A Pan-European Species-directories Infrastructure
PESI WP6 - Integrated e-Services

Deliverable 6.1 Technological Implementation Plan

Milestone 6.3 Web-portal functional description

Table of Contents

Introduction 1
Functional specifications of the PESI web portal preceding implementation 2
Technical aspects, tasks and timeline, including cross-links with other WPs 3
Aspects preceding online dissemination 3
Aspects directly related to online dissemination 4
Aspects related to the use PESI data in the e-science domain 6
Aspects to the future of the PESI e-infrastructure 9

Introduction

PESI aims at building an e-infrastructure for all species in Europe, including a single online portal, which integrates the taxonomic information from the three major European check-lists: Fauna Europaea (FaEu), Euro+Med Plantbase (E+M) and the European Register of Marine Species (ERMS). In future it will possibly take up additional checklists to cover European fungi, algae and bryophytes. In addition to taxonomic information, PESI will harvest additional information on species (common names, images, distributions, literature, conservation status and will provide links to other portals (*e.g.* national check-lists, red species lists) and other bioinformatics databases such as the GenBank sequence database, the Barcode of Life database (BOLD) and the Biodiversity Heritage Library (BHL) for literature data.

Besides an attractive portal with a web page on every species, the portal will provide an intelligent name validation service and webservices to cross match external species lists against names in PESI. This will enable the end-users to standardize the use of organism names and enables them to use the persistent globally unique identifiers (GUIDs) supplied by PESI.

The proto-type portal was deployed in May 2010 and the fully operational version will be ready in March 2011.

Functional specifications of the PESI web portal preceding implementation

The requirements of the PESI portal were discussed during the WP4-5-6 meeting in Ostend (March 2009), and it was agreed that each species page on the PESI web portal should have the following features:

1) Scientific Name and Authority (author + publication year)
2) Name status (nomenclatural and taxonomic status)
3) Taxon rank
4) Classification (parent – child relationships)
5) Taxon relationships such as basionymy, synonymy
6) Common names (if any)
7) Habitat (marine – brackish – freshwater – terrestrial)
8) Distribution maps (occurrence details)
9) Supplemental information (*e.g.* priority status)
10) Bibliographic references and other sources (original description, revisions, databases)
11) External links (deep links to other species web sites, at least one back to the data provider's website)
12) Images (if any)
13) Edit session (from source database: date when record was created or changed and by whom)
14) Source check-list GUID or LSID
15) Page citation

The following search options are required. The visitor should be able to search on a scientific or common name (at all ranks) and then gets a list of results with links to the species pages (if not directly to the species page in case of an exact match). The web portal should also be able to create a list of matching results after setting a number of taxonomic, geographical or other criteria.

For example: create a list of taxa (any rank) based on:

- a higher parent taxon and/or
- geographical zones and/or
- additional data types such as conservation status

For example, this will allow the web visitor to generate a list of all protected bird species occurring in Belgium.

Besides species pages and a search engine, the portal should provide a name validation service and should act as a hub or backbone to link names with other information stored at different locations.

The PESI portal should also provide a platform to find information on local expertise, such as contact details of experts linked to taxa, organisations, literature and websites. The PESI National Focal Points will provide this information.

The PESI web portal should be multilingual (as stated in the Description of Work).

Technical aspects, tasks and timeline, including cross-links with other WPs

Aspects preceding online dissemination

With regards to *taxonomic information*, the procedures on data integration, data quality assurance, merging procedures, database structure and data exchange formats and synchronisation protocols are outlined in **Deliverable 5.2** (report on the criteria, procedures and mechanisms for quality control) and **Deliverable 5.3** (Joint e-infrastructure disseminating Pan-European checklists).

Responsibility: WP 5&6
Timeline: decisions to be taken during 1st and 2nd year, implementation in 3rd year.

Species information provided by the PESI National Focal Points will be stored at VLIZ. Taxon names will be validated against PESI and missing names will be send to the checklist coordinators, for inclusion in the PESI source databases**.**

Responsibility: WP 3&6
Timeline: decisions to be taken during 1st and 2nd year, implementation in 3rd year.

Information on the PESI portal has to be compliant with *international standards*. More information and the decisions taken about the use of standards that apply to taxonomic names, taxon ranks, name author strings, character encoding, commons names, geographical regions, taxon occurrences and legal protection and conservation status are outlined in **Deliverable 4.1** (Report on authoritative taxonomic standards from multiple sources suitable for deployment within European Research Area) and **Deliverable 4.3** (Application and Adoption of Taxonomic Standards). This means that prior to dissemination, the data from the checklists and National Focal Points should be made compliant with these standards either at the source level or a transformation is needed at the data integration and quality assurance level (see **Deliverable 5.2**).

Responsibility: WP 4
Timeline: decisions to be taken during 1st and 2nd year, implementation in 3rd year.

A decision has been made to implement and supply *persistent Globally Unique Identifiers (GUIDs or LSIDs)* for taxonomic names provided by the source checklists. A strategy for the creation and assignment of new GUIDs is provided in **Deliverable 5.5** (Versioning and the use of GUIDs for PESI).

Responsibility: WP 4,5&6
Timeline: decisions to be taken during 1st and 2nd year, implementation in 3rd year.

The contact information related to the expert who provided or validated the taxonomy in the source check-lists is integrated in the online *EDIT expert database* and each expert received a unique identifier. The PESI datawarehouse will store this identifier, so the PESI portal can provide deep links to the expert pages. More information on the EDIT expert database can be found in **Deliverable 2.1** (The European Taxonomic Work force (ETW), its tasks, activities and operational standards inspiration by the Open Source Society).

Responsibility: WP 2&5
Timeline: decisions to be taken during 1st and 2nd year, implementation in 3rd year.

Aspects directly related to online dissemination

Clear information on *copyright, citations and attribution rules* is of utmost importance for the data providers as well as the end-users of the information provided by PESI. Background information and decisions taken within PESI about this topic are available in **Deliverable 2.1** (The European Taxonomic Work force (ETW), its tasks, activities and operational standards inspiration by the Open Source Society) and **Deliverable 2.2** (The Government of IPR of Electronic Biodiversity Data). Each species page on the PESI portal displays the citation, the expert and the date the last time the record was modified.

Responsibility: WP 2&6
Timeline: decisions to be taken during 1st and 2nd year, implementation in 3rd year.

Images are an important source of information on species and they attract more users to the portal. The portal shows thumbnail images, the meta information such as attribution and copyright and provides a link to the webpage that shows the original picture. The PESI web servers will not store a physical copy of the image, but requires the image be stored on a public server, including access to the meta information. The images from the source checklists need to provide a web link to their thumbnail and full image. A successful pilot is the dynamic connection established with the Dutch Species Register (NLSR). The NLSR portal serves the information of species images in an RDF XML format, which we harvest and display on the portal via a POST/GET request.

Responsibility: WP 5&6
Timeline: decisions to be taken during 1st and 2nd year, implementation in 3rd year.

The PESI portal also harvests information from other *bioinformatics databases* such as the GenBank sequence database, the Barcode of Life database (BOLD) and the Biodiversity Heritage Library (BHL) for literature data. Deep links to BOLD and GenBank are generated automatically via a monthly scheduled task that queries all taxon records in PESI and compares them with BOLD/GenBank records. Links from GenBank to PESI will be built using the NCBI LinkOut system. The procedure requires a user to have a "ProviderID". One can then create links to PESI by uploading a special formatted text file on the NCBI FTP server. This is then parsed and displayed on the GenBank website.
BHL links are generated on the fly, when requesting a taxon page. Afterwards the link is cached for one week for subsequent requests and performance reasons.

Responsibility: WP 6
Timeline: decisions to be taken during 1st and 2nd year, implementation in 3rd year.

The PESI portal provides information on the *legal protection and conservation status of European priority species*. The portal displays information from CITES, the IUCN Red list, the Ospar Convention, the EU Bird and Habitat Directives, the EPPO alert and invasive plant species and the pest insect species on the HYPPZ list. Deep links to species pages on the IUCN and HYPPZ websites are included. The information on these priority species is stored in a database hosted by VLIZ, and the species names from these lists need to be re-matched with names in PESI each time an updated and new PESI datawarehouse version is available.

Responsibility: WP 1&6
Timeline: decisions to be taken during 1st and 2nd year, implementation in 3rd year.

The PESI portal provides *species distribution maps* based on occurrence records stored in ERMS, FaEu and E+M. The datawarehouse integrates all available occurrences (based on the proposed standard list of place names). These place names will be mapped to a polygon in GeoServer. The distribution maps will be created on the fly via an Open Layers application.

Responsibility: WP 5&6
Timeline: decisions to be taken during 1st and 2nd year, implementation in 3rd year.

The PESI portal will only show the latest version of the PESI datawarehouse. Previous versions are archived in Miscrosoft SQL server 2008 format. Access to backups will be available in the near future via the VLIZ Marine Data Archive (MDA; http://mda.vliz.be). VLIZ is an official national data centre and any data that are stored on their servers is archived and backed up daily on tapes that are stored at three different physical locations. The data archiving procedures at VLIZ also ensure physical readability of the data through copying archived data to new, modern media every three to five years and through upgrading the files so they are compatible with future reader hardware.

Responsibility: WP 5&6
Timeline: continuous.

The PESI web portal is multilingual. The search terms and standard features are translated into several European languages. It was decided that the PESI National Focal Points will help in the translation. Any free textual phrases can remain in English. Terms that are not yet translated by the focal points are available for download at http://www.eu-nomen.eu/portal/languages/download_translations.html.

Responsibility: WP 3&6
Timeline: decisions to be taken during 1st and 2nd year, implementation in 3rd year.

Aspects related to the use PESI data in the e-science domain

Besides species pages, which are of interest to the general public, the PESI portal also provides services and applications that are targeted towards scientific purposes. In these cases a local copy of the database or dynamic access via a web service is preferred. It was agreed to adapt or build upon existing tools, rather than develop new tools or resources de novo; and to work in a step-wise fashion to judge the success of initial measures, and facilitate adoption of new opportunities as they arose.

The *exchange of data and information* in a structured schema (Darwin Core), a readable and manageable size format (flat text, including meta data) is discussed in **Deliverable 4.3** (Application and Adoption of Taxonomic Standards).

The PESI portal (via the MDA) will provide controlled download access to the taxonomic information in the above mentioned format.

Responsibility: WP 5&6
Timeline: decisions to be taken during 1st and 2nd year, implementation in 3rd year.

The use of GUIDs greatly enhances the potential for database cross-referencing and interoperability between providers and consumers of data. PESI will not assign a new GUID to a PESI record; it will use the existing ones from ERMS, FaEu and E+M. For ERMS, LSIDs have been assigned for all Taxon Names using the LSID standard. The PESI species URLs will contain these source GUIDs, in order to ensure URLs are persistent through time.

GUID (and more specific LSID) resolvers are available on various websites (example: http://lsid.tdwg.org) and can even be embedded in web browsers (http://lsid.silverbiology.com/).

GUID/LSID resolvers should return data in RDF Format. RDF stands for Resource Description Format, and is an XML model for describing and linking data objects. By using the widely used Dublin Core and Darwin Core (for species information) terms, interchange between

other (taxonomic) databases outside PESI is simplified. Both ERMS, FaEu and E+M will need to make their GUIDs/LSIDs resolvable.

Responsibility: WP 5&6
Timeline: decisions to be taken during 1st and 2nd year, implementation in 3rd year.

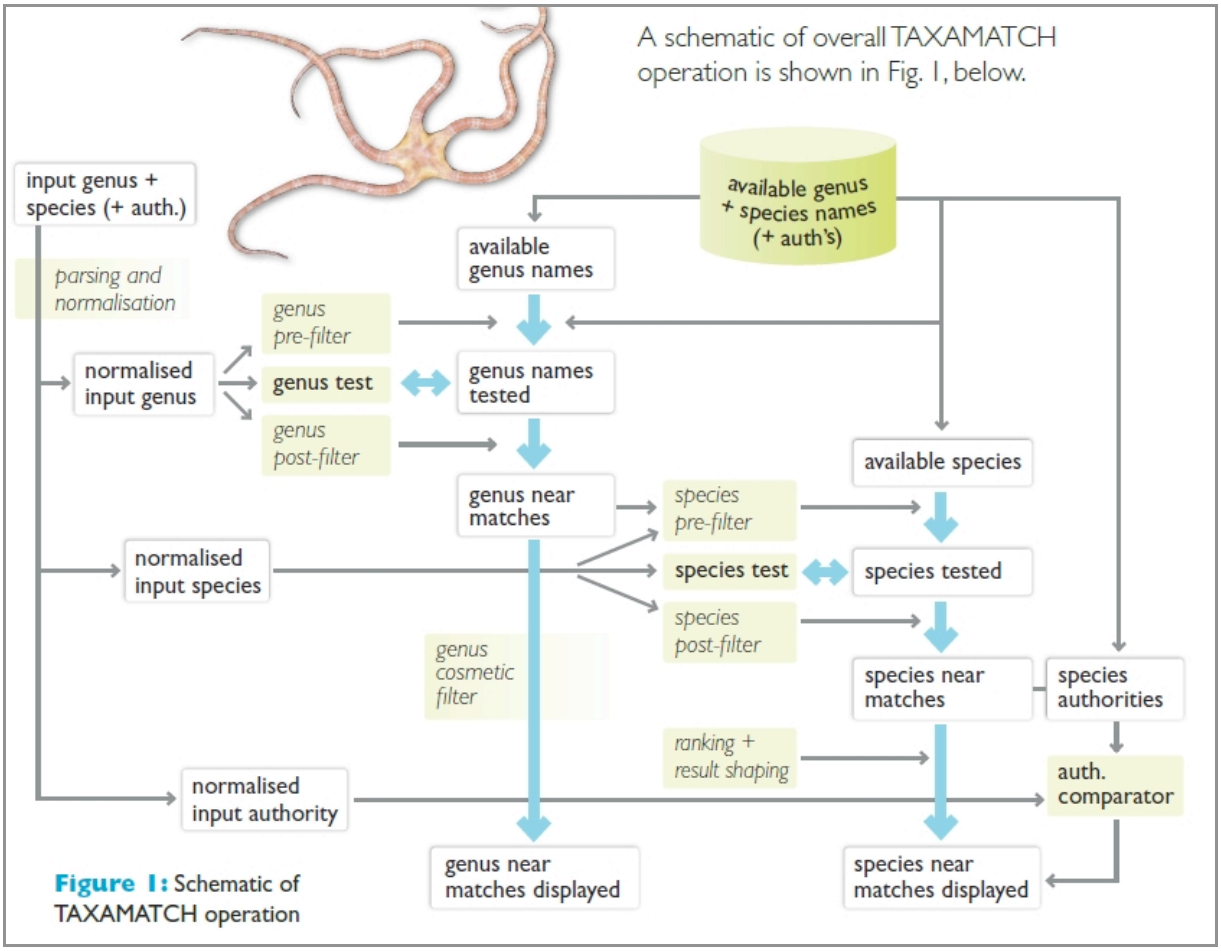
The PESI portal has a *semi-automated name validation tool* (taxon match) to cross-check external species lists. The tool returns standard PESI taxonomic information in a user-friendly format (*e.g.* MS Excel or tab delimited text file). The user needs to upload a list of species names, match the columns with the fields in the PESI data warehouse and the system will return the file with valid names (notifies when the name is an unaccepted synonym), the authority and publication date, the hierarchical classification, quality status (expert validated or not) and the check-list GUIDs. When there are multiple matches the system provides a pick-list. The tool is an implementation of the fuzzy matching algorithm written by Tony Rees (CSIRO, Australia), which comprises a suite of custom filters and tests used in succession on genus, species epithet, plus authority where supplied.

An overview of the tests is given below, and includes the following:

- An exact match test, both before and after minor normalisation
- A phonetic match test, using a custom algorithm “tuned” to the characteristics of taxon scientific names
- A custom “Modified Damerau-Levenshtein Distance” (MDLD) algorithm which looks for possible omitted, inserted, substituted and transposed characters and character blocks
- A modified n-gram comparison of author names and dates where supplied, including expansion of selected known abbreviations of author names as appropriate.

The custom filtering that has been developed at both genus and species epithet levels comprises:

- Genus and species pre-filters, which serve to speed up the algorithm execution by excluding names deemed to be almost certain not to match from being tested
- Genus and species post-filters, which apply a set of rules to assist in the discrimination of likely “true” from “false” near matches
- A genus cosmetic filter, which presents only a subset of “genus near match” search results to the human web interface, while passing a wide range of genera through to the species stage for further testing
- A final result shaping stage (which can be switched out if desired), which masks more distant near matches in the presence of closer ones, but opens automatically to show them when the latter are absent.



The PESI portal also provides a web service that allows users to dynamically link their own applications to the PESI database and will allow them to match a locally stored species list with the PESI check list and add taxonomic and additional information derived from PESI. A few examples of possible applications:

- check the spelling of your taxa
- get the authority for your taxa
- get the full classification for your taxa
- resolve your unaccepted names to accepted ones
- resolve a common name/vernacular to a scientific name
- get the sources/references for a taxon

Responsibility: WP 6
Timeline: decisions to be taken during 1st and 2nd year, implementation in 3rd year.

Aspects to the future of the PESI e-infrastructure

The discussion on this topic is in progress and will be published in **Deliverable 2.3** (Maintenance of the PESI facilities) and **Deliverable 1.3** (Business Plan to guarantee long term continuity of PESI). Technical aspects and implications of the PESI portal as outlined in this document, should be taken into consideration for a future PESI e-infrastructure.

Responsibility: All
Timeline: decisions to be taken during 3rd year, final report due December-January 2011.

Configuration History			
Version No.	**Date**	**Changes made**	**Author**
0.1	11 March 2009	First draft for circulation within WP4, WP5, WP6	WA
0.2	19 March 2009	Updated version of D6.1 (including comments from WP4, WP5, WP6)	WA
1.0	27 May 2009	Final first version of D6.1	WA
1.1	18 September 2009	Revised version of D6.1	WA
1.2	06 October 2009	Revised version of D6.1	WA, BV
1.3	30 October 2009	Revised version of D6.1	FH
1.4	02 December 2009	Revised version of D6.1	WA, LB
1.5	30 September 2010	Final version of D6.1	WA, BV, AG
1.6	13 October 2010	Final version submitted	YdJ