2-O07-2

Origin of isolates - big but obscure data

Andrey Yurkov, Lorenz Reimer, Sabine Gronow
Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, Germany

Purpose: Our understanding of biodiversity heavily relies on specimens preserved in herbaria and culture collections. Collection catalogues provide information about genetic, biochemical and physiological properties of organisms. Furthermore, the deposited material is usually associated with information about country and substrate of isolation. Description of the substrate of isolation does not follow any standard and varies in terms of complexity of the provided information, language, wording, spelling and heterogeneity of recorded environmental data. This makes any analysis of thousands of records extremely difficult, if possible at all.
Methods: Here we report a novel approach to classify and analyse data of the source of isolation in the DSMZ culture collection. We allowed a record to be described with several keywords, analogous to hashtags used in social media. Each keyword was characterised within an own hierarchical ontology-like structure. Additionally, we extracted host data and included Latin names.
Results: Strains were associated with 1-6 (out of 370) keywords that were used to classify sources of isolation into 8 categories and 61 subcategories. Using the advantage of the ontology-like structure, we were able analyse the linking of strains based on their origin on different hierarchical levels, e.g. Environmental > Aquatic > Marine.
Conclusions: We explored the application of this approach on collections of fungi, yeasts and bacteria in a few institutions and collections. With these results, we provide examples on how advanced classification of substrates can improve the use of data from culture collections and other repositories for research in microbial ecology, medicine and biotechnology.