Standardising taxonomic data

Dr Yde de Jong offers insights into a project enhancing taxonomic information; work that is vital to supporting scientific research and real-life biodiversity management applications throughout Europe

Can you give us a brief overview of the main goals of your work?

The overall objective of the Pan-European Species-Directories Infrastructure (PESI) is to define and coordinate strategies to enhance the quality and reliability of European biodiversity information and information services. In particular, we aim to establish an authoritative taxonomic reference list for Europe, in order to standardise the use of species names both in biodiversity research and practical applications.

It is almost a year since PESI was first featured in _International Innovation_. What has the project achieved in this period and have you faced any significant challenges during this time?

After a slow start the Focal Points accelerated the execution of their work plan. This resulted in, among other things, the delivery of 130,000 vernacular species names. In addition, the pan-European checklists have been updated for most of the groups in the last project year, providing an up-to-date version for the initial PESI Portal release, including around 250,000 scientific species names and 400,000 taxon names. This has been a major achievement for the coordinators and custodians of the checklists, taking into account the limited capacity and resources for managing those huge networks of specialists.

Lastly, the design and functionalities of the PESI Portal have been significantly improved during the last year and the PESI Portal is now ready for release.

How challenging has it been to coordinate a cross-border collaborative project of this scale? Could you briefly outline the biggest challenges and benefits from attempting to unify so many disparate groups?

Apart from the administrative management, the major challenges of PESI result from the size and complexity of the project. The project was both complex – in terms of the number and kind of social and information networks to be developed and connected – and large, due to the number of partners and the quantity of information feeding into the project. As an example, more than 700 experts, organised in the Society for the Management of European Biodiversity Data, directly contribute to the updating of the pan-European checklists. In parallel, the Network of National Focal Points, supported by around 1500 Focal Point experts, contributes to the validation of the pan-European checklists. To effectively merge the results of both efforts will certainly exceed the lifespan of the PESI project and will require a long-term, sustainable work plan.

The long-term positive outcome for Europe could be the establishment of an extended, integrated taxonomic work force that is able to work collaboratively on certain tasks under common research standards and formats. This could notably contribute to next generations of taxonomic issues that may require a more concerted research approach to be resolved. From an informatics perspective, PESI is performing very well with regard to developing and connecting the scheduled information services. Nevertheless, we are having to deal with a relatively limited ICT capacity which hinders a larger, more significant, outreach in the information and technical domains, such as the further development of advanced, virtual workbenches for taxonomists to manage their data. We expect that PESI in this regard could benefit from some of the recently funded EC FP7 projects, like ViBRANT.

Could you explain how PESI aims to release its information as Linked Data and what the benefits are for end users?

PESI defines and implements strategies to optimise the sharing of taxonomic data. Linked Data is a mark-up using the HTTP URIs as identifiers to ease the semantic web application. The Linked Data concept will be exploited by PESI as a minimal ontology to share and integrate data across web accessible resources, beside other strategies to implement persistent identifiers, like GUIDs.

PESI PORTAL SCREEN SHOT

Your project is due to end shortly. What plans are in place to continue the work of the programme after this period?

Critical for the continuation of the PESI project will be the maintenance of relevant staff positions at the involved institutes to secure a basic management of the different infrastructural components. In addition, new funding needs to be raised to further develop the PESI infrastructural architecture to a higher functional level. Both objectives will be addressed partly by means of in-kind institutional or national contributions to the LifeWatch construction plan, and partly by drafting new applications, for instance for the Eighth Framework Programme (FP8) or national science foundations.

The backbone of Europe's biodiversity management

PESI, the ambitious project to standardise taxonomic information, is being coordinated by the Zoological Museum Amsterdam and is proving essential for supporting European biodiversity research and management initiatives

GIVEN THE PERVASIVE issues related to climate change and global warming, and the resulting changes to ecosystems, the need to improve our management of the Earth's biodiversity has never been more pressing. However, before we can properly tackle this mammoth task, we need to make sure our taxonomic information is accurate. The correct use of species names is absolutely essential for proper biodiversity management, and a fundamental component of this is the availability of taxonomically validated standardised nomenclators (or name databases), so that researchers and policy makers alike can be assured that they are working with the correct species.

There are a number of initiatives happening within the European Research Area (ERA) to help develop systems that will assemble and integrate taxonomic information. For example, ViBRANT is attempting to accumulate species datasets into standardised web interfaces (so-called 'Scratchpads'), and 4D4Life is contributing to the preparation of a global catalogue of species names by bringing together a number of separate taxonomic databases into the Catalogue of Life initiative. This work is very valuable, but adequate access to taxonomic information remains scarce in terms of availability (lacking completeness), connectedness (lacking effective integration), and reliability (lacking consistency), hindering a synthesised use of biodiversity information on the Internet for European users.

OPEN-ACCESS INFRASTRUCTURE

Networks of taxonomic and biological experts have been working on a solution to this problem. The Pan-European Species-Directories Infrastructure (PESI) project is focused on coordinating the delivery of a comprehensive Taxonomic Backbone for Europe through the inter-operation of existing data infrastructures. The PESI initiative, funded by the European Union under the Research Infrastructures division of the Seventh Framework Programme (FP7), is being led by the University of Amsterdam and involves 40 partner organisations from 26 countries.

PESI is a three-year project that kicked off in May 2008 and is now close to completion. Overseeing this bold initiative is Dr Yde de Jong, who is based at the Zoological Museum in Amsterdam. The initiative represents the next step in integrating and securing taxonomically authoritative species name registers that underpin the management of biodiversity in Europe. PESI seeks to integrate the three main all-taxon registers in Europe, namely the European Register of Marine Species, Fauna Europaea, and Euro+Med PlantBase, in coordination with EU-based nomenclators and the network of EU based Global Species Databases. As such, it is a standards-based, quality controlled, expertly validated, open-access infrastructure for research, education and data management.
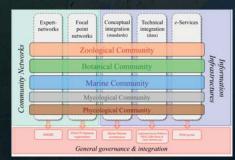
A TAXONOMIC HUB OF EUROPE

PESI coordinates strategies designed to build the quality and consistency of European biodiversity information, by integrating the infrastructural components of five major community networks on taxonomic indexing into a joint work programme. This will result in functional knowledge networks of taxonomic experts and regional focal points, which will collaborate on the establishment of standardised and authoritative taxonomic data.

One of the major components of the PESI project is the 'PESI data warehouse' which hosts the integrated Taxonomic Backbone for Europe. This data warehouse includes an index of species names associated with a number of attributes, such as synonyms, their place in the management phylogeny and their geographical distribution. The data content results from the integration of the earlier pan-European checklists into a unified directory, following advanced routines on data verification and harmonisation. These routines are laid down in the so-called 'PESI Common Data Model store' (PESI CDM-store), a component of EDIT's Cybertaxonomy Platform. "The PESI data warehouse is also the decisive layer to optimise and implement mechanisms," explains de Jong, "allowing unambiguous cross-links to, and compatibility with, other biodiversity information services, like persistent identifiers, standardised vocabularies and exchange formats." He is convinced that this component is crucial to secure the stability of the European Taxonomic Backbone and its interoperability with other data sources. From the 'PESI data warehouse', datasets are disseminated to various external services, like

PESI INFRASTRUCTURAL COMPONENTS

PESI

A PAN-EUROPEAN SPECIES-DIRECTORIES INFRASTRUCTURE

OBJECTIVES

PESI will contribute to the strengthening of the respective scientific, social, political, technological, and information capacities in Europe, needed for a proper biodiversity assessment. The project is the next step in integrating and securing taxonomically authoritative species name registers that underpin the management of biodiversity in Europe.

PARTNERS

UvA, NL • **UKBH**, DK • **TU**, TR • **NHM**, UK • **FUB-BGBM**, DE • **VLIZ**, BE • **EcoServe**, IE • **SMEBD**, IE • **MNHN**, FR • **RBG Kew (IPNI)**, UK • **ITZN**, UK • **CABI**, UK • **NUIG**, IE • **UH**, FI • **UNIPA Palermo**, IT • **US Sevilla**, ES • **IBSAS**, SK • **NKMA**, GR • **NNM**, NL • **IEVU**, LT • **CSFI**, IT • **NRM**, SE • **CUB**, SK • **NTNU**, NO • **SMNH**, UA • **PAS-MIZPAN**, PL • **SSS**, CH • **ILAUNI**, GE • **CSIC**, ES • **NIB**, SL • **NMNHS**, BG • **MyNA**, RO • **LU**, LV • **HCMR**, GR • **IOLR**, IL • **IOPAS**, PL • **ZIN RAS**, RU • **IBSS**, UA • **MBA**, UK • **USTL**, FR

FUNDING

Funded by the European Commission's Seventh Framework Programme (FP7) within the Research Infrastructures programme. Contract no. RI-223806

CONTACT

Dr Yde de Jong
Project Coordinator

Netherlands Centre for Biodiversity Naturalis (section Zoological Museum Amersterdam)
University of Amsterdam
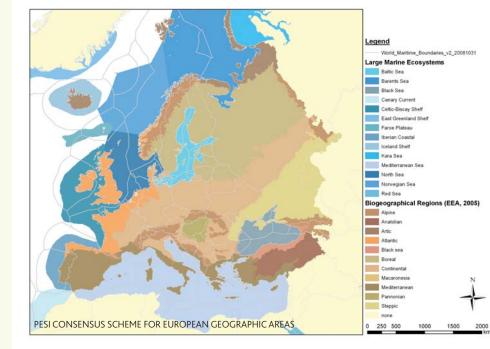PO Box 94766
1090 GT Amsterdam, The Netherlands

T + 31 20 525 7191
F + 31 20 525 7780
E yjong@uva.nl

www.eu-nomen.eu/pesi

· ·

DR YDE DE JONG has an academic background in systematics, evolutionary biology, bioinformatics, epistemology, and marine botany. He is Head of the Department Biodiversity Informatics of the Zoological Museum Amsterdam, coordinator of the PESI project and EDIT's task leader on integrating Europe's taxonomic information infrastructures. Further to this, he is the executive manager of the Fauna Europaea project, chair of GBIF's ECAT work area, and involved in several biodiversity informatics initiatives and work groups.

PESI CONSENSUS SCHEME FOR EUROPEAN GEOGRAPHIC AREAS

the Global Names Architecture and the PESI Portal. This makes the content of the major pan-European taxonomic infrastructures available on the web, and supports its use in other e-science applications.

One of the important components of PESI is to significantly expand the network of end-users adopting the pan-European checklists as their taxonomic point of reference. This has been achieved through the broad dissemination model that PESI has adopted, which addresses scientific communities, applied biodiversity researchers, and national biodiversity networks and services as well as the general public. In addition, PESI provides a baseline infrastructure that contributes to large-scale biodiversity research initiatives, supporting tens of thousands of users working with biodiversity data. Targeting end-users as adopters of the PESI taxonomic standard required a differentiated approach. One example that de Jong highlights is how PESI joined the development of a common agenda on managing prioritised taxa for Europe, meaning the set-up of a common policy on those targeted species that play a role in Europe's regulations: "PESI scrutinised the European prioritised species lists by comparing these to the PESI taxonomic standard and then disseminated the outcomes to our network through the PESI portal," he recalls. For individual users, the PESI Portal offers easy validation routines, so they can check their own lists against PESI. In addition, the species pages include a feedback form that allows users to discuss their findings with the PESI taxonomic experts.

DELIVERING PRACTICAL RESULTS

Another area the PESI network is working on is providing species lists based on geographic regions. In de Jong's opinion, agreeing the best practices for delineating these regions has at times been a challenge: "At the moment, providing species occurrence details at the country level is the best we can offer taking into account all taxonomic groups," he says. "For some groups,

evidently sub-country level details could easily be delivered, but for other groups even providing country level details is a hard job." The project team are considering showing geo-reference data from GBIF or BioCASE in the PESI Portal in the near future and this would be an optional function, overlaying the distribution maps and enabling users to compare the occurrence details from two information sources.

The PESI project is working in collaboration with a number of other FP7 projects such as LifeWatch. In general, the maintenance of most of PESI's infrastructural components is incorporated in the construction phase of the latter project. In addition, PESI provides a supporting infrastructure for some of the LifeWatch-initiated projects, like BioVel, which establishes a virtual biodiversity laboratory: "PESI will provide a liaison to the taxonomic research community, as well as a connection to the PESI information services," points out de Jong. In EcoBOS, which builds a network of marine research stations for biodiversity monitoring and study, PESI will provide a Clearing House to support the continuing efforts on building a taxonomic reference list on marine species.

PESI also provides core infrastructural services to some of the recently funded EU FP7 projects. In ViBRANT, PESI will, for instance, contribute to the set-up of an architecture supporting the access to, and exchange of, associated vocabulary services (like the PESI CDM-store and GBIF's Vocabulary Service). This facilitates the management and dissemination of standardised ontologies for Scratchpad users.

As this current stage draws to a close, de Jong is looking forward to a bright future for PESI. He hopes to maintain critical staff positions at a number of institutes, as well as taking the PESI project to the next level of functionality so it can continue to support the management of biodiversity within the European Union.