

3.5 Migration of Euro+Med and Fauna Europaea

Validating the result of the import scripts was not a trivial task due to the amount and complexity of the data. By implementing validation queries during the last import in 2014 many hidden errors were found that led to missing data, duplicate records, or incorrectly formatted names, to name only a few. However, validation is difficult to implement with SQL Scripts only and therefore the validation framework was not comprehensive, yet.

Funded by the European Union

3.6 Migration of Euro+Med and Fauna Europaea

The last import to EU-Nomen did run in 2014. In the meanwhile two of the four original source databases (Euro+Med and Fauna Europaea) migrated to the EDIT Platform for Cybertaxonomy (Ciardelli et al. 2009), which uses an object-oriented data access layer. The old SQL scripts could therefore no longer be used without adding new scripts to transform the new format (Common Data Model - CDM) into the old one.

Thus, it was decided to revitalize the former workflow using the EDIT Platform to integrate and merge the data.

4. The new workflow

Even though the EDIT Platform in general performs much better in the meanwhile, and the scripts from the old pilot implementation did run much faster already, it was decided in the context of TETTRIs to fully refactor or replace the old workflows and algorithms by new ones.

The new workflow includes the following steps:

- Euro+Med, extended by source flag for all data, taken as starting point for the merging instance (see 4.1)
- Fauna Europaea added via CDM-2-CDM import to the merging instance (see 4.2)
- ERMS imported to CDM (including validation) and imported via CDM-2-CDM import to the merging instance (see 4.3)
- Index Fungorum imported to CDM (including validation) and imported via CDM-2-CDM import to the merging instance (see 4.4)
- Merging and deduplicating the taxonomic classifications where overlapping (see 5)
- Exporting the data from the merging instance to the target data warehouse (including validation)(see 6)

4.1 Euro+Med

A copy of Euro+Med is taken as the starting point for the new merged instance. However, to later distinguish Euro+Med data from other data, in a first step all Euro+Med data is flagged as such. This is necessary to provide backlinks to the original database.

4.2 Adding Fauna Europaea

A completely new import routine was implemented to import Fauna Europaea data from its own CDM-based database instance to the common merging CDM instance.

For this, a full classification was needed to be transported from one database into the merge instance. The problem here is that database identifiers may need to be changed as the same identifier value may be used in both databases. At the same time, some data exists in each CDM instance and must not be duplicated. Such data are terms for e.g. areas, languages, or taxon relationship types.

Distinguishing records that should not be deduplicated became possible as the CDM provides local database identifiers, which can change when moving to a new instance, as well as universally unique Identifiers (UUIDs). The latter allows to recognize commonly used terms.

4.3 Adding ERMS

For ERMS the original import algorithm from the pilot implementation could be taken as a base routine as the original format did not change substantially in the meanwhile.

Funded by the European Union

However, a couple of adaptations were needed to improve the performance of the routines, e.g. the classification graphs for saving were cut into smaller pieces.

In a second step the import routines were extended by intensive logging and validation. This helped to handle multiple issues: (1) finding errors in the original data (e.g. genus-species combinations not matching), (2) solving problems in the import routines, and (3) improving handling of exceptional data in the original data set, e.g. extraordinary long reference titles that appear only a few times.

Erroneous data (and sometimes even unwanted exceptional data) could be reported back to the original source for correction in the original database.

By adding validating logging the data quality could be greatly improved already. However, to further improve and ensure high quality, additional round-trip validation routines were written. Round-trip validation checks if the imported data can be exported again into the original format and the result is the same or similar to the original one. This way unintended data loss or incorrect data transformation can be recognized, and it can be validated that the import is correct and complete.

However, as some data is neglected on purpose during the import, not all data can be restored exactly as in the original data. The validation routines have to take this into account. Similarly, not all data transformations are 100% reversible and therefore cannot be restored in the original format. This means that the validation routines have to be adapted such that they allow handling these differences while still validating the correctness as far as possible. Even if 100% restoring was not possible, the round-trip validation helped to find a number of hidden issues in the import routines that could be fixed or improved.

In general, the ERMS import could run directly into the merging CDM instance, including validation. However, as validation results may make it necessary to run an import more than once, the workflow was designed such that it first runs into a separate instance, and only if this import was sufficiently successful an additional CDM-2-CDM import (see 4.2) is triggered to import the data to the merging CDM instance. This is faster than running the original import from scratch.

4.4 Adding Index Fungorum

The Index Fungorum import improvements were similar to those for ERMS (see 4.3). The old pilot implementation routines could be used, and improved and validating logging was added. Additionally round-trip validation was enclosed and CDM-2-CDM import appended.

As the Index Fungorum data is a mix of data from different sources (names from Index Fungorum, classification from Species Fungorum, and distribution data from FRDBI) and the collaboration between Index Fungorum and FRDBI stopped many years ago, currently fungi data to be imported to EU-Nomen does not change over time. As long as this is the case, the fungi import can be short-cutted as only the CDM-2-CDM import needs to run where the existing CDM instance from the previous import can be taken as source.

5. Merging

The core module of the EU-NOMEN process is the merge module. It merges the four separate source classifications into one consensus classification. A top down approach is followed here which first merges highest rank taxa (e.g. rank of kingdom) and then works further down to the lowest ranks.

5.1 General considerations

Funded by the European Union

For this, in a first step names are matched against each other to find potential merge candidates. This can be done mostly automatically but in some cases, e.g. names that differ slightly in authorship, user feedback is needed to avoid matching homonyms.

5.1.1 Priorities

To allow semi-automated merging it is important to define priorities for the sources for certain taxonomic groups. For plants, for example, Euro+Med was given priority over ERMS and Index Fungorum, which also include some plants. In the following we distinguish between the *priority classification* and the *non-priority classification* for a given name or taxon.

5.1.2 Name merging

To ease the process of taxon merging, it is split into two parts, pure name merging and taxon merging. Name merging results in a single name record, which includes all data from the priority name and, if it exists and is not contradictory, additional data from the non-priority name. Additionally, the non-priority classification's flag is added to the name to indicate that it originates from both classifications. Manual feedback is only needed if potentially contradictory information exists (e.g. spelling of authorship or differing nomenclatural status). The priority information is taken by default, but in some cases the default can be overwritten. E.g. for the fungi name Claviceps purpurea it was needed to decide if the author and reference information retrieved from Index Fungorum ("(Fr.) Tul., Annls Sci. Nat., Bot., sér. 3 20. 1853") or the one retrieved from ERMS ("(Fr.:Fr.)Tul., 1883") or a merge of both should be taken.

Name matching does not include the taxonomic status of the name.

5.1.3 Taxon merging

Taxon merging is more complex for multiple reasons. Work on the correct classification of species is ongoing and for many taxa differences in opinions regarding classification and names exist between scientists.

Additionally, the associated data structure of taxa is much more complex. Taxa are organized into hierarchical trees of accepted names, which may also include synonyms. These synonyms may be accepted names in other classifications. Merging two or more taxa may have various consequences for the related child or parent taxa, as well as for the associated synonyms. The fact that classifications may be incomplete or exclude certain ranks—ranks that could be included in another classification—adds complexity to the merging process..

Last but not least, taxa are associated with factual data (such as distribution data and common names), which requires careful decisions on the correct handling and deduplication of these factual data during the merge process.

The following describes typical situations encountered during the merging process, along with their default resolution strategies.

- Both names are accepted and have the same parent

In this case only internal taxon information needs to be merged. This includes name merging, synonym merging, and full factual data merging. Finally all children are to be moved to the new taxon without checking for duplicates as this is part of the next iteration.

- Both names are accepted but have different parents

This is often the case if one of the classifications uses intermediate ranks which do not exist in the other classification. In this case the taxon that is not yet attached to the

intermediate rank taxon needs to be moved there as a child. Everything else is similar to merging accepted names with the same parent.

If the reason that there are two different parents is due to differing taxonomic opinions rather than intermediate ranks, a manual decision is required. By default, the priority classification's approach will be taken and the non-priority classification will be adapted such that it fits into the priority classification. The adaptation will include the moving of child taxa, synonyms, and factual data. In certain cases, e.g. when it becomes clear that the non-priority classification follows a more recent taxonomic approach, the adaptation will be done the other way round.

- Only the non-priority name is a synonym

This indicates that the taxonomic concepts differ between the two classifications. A taxon in the non-priority classification may have been split in the priority classification. To solve this, the synonym status of the name needs to be removed as the consensus classification only allows one taxonomic status for each name. This case may also give some indication that the accepted taxon of the synonym in the non-priority classification may have child taxa that also need to be moved as children to the accepted taxon in the priority classification. However, if these child taxa use names that do not exist in the priority classification, it can not be decided from the data, which child taxa have to be moved. These cases need either to be solved manually or that a certain degree of inconsistency is accepted

- Only the priority name is a synonym

This also indicates that taxonomic concepts differ between the two classifications. Two taxa in the non-priority classification may have been unified in the priority classification. This can be easily solved if the accepted taxon of the synonym in the priority classification also exists as an accepted taxon in the non-priority classification. In this case the accepted status of the name in the priority classification can be removed. All children of the taxon can then be moved to the synonym's accepted taxon in the priority classification. If the accepted taxon of the synonym in the priority classification does not exist in the non-priority classification, the case is more complex and needs to be solved manually.

- Both names are synonyms for the same accepted taxon

This case is similar to "both are accepted and have the same parent" and can be handled accordingly. However, no merging of factual data and moving of child taxa is needed.

- Both names are synonyms for different accepted taxa

This also indicates that the taxonomic concepts differ between the two classifications. Often the border between two neighboring taxonomic concepts may have shifted, so the concepts are overlapping in certain parts. To solve this nothing needs to be done but the relationship to the formerly accepted taxon of the non-priority classification will get lost. As in the case of "Only the non-priority name is a synonym", this may also give some indication that the accepted taxon of the synonym in the non-priority classification may have child taxa that need to be moved as children to the accepted taxon of the synonym in the priority classification.

- Complex cases

Some cases are more complex and can only be decided and resolved manually

5.2 Algorithm

As mentioned above, a top-down approach is followed here, starting with the highest ranks. As expected, name-matching showed that overlaps occur almost exclusively between ERMS and each of the other classifications. Therefore, and since ERMS is the only classification that contains the highest taxon Biota, the full taxon trees of the other

Funded by the European Union

classifications are first moved under ERMS Biota without merging them. This results in sibling taxa like Animalia sensu ERMS and Animalia sensu Fauna Europaea and is the starting point for the merge.

Subsequently, the following step is executed starting with the highest rank down to the lowest. For each matching name of the given rank both classifications are analyzed and the most suitable merging strategy is chosen, depending on the position of the name in both classifications and depending on related taxa (parents, children, synonyms).

If the situation is unambiguous the merge is done automatically, if not, user feedback is requested.

To be able to run the merge frequently, the results of manual decisions are stored as far as possible in a merge decision repository. Then, the next time that the algorithm is running the repository is consulted first before asking for manual feedback. However, this does not (yet) work in complex cases.

5.3 Implementation

As the EDIT Platform core components use Java as programming language, the merging was also written in Java. Currently it is still a command line tool. When user feedback is needed it provides information on the situations and offers options to resolve them.

In the future, there are plans to develop a visual user interface that will offer the same functionality while providing improved visualization for the corresponding problem.

6. Export to the Data Warehouse

Once the data is fully merged it still needs to be exported to the final data warehouse (see 3.2). A first but very inefficient export existed already from the first pilot implementation. This export was fully refactored and adapted to use the same framework that was developed in the meanwhile for other EDIT Platform exports such as Darwin Core Archive, World Flora Online Backbone and CDM light (Luther 2020).

Same as the imports, the export is flanked by validation measures that run validation during the export (via logging) and after the export as round-trip validation routines.

7. New Data

Within TETTRIs' Task 2.1, entirely new data was either added or prepared for inclusion in EU-Nomen or its source databases. These data include:

- Bryophytes
- Lichens
- Common names for Animalia

7.1 Bryophytes

In 2020 Hodgetts et al. (2020) published a new European checklist for Bryophytes. These data were fully included into Euro+Med as a new taxonomic group. Additionally, Mediterranean distributions were added from Ros et al. (2013) . Synonyms were taken from multiple sources (e.g. Koperski (2000)). The inclusion of synonyms is not fully completed yet and still needs review by relevant experts. However, as Bryophytes became part of Euro+Med they will automatically also become part of the upcoming new EU-Nomen release.

7.2 Lichens

During the refactoring of the EU-Nomen process it was realized that the situation for Fungi is unsatisfactory. Not only that the existing data has a strong focus on Great Britain

Funded by the European Union

and is not sufficiently covering other areas but also the cooperation between Index Fungorum and FRDBI which provided the required distributions has been stopped. This way it is not possible to provide a data update anymore.

As a first step to overcome this, and as expertise on European lichens exists at FUB-BGBM, it was decided to set up a European lichens checklist similar to the plant checklist Euro+Med. The data gathering is still under way but data for most areas already exists, and it is expected that a first version of the checklist will be available still in 2025.

The European lichens checklist will also run on the EDIT Platform and therefore integration into the EU-Nomen process will be relatively easy as it will be analogous to Euro+Med.

Once the European lichens checklist is up and running, an investigation will be undertaken if further fungi data can be included.

7.3 Common names for Animalia

Common names are part of EU-Nomen. However, only two source databases (Euro+Med and ERMS) provide common name information on their own.

In former EU-Nomen versions a minor set of common names were attached to zoological data directly into the final data warehouse. For the new EU-Nomen process it was decided to use the common name database located at NHMW to enrich the data with further zoological common names. In an initial step NHMW provided common names for more than 14,000 zoological taxa, which were directly imported to Fauna Europaea before the import to EU-Nomen. As NHMW also provides web services for their common name data, these services will be used in the future to check for data updates.

8. Automation

To simplify the execution of the EU-Nomen workflow, the continuous integration framework Jenkins (https://www.jenkins.io/) is used. Dedicated jobs have been set up to automatically execute individual steps of the workflow, such as installing the databases, merging them into a single database, and exporting the merged data into the final data warehouse. These jobs can be started via a user interface (see figure 5). Should any step be unsuccessful, the system prompts the user for necessary manual feedback.

Only the merge process is not yet included into this pipeline as it currently still requires more user feedback and therefore cannot yet run fully automatically.

Funded by the European Union

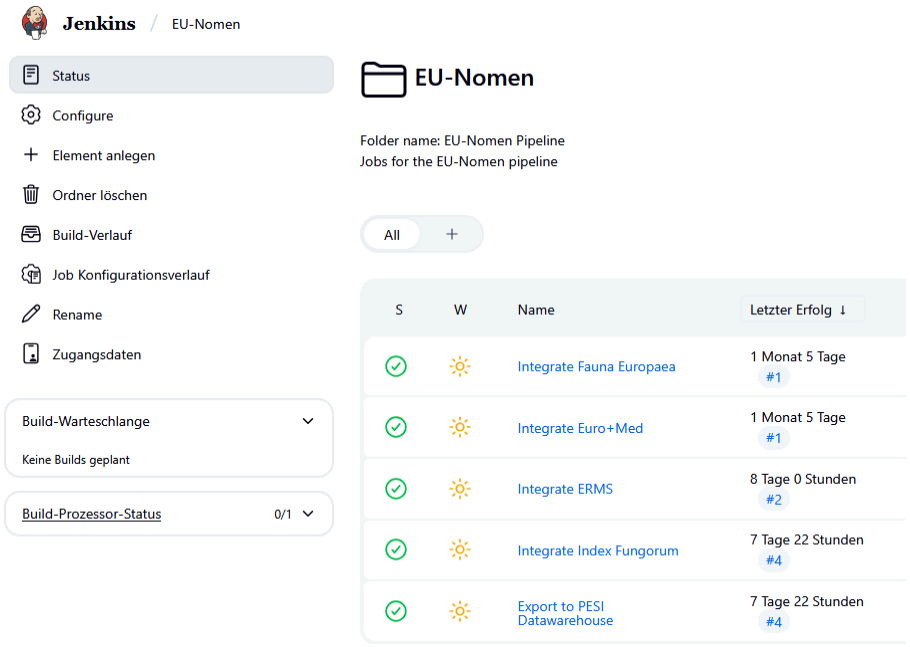

Figure 5: Jenkins UI to run the EU-Nomen workflow

9. Outlook

The improved EU-Nomen process described above makes it much easier to update the data on a more regular basis. An annual update of the data is therefore envisaged. It will be important to what extent Fauna Europaea, as the largest data source, is actively curated again.

Fauna Europaea currently lacks stable funding and structures to ensure the sustainable operation of the platform. To secure its future, a host has been found in the Natural History Museum in Stuttgart, which has taken on the task of hosting the data and acting as a point of contact for the curatorial network. In parallel, an application for third-party funding has been prepared for the German Research Foundation (DFG), with the central aim of ensuring the sustainability of the infrastructure and its integration into the international data landscape.

10. ACKNOWLEDGEMENTS

We would like to thank the Flanders Marine Institute (VLIZ) for the good cooperation in making EU-nomen available online and in providing marine database ERMS. We would also like to thank the Natural History Museum Vienna (NHMW) for providing the Common Names, the Natural History Museum Stuttgart (SMNS) for their willingness to host and further develop Fauna Europaea in the future, and Robert Lücking (FUB-BGBM) for his commitment to the creation of a European lichen database.

11. REFERENCES

Ciardelli P, Kelbert P, Kohlbecker A, Hoffmann N, Güntsch A, Berendsohn WG, The EDIT Platform for Cybertaxonomy and the taxonomic workflow: selected Components, Lecture Notes in Informatics (LNI), vol. 154, pp. 625-638, 2009

Güntsch A, Overmann J, Ebert B, Bonn A, Le Bras Y, Engel T, Anders Hovstad, K Lange Canhos DA, Newman P, van Ommen Kloeke E, Ratcliffe S, le Roux M, Smith VS, Triebel D, Fichtmueller D, Luther K, National biodiversity data infrastructures: ten essential functions for science, policy, and practice, *BioScience*, Volume 75, Issue 2, February 2025, Pages 139–151, https://doi.org/10.1093/biosci/biae109

Hodgetts N. G., Söderström L., Blockeel T. L., Caspari S., Ignatov M. S., Konstantinova N. A., Lockhart N., Papp B., Schröck C., Sim-Sim M., Bell D., Bell N. E., Blom H. H., Bruggeman-Nannenga M. A., Brugués M., Enroth J., Flatberg K. I., Garilleti R., Hedenäs L., Holyoak D. T., Hugonnot V., Kariyawasam I., Köckinger H., Kučera J., Lara F., Porley R. D. 2020: An annotated checklist of bryophytes of Europe, Macaronesia and Cyprus. Journal of Bryology 42: 1-116. https://dx.doi.org/10.1080/03736687.2019.1694329

de Jong Y, Kouwenberg J, Boumans L, Hussey C, Hyam R, Nicolson N, Kirk P, Paton A, Michel E, Guiry MD, Boegh PS, Pedersen HÆ, Enghoff H, von Raab-Straube E, Güntsch A, Geoffroy M, Müller A, Kohlbecker A, Berendsohn W, Appeltans W, Arvanitidis C, Vanhoorne B, Declerck J, Vandepitte L, Hernandez F, Nash R, Costello MJ, Ouvrard D, Bezard-Falgas P, Bourgoin T, Wetzel FT, Glöckler F, Korb G, Ring C, Hagedorn G, Häuser C, Aktaç N, Asan A, Ardelean A, Borges PAV, Dhora D, Khachatryan H, Malicky M, Ibrahimov S, Tuzikov A, De Wever A, Moncheva S, Spassov N, Chobot K, Popov A, Boršić I, Sfenthourakis S, Kõljalg U, Uotila P, Gargominy O, Dauvin J-C, Tarkhnishvili D, Chaladze G, Tuerkay M, Legakis A, Peregovits L, Gudmundsson G, Ólafsson E, Lysaght L, Galil BS, Raimondo FM, Domina G, Stoch F, Minelli A, Spungis V, Budrys E, Olenin S, Turpel A, Walisch T, Krpach V, Gambin MT, Ungureanu L, Karaman G, Kleukers RM.J.C, Stur E, Aagaard K, Valland N, Moen TL, Bogdanowicz W, Tykarski P, Węsławski JM, Kędra M, M. de Frias Martins A, Abreu AD, Silva R, Medvedev S, Ryss A, Šimić S, Marhold K, Stloukal E, Tome D, Ramos MA, Valdés B, Pina F, Kullander S, Telenius A, Gonseth Y, Tschudin P, Sergeyeva O, Vladymyrov V, Rizun VB, Raper C, Lear D, Stoev P, Penev L, Rubio AC, Backeljau T, Saarenmaa H, Ulenberg S (2015) PESI - a taxonomic backbone for Europe. Biodiversity Data Journal 3: e5848. https://doi.org/10.3897/BDJ.3.e5848

Koperski M, Sauer M, Braun W, Gradstein S.R. (2000). Referenzliste der Moose Deutschlands. Schriftenreihe für Vegetationskunde 34. Bundesamt für Naturschutz Bonn: 1-519.

Luther K, Müller A, Kohlbecker A, Güntsch A, Berendsohn WG (2020), EDIT Platform output model for botany v. 2.05, Freie Universität Berlin, http://dx.doi.org/10.17169/refubium-29536

Ros RM, Mazimpaka V, Abou-Salama U, Aleffi M, Blockeel TL, Brugués M, Cros RM, Dia MG, D GM, Draper I, El-Saadawi W, Erdağ A, Ganeva A, Gabriel R, González-Mancebo JM, Granger C, Herrnstadt I, Hugonnot V, Khalil K, Kürschner H (2013). Mosses of the Mediterranean, an annotated checklist. – Cryptogamie, Bryologie 34: 99-283. http://dx.doi.org/10.7872/cryb.v34.iss2.2013.99

Vandepitte L., Vanhoorne B., Decock W., Dekeyzer S., Trias Verbeeck A., Bovit L., Hernandez F., Mees J. (2015). How Aphia - the platform behind several online and taxonomically oriented databases - can serve both the taxonomic community and the field of biodiversity informatics. Journal of Marine Science and Engineering 3(4): 1448-1473. https://dx.doi.org/10.3390/jmse30

Funded by the European Union