Response to the editorial team

JSS 2581: Szocs, Schafer webchem: An R Package to Retrieve Chemical Information from the Web

Eduard Szöcs and Ralf B. Schäfer

September 20, 2016

Dear editorial team and reviewers,

We thank the reviewer for checking the functionality of our package and also checking the appropriateness of the documentation and providing valuable comments on our manuscript. We incorporated all the suggested changes and proof read the manuscript. Please find below detailed description of the changes made and responses to specific comments.

Kind regards,
Eduard Szöcs and Ralf B. Schäfer

Reviewer A

We are thankful, for providing comments, checking the functionality of our package and also checking the appropriateness of the documentation.

Comment 1: *"The article should be proof read to correct grammatical mistakes. page 1, "ensure a good data quality" should read "ensure good data quality". A few other similar errors were scattered elsewhere"*

Response: We proof read the manuscript and removed remaining spelling errors.

Comment 2: *"On page 2 the authors note that the query methods implement a time-delay. Is this user configurable? It doesn't appear to be and this might be a parameter to consider providing to the user (acknowledging the fact that a user could mis-use it!)"*

Response: The time delay is currently hard-coded and adjusted to the capabilities of the data sources. Some providers block IP addresses if to many request arrive. Therefore, we do not intend to make this user configurable, as misuse might affect whole institutions. However, we might adapt the delays if data sources increase their capabilities.

Comment 3: *"Some of the man pages I looked at indicated that the query might fail if the API is unavailable. It might be useful to have a helper method to test the availability of the remote resource before the query is run."*

Response: CRAN runs automated tests and all examples in the documentation on different operating systems. Sometimes theses might fail, because an API is not available. From a related project (taxize package to handle taxonomic data in R) we made the experience that APIs are often down only for a short period (updates, maintenance etc). To avoid triggering a failed build on CRAN, we skip running the examples on CRAN and provide the reason why we omitted the examples. We are thankful for the suggesting and will add ping_*() functions to check if a data-source is up and running in a future release.

Reviewer B

Comment 4: *"The paper strikes this reviewer as a little "light" compared to most papers in JSS. This is not to say that the information presented is not useful, but that it is not very elaborate. I recommend that the authors consider either extending one or more of the examples, or adding an additional example or two. "*

Response: We added an example how to query legal information in compounds, how to reuse already queried identifiers to query more information and how to join the retrieved information with the original data.

Comment 5: *"Please clarify what the differences are between ChemmineR and this package. Is the main unique strength of this package that it allows use of more data sources? If so, please make this more clear. "*

Response: The main difference to all mentioned packages is that webchem integrates access to many data sources. We rephrased this paragraph and it now reads: *"Within the R ecosystem, there are only few similar projects: rpubchem (Guha 2014) provides an interface PubChem. Similarly, ChemmineR (Cao et al. 2008), a mature chemoinformatics package, also provides an interface to Pubchem. webchem does not provide any chemoin-formatic functionality, but integrates access to many data sources.[...]"*

Comment 6: *"However, data quality is also crucial for data analysis (Stieger et al. 2014). Ensuring good quality requires additional effort and methods to be developed. - Yes, garbage in, garbage out. But just post-processing cannot make bad data good. I would rather you said "validating the quality of data" or something similar, and rephrased this paragraph. "*

Response: We agree and rephrased this paragraph and merge it with the previous paragraph. It now reads *"However, good quality of data is crucial for every analysis (Stieger et al. 2014) and additional effort and methods are needed to validate data quality."*

Comment 7: *"Recommend to put the abstract into a single paragraph, (same text is fine)."*

Response: We agree and changed accordingly.

Comment 8: *"R is one of the most widely used software for data cleaning,- R is one of the most widely used software enviroments for data cleaning,"*

Response: We agree and changed accordingly.

Comment 9: *"For all URLS (to CRAN, Github, etc): Do not insert the URL in the text. Instead, create a reference in the bibliography, and add a citation in the text."*

Response: We replaced all URLs in the text with citations.

Comment 10: *"Add references for Travis-CI and AppVeyor."*

Response: We added references to both.

Comment 11: *"'best' match - 'best' match (fix the quote here and in other places)"*

Response: We are thankful for catching this typographical error and changed through the manuscript.

Comment 12: *"In sentence 1 of Section 5.1: change "both" to "and both""*

Response: We changed this paragraph. See also comment number 5.

Comment 13: *"Go through all the references and ensure that titles are in title style, e.g., "ChemmineR: a compound mining framework for R." – "ChemmineR: A Compound Mining Framework for R.""*

Response: We checked all references and changed all titles to title style.