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

séquençage à haut-débit génère un très grand nombre de séquences de microorganismes associées ici encore à leur lieu d'isolation et disponibles publiquement dans des bases de données comme GenBank.

Parallèlement, l'abondance de ces descriptions favorise l'émergence en biologie de questions transversales aux différents milieux, telles que les questions relatives à la provenance des organismes et aux parcours de contamination, à l'adaptation des microorganismes à différents milieux en lien avec des questions évolutives et génétiques. L'exploitation d'un tel volume de données requiert l'utilisation de méthodes automatiques. Les descriptions des milieux d'échantillonnage microbiens restent largement sous-exploitées par manque de solutions automatisées. La raison est double. L'analyse à grande échelle des descriptions des milieux de vie des microorganismes requiert (1) une classification de référence aux catégories de laquelle attacher les descriptions et (2) un moyen automatique d'associer les descriptions des habitats à ces catégories. Nous proposons dans cet article une solution qui répond à ces deux objectifs.

Le deuxième objectif relève de la fouille de texte pour extraire finement les informations, les catégoriser et les relier. Les progrès récents des méthodes permettent d'atteindre des performances qui les rendent aujourd'hui exploitables pour ce type de tâche mesurées par des compétitions internationales comme *BioNLP Shared Task Bacteria Biotope* (Bossy et al., 2015).

Le premier objectif relatif à la disponibilité d'une classification de référence, est un point critique. Pour être utilisable, elle doit répondre à plusieurs critères. Elle doit être suffisamment riche pour rendre compte de la grande diversité des habitats et permettre ainsi de distinguer des habitats dont les propriétés physico-chimiques diffèrent, mais sans être trop vaste, ce qui nuirait à sa maintenance et à son utilisation manuelle. Sa structure doit à la fois refléter les domaines d'études de biodiversité microbienne pour faciliter son appropriation par les utilisateurs microbiologistes, mais également regrouper les milieux très similaires de manière à en faciliter les traitements. Son organisation doit être hiérarchique pour permettre son utilisation à différents niveaux de précision.

Les classifications d'habitats de microorganismes sont peu nombreuses et ne répondent pas à ces critères. Par exemple, la classification ATCC (Floyd et al., 2005) est une liste de 37 entrées d'habitat environnemental, insuffisante de par sa petite taille et sa structure à plat. GOLD (*Genome OnLine Database*) utilise un vocabulaire contrôlé plus riche, mais non hiérarchisé pour indexer l'information d'isolation des échantillons biologiques (Reddy et al., 2014).

EnvO (*Environment Ontology project* : `https://bioportal.bioontology.org/ontologies/ENVO`) est une ontologie hiérarchique de 7000 classes, soutenue par le *Genomics Standards Consortium* (GSC) destinée à l'annotation manuelle des environnements des organismes et des échantillons biologiques (Buttigieg et al., 2013), mais elle souffre de limitations pour la description des habitats d'organismes microscopiques. Ratkovic et al. (2012) ainsi que Cook et al. (2016) ont montré qu'EnvO n'était pas bien adaptée à l'extraction d'information en microbiologie. Le développement d'EnvO repose sur la réutilisation de classifications connues qui ont été conçues pour d'autres objectifs. La conséquence en est qu'elles ne sont généralement pas adaptées à la description des habitats microbiens. Elle sous-représente certains domaines importants en recherche microbienne comme la transformation des aliments. La classe des aliments réutilise FoodON, *the United Nations Food classification* où par exemple les fromages sont distingués par leur couleur (*red marbled, white*) ou leur présentation (*sliced, dip*). La transformation (cuisson, lavage) ou l'animal dont le lait est utilisé (vache, brebis) sont des concepts plus pertinents pour l'étude écologique. Un autre exemple est la classe des sols

issue de la classification pédologique de *Agriculture Organization soil classification*. Elle n'est pas structurée selon les propriétés principales des sols comme l'acidité ou l'humidité qui sont critiques pour les microorganismes.

L'absence d'ontologie adaptée à la catégorisation automatique des descriptions d'habitats microbiens en microbiologie a motivé la construction de l'ontologie OntoBiotope.

3 Construction et principes de l'ontologie OntoBiotope

La partie Habitat qui fait l'objet de cet article est la partie principale de l'ontologie OntoBiotope. Elle respecte les critères définis ci-dessus. L'approche suivie pour sa construction est assistée par des outils automatiques et des outils d'édition. Une attention particulière est apportée à la terminologie pour permettre son usage à des fins d'extraction fine de l'information.

La construction s'est faite de manière ascendante et descendante. L'approche descendante structure successivement la classification en fonction des grands domaines d'étude de la microbiologie et de leurs subdivisions successives pour faciliter son appropriation par les microbiologistes. L'approche ascendante part de l'ensemble des termes particuliers qui dénotent des habitats pour les regrouper itérativement et hiérarchiquement. Ces termes ont été extraits automatiquement du champ «Habitat» de la base de données GOLD et du champ «Source» de la base de données GenBank par l'extracteur de terme BioYaTeA (Golik et al., 2013) intégré dans la suite Alvis (Ba et Bossy, 2016). L'analyse terminologique manuelle des termes extraits a été assistée par l'outil TyDI (*Terminology Design Interface*) suivant la méthode décrite par Nédellec et al. (2010). La formalisation en sous-arbres assure que les propriétés d'une classe sont partagées par toutes les sous-classes, afin de garantir que les informations indexées par une classe particulière pourront être retrouvées par l'interrogation par des classes parentes. OntoBiotope Habitat ne décrit pas les lieux géographiques, d'autres classifications pertinentes comme la base GeoNames leur sont dédiées. Le sous-arbre *Food* a été construit en adaptant FoodEx2, la nouvelle classification des aliments de l'EFSA (l'Autorité européenne de sécurité des aliments) et grâce à l'expertise des microbiologistes du projet Florilège sur la flore positive des aliments (Falentin et al., 2017).

Le format initial choisi est le format *Open Biomedical Ontologies* (OBO) développé par OBO Foundry dont l'éditeur Obo-Edit présentait au début du projet en 2010 de bonne propriétés d'utilisabilité pour des non-spécialistes. L'expressivité du format OBO est adaptée aux besoins du projet permettant la représentation du niveau lexical (synonymes) et du niveau conceptuel (classes et relations). Trois types de synonymes sont considérés, les synonymes exacts (*exact synonym*) comme les acronymes (*perchloroethylene contaminated site / PCE contaminated site*) ou les variations typographiques, les synonymes proches (*close synonym*) (*polluted site / contaminated site*) et les synonymes associés (*related synonym*) (*PCP percolated soil / PCP contaminated soil*).

OntoBiotope Habitat est distribuée par le portail d'ontologie AgroPortal (`http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE`), sous la licence *Creative Commons with Attribution* (CC-BY). La version publique ne contient pas d'autre relation que la relation hiérarchique. Elle est explorable en ligne et téléchargeable aux formats standards, OBO (format d'origine) et traduit en RDF/XML. Elle contient 2320 classes et 492 synonymes, elle est organisée dans une hiérarchie d'une profondeur de 13. La racine se divise en 11 grands domaines (figure 1).

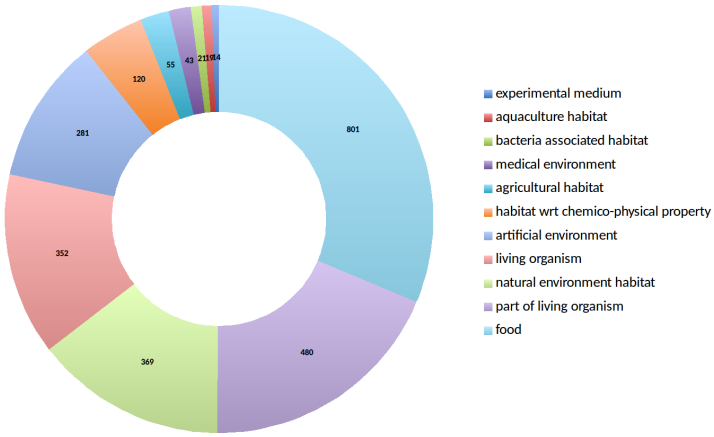

FIG. 1 – *Distribution des classes Habitat dans les branches de l'ontologie OntoBiotope.*

4 Exemples d'utilisation

Nous avons proposé Ontobiotope Habitat comme ontologie de référence pour catégoriser les habitats microbiens et les relier aux bactéries par les trois dernières éditions de la compétition internationale BioNLP Shared Task Bacteria Biotope (Bossy et al., 2015), organisées avec l'objectif de stimuler la recherche en extraction d'information dans le domaine de l'écologie microbienne. Ontobiotope a démontré son utilité pour annoter automatiquement des textes par les différents algorithmes participant aux compétitions.

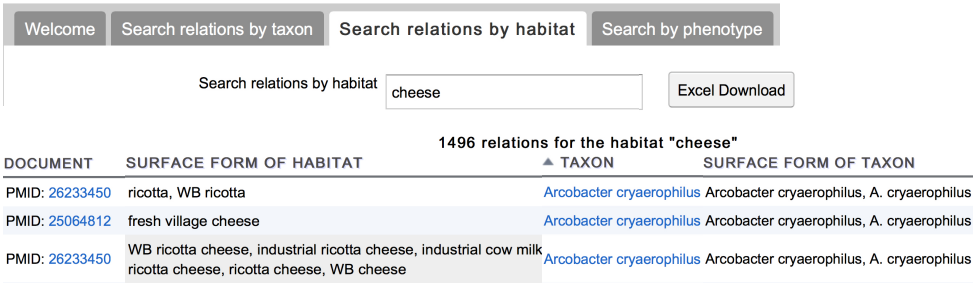

FIG. 2 – *Exemple de l'habitat "cheese" dans la base de données Florilège.*

Nous avons développé une telle application en utilisant la Suite Alvis de *text-mining* pour analyser et catégoriser à grande échelle les habitats de microorganismes. Grâce à la catégorisation, les données du texte deviennent comparables et peuvent être croisées avec d'autres données, en particulier les données génétiques. Le projet Florilège en est un exemple, il a pour objectif la mise à disposition de données relatives à la flore positive des aliments, espèces, habitats, phénotypes, usages, *etc.*, dans une base de données structurée (Falentin et al., 2017). Elle intègre les informations de l'ensemble des résumés de la base bibliographique PubMed en

microbiologie des aliments, analysée par la suite Alvis. L'ontologie OntoBiotope est utilisée pour indexer les données d'habitats. Les taxa sont normalisés par la taxonomie de référence du NCBI. La figure 2 donne un exemple de l'interface d'interrogation : le taxon *Arcobacter cryaerophilus* et ses habitats dont *fresh village cheese* par exemple ont été identifiés dans des textes, catégorisés et reliés par la relation *lives_in*. *fresh village cheese* est normalisé par *cheese* qui est l'objet de la requête. La Suite Alvis et l'application sont en cours de déploiement sur l'infrastructure européenne OpenMinTeD, ce qui la rendra publiquement utilisable (Ba et Bossy, 2016) sur différents corpus. A terme, l'application Florilège rendra accessible dans une interface unifiée les données de microbiologie de la littérature et des sources de données expertisées comme la collection allemande de bactéries DSMZ et le Centre International de Ressources Microbiennes (CIRM) dédié aux bactéries de l'INRA. Toutes les données extraites des textes sont également accessibles par le moteur de recherche sémantique AlvisIR (`http://bibliome.jouy.inra.fr/demo/ontobiotope/alvisir2/webapi/search`). L'interface du moteur comme celle la base de données Florilège (`http://genome.jouy.inra.fr/Florilege`) permettent d'exprimer des requêtes pour une recherche hiérarchique et relationnelle (quel organisme vit où ?), c'est-à-dire que les résultats contiennent les relations correspondant à la classe recherchée, ainsi que les relations faisant intervenir les entités associées aux classes plus spécifiques suivant la hiérarchie de l'ontologie.

5 Perspectives

Les branches d'OntoBiotope sont à différents stades de développement en fonction des ontologies réutilisables répondant aux besoinx et des collaborations avec des experts. La partie être vivants et anatomie est ainsi une des parties les plus riches en nombre de concepts avec 51 classes dans la branche *gastrointestinal part*. Sa structuration nécessite une réflexion approfondie. Il ne serait pas judicieux de reprendre en l'état les classifications anatomiques médicales qui sont regroupées en grands systèmes ou par fonction, plutôt que structurées en fonction des propriétés physico-chimiques des milieux. Par contre, le lien vers les concepts d'anatomie médicale devra être préservé de façon à permettre l'intégration de données annotées par ces différentes ressources. Plusieurs sous-arbres complémentaires des habitats, notamment les propriétés des habitats, les phénotypes microbiens et leurs usages technologiques complètent OntoBiotope et seront publiés prochainement.

6 Remerciements

Ce travail a été financé par le programme Quaero d'Oséo, par le métaprogramme INRA MEM et par le projet européen H2020 OpenMinTeD (EC/H2020-EINFRA 654021).

Références

Ba, M. et R. Bossy (2016). Interoperability of corpus processing workflow engines : the case of AlvisNLP/ML in OpenMinTeD. In *INTEROP 2016, LREC*, Portoroz, Slovenia.

Bossy, R., W. Golik, Z. Ratkovic, D. Valsamou, P. Bessieres, et C. Nédellec (2015). Overview of the Gene Regulation Network and the Bacteria Biotope tasks in BioNLP'13 shared task. *BMC bioinformatics 16*(10), S1.

Buttigieg, P. L., N. Morrison, B. Smith, C. J. Mungall, et S. E. Lewis (2013). The environment ontology : contextualising biological and biomedical entities. *Journal of biomedical semantics 4*(1), 43.

Cook, H. V., E. Pafilis, et L. J. Jensen (2016). A dictionary-and rule-based system for identification of bacteria and habitats in text. *ACL 2016 50*.

Deléger, L., R. Bossy, E. Chaix, M. Ba, A. Ferré, P. Bessieres, et C. Nédellec (2016). Overview of the Bacteria Biotope task at BioNLP Shared Task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pp. 12–22.

Falentin, H., E. Chaix, S. Derozier, M. Weber, S. Buchin, B. Dridi, S.-M. Deutsch, F. Valence-Bertel, S. Casaregola, P. Renault, et al. (2017). Florilege : a database gathering microbial phenotypes of food interest. In *4. International Conference on Microbial Diversity 2017*.

Floyd, M. M., J. Tang, M. Kane, et D. Emerson (2005). Captured diversity in a culture collection : case study of the geographic and habitat distributions of environmental isolates held at the American Type Culture Collection. *Applied and Environmental Microbiology 71*(6), 2813–2823.

Golik, W., R. Bossy, Z. Ratkovic, et C. Nédellec (2013). Improving term extraction with linguistic analysis in the biomedical domain. *Research in Computing Science 70*, 157–172.

Ivanova, N., S. G. Tringe, K. Liolios, W.-T. Liu, N. Morrison, P. Hugenholtz, et N. C. Kyrpides (2010). A call for standardized classification of metagenome projects. *Environmental microbiology 12*(7), 1803–1805.

Nédellec, C., W. Golik, S. Aubin, et R. Bossy (2010). Building large lexicalized ontologies from text : a use case in automatic indexing of biotechnology patents. *Knowledge Engineering and Management by the Masses*, 514–523.

Ratkovic, Z., W. Golik, et P. Warnier (2012). Event extraction of Bacteria Biotopes : a knowledge-intensive NLP-based approach. *BMC bioinformatics 13*(11), S8.

Reddy, T. B., A. D. Thomas, D. Stamatis, J. Bertsch, M. Isbandi, J. Jansson, J. Mallajosyula, I. Pagani, E. A. Lobos, et N. C. Kyrpides (2014). The Genomes OnLine Database (GOLD) v. 5 : a metadata management system based on a four level (meta) genome project classification. *Nucleic acids research 43*(D1), D1099–D1106.

Summary

The integration of heterogeneous data is a major research challenge in Life Sciences. The importance and the volume of information on the environments of microorganisms in all areas such as health, agriculture or environment has prompted the development of automated processes. We describe the design principles of the OntoBiotope ontology, and its use for the automatic annotation and indexing of microbial habitats described in natural language in scientific documents.