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

Titre du fait marquant : Florilège, un service d'agrégation de connaissance en microbiologie des aliments

Catégorie: ***Publications, colloques, logiciels, base de données et méthodes***
Autres : service

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

Priorité du Document d'Orientation: OpenScience, #OpenScience-1, 2, 3 et 4

- #OpenScience-1 : Des infrastructures de recherche connectées
- #OpenScience-2 : Une organisation des données pour le partage et la réutilisation
- #OpenScience-3 : Des approches prédictives en biologie
- #OpenScience-4 : De nouveaux modes de diffusion de la connaissance

Méta-programme (si adapté): MEM
Mots-clés (rubrique libre) : microbiologie des aliments, phénotypes microbiens, intégration de données, extraction d'information à partir de textes, text mining, ontologies

Résumé (5 lignes) :
Les équipes Bibliome et Migale de l'unité MaIAGE en collaboration avec le GT Food du metaprogramme MEM ont développé le service en ligne Florilège destiné aux microbiologistes pour l'étude des phénotypes et habitats microbiens. Florilège donne accès sur la platforme Migale à un ensemble d'informations extraites, structurées, agrégées et normalisées à partir de sources externes : publications scientifiques, bases de données génétiques et centres de ressource biologiques. L'analyse sémantique de ces sources est réalisée à grande échelle par des algorithmes d'IA et de fouille de texte de l'équipe Bibliome.

Contexte et enjeux :
Dans le domaine de l'alimentation, la connaissance des informations sur les milieux de croissance des microbes est critique, qu'elle porte sur la flore positive pour la conception et le contrôle d'aliments innovants, la production de nutriments ou la biopréservation, ou sur la flore pathogène, pour contrôler la dissémination tout au long de la chaîne alimentaire. Malgré leur intérêt, les informations d'habitats et de phénotypes de ces microorganismes ne faisaient jusqu'ici l'objet d'aucun recensement systématique parce qu'elles sont exprimées en langue naturelle non structurée dans des millions de sources publiques, articles scientifiques ou bases de données. Leur exploitation automatique requiert d'identifier et de télécharger automatiquement les données, d'en traiter les formats hétérogènes et évolutifs, d'en analyser le contenu et de le formaliser grâce à une ontologie de référence qui permet de catégoriser et de relier les informations.

Résultats : les équipes Migale et Bibliome avec le GT Food MEM ont développé Florilège. Elle constitue la source en ligne la plus riche d'information structurées sur les microorganismes, leurs habitats et leur phénotypes. Le service de Florilège permet d'interroger les relations entre Taxa et Habitats (820,000 relations) et entre Taxa et Phenotype (86,000 relations) à tous les niveaux de généralité, du plus précis au plus général, à l'aide de la taxinomie de référence du NCBI[1] pour les taxa et de l'ontologie OntoBiotope[2] pour les habitats et phénotypes (>3000 catégories). Elle présente l'originalité de tenir la majorité des informations de textes : articles scientifiques (2,3 millions

1 URL
2 http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE/?p=classes&conceptid=root

d'entrées), champs textuels des bases de données génétiques (GenBank) et collections microbiennes de référence (BacDive DSMZ, CIRM (Centre International de Ressources Microbiennes) INRA), traitées par les algorithmes originaux d'analyse sémantique profonde de l'équipe Bibliome. Ils réalisent l'identification dans les textes de termes complexes, l'assignation à des catégories d'ontologie et de nomenclatures et la mise en relation. Ils sont intégrées dans un *workflow*, déployé et accessible publiquement sur la plateforme européenne de text mining, OpenMinTeD.
Les millions d'informations hiérarchisées ainsi produites sont interrogeables dans Florilège par l'utilisateur microbiologiste par des requêtes d'une grande expressivité combinant catégories et de relations, adaptées aux besoins récurrents, mais très divers des microbiologistes : listes des microorganismes connus pour croître dans un milieu donné (projet CNIEL Food Microbiome Transfert), provenance de microorganismes identifiés dans des échantillons, conception de nouveaux produits alimentaires fermentés (projet MEM ENovFood), etc.

Perspectives :
A court terme nous intègrerons dans le traitement sémantique la nouvelle méthode CONTES d'apprentissage développée dans l'équipe Bibliome pour l'assignation automatique de catégories aux aliments avec peu d'exemples d'entraînement [REF]. Les perspectives en recherche en microbiologie et en bioinformatique sont nombreuses pour l'analyse des écosystèmes, la production des hypothèses de provenance des microorganismes isolés ou encore l'étude du rôle des gènes dans l'adaptation au milieu. Notre publication dans Food Microbiology illustre sur plusieurs exemples frappants ce potentiel [REF].
Florilège sera aisément étendu à de nouvelles sources grâce à son approche générique, dont d'autres bases des CIRM Inra dans le projet ENovFood ou la base GOLD : Genomes Online Database au JGI. Notre 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, dont métagénomique.
La généralité de nos services d'analyse sémantique et l'ouverture des données permettra à court terme leur exploitation sur d'autres types de documents y compris non scientifiques, par exemple, notices de retrait de produit alimentaire, bulletins d'alerte dans une perspective de surveillance sanitaire, facilitée par notre intégration de la classification FoodEx2 de l'EFSA dans OntoBiotope [REF].

Valorisation :
Les deux équipes et le GT Food du programme MEM ont été invités à de très nombreuses reprises à présenter ces résultats à des publics nationaux et internationaux [REF].
Les services d'analyse sémantique s'exécutent sur OpenMinTeD et les services de Florilège s'exécutent sur la plateforme Migale et les données produites sont des formats standards (XML, RDF). Ils sont accessibles par des interfaces de programmation (API) qui permettent très simplement et efficacement leur intégration avec d'autres services dans des infrastructures de données et de service. Nous étudions dans le programme EOSC Food Cloud au niveau européen et dans le projet Visa TM au niveau français comment étendre les services avec d'autres partenaires et diversifier les applications microbiologiques de ces résultats.

Financements : réseau MEM (INRA), OpenMinTeD (H2020), thèses IDI IDEX Paris-Saclay, projet CoSO Visa TM.

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

[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.