

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

2   Scott Chamberlain<sup>\*,a</sup>

3   <sup>a</sup>*University of California, Berkeley, CA, USA*

## 4   **Abstract**

- 5       1. xxx
- 6       2. xxx
- 7       3. xxx
- 8       4. xxxx

## 9   **Introduction**

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

22   The exchange problem for phylogenetic data is particularly acute in light of the challenges in finding  
23   and sharing phylogenetic data without the otherwise common loss of most data and metadata semantics  
24   (Stoltzfus *et al.* 2012; Drew *et al.* 2013; Cranston *et al.* 2014). For example, the still popular NEXUS  
25   file format (Maddison *et al.* 1997) cannot consistently represent horizontal gene transfer or ambiguity  
26   in reading a character (such as a DNA sequence base pair). This and other limitations have led to  
27   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)

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

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

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

## 50 The `rgbif` package

51 The `rgbif` package ...

## 52 Conclusions and future directions

53 `rgbif` ...

## 54 *Acknowledgements*

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

## 56 *Data Accessibility*

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

## 60 **References**

- 61 Cranston, K., Harmon, L.J., O’Leary, M.A. & Lisle, C. (2014). Best practices for data shar-  
62 ing in phylogenetic research. *PLoS Curr.* Retrieved from [http://dx.doi.org/10.1371/currents.tol.](http://dx.doi.org/10.1371/currents.tol.bf01eff4a6b60ca4825c69293dc59645)  
63 [bf01eff4a6b60ca4825c69293dc59645](http://dx.doi.org/10.1371/currents.tol.bf01eff4a6b60ca4825c69293dc59645)
- 64 Drew, B.T., Gazis, R., Cabezas, P., Swithers, K.S., Deng, J., Rodriguez, R., Katz, L.A., Crandall,  
65 K.A., Hibbett, D.S. & Soltis, D.E. (2013). Lost branches on the tree of life. *PLoS Biol.* **11**, e1001636.  
66 Retrieved from <http://dx.doi.org/10.1371/journal.pbio.1001636>
- 67 Maddison, D., Swofford, D. & Maddison, W. (1997). NEXUS: An extensible file format for systematic  
68 information. *Syst. Biol.*, **46**, 590–621. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11975335>
- 69 Midford, P., Dececchi, T., Balhoff, J., Dahdul, W., Ibrahim, N., Lapp, H., Lundberg, J., Mabee, P.,  
70 Sereno, P., Westerfield, M., Vision, T. & Blackburn, D. (2013). The vertebrate taxonomy ontology: A  
71 framework for reasoning across model organism and species phenotypes. *J. Biomed. Semantics*, **4**, 34.  
72 Retrieved from <http://dx.doi.org/10.1186/2041-1480-4-34>
- 73 O’Meara, B. (2014). CRAN task view: Phylogenetics, especially comparative methods. Retrieved from  
74 <http://cran.r-project.org/web/views/Phylogenetics.html>
- 75 Parr, C.S., Guralnick, R., Cellinese, N. & Page, R.D.M. (2011). Evolutionary informatics: unifying  
76 knowledge about the diversity of life. *Trends in ecology & evolution*, **27**, 94–103. Retrieved from  
77 <http://www.ncbi.nlm.nih.gov/pubmed/22154516>
- 78 R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for  
79 Statistical Computing, Vienna, Austria. Retrieved from <http://www.R-project.org/>

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

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