L'ontologie OntoBiotope pour l'étude de la biodiversité microbienne

Claire Nédellec, Estelle Chaix, Robert Bossy, Louise Deléger
Sandra Dérozier, Jean-Baptiste Bohuon, Valentin Loux

MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France
prénom.nom@inra.fr,
http://maiage.jouy.inra.fr/

Résumé. L'intégration des données hétérogènes en Sciences de la Vie est un sujet de recherche majeur. L'importance et le volume considérable des informations sur les milieux de vie des microorganismes dans tous les domaines tels que la santé, l'agriculture ou l'environnement justifie le développement de traitements automatisés. Nous proposons ici l'ontologie OntoBiotope dont nous décrivons les principes de construction ainsi que des exemples d'utilisation pour l'annotation et l'indexation sémantique des habitats microbiens décrits en langue naturelle dans les documents scientifiques.

1 Introduction

La recherche en microbiologie dispose aujourd'hui de très grandes quantités de données sur les habitats des microorganismes en raison de l'expansion des technologies de séquençage à haut-débit et de la croissance du volume des publications et des bases de données. De nombreux domaines de recherche en microbiologie ont l'usage de cette information, dont, en premier lieu, l'étude de la diversité microbienne. L'expression en langue naturelle de l'information sur les habitats microbiens est un frein majeur à son exploitation. Il est très fréquent que des habitats similaires soient décrits par des termes différents, ce qui rend difficile leur comparaison automatique. (Ivanova et al., 2010) souligne l'importance de la construction d'un référentiel commun pour standardiser les descriptions de ces habitats, nous proposons ici un tel référentiel sous la forme d'une ontologie, appelée OntoBiotope.

2 Contexte et motivation

Tous les domaines de la microbiologie produisent des descriptions d'habitat, en premier lieu sous forme d'articles – près de 7 millions d'habitats de microorganismes sont mentionnés dans PubMed selon Deléger et al. (2016). Les bases de données de ressources biologiques comportent toujours un champ «isolation», plus ou moins structuré et détaillé qui décrit le site où l'échantillon a été prélevé, comme BacDive, the *Bacterial Diversity Metadatabase* de DSMZ (`https://bacdive.dsmz.de`). Plus récemment, l'utilisation des technologies de