Architecture, concepts et services d'un système d'indexation de données distribuées pour l'observation à large échelle en écologie marine

Romain David, Jean-Pierre Féral, Thierry Tatoni

Architecture, concepts et services d'un système d'indexation de données distribuées pour l'observation à large échelle en écologie marine

Dans le cadre du consortium IndexMeed (Indexing for Mining Ecological and Environmental Data)

R. David*, J.-P. Féral#, T. Tatoni$
Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale (IMBE)
Aix Marseille Université, CNRS, IRD, Université d'Avignon
Station Marine d'Endoume, 13007 Marseille, France. *romain.david@imbe.fr*

RESUME

Cet article présente la partie *architecture et services* de la thèse de R. DAVID, intitulé *Méthodes et outils pour le suivi à large échelle des habitats coralligène en Méditerranée : du système d'observation au système d'information*. Il présente les processus d'indexation et de qualification de données hétérogènes et distantes, basés sur des services WEB. Ceux-ci permettent de construire des graphes à partir de données et de thesaurus concernant les habitats coralligènes en Méditerranée orientale et occidentale et de les exploiter avec des objectifs d'indication et d'aide à la décision.

* étudiant en thèse, # encadrant principal, $ responsable du projet

Mots clefs

Indexation, qualification de données, traçabilité, système d'information décentralisé, fouille de donnée, écologie, graphes.

1. CONTEXTE

La révolution du *Big Data* en écologie tarde, alors qu'elle est considérée par la plupart des disciplines scientifiques et des industries produisant et utilisant de l'information comme la plus prometteuse des pistes de progrès et de découvertes. Les données utilisées par les scientifiques dans le domaine de l'écologie continuent de se diversifier et ne sont plus principalement produites par les institutions scientifiques, mais par des réseaux extérieurs d'acteurs et de compétences. De nouvelles disciplines (protéomique, métabolimique, métagénomique) complètent les prismes d'observation déjà multiples de la biodiversité. L'accessibilité de ces données est variable, et les processus de qualification qui évaluent leur utilisabilité et leur efficacité sont encore rares, a fortiori dans le domaine marin [1]. *A contrario*, la production de données scientifiques est de plus en plus financée sous condition de leur mise à disposition (depuis plusieurs décennies pour les données de biologie moléculaire, mais encore de façon balbutiante pour les données écologiques et environnementales), sans que des outils appropriés à une analyse intégrative soient proposés.

2. PROBLEMATIQUES

Des formats et des protocoles standards permettent d'interconnecter les bases de données dans le domaine de l'environnement. Des approches intégratives notamment en « écologie statistique » permettent des analyses à l'échelle globale [4]. Les approches sémantiques contribuent largement à améliorer leur interopérabilité. Il reste que les objectifs scientifiques spécifiques, les logiques d'organisation de projets et de collecte d'informations conduisent à une distribution décentralisée des données qui peut freiner la recherche dans le domaine de l'écologie. Dans ce contexte, comment i) analyser des données hétérogènes situées dans des bases de données distantes ? ii) inter-calibrer des systèmes d'observations différents, créer des correspondances et intégrer certaines approximations dans les correspondances entre systèmes descriptifs différents ? iii) mettre en évidence des relations entre des données d'observation les mettre en relation avec des patrons contextuels ? iv) favoriser l'ouverture et le partage et la valorisation des données ?

Les verrous scientifiques identifiés dans le cadre de ces travaux de thèse concernent notamment i) l'augmentation des fréquences et de la densité d'acquisition des observations (méthodes de reconnaissance automatique, outils d'acquisition moins onéreux), ii) la diversification des objets et des descripteurs d'objet intégrés dans les graphes, iii) la normalisation des descripteurs de la donnée et les méthodes permettant d'intégrer les différents niveaux de qualité des données. De nouvelles approches, basées sur une architecture répartie d'informations normalisées, mettant en rapport des données quantitatives et qualitatives, rendent possible l'investigation de questions de recherche complexes. Le programme européen CIGESMED (France, Grèce et Turquie - <www.cigesmed.eu>) et la récente structuration de l'observation des habitats méditerranéens dénommés *coralligène* sert de cadre à cette étude.

3. DEVELOPPEMENTS

Un outil de représentation sous forme de graphes des données de différents champs disciplinaires (écologie, sociologie, économie) a été développé en vue d'élaborer des méthodes de création de scénarios par approches successives (coévolution de facteurs), basée sur des concepts actuellement décrits par les approches globales en écologie. L'objet de ce prototype est de construire des graphes paramétrables avec des données hétérogènes (de la molécule à l'écosystème, en passant par les traits de vie, jusqu'aux paysages et aux interactions homme-milieux) concernant l'écologie de cet habitat et d'analyser les données grâce à des algorithmes utilisés dans d'autres disciplines. Il est alimenté par des systèmes d'information distants (10 laboratoires), dont chaque enregistrement est indexé par le serveur, ainsi que les qualifications nécessaires pour créer des relations entre chaque type d'objets indexés. Ces qualifications sont décrites sous forme de micro-thésaurus.

4. RESULTATS PRELIMINAIRES

Les graphes sont construits *via* le prototype de visualisation de graphe (serveur WEB) à partir d'informations agrégées grâce à des points nodaux d'indexation et de qualification des données de contexte sur l'environnement littoral et marin méditerranéen [3], dans différentes disciplines et à l'échelle méditerranéenne. Ces graphes sont paramétrables pour fouiller et visualiser ces données pluri-

disciplinaires en mettant sur le même plan des données de types écologiques, moléculaires et fonctionnelles (relations trophiques, traits fonctionnels), et le seront pour des aspects socio-écologiques, économiques. Les questionnements scientifiques possibles concernent l'écologie des systèmes observés (bon état écologique, correspondance de patrons de contextes et de données concernant les abondances relatives d'espèces comme dans la Figure 1), ou des systèmes d'observations (détection des biais dans la formation des observateurs, expertise partielle dans les jeux de données, définition de la puissance de l'échantillonnage nécessaire, gestion des coûts associés).

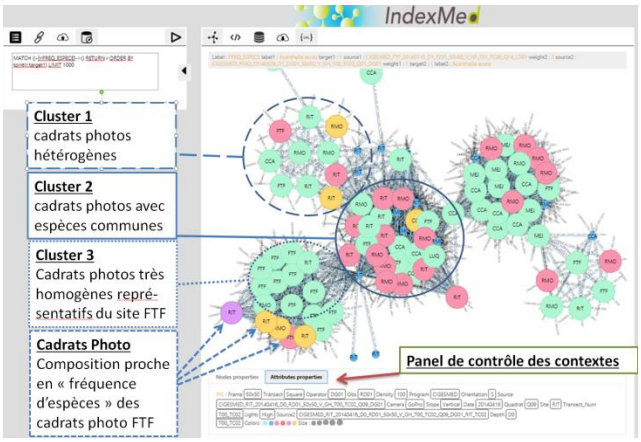

Figure 1 : Prototype de visualisation des données d'IndexMed utilisant dans cet exemple des fréquences d'espèces par photo. Les nœuds représentent les photos, les liens correspondent aux fréquences d'espèces, la grosseur des nœuds à différents paramètres environnementaux. Les photos proviennent de différents sites et les données indexées sont accessibles sur les systèmes d'information des partenaires, interrogées à distance par le prototype (JSON ou XML).

Les points nodaux d'indexation créés par le prototype sont *clonables* à volonté avec des règles d'enrichissement et de partage correspondant aux licences *creative common* du type «partage dans les mêmes conditions», autorisant les autres à reproduire, diffuser et modifier l'index, à condition qu'ils publient toute adaptation de l'index sous les mêmes conditions (open-source, open data). Ces règles devront favoriser l'alimentation de standards améliorant l'interopérabilité des données et favorisera la participation de nouveaux laboratoires en tenant compte de leurs capacités.

5. CONCLUSION ET PERSPECTIVES

L'architecture définie dans le cadre de ce travail permet de répliquer des « points nodaux d'indexation » dans différents domaines thématiques de l'écologie marine, permettant une qualification itérative des données compatible avec les systèmes normatifs internationaux (TDWG Taxonomic Data-base Working Group, recommandation INSPIRE), mais intégrant de manière générique tout nouveau système de qualification dit *métier* sous la forme de micro-thésaurus. La polysémie générée par l'adjonction de nouvelles disciplines doit être gérée par consensus sous la forme d'une interface (en cours de déploiement). De nouveaux champs disciplinaires sont intéressés au projet (EPD (European Pollen Database), ArkéoGIS (GIS en archéologie) ou GBIF (occurrences d'espèces terrestres) et renforcent l'aspect multidisciplinaire du projet.

Cette thèse a permis la création d'un consortium qui développe la culture des bases de données et leur utilisation efficace dans le milieu de la recherche en écologie. Sous l'impulsion du doctorant, il s'est étendu à plusieurs UMR de disciplines différentes. Lors du dernier séminaire (Journées du GRAAL - GRAphs and datamIning for environmentaL research), dans l'optique d'une internationalisation, la communauté IndexMed (Interopérabilité des bases de données en écologie) forte de plusieurs dizaines de nouveaux membres est devenu IndexMeed (Indexing for Mining Ecological and Environmental Data). La prochaine étape sera la co-élaboration d'un projet de recherche (BiodivERsA, SeasEra, H2020).

6. REMERCIEMENTS

Ce travail est réalisé à l'IMBE avec le soutien de France Grille et de la FRB (Fondation pour la Recherche sur la Biodiversité). La construction du premier prototype du consortium IndexMed a été est financé par le *défi CNRS* "VIGI- GEEK[1]" et le PEPS Blanc CNRS INEE avec le projet "Charliee[2]". Les données utilisées pour cet article ont été obtenues par le biais du projet CIGESMED (www.cigesmed.eu) dont nous remercions chaque partenaire. L'architecture a été débattu en groupe de travail lors du séminaire «design your infrastructure» organisé par European Grid Infrastructure http://www.egi.eu/ à Amsterdam (avril 2016) https://indico.egi.eu/indico/event/3025/. Et a été présentée lors du congrès IEMSs à Toulouse en Juillet 2016 [2]. http://www.iemss.org/sites/iemss2016. Nous remercions tous les membres actifs du consortium IndexMed pour leurs contributions et les GDR MaDICS et EcoStat pour leurs labellisations et leurs soutiens.

7. REFERENCES

[1] J. Barde, Mutualisation de données et de connaissances pour la Gestion Intégrée des Zones Côtières. Application au projet SYSCOLAG. Mathématiques [math]. U. Montpellier II - *Sciences et Techniques du Languedoc*, Nov 2006

[2] R. David, J.P. Féral, A-S. Archambeau, N. Bailly, C. Blanpain, V. Breton, A. De Jode, A. Delavaud, A. Dias, S. Gachet, D. Guillemain, J. Lecubin, G. Romier, C. Surace, L. Thierry de Ville d'Avray, C. Arvanitidis, A. Chenuil, M.E. Çinar, D. Koutsoubas, S. Sartoretto, T. Tatoni ; IndexMed projects : new tools using the CIGESMED DataBase on Coralligenous for indexing, visualizing and data mining based on graphs. In : Sauvage S, Sánchez-Pérez J-M., Rizzoli, AE (Eds.), *Proceedings of the 8th International Congress on Environmental Modelling and Software, Environmental modelling and software for supporting a sustainable future*, Vol. 3, pp.656-665, Toulouse, France. July 2016. ISBN : 978-88-9035-745-9.

[3] R. David, J.P. Féral, C. Blanpain, C. Diaconu, A Dias, S. Gachet, K. Gibert, J. Lecubin, C Surace, A first prototype for indexing, visualizing and mining heterogeneous data in Mediterranean ecology within the IndexMed consortium interdisciplinary framework. In: *SITIS 2015, 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, Bangkok, Thailand, pp. 232-239, nov. 2015 doi: 10.1109/SITIS.2015.119.

[4] O. Gimenez, S.T. Buckland, B.J.T. Morgan, N. Bez, S. Bertrand, R. Choquet, S. Dray, M.P. Etienne, R. Fewster, F. Gosselin, B. Mérigot, P. Monestiez, J. Morales, F. Mortier, F. Munoz, O. Ovaskainen, S. Pavoin, R. Pradel, F.M. Schurr, L. Thomas, W. Thuiller, V. Trenkel, P. de Valpine, E. Rexstad. Statistical ecology comes of age. (2014) *Biology Letters* 10: 20140698.

[1] VIGI-GEEK : VIsualisation of Graph In transdisciplinary Global Ecology, Economy and Sociology data-Kernel

[2] CHARLIEE : CHAnger de Regard En Liant dans Indexmed l'Environnement et les Etoiles