iPhylo

Revisiting RSS to monitor the latest taxonomic research

Roderic Page

Published November 23, 2021

Citation

Page, R. (2021). Revisiting RSS to monitor the latest taxonomic research. *Iphylo*. https://doi.org/10.59350/ndtkv-6ve80

Keywords

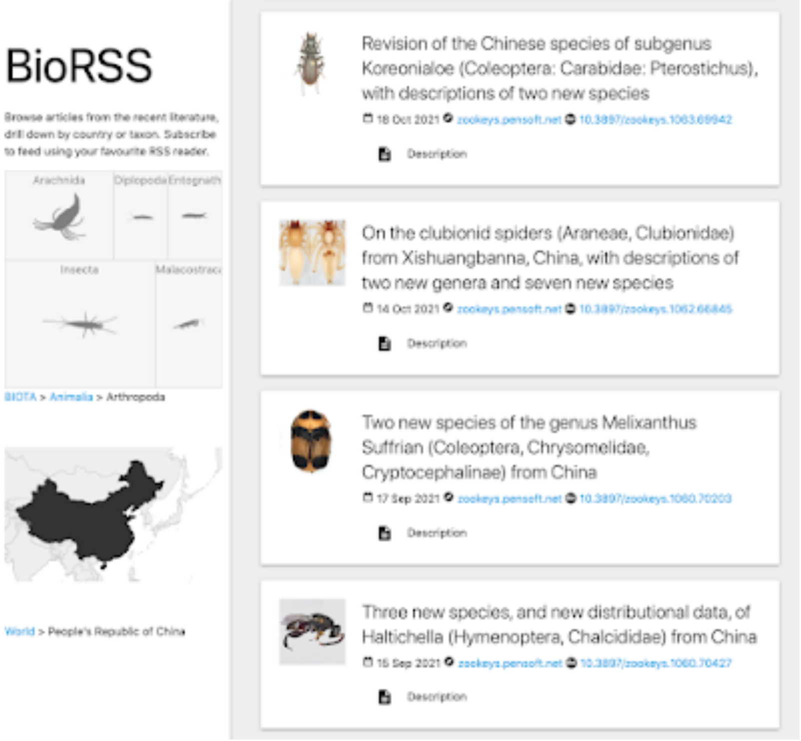
Geocoding, NoCode, RSS

Copyright

Copyright © Roderic Page 2021. Distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Over a decade ago RSS (RDF Site Summary or Really Simple Syndication) was attracting a lot of interest as a way to integrate data across various websites. Many science publishers would provide a list of their latest articles in XML in one of three flavours of RSS (RDF, RSS, Atom). This led to tools such as uBioRSS [1] and my own e-Biosphere Challenge: visualising biodiversity digitisation in real time. It was a time of enthusiasm for aggregating lots of data, such as the ill-fated PLoS Biodiversity Hub [2].

Since I seem to be condemned to revisit old ideas rather than come up with anything new, I've been looking at providing a tool like the now defunct uBioRSS. The idea is to harvest RSS feeds from journals (with an emphasis on taxonomic and systematic journals), aggregate the results, and make them browsable by taxon and geography. Here's a sneak peak:



BioRSS

Browse articles from the recent literature, drill down by country or taxon. Subscribe to feed using your favourite RSS reader.

BIOTA > Animalia > Arthropoda

World > People's Republic of China

Revision of the Chinese species of subgenus Koreonialoe (Coleoptera: Carabidae: Pterostichus), with descriptions of two new species
18 Oct 2021 zookeys.pensoft.net 10.3897/zookeys.1063.69942
Description

On the clubionid spiders (Araneae, Clubionidae) from Xishuangbanna, China, with descriptions of two new genera and seven new species
14 Oct 2021 zookeys.pensoft.net 10.3897/zookeys.1062.66845
Description

Two new species of the genus Melixanthus Suffrian (Coleoptera, Chrysomelidae, Cryptocephalinae) from China
17 Sep 2021 zookeys.pensoft.net 10.3897/zookeys.1060.70203
Description

Three new species, and new distributional data, of Haltichella (Hymenoptera, Chalcididae) from China
15 Sep 2021 zookeys.pensoft.net 10.3897/zookeys.1060.70427
Description

What seems like a straightforward task quickly became a bit of a challenge. Not all journals have RSS feeds (they seem to have become less widely supported over time) so I need to think of alternative ways to get lists of recent articles. These lists also need to be processed in various ways. There are three versions of RSS, each with their own idiosyncracies, so I need to standardise things like dates. I also want to augment them with things like DOIs (often missing from RSS feeds) and thumbnails for the articles (often available on publisher websites but not the feeds). Then I need to index the content by taxon and geography. For taxa I use a version of

Patrick Leary's "taxonfinder" (see https://right-frill.glitch.me) to find names, then the Global Names Index to assign names found to the GBIF taxonomic hierarchy.

Indexing by geography proved harder. Typically geoparsing involves taking a body of text and doing the following:

- Using named-entity recognition NER to identity named entities in the text (e.g., place names, people names, etc.).
- Using a gazetteer of geographic names GeoNames to try and match the place names found by NER.

An example of such a parser is the Edinburgh Geoparser. Typically geoparsing software can be large and tricky to install, especially if you are looking to make your installation publicly accessible. Geoparsing services seem to have a short half-life (e.g., Geoparser.io), perhaps because they are so useful they quickly get swamped by users.

Bearing this in mind, the approach I've taken here is to create a very simple geoparser that is focussed on fairly large areas, especially those relevant to biodiversity, and is aimed at geoparsing text such as abstracts of scientific papers. I've created a small database of places by harvesting data from Wikidata, then I use the "flash text" algorithm [3] to find geographic places. This approach uses a trie to store the place names. All I do is walk through the text seeing whether the current word matches a place name (or the start of one) in the trie, then moving on. This is very quick and seems to work quite well.

Given that I need to aggregate data from a lot of sources, apply various transformations to that data, then merge it, there are a lot of moving parts. I started playing with a "NoCode" platform for creating workflows, in this case n8n (in many ways reminiscent of the now defunct Yahoo Pipes). This was quite fun for a while, but after lots of experimentation I moved back to writing code to aggregate the data into a CouchDB database. CouchDB is one of the NoSQL databases that I really like as it has a great interface, and makes queries very easy to do once you get your head around how it works.

So the end result of this is "BioRSS" https://biorss.herokuapp.com. The interface comprises a stream of articles listed from newest to oldest, with a treemap and a geographic map on the left. You can use these to filter the articles by taxonomic group and/or country. For example the screen shot is showing arthropods from China (in this case from a month or two ago in the journal *ZooKeys*). As much fun as the interface has been to construct, in many ways I don't really want to spend time making an interface. For each combination of taxon and country I provide a RSS feed so if you have a favour feed reader you can grab the feed and view it there. As BioRSS updates the data your feed reader should automatically update the feed. This means that you can have a feed that monitors, say, new papers on spiders in China.

In the spirit of "release early and release often" this is an early version of this app. I need to add a lot more feeds, back date them to bring in older content, and I also want to make use of aggregators such as PubMed, CrossRef, and Google Scholar. The existence of these tools is, I suspect, one reason why RSS feeds are less common than they used to be.

So, if this sounds useful please take it for a spin at https://biorss.herokuapp.com. Feedback is welcome, especially suggestions for journals to harvest and add to the news feed. Ultimately I'd like to have sufficient coverage of the taxonomic literature so that BioRSS becomes a place where we can go to find the latest papers on any taxon of interest.

References

> 1. *Patrick R. Leary, David P. Remsen, Catherine N. Norton, David J. Patterson, Indra Neil Sarkar, uBioRSS: Tracking taxonomic literature using RSS, Bioinformatics, Volume 23, Issue 11, June 2007, Pages 1434–1436,* https://doi.org/10.1093/bioinformatics/btm109

> 2. *Mindell, D. P., Fisher, B. L., Roopnarine, P., Eisen, J., Mace, G. M., Page, R. D. M., & Pyle, R. L. (2011). Aggregating, Tagging and Integrating Biodiversity Research. PLoS ONE, 6(8), e19491.* doi:10.1371/journal.pone.0019491

> 3. *Singh, V. (2017). Replace or Retrieve Keywords In Documents at Scale. CoRR, abs/1711.00046.* http://arxiv.org/abs/1711.00046