Journées SUCCES, rencontres Scientifiques des Utilisateurs de Calcul intensif, de Cloud Et de Stockage

Fouille de données hétérogènes en écologie :

les besoins émergeants en infrastructures de calcul du consortium IndexMed

http://www.indexmed.eu

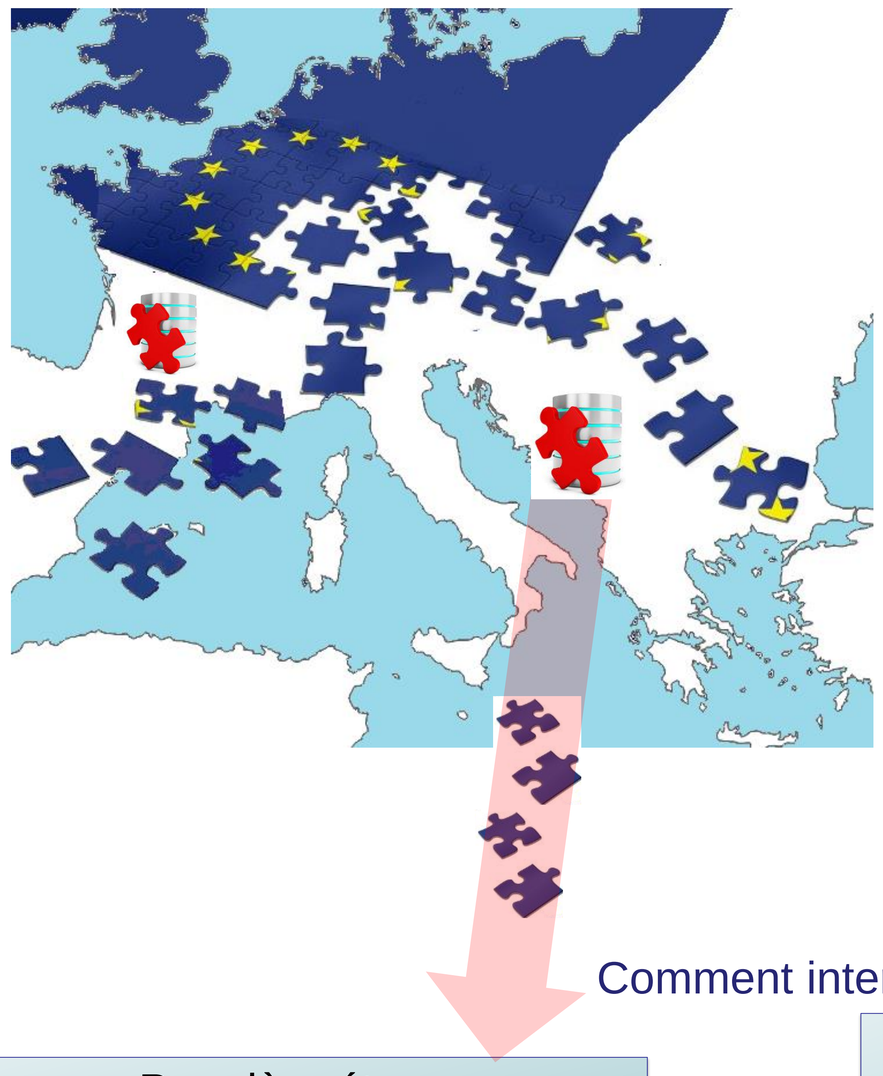
R. David [1], J.-P. Féral [1], A.S. Archambeau [2], L. Bernard [3], C. Blanpain [4], A. Dias [1], S. Gachet [1], K.Gibert [5], J. Lecubin [4], M. Leydet [1], I.Mougenot [6], C. Surace [7] and IndexMed Community

Animé par : Supporté par :

Données environnementales : quel contexte ?

Que ce soit pour les études sur la biodiversité ou pour la connaissance des systèmes socio-écologiques, la production de données s'intensifie et se diversifie. Les longues séries temporelles et/ou les études à large emprise spatiale sont difficiles à analyser. Dès-lors qu'il faut recourir à plusieurs observateurs la robustesse de l'observation est plus difficile à obtenir. La reproductibilité est aujourd'hui fréquemment mise en doute, voire réfutée.

Dans un cadre de production de données multi-sources, les problématiques d'équivalence de systèmes d'observations et d'inter-calibration d'observateurs deviennent alors cruciales. De plus en plus, **le besoin d'approches intégratives pluri- voire transdisciplinaires** devient nécessaire, dans l'étude de systèmes où la production de données dans chaque discipline est discontinue, peu précise et mal répartie. Toutes les variables de ces systèmes **interagissent dans le temps et à chaque échelle spatiale** (variables biotiques, abiotiques, pressions anthropiques et naturelles telles que décrites par les biologistes mais aussi les archéologues, historiens etc., services rendus et ressentis, image sociétale). Pour développer de nouvelles approches**, elles doivent être prise en compte simultanément**.



Le défi scientifique d'une meilleure compréhension globale des équilibres des Systèmes Socio-Ecologiques (S.S.E.) et de leur influence sur la biodiversité doit passer par la construction et **le test de méthodes de co-interprétation de ces données hétérogènes**. La science de la fouille de données doit pouvoir apporter de nouvelles perspectives aux recherches disciplinaires sur ces systèmes complexes qui finalement prennent en compte des objets intimement liés (chimie environnementale, génomique, transcriptomique, protéomique, métabolomique, écologie des peuplements, systèmes socio-écologiques, écologie du paysage, palynologie, archéologie et histoire en sont quelques exemples).

Comment interpréter des données hétérogènes et non liées?

Première étape :
L'indexation et la qualification des données (objets et attributs)

Deuxième étape :
La conception de graphes multi-thématiques à partir de questions scientifiques

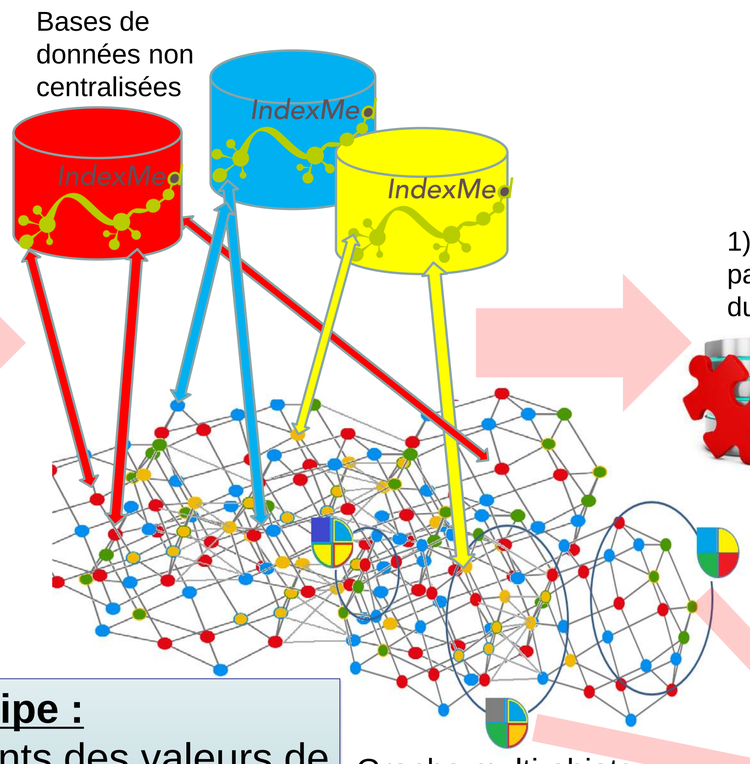
Troisième étape :
La fouille de graphes grâce aux infrastructures de calcul

L'indexation permet de laisser les données chez le producteur. Elle est associé à un travail de requalification des données afin de décrire les éléments communs, et ce qui les différencie (noms de champs, formats, rythme de mise à jour, précision, mesure ou observation, etc.). Ces descripteurs sont ajoutés aux données et constituent un corpus de liens possibles utilisables pour la fouille de données.

Prérequis techniques:

- Traçabilité
- Accessibilité
- ID Pérennes
- Modèles de données simples
- Bases de données ouvertes

Résolution (*via l'index*)
Ces index sont réplicables, open source, open data !



Principe :
Tous les objets ayants des valeurs de facteurs proches dans des contextes comparables s'attirent.
Plus il y a de liens entre données, **plus le graphe est précis.**

Graphe multi-objets :
les objets (observateurs, espèces, habitats…) sont des nœuds et les modalités d'attributs forment les liens

We need you!

1) Génération d'un flux paramétré contenant les objets du graphe et leurs liens

XML, JSON…

2) Validation des motifs de contextes significatifs par des méthodes de fouille de données

3) Conservation des motifs de contextes significatifs et de leur fréquence

Nous avons besoin de vous!

IndexMed renforce son réseau: manifestez votre intérêt!

Ce travail interdisciplinaire doit permettre une réponse concertée aux appels à projets européens s'intéressant aux problématiques suivantes : comment identifier puis lever les verrous scientifiques liés à la qualité des données et leur hétérogénéité ? Comment les considérer, malgré leur disparité, à un niveau équivalent dans le cadre de ces nouvelles approches ? Quelles nouvelles questions scientifiques en écologie peuvent être résolues par les nouvelles méthodes émergentes de datamining (clustering collaboratif, machine-learning, fouille de graphes, intelligence artificielle...) ? En fin de compte, quelles nouvelles perspectives pour la connaissance des S.S.E. et l'aide à la décision ? Le passage du prototype à la mise en place d'un programme de recherche à long terme et à large échelle nécessitera de **nouveaux moyens humains et de nouvelles compétences**. Il nécessitera d'accéder avec discernement, en l'adaptant aux besoins au cours du temps, **aux infrastructures de calcul à la bonne échelle et aux compétences associées**.

Journées SUCCES 2015

(1) Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale (IMBE), CNRS, IRD, Aix Marseille Université and Université d'Avignon, Station Marine d'Endoume, Chemin de la Batterie des Lions, 13007 Marseille, France romain.david@imbe.fr , jean-pierre.feral@imbe.fr , alrick.dias@imbe.fr , sophie.gachet@imbe.fr , michelle.leydet@imbe.fr (2) GBIF-France, MNHN, CP48, 43 rue Buffon, 75005 Paris, France archambeau@gbif.fr, gbif@gbif.fr (3) UMR ArcHiMedE, UFR des Sciences Historiques, Palais Universitaire, 9 Place de l'Université, 67084 Strasbourg Cedex, France. loup.bernard@unistra.fr (4) Service informatique (SIP), OSU Pythéas, CNRS, Aix Marseille Université, 13007 Marseille, France cyrille.blanpain@osupytheas.fr, julien.lecubin@osupytheas.fr (5) Department of Statistics and Operations Research, Universitat Politecnica de Catalunya, Barcelona, Spain, karina.gibert@upc.edu (6) Espace-Dev UM, Rue Jean-François Breton, 34000 Montpellier, France isabelle.mougenot@umontpellier.fr (7) Laboratoire d'Astrophysique de Marseille (LAM), CNRS, Aix Marseille Université, rue Frédéric Joliot-Curie, 13013 Marseille, France christian.surace@lam.fr