Data mining applied to scholarly publications to finally reveal Earth's biodiversity

October 21 2019

A composite image made of taxonomic treatments as extracted by the Biodiversity Literature Repository (BLR) from scholarly publications. Credit: P Kishor (Plazi)

At a time when a million species are at risk of extinction, according to a recent UN report, ironically, we don't know how many species there are on Earth, nor have we noted down all those that we have come to know on a single list. In fact, we don't even know how many species we would have put on such a list.

The combined research including over 2,000 natural history institutions

worldwide, produced an estimated ~500 million pages of scholarly publications and tens of millions of illustrations and species descriptions, comprising all we currently know about the diversity of life. However, most of it isn't digitally accessible. Even if it were digital, our current publishing systems wouldn't be able to keep up, given that there are about 50 species described as new to science every day, with all of these published in plain text and PDF format, where the data cannot be mined by machines, thereby requiring a human to extract them. Furthermore, those publications would often appear in subscription (closed access) journals.

The Biodiversity Literature Repository (BLR), a joint project of Plazi, Pensoft and Zenodo at CERN, takes on the challenge to open up the access to the data trapped in scientific publications, and find out how many species we know so far, what are their most important characteristics (also referred to as descriptions or taxonomic treatments), and how they look on various images. To do so, BLR uses highly standardised formats and terminology, typical for scientific publications, to discover and extract data from text written primarily for human consumption.

By relying on state-of-the-art data mining algorithms, BLR allows for the detection, extraction and enrichment of data, including DNA sequences, specimen collecting data or related descriptions, as well as providing implicit links to their sources: collections, repositories etc. As a result, BLR is the world's largest public domain database of taxonomic treatments, images and associated original publications.

Once the data are available, they are immediately distributed to global biodiversity platforms, such as GBIF—the Global Biodiversity Information Facility. As of now, there are about 42,000 species, whose original scientific descriptions are only accessible because of BLR.

The very basic principle in science to cite previous information allows us to trace back the history of a particular species, to understand how the knowledge about it grew over time, and even whether and how its name has changed through the years. As a result, this service is one avenue to uncover the catalogue of life by means of simple lookups.

So far, the lessons learned have led to the development of TaxPub, an extension of the United States National Library of Medicine Journal Tag Suite and its application in a new class of 26 scientific journals. As a result, the data associated with articles in these journals are machine-accessible from the beginning of the publishing process. Thus, as soon as the paper comes out, the data are automatically added to GBIF.

While BLR is expected to open up millions of scientific illustrations and descriptions, the system is unique in that it makes all the extracted data findable, accessible, interoperable and reusable (FAIR), as well as open to anybody, anywhere, at any time. Most of all, its purpose is to create a novel way to access scientific literature.

To date, BLR has extracted ~350,000 taxonomic treatments and ~200,000 figures from over 38,000 publications. This includes the descriptions of 55,800 new species, 3,744 new genera, and 28 new families. BLR has contributed to the discovery of over 30% of the ~17,000 species described annually.

Prof. Lyubomir Penev, founder and CEO of Pensoft says: "It is such a great satisfaction to see how the development process of the TaxPub standard, started by Plazi some 15 years ago and implemented as a routine publishing workflow at Pensoft's journals in 2010, has now resulted in an entire infrastructure that allows automated extraction and distribution of biodiversity data from various journals across the globe. With the recent announcement from the Consortium of European Taxonomic Facilities (CETAF) that their *European Journal of*

Taxonomy is joining the TaxPub club, we are even more confident that we are paving the right way to fully grasping the dimensions of the world's biodiversity."

Dr. Donat Agosti, co-founder and president of Plazi, adds: "Finally, information technology allows us to create a comprehensive, extended catalogue of life and bring to light this huge corpus of cultural and scientific heritage—the description of life on Earth—for everybody. The nature of taxonomic treatments as a network of citations and syntheses of what scientists have discovered about a species allows us to link distinct fields such as genomics and taxonomy to specimens in natural history museums."

Dr. Tim Smith, Head of Collaboration, Devices and Applications Group at CERN, comments: "Moving the focus away from the papers, where concepts are communicated, to the concepts themselves is a hugely significant step. It enables BLR to offer a unique new interconnected view of the species of our world, where the taxonomic treatments, their provenance, histories and their illustrations are all linked, accessible and findable. This is inspirational for the digital liberation of other fields of study!"

Provided by Pensoft Publishers