Integrating semantic isolation habitat and location information in StrainInfo

Author(s) Bert Verslyppe, Wim De Smet, Paul De Vos, Bernard De Baets, Peter Dawyndt

Institution(s) 1. UGent, Ghent University, Krijgslaan 281

Abstract:

StrainInfo (http://www.straininfo.net) is a global catalog of microbial material, building upon the catalogs of Biological Resource Centers (BRCs) by integrating catalog entries of equivalent microbial material. The adoption of Microbiological Common Language (MCL) XML synchronization quickly increased the volume of semantic information in StrainInfo. Semantic information is information corresponding with fine-grained, precisely defined fields such as for example isolation sample location and habitat, oxygen relationship and optimal growth temperature. As the effective data values of the different semantic fields entering StrainInfo are raw textual entries, are of varying detail, can have different forms or languages, and sometimes contain inconsistencies, they need to be converted to a semantic representation based on ontologies. Using a specialized semantic integration algorithm, these values then can be converted to a strain level consensus value for each field. As a case study, the focus was put on the isolation habitat and location fields. These strain level consensus values allow to use ontological knowledge when searching and therefore increase precision and recall compared to full text search. For example, it allows searching for all strains isolated in a given continent or in the neighborhood of a particular place, even if this additional information is not mentioned in the original catalog entries. The Environment Ontology can be used to immediately retrieve all different types of diary products when searching using the general term. In addition, the ontologies enrich the data by providing or linking additional information (e.g. GPS coordinates for geographical locations). The consensus values are made available to end-users by displaying them on the corresponding strain passports. Geographical locations can be visualized on a map. Advanced search functionality is made available to allow users to perform true semantic search based on ontologies. The integration results are also available for electronic processing from the MCL XML exports. As the coverage and the quality of the system improves with the addition of more semantic information, we invite users administering datasets containing semantic information to consider making this information universally available to the complete microbiological community through StrainInfo!

Key words: isolation sample habitat and location, semantic information, strain level integration, StrainInfo