**FICHE TYPE DE RECUEIL DES FAITS MARQUANTS 2016 DES DEPARTEMENTS/CENTRES**

**Titre du fait marquant : OntoBiotope, une révolution numérique en marche pour l'étude des habitats microbiens**

**Catégorie:** ***Publications, colloques, logiciels, base de données et méthodes***
**(Publication –en indiquant le doi) , Colloque, Partenariat, ESCO, Prospectives, Etudes, Brevets, Lancements/Inaugurations, Autres : précisez)**

**Contact : Claire Nédellec**
**Unité : MaIAGE**
**Département : MIA et MICA**
**Centre INRA de Recherche : Jouy-en-Josas**

**Axe du document d'orientation 2015-2025** : Approches prédictives pour la biologie
**Axe du tripode** : tous
**Domaine d'activité**: Mathématique - modélisation - informatique
**Méta-programme (si adapté)**: MEM
**Mots-clés (rubrique libre)** : habitats microbiens, extraction d'information à partir de textes, ontologies, compétition internationale.

**Résumé (5 lignes)** :
L'équipe Bibliome a produit un ensemble de résultats pour l'étude des habitats des microorganismes à partir des publications scientifiques en s'appuyant sur le réseau MEM OntoBiotope et les résultats de l'organisation de trois compétitions internationales. Les résultats font l'objet de services publics en ligne destinés aux biologistes et notamment les microbiologistes. Ces services exploitent un ensemble de méthodes et logiciels, une ontologie et une base de connaissances dont il n'existe pas d'équivalent au monde.

**Contexte et enjeux** :
L'information d'habitat des organismes est critique dans tous les domaines de la microbiologie, notamment pour l'analyse systématique des écosystèmes, la production des hypothèses de provenance des microorganismes isolés ou encore l'étude du rôle des gènes dans l'adaptation au milieu. Les habitats des microorganismes sont décrits en langue naturelle non structurée dans des millions de sources publiques, articles scientifiques ou bases de données (séquences, ressources biologiques). Pour être exploitable, cette information doit être extraite et reliée à une classification de référence qui permette de croiser les informations provenant de sources diverses. Les ontologies sont la réponse reconnue au besoin des chercheurs de standardiser ces informations et de les interpréter. Le développement de méthodes informatiques d'extraction d'information à partir de texte à l'aide d'ontologie connaît des succès récents et spectaculaires dont nous sommes partie prenante. Un des moteurs en est l'organisation de compétitions internationales sur des corpus de référence qui encouragent le développement de méthodes génériques rapidement adaptables, leur comparaison et leur partage.

**Résultats** :
Nous avons développé l'ontologie OntoBiotope pour la description des habitats et phénotypes microbiens pour tous les écosystèmes, alimentaire, hôtes animal, plante et homme, milieux aquatiques, traitement des effluents, *etc*. qui intéressent potentiellement tous les microbiologistes de l'INRA [1]. Nous avons conçu un « workflow » de traitement automatique avec la suite logicielle Alvis [2] ; celle-ci extrait les taxa et les habitats de toute collection de textes en anglais, leur assigne une catégorie, respectivement taxinomique et d'OntoBiotope et extrait la relation entre habitat et taxon [3]. Plusieurs services en ligne permettent d'interroger la base de connaissances (3,63 millions de relations) générée automatiquement à partir de l'ensemble des 1,16 millions de références

Pubmed du domaine : un moteur de recherche sémantique [4] et un « treemap» [5] pour un affichage synthétique des informations. Pour l'extraction d'information, nous avons développé un ensemble de méthodes originales basées sur l'apprentissage et le traitement automatiques de la langue (TAL) et que nous avons adaptées, par exemple : ToMap [6] pour les termes et catégories et AlvisRE [7, 8] pour les relations. Nous avons entraîné et évalué ces méthodes grâce à des corpus de référence d'articles scientifiques [9] et de centres de séquençage [10, 11] annotés manuellement en utilisant l'application web AlvisAE [12]. Enfin, nous avons organisé trois compétitions internationales intitulées Bacteria Biotope [13, 14, 15], trois ateliers de restitution associés à ACL (Association for Computational Linguistics), la principale conférence internationale de TAL [16, 17, 18] et édité trois suppléments du journal BMC Bioinformatics [19, 20, 21].

**Perspectives** :
L'autorisation légale à court terme de l'accès aux articles complets pour la fouille de texte va permettre d'accroître considérablement l'application de nos résultats et la portée de nos recherches dans ce domaine. Nous collaborons avec la DIST (Délégation Information Scientifique et Technique) pour identifier et intégrer les sources les plus pertinentes.
Les méthodes sont en cours d'extension pour l'extraction et la formalisation des propriétés et des phénotypes des microorganismes, y compris les molécules dégradées et produites. La première étape est le déploiement de bases de données publiques sur la plateforme IFB-Migale : Florilège action MEM sur la flore positive des aliments et FoodMicrobiome sur le metagénome du fromage avec l'institut INRA Micalis et le CNIEL (Centre National Interprofessionnel de l'Economie Laitière). Elles intègreront des informations textuelles d'habitats des bases de données du domaine (dont CIRM : Centre International de Ressources Microbiennes, GOLD : Genomes Online Database, et GenBank). L'objectif sera de favoriser, par le croisement d'informations génétiques et de milieu, l'analyse et l'interprétation des expériences en microbiologie produisant des données à haut débit.

**Valorisation** : la Suite Alvis et son instance OntoBiotope font l'objet de valorisation sur l'infrastructure H2020 OpenMinTeD (OMTD) de "text-mining" à laquelle nous contribuons [22]. OMTD apporte une visibilité internationale et rend les utilisateurs (micro)biologistes autonomes dans l'application de ces technologies aux collections de documents, aux milieux et aux souches de leur choix. OMTD facilitera l'extension du "workflow" logiciel OntoBiotope par l'ajout de nouvelles fonctions basées sur de nouveaux composants, comme l'extraction des molécules.
L'ontologie OntoBiotope est distribuée publiquement et son utilisation par d'autres infrastructures pour l'indexation d'habitats microbiens est en cours d'étude notamment pour GOLD et GBIF France (Global Biodiversity Information Facility). Les corpus de référence des compétitions « Bacteria Biotope » et les modules automatique d'évaluation, accessibles en ligne, rencontrent un succès croissant (68 citations dans Google Scholar).

**Financements** : réseau MEM (INRA), projets Alvis (FP6), Quaero (BPI), OpenMinTeD (H2020), thèses IDI IDEX Paris-Saclay.

**Références bibliographiques** :
[1] Ontologie OntoBiotope (partie Habitats) sur AgroPortal :
http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE/?p=classes&conceptid=root
[2] Ba M. and Bossy R. Interoperability of corpus processing workflow engines: the case of AlvisNLP/ML in OpenMinTeD. *LREC Workshop on Cross-Platform Text Mining and Natural Language*, Portoroz May 2016.
[3] Ratkovic Z, Golik W, Warnier P (2012). Event extraction of bacteria biotopes: a knowledge-intensive NLP-based approach. *BMC Bioinformatics*, 13(Suppl 11):S8. 10.1186/1471-2105-13-S11-S8
[4] AlvisIR Pubmed OntoBiotope :
http://bibliome.jouy.inra.fr/demo/ontobiotope/alvisir2/webapi/search?
[5] Treemap Pubmed OntoBiotope : http://bibliome.jouy.inra.fr/demo/alvisdb/obt/browse2
[6] Golik W., Warnier P., Nédellec C.. Corpus-based extension of termino-ontology by linguistic analysis: a use case in biomedical event extraction. Ontology and Lexicon: new insights. *Actes du*

*workshop TIA 2011 : 9th International Conference on Terminology and Artificial Intelligence*, M. Slodzian et al., (eds), Paris, novembre 2011.

[7] Ratkovic Z. *Analyse prédicative pour l'extraction d'information : application au domaine de la biologie*. Thèse de doctorat en Sciences du Langage. Encadrée par T. Poibeau et C. Nédellec. ED 268 Langage et langues, Université Sorbonne Nouvelle - Paris III, 11 décembre 2014.

[8] Valsamou D., *Information Extraction for the Seed Development Regulatory Networks of* Arabidopsis Thaliana. Encadrée par P. Zweigenbaum et C. Nédellec, thèse de doctorat en Informatique, ED STIC, Université Paris-Sud, soutenance prévue en janvier 2107.

[9] Corpus Bacteria Biotope of BioNLP-ST'11 : https://sites.google.com/site/bionlpst/home/bacteria-biotopes

[10] Corpus Bacteria Biotope of BioNLP-ST'13 : http://2013.bionlp-st.org/tasks/bacteria-biotopes

[11] Corpus Bacteria Biotope of BioNLP-ST'16 : http://2013.bionlp-st.org/tasks/bacteria-biotopes

[12] Papazian F., Bossy R., Nédellec C. (2012). AlvisAE: a collaborative Web text annotation editor for knowledge acquisition. *Proc. Sixth Linguistic Annotation Workshop (Association for Computational Linguistics)*, 149-152.

[13] Bossy R., Jourde J., Manine A.-P., Veber P., Alphonse E., van de Guchte M., Bessières P., Nédellec C. (2012). BioNLP Shared Task – The Bacteria Track. *BMC Bioinformatics*, 13(Suppl 11):S3. doi: 10.1186/1471-2105-13- S11-S3

[14] Bossy R., Golik W., Ratkovic Z., Valsamou D., Bessières P., Nédellec C.. Overview of the Gene Regulation Network and the Bacteria Biotope Tasks in BioNLP'13 shared task. *BMC Bioinformatics*, 16(Suppl 10):S1, 2015. doi:10.1186/1471-2105-16-S10-S1.

[15] Deléger L., Bossy R., Chaix E., Ba M., Ferré A., Bessières P., Nédellec C., Overview of the Bacteria Biotope Task at BioNLP Shared Task. In *Proceedings of the BioNLP Shared Task 2016 Workshop*, Association for Computational Linguistics, Berlin, Allemagne 2016.

[16] Proceedings of the BioNLP Shared Task 2011 workshop, joint with The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, USA, 2011. http://aclweb.org/anthology/W/W11/W11-18.pdf

[17] Proceedings of the BioNLP Shared Task 2013 workshop, joint with The 51th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Sofia, Bulgaria 2013. https://aclweb.org/anthology/W/W13/W13-2000.pdf

[18] Proceedings of the 4th BioNLP Shared Task 2016 workshop, joint with The 54th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portoroz, Slovénie, 2016. https://aclweb.org/anthology/W/W16/W16-3000.pdf

[19] Jin-Dong Kim, Sampo Pyysalo, Claire Nédellec, Sophia Ananiadou and Jun'ichi Tsujii. Selected articles from the BioNLP Shared Task 2011. *BMC Bioinformatics* 13(Suppl 11):S3, juin 2012.

[20] Claire Nédellec, Jin-Dong Kim, Sampo Pyysalo, Sophia Ananiadou, Pierre Zweigenbaum. BioNLP Shared Task 2013: Part 2. *BMC Bioinformatics*, Vol 16 Suppl 16, 2015.

[21] Claire Nédellec, Jin-Dong Kim, Sampo Pyysalo, Sophia Ananiadou, Pierre Zweigenbaum. BioNLP Shared Task 2013: Part 1. *BMC Bioinformatics*, Vol 16 Suppl 10, 2015.

[22] Przybyła P., Shardlow M., Aubin S., Bossy R., Eckart de Castilho R., Piperidis S., McNaught J., and Ananiadou S., Text mining resources for the life science, *Database* 2016: baw145 doi:10.1093/database/baw145 published online November 25, 2016.

**Annexe à la fiche type de recueil des faits marquants 2016 des départements/centres**

**CLASSIFICATION**

**Axes du document d'orientation**

- Intégration des performances économiques, sociales et environnementales de l'agriculture
- Développement de systèmes alimentaires sains et durables
- Atténuation de l'effet de serre et adaptation de l'agriculture et de la forêt au changement climatique
- Valorisation de la biomasse pour la chimie et l'énergie
- Sécurité alimentaire mondiale et changements globaux
- Approches prédictives pour la biologie
- Agro-écologie

**Tripode**

- Alimentation
- Agriculture
- Environnement

**Domaine d'activités**

- Animaux
- Végétaux
- Micro-organismes
- Procédés agro-industriels
- Mathématique - modélisation - informatique

**Méta-programmes**

- SMACH
- M2E-MEM
- GISA
- SELGEN
- DID'IT
- ACCAF
- EcoServ
- Glofoods