rgbif: a package for working with species occurrence data in R

Scott Chamberlain*,[a]

[a] *University of California, Berkeley, CA, USA*

Abstract

1. xxx
2. xxx
3. xxx
4. xxxx

Introduction

Users of the popular statistical and mathematical computing platform R (R Core Team 2014) enjoy a wealth of readily installable comparative phylogenetic methods and tools (O'Meara 2014). Exploiting the opportunities arising from this wealth for complex and integrative comparative research questions relies on the ability to reuse and integrate previously generated or published data and metadata. The expanding data exchange needs of the evolutionary research community are rapidly outpacing the capabilities of most current and widely used data exchange standards (Vos *et al.* 2012), which were all developed a decade or more ago. This has resulted in a radiation of different data representations and exchange standard "flavors" that are no longer interoperable at the very time when the growth of available data and methods has made that interoperability most valuable. In response to the unmet needs for standardized data exchange in phylogenetics, a modern XML-based exchange standard, called NeXML, has recently been developed (Vos *et al.* 2012). NeXML comprehensively supports current data exchange needs, is predictably machine-readable, and is forward compatible.

The exchange problem for phylogenetic data is particularly acute in light of the challenges in finding and sharing phylogenetic data without the otherwise common loss of most data and metadata semantics (Stoltzfus *et al.* 2012; Drew *et al.* 2013; Cranston *et al.* 2014). For example, the still popular NEXUS file format (Maddison *et al.* 1997) cannot consistently represent horizontal gene transfer or ambiguity in reading a character (such as a DNA sequence base pair). This and other limitations have led to modifications of NEXUS in different ways for different needs, with the unfortunate result that NEXUS

*Corresponding author
Email address: scott(at)ropensci.org (Scott Chamberlain)

files generated by one program can be incompatible with another (Vos *et al.* 2012). Without a formal grammar, software based on NEXUS files may also make inconsistent assumptions about tokens, quoting, or element lengths. Vos et al. (2012) estimates that as many as 15% of the NEXUS files in the CIPRES portal contain unrecoverable but hard to diagnose errors.

A detailed account of how the NeXML standard addresses these and other relevant challenges can be found in Vos et al. (2012). In brief, NeXML was designed with the following important properties. First, NeXML is defined by a precise grammar that can be programmatically **validated**; i.e., it can be verified whether a file precisely follows this grammar, and therefore whether it can be read (parsed) without errors by software that uses the NeXML grammar (e.g. RNeXML) is predictable. Second, NeXML is **extensible**: a user can define representations of new, previously unanticipated information (as we will illustrate) without violating its defining grammar. Third and most importantly, NeXML is rich in **computable semantics**: it is designed for expressing metadata such that machines can understand their meaning and make inferences from it. For example, OTUs in a tree or character matrix for frog species can be linked to concepts in a formally defined hierarchy of taxonomic concepts such as the Vertebrate Taxonomy Ontology (Midford *et al.* 2013), which enables a machine to infer that a query for amphibia is to include the frog data in what is returned. (For a more broader discussion of the value of such capabilities for evolutionary and biodiversity science we refer the reader to Parr et al. (2011).)

To make the capabilities of NeXML available to R users in an easy-to-use form, and to lower the hurdles to adoption of the standard, we present RNeXML, an R package that aims to provide easy programmatic access to reading and writing NeXML documents, tailored for the kinds of use-cases that will be common for users and developers of the wealth of evolutionary analysis methods within the R ecosystem.

The rgbif package

The `rgbif` package ...

Conclusions and future directions

`rgbif` ...

Acknowledgements

This project was supported in part by the Alfred P Sloan Foundation (Grant 2013-6-22).

Data Accessibility

All software, scripts and data used in this paper can be found in the permanent data archive Zenodo under the digital object identifier (DOI). This DOI corresponds to a snapshot of the GitHub repository at github.com/ropensci/rgbif.

References

Cranston, K., Harmon, L.J., O'Leary, M.A. & Lisle, C. (2014). Best practices for data sharing in phylogenetic research. *PLoS Curr*. Retrieved from http://dx.doi.org/10.1371/currents.tol.bf01eff4a6b60ca4825c69293dc59645

Drew, B.T., Gazis, R., Cabezas, P., Swithers, K.S., Deng, J., Rodriguez, R., Katz, L.A., Crandall, K.A., Hibbett, D.S. & Soltis, D.E. (2013). Lost branches on the tree of life. *PLoS Biol*, **11**, e1001636. Retrieved from http://dx.doi.org/10.1371/journal.pbio.1001636

Maddison, D., Swofford, D. & Maddison, W. (1997). NEXUS: An extensible file format for systematic information. *Syst. Biol.*, **46**, 590–621. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11975335

Midford, P., Dececchi, T., Balhoff, J., Dahdul, W., Ibrahim, N., Lapp, H., Lundberg, J., Mabee, P., Sereno, P., Westerfield, M., Vision, T. & Blackburn, D. (2013). The vertebrate taxonomy ontology: A framework for reasoning across model organism and species phenotypes. *J. Biomed. Semantics*, **4**, 34. Retrieved from http://dx.doi.org/10.1186/2041-1480-4-34

O'Meara, B. (2014). CRAN task view: Phylogenetics, especially comparative methods. Retrieved from http://cran.r-project.org/web/views/Phylogenetics.html

Parr, C.S., Guralnick, R., Cellinese, N. & Page, R.D.M. (2011). Evolutionary informatics: unifying knowledge about the diversity of life. *Trends in ecology & evolution*, **27**, 94–103. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22154516

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Stoltzfus, A., O'Meara, B., Whitacre, J., Mounce, R., Gillespie, E.L., Kumar, S., Rosauer, D.F. & Vos, R.A. (2012). Sharing and re-use of phylogenetic trees (and associated data) to facilitate synthesis. *BMC Research Notes*, **5**, 574. Retrieved from http://dx.doi.org/10.1186/1756-0500-5-574

Vos, R.A., Balhoff, J.P., Caravas, J.A., Holder, M.T., Lapp, H., Maddison, W.P., Midford, P.E., Priyam, A., Sukumaran, J., Xia, X. & Stoltzfus, A. (2012). NeXML: Rich, extensible, and verifiable representation of comparative data and metadata. *Systematic Biology*, **61**, 675–689. Retrieved from http://dx.doi.org/10.1093/sysbio/sys025