Evolutionary Data Leaping to Web 3.0: Some Highlights From NESCent’s Third Hackathon

Hilmar Lapp1, Arlin Stoltzfus2,3, Todd J. Vision1,4, Rutger A. Vos5

1*US National Evolutionary Synthesis Center (NESCent), Durham, NC, USA (2) Center for Advanced Research in Biotechnology, University of Maryland, Rockville, MD, USA (3) National Institute for Standards and Technology (NIST), Gaithersburg, MD, USA (4) Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA (5) Departments of Zoology and Botany, University of British Columbia, Vancouver, BC, Canada

Motivation

There is a rich and meticulously curated variety of evolutionary data available through online databases, yet most of these do not support standard formats or explicit semantics, and can only be queried through interfaces intended for humans. This presents barriers to effective reuse of the data, such as combining, connecting, and visualizing data from different sources in ways not often imagined by the original database creators. Removing these obstacles to data interoperability can facilitate a much broader utilization of the already existing wealth of data, which in turn may further spur efforts to bring an increasing amount of the evolutionary data isolated in silos or still hidden offline into an online web of interoperable data.

To promote the adoption and refinement of standards towards this vision, NESCent sponsored the Evolutionary Database Interoperability Hackathon.

Data Annotation With Rich Semantics

In an intensive collaboration, a subgroup of participants drafted a RDFa-compliant specification for how to embed annotation in NeXML with ontology-anchored semantics. They also created a parser and writer toolkit in the Java programming language to implement the specification. The semantics of the annotation can be extracted as an RDF graph using general-purpose semantic web tools.

Phyloinformatics Standards

To address the interoperability obstacles, the hackathon aimed to bring to bear three emerging standards for phylogenetics data and databases:

1) NeXML (http://nexml.org) as a formal, unified syntax that can be validated and embedded in arbitrary metadata
2) Comparative Data Analysis Ontology (CDAO) for explicit, computable semantics of evolutionary data and analyses (http://www.evolutionaryontology.org)
3) PhyloWS as a standard programmable interface for phylogenetic data providers (http://evoinfo.nescent.org/PhyloWS)

All three standards emerged from the Evolutionary Informatics Working Group sponsored by NESCent.

Data Flow and Use Cases

The collaborative efforts at the hackathon focused on three main objectives towards a seamless data flow:

1) Remove the barriers to data providers making their holdings searchable and accessible through web-services.
2) Provide canonical and globally unique identifiers for data items.
3) Allow automatic extraction of semantic web-compliant RDF graphs of data entities as well as of metadata.

The combination of web-services and semantic web is often referred to as Web 3.0.

Phylogenetic data that is accessible through web-services in standard formats facilitates mash-up applications that combine different kinds of data. As an example, a group of participants created an application that can load a NeXML-formatted tree from any URL and decorate it with images pulled from Morphbank for each species in the tree.

Web-Services for Data Access

One subgroup launched a generic publishing toolkit that allows data providers to make their data holdings searchable and accessible through the PhyloWS programmable interface standard. Participants also created a web-service prototype for TreeBASE that demonstrates the resolution of canonical PhyloWS-based URLs for trees to NeXML documents.

Data Visualization and Mash-Up

Phylogenetic data can be visualized in ways often not intended for humans. This presents barriers to effective reuse of the data, such as combining, connecting, and visualizing data from different sources in ways not often imagined by the original database creators. Removing these obstacles to data interoperability can facilitate a much broader utilization of the already existing wealth of data, which in turn may further spur efforts to bring an increasing amount of the evolutionary data isolated in silos or still hidden offline into an online web of interoperable data.

To promote the adoption and refinement of standards towards this vision, NESCent sponsored the Evolutionary Database Interoperability Hackathon.

Data Annotation With Rich Semantics

In an intensive collaboration, a subgroup of participants drafted a RDFa-compliant specification for how to embed annotation in NeXML with ontology-anchored semantics. They also created a parser and writer toolkit in the Java programming language to implement the specification. The semantics of the annotation can be extracted as an RDF graph using general-purpose semantic web tools.

Phyloinformatics Standards

To address the interoperability obstacles, the hackathon aimed to bring to bear three emerging standards for phylogenetics data and databases:

1) NeXML (http://nexml.org) as a formal, unified syntax that can be validated and embedded in arbitrary metadata
2) Comparative Data Analysis Ontology (CDAO) for explicit, computable semantics of evolutionary data and analyses (http://www.evolutionaryontology.org)
3) PhyloWS as a standard programmable interface for phylogenetic data providers (http://evoinfo.nescent.org/PhyloWS)

All three standards emerged from the Evolutionary Informatics Working Group sponsored by NESCent.

Data Flow and Use Cases

The collaborative efforts at the hackathon focused on three main objectives towards a seamless data flow:

1) Remove the barriers to data providers making their holdings searchable and accessible through web-services.
2) Provide canonical and globally unique identifiers for data items.
3) Allow automatic extraction of semantic web-compliant RDF graphs of data entities as well as of metadata.

The combination of web-services and semantic web is often referred to as Web 3.0.

Phylogenetic data that is accessible through web-services in standard formats facilitates mash-up applications that combine different kinds of data. As an example, a group of participants created an application that can load a NeXML-formatted tree from any URL and decorate it with images pulled from Morphbank for each species in the tree.