Nature Precedings : hdl:10101/npre.2010.4588.1 : Posted 29 Jun 2010

EvoIO: Community-driven standards for sustainable interoperability

Arlin Stoltzfus[1], Karen Cranston[2], Hilmar Lapp[3], Sheldon McKay[4], Enrico Pontelli[5], Rutger Vos[6], Nico Cellinese[7]

1 National Institute of Standards and Technology, Gaithersburg, MD, 20899 (stoltzfu@umbi.umd.edu); 2 Field Museum of Natural History, Chicago, IL 50506; 3 National Evolutionary Synthesis Center (NESCent), Durham, NC 27705; 4 Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724; 5 Department of Computer Science, New Mexico State University, Las Cruces, NM 88003; 6 University of Reading, Reading, RG6 6BX, United Kingdom; 7 University of Florida, Florida Museum of Natural History, Gainesville, FL 32611

Interoperability is the property that allows systems to work together independent of who created them, or how or for what purpose they were implemented. Data that is interoperable can be reliably read, written, and interpreted by a broad range of tools and databases, independent of who produces and who consumes the data. Interoperability is crucial for aggregating data from different online resources and for integrating different kinds of data. Interoperability is also a key ingredient of automated analysis workflow construction and discovery, which is increasingly important to make sense out of high-throughput data. As an example from evolutionary informatics, the ability to retrieve and display a phylogenetic tree by specifying solely the tree's unique identifier, followed by resolving the tip labels to taxon names and decorating the tree with geographic ranges pulled from specimen-based occurrence repositories, requires interoperability of data resources, data formats, and data semantics for trees, taxonomic names, and geo-references. Interoperability is based on effective standards that become and remain broadly adopted. We argue that to develop and apply such standards for evolutionary and biodiversity data sustainably, we need a community-driven, open, and participatory approach.

With the goal to build such an approach, the EvoIO collaboration emerged in 2009 from several NESCent-sponsored activities towards software and data interoperability for evolutionary analysis, including the Evolutionary Informatics working group (2006-2009), and the Evolutionary Database Interoperability hackathon (2009). EvoIO aims to be a nucleating center for developing, applying and disseminating interoperability technology that connects and coordinates between stakeholders, developers, and standards bodies.

Members of the EvoIO group, which include biologists and computer scientists, have over the past 3 years harnessed a variety of collaborative events to successfully build an initial stack of interoperability technologies that is owned by the community and open to participation. The stack addresses syntax, semantics, and programmable services, and at present includes the following components: NeXML (http://nexml.org, LGPL), a NEXUS-inspired XML format that is validatable yet extensible; CDAO (http://www.evolutionaryontology.org, GPL), an ontology of comparative data analysis formalizing the semantics of evolutionary data and metadata; and PhyloWS (http://evoinfo.nescent.org/PhyloWS), a web-services interface standard for querying, retrieving, and referencing phylogenetic data on the web. Beyond demonstration prototypes, reference implementations of EvoIO stack technologies are starting to appear in production use. Examples include the TreeBASE API, and the exchange format for computable phenotype annotations attached to character states in phylogenetic data matrices.

Aside from producing such information artefacts, EvoIO devotes much of its energy to applying principles of communication and organization that result in open and inclusive processes of community science. One of the key tools employed by EvoIO is the hackathon event format. Hackathons are highly collaborative, hands-on working meetings that catalyze practical innovation, train researchers, and foster cohesion as well as a sense of shared ownership in the results. Though participants in such events self-organize organically, they have been effective at creating community-owned standards with broad buy-in and shared maintenance. As an example, two components of the EvoIO technology stack are direct results of such events: PhyloWS was initiated by an EvoIO group member at the 2008 BioHackathon in Tokyo, Japan; and the 2009 Phyloinformatics VoCamp gave rise to a standard for referencing nodes in a phylogenetic tree by PhyloCode-inspired phylogenetic definitions. Similarly, hackathons triggered all current reference implementations of EvoIO stack technologies.

In summary, we find that broad community participation, buy-in, and ownership are critical for developing interoperability in a sustainable fashion, and there are approaches and tools that can foster these effectively.