FICHE DE RECUEIL DES FAITS MARQUANTS DES DEPARTEMENTS/CENTRES

Année concernée : 2017

Fiche envoyée par : *Département MIA*

Titre du fait marquant : OpenMinTeD, une plateforme européenne pour des services de fouille de textes

Catégorie: Autres : service, e-infrastructure scientifique

Contact : Claire Nédellec

Unité : MaIAGE

Département : MIA

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 (cas d'usage)

Mots-clés (rubrique libre) : e-infrastructure, interopérabilité, réutilisation, fouille de textes, service en ligne, bibliothèque numérique, intégration des données, ontologie, phénotypage, régulation génique, écologie microbienne.

Résumé (5 lignes) :
L'équipe Bibliome en collaboration avec la DIST a contribué au développement de l'e-infrastructure européenne OpenMinTeD. OpenMinTeD facilite l'appropriation des technologies de *text mining* pour la recherche scientifique. Elle s'appuie sur des outils et plates-formes de *text mining* existants et les rend identifiables automatiquement grâce à des catalogues, et interopérables grâce à des standards. Des services développés en sciences de la vie, en alimentation et en agriculture démontrent les mérites de l'approche.

Contexte et enjeux :
L'abondante production académique textuelle est une source de données accessible aux outils modernes de recherche bibliographique et de fouille de documents. Pour être exploitée conjointement avec des informations provenant de sources expérimentales ou analytiques, l'information des textes doit être extraite et formalisée à l'aide d'ontologies. Le développement de méthodes informatiques d'extraction d'information à partir de textes à l'aide d'ontologie connaît des progrès récents et des succès spectaculaires. En dépit des besoins considérables, leur exploitation

par différentes communautés scientifiques, IST et chercheurs reste faible. La capacité à développer rapidement de nouvelles applications à partir des composants existants est identifié comme le chaînon manquant. Différentes équipes européennes ont développé des plateformes de conception et de réutilisation d'extraction d'information à partir de textes, dont l'équipe Bibliome avec *Alvis*. Ces équipes ont mis en commun leurs développements et leur expertise dans le projet d'e-infrastructure H2020 OpenMinTeD (2015-2018).

Résultats :
L'e-infrastructure européenne de *text-mining* OpenMinTeD offre à différents types d'utilisateurs les moyens d'exploiter des services de manière autonome dans un cadre de mutualisation ouvert, unifié et juridiquement sûr. Les chercheurs du domaine peuvent y valoriser leurs méthodes et réutiliser des composants à des fins expérimentales. Les développeurs d'applications et de services y composent leurs traitements (*workflows*) en sélectionnant les collections documentaires, les ressources sémantiques et les composants de service pertinents. Enfin les chercheurs, "utilisateurs finaux" réutilisent les traitements partagés et les exécutent sur les collections documentaires de leur choix. L'équipe Bibliome a contribué aux différents volets stratégiques du développement de l'infrastructure par, (1) le développement des technologies qui permettent l'interopérabilité des composants (langage, encapsulation) et leur identification dans le catalogue [1, 2], (2) la connexion avec d'autres e-infrastructures, bibliothèques numérique (OpenAire, Istex), portail d'ontologie (AgroPortal) et services en bioinformatique (réseau européen Elixir) et (3) l'offre de nouveaux services d'extraction d'information dans trois domaines identifiés comme stratégiques par l'Inra [6], l'écologie microbienne (focus sur la flore positive des aliments) [3, 4], le phénotypage du blé [5] et le développement des plantes (focus sur la graine d'arabette) [7], en collaboration étroite avec de nombreuses unités Inra. Ces trois services sont intégrés avec des données externes, et exploités par des e-infrastructures bioinformatique de l'Inra (Migale à MaIAGE, WheatIS à l'URGI et FlagDB++ à IPS2), où ils démontrent la valeur ajoutée de la modélisation de connaissance à partir de textes à destination des chercheurs du domaine.

Perspectives :
L'équipe Bibliome et la DIST de l'Inra travaillent à l'étude des conditions de mise en œuvre d'OpenMinTeD pour les scientifiques français dans le projet Visa TM (*Vers une infrastructure de services avancés pour le text-mining*) sous l'égide de la BSN (*Bibliothèque Scientifique Numérique*), avec l'INIST (*Institut de l'information scientifique et technique*) et le LIRMM (*Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier*). Le projet examine la question du développement de nouvelles offres de service selon trois dimensions interdépendantes, organisationnelle et économique, technologique et usage. L'interconnexion informatique des trois types d'e-infrastructures : bibliothèque numérique (ISTEX), ressources sémantiques (AgroPortal) et *text-mining* (OpenMinteD) et l'authentification des utilisateurs sont au coeur de la question technologique. La dimension de l'usage est examinée à travers des applications pilotes pour trois types d'utilisation, IST, développement d'application et recherche.
L'équipe Bibliome explore en particulier avec la plateforme Migale dans le domaine de l'alimentation et de l'écologie microbienne, les conditions pour une meilleure offre de service par l'intégration de de l'extraction d'information textuelle dans les services d'analyse développés par les plateformes de l'Inra.
Parallèlement, la DIST et Bibliome étudient les perspectives de développement d'OpenMinTeD au niveau européen par l'exploitation des services d'OpenMinTeD dans des e-infrastructures spécialisées comme Elixir en bioinformatique, AGInfra+ en agriculture, ou généralistes comme OpenAire.

Valorisation :
L'ensemble des développements logiciels (GitHub) et des documents du projet OpenMinTeD sont disponibles librement. Nous promouvons la plateforme par des formations en ligne sur FOSTER et en présentiel (formations bioinformatique par la pratique). L'équipe Bibliome contribue à la valorisation principale de la plateforme, par son appropriation par son réseau de collaborateurs français dans Visa TM et international, dont le DBCLS (Tokyo), BioASQ (Univ. Athènes).

Références :

Logiciels et ressources

OpenMinTeD : http://openminted.eu
OpenMinTeD sur GitHub : https://github.com/openminted
OpenMinTeD sur Foster : https://www.fosteropenscience.eu/openminted
Bibliome sur GitHub : https://github.com/Bibliome

Services en ligne

Biodiversité microbienne (2,3 millions documents) :

Moteur de recherche bibliographique : http://bibliome.jouy.inra.fr/demo/ontobiotope/alvisir2/webapi/search

Base de donnée Florilège (MEM) : http://genome.jouy.inra.fr/FlorilegeDemo

Les données du blé (3881 documents) :

Moteur de recherche bibliographique : http://bibliome.jouy.inra.fr/demo/wheat/alvisir/webapi/search

Système d'information WheatIS (Wheat Initiative) (lien temporaire) : https://urgi.versailles.inra.fr/beta/wheatis-gwt/ - result/term=mildew

Développement de la graine (documents) :

Moteur de recherche bibliographique (2046 documents) : http://bibliome.jouy.inra.fr/demo/seedev/alvisir/webapi/search

Système d'information FlagDB++ : http://tools.ips2.u-psud.fr/projects/FLAGdb++/HTML/NewCPG/index.shtml

Publications

[1] 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.

[2] Bohuon, J.-B. ; BA, M. ; Chaix, E. ; Bossy, R. ; Nédellec, C.. Integration of Alvis within the OpenMinTeD platform as Galaxy utilities [Poster]. *Galaxy Community Conference* (2017-06-26-2017-06-30) Montpellier (FRA).

[3] Chaix E., Aubin S., Deléger L., and Nédellec C. Text-mining needs of the food microbiology research community, In *Ovive workshop joint to European Federation for Information Technology in Agriculture, Food and the Environment (EFITA) conference*. Montpellier, 12 pages, juillet 2017.

[4] Nédellec C., Bossy R., Chaix E., Deléger L. Text-mining and ontologies: new approaches to knowledge discovery of microbial diversity. In *Proceedings of the 4th International Microbial Diversity Conference*. pp. 221-227, ed. Marco Gobetti. Pub. Simtra. ISBN 978-88-943010-0-7, Bari, October 2017.

[5] Nédellec C., Bossy R., Valsamou D., Ranoux M., Golik W., Sourdille P.. Information Extraction from Bibliography for Marker Assisted Selection in Wheat. In *proceedings of Metadata and Semantics for Agriculture, Food & Environment (AgroSEM'14), special track of the 8th Metadata and Semantics Research Conference (MTSR'14),* Springer Communications in Computer and Information Science, Series Volume 478, Karlsruhe, pp 301-313, Allemagne, 2014. DOI: 10.1007/978-3-319-13674-5_28

[6] 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* journal 2016: baw145 doi:10.1093/database/baw145 published online November 25, 2016.

[7] 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 17 janvier 2107.

Annexe à la fiche type de recueil des faits marquants 2017 des départements/centres

CLASSIFICATION

Priorités du Document d'Orientation (voir http://2025.inra.fr/)

[#Global] L'ambition globale d'atteindre la sécurité alimentaire dans un contexte de transitions

- **#Global-1** : Des transitions globales assumées
- **#Global-2** : La disponibilité des bio-ressources gérée aux différentes échelles
- **#Global-3** : Une vision intégrée des comportements, des marchés et des échanges
- **#Global-4** : Des approches territorialisées au service d'une compréhension générique des performances des systèmes alimentaires

[#3Perf] Des agricultures diverses et multi-performantes

- **#3Perf-1** : L'agro-écologie mobilisée au service de la multi-performance des agricultures
- **#3Perf-2** : D'autres leviers biologiques et technologiques pour la multi-performance
- **#3Perf-3** : L'évaluation multicritère pour objectiver les performances
- **#3Perf-4** : Des transitions comprises et facilitées

[#Climat] Les systèmes agricoles et forestiers face au défi climatique

- **#Climat-1** : L'adaptation de l'agriculture et de la forêt au changement climatique
- **#Climat-2** : La maîtrise de la contribution de l'agriculture et de la forêt à l'effet de serre
- **#Climat-3** : La conservation de la biodiversité et la valorisation des services
- **#Climat-4** : La préservation et la valorisation des ressources en eau et en sol

[#Food] Une alimentation saine et durable

- **#Food-1** : De nouveaux systèmes alimentaires territorialisés, notamment urbains
- **#Food-2** : Les systèmes alimentaires alliés de la santé
- **#Food-3** : Les qualités des aliments élaborées dès l'amont

[#BioRes] Des bio-ressources aux usages complémentaires

- **#BioRes-1** : Le développement des biotechnologies vertes et blanches
- **#BioRes-2** : L'apport des biotechnologies et des procédés pour de nouvelles ressources adaptées aux usages
- **#BioRes-3** : La conception de systèmes bioéconomiques

[#OpenScience] Une science ouverte grâce au numérique

- **#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
- **#OpenScience-5** : Le métier et l'environnement du chercheur adaptés au numérique

[#OpenInra] Un acteur national de l'innovation ouvert dans les territoires

- **#OpenInra-1** : Une ouverture vers l'enseignement supérieur et un partenariat territorial renforcés
- **#OpenInra-2** : La mobilisation de toute l'expertise de l'Inra en appui aux politiques publiques
- **#OpenInra-3** : Le chemin vers l'innovation bénéficie d'un pilotage renforcé
- **#OpenInra-4** : La Science ouverte aux acteurs non-marchands de la société

[#Appui] Anticiper et accompagner les évolutions

- **#Appui-1 :** Une organisation efficiente, agile, résiliente
- **#Appui-2 :** Une stratégie de financement fiable et solidaire
- **#Appui-3 :** Un Institut attractif et motivant pour ses agents
- **#Appui-4 :** Les actions et les valeurs de l'Institut visibles et partagées par une communication externe et interne active
- **#Appui-5 :** Un pilotage institutionnel efficace et partagé

Plans d'action

- **Ressources humaines et communication interne** : pour assurer l'attractivité et la cohésion d'une communauté de travail chargée d'une mission majeure de service public, en veillant à la motivation et à la qualité de vie au travail des agents titulaires, contractuels ou partenaires
- **Coopération avec l'enseignement supérieur** : pour décliner les thématiques prioritaires de l'Inra en stratégies scientifiques de sites, partagées avec nos partenaires dans les territoires, contribuant à faire de chaque grand site universitaire un pôle de rayonnement international sur les thématiques d'excellence de l'Inra
- **Innovation** : pour valoriser et élargir le formidable potentiel d'innovation de l'Institut, en combinant les disciplines, en co-construisant avec les acteurs des filières et des territoires, en valorisant nos infrastructures et en ciblant des domaines d'innovation prioritaires
- **Stratégie européenne et internationale** : pour décliner la stratégie scientifique de l'Inra avec un plan d'action visant à mobiliser nos principaux partenaires sur nos priorités au sein d'un réseau mondial de la recherche agronomique et alimentaire, et à assurer notre présence dans les institutions internationales
- **Prospective scientifique interdisciplinaire** : pour éclairer les futurs fronts de science, enrichir nos orientations, développer des actions incitatives, favoriser des partenariats scientifiques, économiques, disciplinaires ou de formation
    - ✓ Sciences pour les élevages de demain
    - ✓ Intégration des recherches (nexus) santé-alimentation-élevage
    - ✓ Agro-écologie
    - ✓ Approches prédictives en biologie et en écologie

Méta-programmes

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