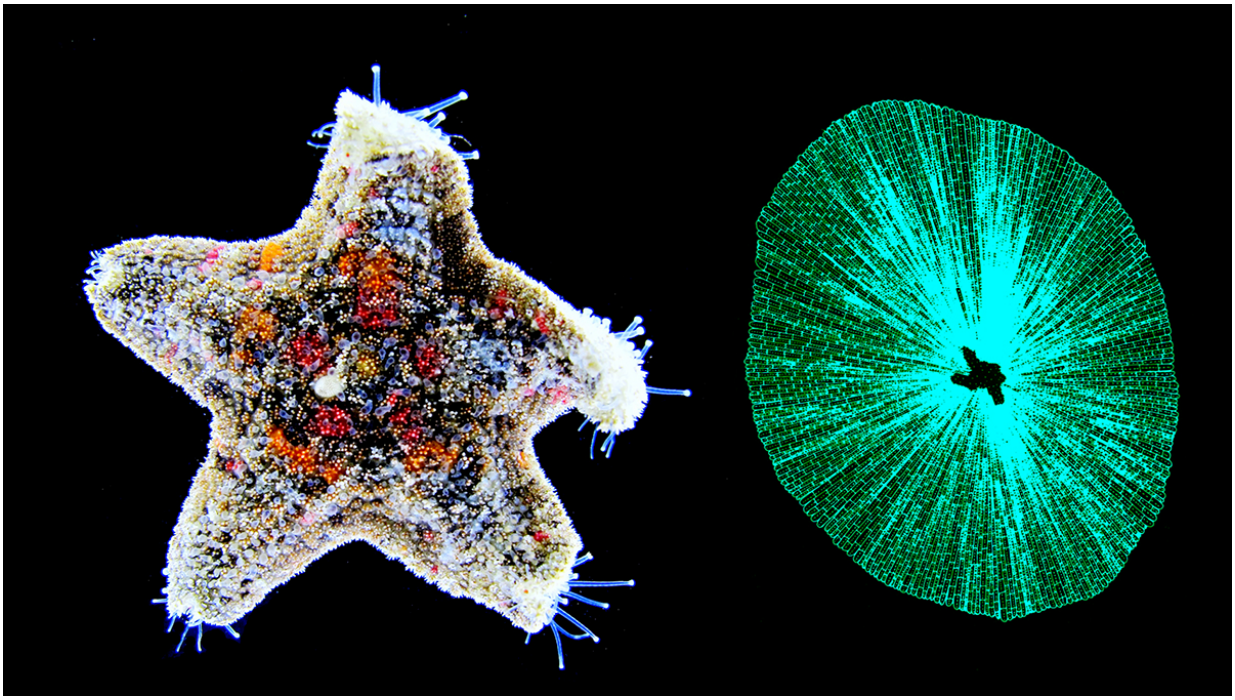
How the names of organisms help to turn 'small data' into 'Big Data'

June 1 2016



The genera of both this starfish (left) and this fungus (right) are called *Asterina*, making them confusing homonyms. Credit: Dr David Patterson

Innovation in 'Big Data' helps address problems that were previously overwhelming. What we know about organisms is in hundreds of millions of pages published over 250 years. New software tools of the Global Names project find scientific names, index digital documents quickly, correcting names and updating them. These advances help

"Making small data big" by linking together to content of many research efforts. The study was published in the open access journal *Biodiversity Data Journal*.

The 'Big Data' vision of science is transformed by computing resources to capture, manage, and interrogate the deluge of information coming from new technologies, infrastructural projects to digitise physical resources (such as our literature from the Biodiversity Heritage Library), or digital versions of specimens and records about specimens by museums.

Increased bandwidth has made dialogue among distributed data centres feasible and this is how new insights into biology are arising. In the case of biodiversity sciences, data centres range in size from the large GenBank for molecular records and the Global Biodiversity Information Facility for records of occurrences of species, to a long tail of tens of thousands of smaller datasets and web-sites which carry information compiled by individuals, research projects, funding agencies, local, state, national and international governmental agencies.

The large biological repositories do not yet approach the scale of astronomy and nuclear physics, but the very large number of sources in the long tail of useful resources do present biodiversity informaticians with a major challenge - how to discover, index, organize and interconnect the information contained in a very large number of locations.

In this regard, biology is fortunate that, from the middle of the 18th Century, the community has accepted the use of latin binomials such as *Homo sapiens* or *Ba humbugi* for species. All names are listed by taxonomists. Name recognition tools can call on large expert compilations of names (Catalogue of Life, Zoobank, Index Fungorum, Global Names Index) to find matches in sources of digital information.

This allows for the rapid indexing of content.

Even when we do not know a name, we can 'discover' it because scientific names have certain distinctive characteristics (written in italics, most often two successive words in a latinised form, with the first one - capitalised). These properties allow names not yet present in compilations of names to be discovered in digital data sources.

The idea of a names-based cyberinfrastructure is to use the names to interconnect large and small distributed sites of expert knowledge distributed across the Internet. This is the concept of the described Global Names project which carried out the work described in this paper.

The effectiveness of such an infrastructure is compromised by the changes to names over time because of taxonomic and phylogenetic research. Names are often misspelled, or there might be errors in the way names are presented. Meanwhile, increasing numbers of species have no names, but are distinguished by their molecular characteristics.

In order to assess the challenge that these problems may present to the realization of a names-based cyberinfrastructure, we compared names from GenBank and DRYAD (a digital data repository) with names from Catalogue of Life to assess how well matched they are.

As a result, we found out that fewer than 15% of the names in pair-wise comparisons of these data sources could be matched. However, with a names parser to break the scientific names into all of their component parts, those parts that present the greatest number of problems could be removed to produce a simplified or canonical version of the name. Thanks to such tools, name-matching was improved to almost 85%, and in some cases to 100%.

The study confirms the potential for the use of names to link distributed data and to make small data big. Nonetheless, it is clear that we need to continue to invest more and better names-management software specially designed to address the problems in the biodiversity sciences.

More information: David Patterson et al, Challenges with using names to link digital biodiversity information, *Biodiversity Data Journal* (2016). DOI: 10.3897/BDJ.4.e8080

Provided by Pensoft Publishers