9.26 Reusable biodiversity informatics tools

Dmitry Mozzherin, Patrick Leary, Anna Shipunov, Alexey Shipunov
Encyclopedia of Life, <dmozzherin[at]eol.org

Many biodiversity informatics web-based projects provide Application Programming Interfaces (APIs), which allow outside users to take advantage of the projects' functionality. However when operating on large datasets, network communication often becomes a bottleneck. In such cases, using software installed on the client side can significantly improve processing times.

During the development of some Global Names Architecture (GNA) applications such as Global Names Index (GNI) or Global Names Integrated Taxonomic Editor (GNITE) the authors have created a set of components for processing scientific names. These components are created in the Ruby programming language (they are called 'gems' in Ruby) and are very easy to install on various platforms and operating systems. We also provide wrappers that allow these components to be accessed from other programming languages.

Scientific Name Parser (biodiversity gem, rubygems.org/gems/biodiversity): It is unthinkable to start any automatic treatment of biodiversity information without first finding the semantic elements of scientific names. Scientific names often have a complex structure that can be hard to parse in an automated fashion. This component is able to evaluate scientific names with complex structure including multiple authorships, names of hybrids, etc. The underlying technology used in the parser could be adapted to strictly enforce the codes in order to create a nomenclatural code verification tool to support submission of newly described species. In addition to the standard Ruby interface, there are also command line and socket interfaces which allow it to be used with other languages.

Taxonomic matcher (taxamatch_rb gem, rubygems.org/gems/taxamatch_rb): This component uses algorithms developed by Tony Rees to decide if two name strings are variants of the same scientific name. The core of this gem is written in the C programming language to increase performance. This component can either be optimized for convenience (rapid development) or for performance (processing large datasets and caching the results).

Darwin Core Archive Tool (dwc-archive gem, rubygems.org/gems/dwc-archive): A component to handle creation of new or reading of existing Darwin Core Archive files.

Developers working on biodiversity related projects can use flexible, reusable, open source components such as these to quickly recombine the functionality in various meaningful ways. These components can be published in online software repositories, allowing public discovery of and access to these tools. Such 'small' tools are being developed by many biodiversity groups in several programming languages. One downside of this approach is that currently finding such tools is not a trivial task. The authors are interested in helping to develop a shared registry or repository of such tools for the benefit of the broader biodiversity informatics community.

This work was sponsored by grants for the Encyclopedia of Life (MacArthur & Sloan) and the Data Conservancy (NSF).

An example of reusing the parser tool: www.marinespecies.org/aphia.php?p=match
biodiversity parser online: gni.globalnames.org/parsers/new
taxamatch original: www.cmar.csiro.au/datacentre/taxamatch.htm
taxamatch-webservice by GBIF: code.google.com/p/taxamatch-webservice/