Managing Biodiversity Knowledge in the Encyclopedia of Life

Jennifer M. Schopf, Sarah Bordenstein, Patrick Leary, Peter Mangiafico, David J. Patterson, Alexey Shipunov, David Shorthouse

Encyclopedia of Life Informatics, Marine Biological Laboratory, Woods Hole, MA 02543
Email: {jms, srbordenstein, pleary, peter, dpatterson, ashipunov, dshorthouse}@eol.org

Abstract. The Encyclopedia of Life is currently working with hundreds of Content Providers to create 1.8 million aggregated species pages, consisting of tens of millions of data objects, in the next ten years. This article gives an overview of our current data management and Content Provider interactions.

Keywords: Encyclopedia of Life (EOL), Content management, Biodiversity

1 Introduction

The Encyclopedia of Life (EOL) project (http://www.eol.org) aims to create Web pages for each of the 1.8 million known species. We acquire data objects about organisms from an array of Content Providers, index those objects in a biologically meaningful fashion using a names-based cyberinfrastructure, and make them available through a variety of means including species Web pages and application programming interfaces (APIs).

For the release of EOL v1.0 in February 2008 we worked with a dozen formal data providers and hundreds of additional institutions and individuals to serve over 30,000 species pages and 24 fully developed exemplar pages. To proceed in a scalable fashion, we are developing innovative solutions to interact with a wide variety of providers. The goal of the EOL project is to deliver Web pages for all species in 10 years time, which means continually adding content on the order of 500 species a day. This can only be done through careful management and interactions with data providers and community members.

2 Current Approach

Our approach has three main facets: the use of scientific and vernacular names to index and connect data objects; a set of interfaces to Content Providers with internet-accessible data; and a flexible content management system for other content.

We use the name of an organism as a universal identifier for species data. However, names are imperfect in this role because species can have more than one name, the same name can be used for more than one species, and there can be disagreement as to what a name refers to or how organisms should be grouped together. Therefore, we have developed algorithms to disambiguate and reconcile names, and improve our ability to index data objects based on names alone.

As our partnerships with Web-enabled stewards of biodiversity data mature, we will increasingly focus on standardized data structures and transport mechanisms. We have defined *connector* consisting of APIs and schemas that allow Content Providers to share their information with the EOL. Providers with structured data that use standardized schemas and APIs can be easily incorporated, and we strongly encourage the adoption of these standards. The ideal Content Provider structures their data in a consistent fashion for all species, uses accepted standard data schemas, provides open access to their data with some kind of Web service or data transfer protocol, includes a resolvable link that allows us to embed links back to the original data set, and has adopted a Creative Commons license that allows unimpeded re-use of content. Unfortunately, the largest pool of information – at this time – does not comply with all of these factors.

In parallel with the adoption of standard processes and connectors, we are developing Drupal-based content management systems (CMS), currently referred to internally as *LifeDesk*. These initially hosted environments allow clade-specific special interest groups to compile structured and compliant data for eventual aggregation on species pages. Data may be entered, identified, and curated through a set of graphically-rich tools that interface to a relational database. In addition, APIs developed within the LifeDesk environments act to solidify the schema under development for structured, compliant Content Providers. Our approach has been heavily influenced by the EDIT project ScratchPads and we will be developing compatible modules to interface with their existing clade-specific interest groups.

Initial implementation of the LifeDesk environment will focus on tools for the Expert user, although citizen scientist and teacher/classroom versions will follow. Initial functionality will include the creation of a set of "stub" species pages given a list of names, inclusion of text and image, and literature tools. We aim to have hosted LifeDesk services through alpha testing and available to selected community members by Fall 2008. Policy and implementation to include LifeDesk data as part of the EOL will follow shortly thereafter.

EOL is committed to providing services for its Content Partners. We currently offer a series of benefits that include:

- Visibility: EOL will showcase efforts of its contributors.
- Improved data usage: With increased visibility, data will be more often accessed and used. EOL can report how often we have re-served their data, to whom, and where they stand with respect to other providers.
- Feedback: User feedback regarding the accuracy, completeness, and appropriateness of information will be directed back to content.
- Community building: The EOL is an integrative environment that has the capacity to create communal data pools and consensus classifications.

3 Conclusion

In summary, EOL has the potential to use novel nomenclatural approaches and state of the art technology to index a wide variety of biodiversity information for 1.8 million species. Managing data and communicating with Content Providers is a key aspect to our approach, and we are developing software to interact with both existent and future sources of content.