Visualisation de données sous forme de graphes en archéologie. Rencontre opérationnelle des archéologues d'ArkeoGIS et des écologues d'IndexMed

Data visualisation in archaeology based on graph approach. Operational meeting of ArkeoGIS archaeologists and IndexMed ecologists

Romain David[1], Loup Bernard[2], Cyrille Blanpain[3], Alrick Dias[1], Jean-Pierre Féral[1], Sophie Gachet[1], Julien Lecubin[3], Christian Surace[4], Thierry Tatoni[1]

[1] Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale (IMBE), CNRS, Aix Marseille Université, IRD, et Université d'Avignon, Station Marine d'Endoume, romain.david@imbe.fr, alrick.dias@imbe.fr, jean-pierre.feral@imbe.fr, sophie.gachet@imbe.fr, thierry.tatoni@imbe.fr
[2] Université de Strasbourg, Université de Haute-Alsace, CNRS, Archimède UMR 7044, Strasbourg, loup.bernard@unistra.fr
[3] Service informatique (SIP), OSU Pythéas, CNRS, Aix Marseille Université, Marseille, cyrille.blanpain@osupytheas.fr, julien.lecubin@osupytheas.fr
[4] Laboratoire d'Astrophysique de Marseille (LAM), CNRS, Aix Marseille Université, Marseille, christian.surace@lam.fr

RÉSUMÉ. Un point commun des études en archéologie, en écologie ou sur les systèmes sociaux est que la production de données est à la fois coûteuse et peu automatisée. Les suivis de longues séries temporelles et/ou à larges emprises spatiales sont difficiles à mener, dès lors qu'il faut recourir sur une longue durée à plusieurs observateurs. La robustesse et la reproductibilité de l'observation sont aussi plus difficiles à obtenir, voire impossibles en archéologie, même si les méthodes de modélisation se développent.

Dans un cadre de production de données multi-sources, l'équivalence des systèmes d'observations et l'inter-calibration d'observateurs deviennent cruciales. Des approches intégratives, pluri- ou trans- disciplinaires, deviennent nécessaires à l'étude de systèmes où la production de données dans chaque discipline est discontinue, plus ou moins précise et mal répartie. Pourtant, toutes les variables (caractérisation des activités économiques, des installations humaines, études des productions, objets reconstitués ou découverts, données biotiques et abiotiques, cartographies des pressions anthropiques et naturelles, services rendus et ressentis, image sociétale...) de ces systèmes interagissent dans le temps et à chaque échelle spatiale.

Après quelques années d'existence, ArkeoGIS agrège aujourd'hui 67 bases de données représentant plus de 50 000 objets (sites, analyses...). Fort de cette normalisation de l'information archéologique et paléo-environnementale, il nous a semblé important de tester de nouvelles méthodes de fouille de données, afin de mettre en évidence de possibles données « connexes » et complexes possiblement reliables à ces jeux de données. Le lien entre les extraits des bases agrégées au sein d'ArkeoGIS nous a permis de tester ces approches grâce à un prototype "open source" développé par le consortium IndexMed. Ce prototype permet la mise en place de liens entre objets de bases de données différentes.

Le consortium IndexMed a pour objectif d'identifier puis de lever les verrous scientifiques liés à la qualité des données et à leur hétérogénéité. La représentation de l'information sous forme de graphe rend possible la prise en compte des données malgré leur disparité et sans les hiérarchiser, et permet d'améliorer la précision des outils d'aide à la décision utilisant des méthodes émergentes d'analyse de données (*clustering* collaboratif, classification collective, fouille de graphes, analyse de réseau, extraction de communautés). Adapter ces méthodes à l'archéologie nous permet d'aller au-delà de la « simple » agrégation de données : ArkeoGIS peut donc aussi servir à alimenter les outils de fouille utilisés au sein de nos données et métadonnées.

ABSTRACT. The one thing in common "archaeological", "biodiversity" or "social systems" studies share is that data production is both expensive and few automated. Long time series and / or large spatial surveys are difficult to conduct, since it is necessary to use several observers. The robustness and reproducibility of the observation are also harder to get and is obviously impossible in archaeological sciences, even if modeling methods are improved.

In a context of multi-source data production, the equivalence of observation systems and the inter-calibration of the observers become crucial. Multi-disciplinary integrative approaches become necessary to study systems where the output

of data, in each discipline, is discontinuous, imprecise and poorly distributed. Yet, all variables (characterization of economic activities and human installation, productions studies, characteristics of the discovered or reconstituted objects, biotic or abiotic data, maps of anthropogenic and natural pressures, rendered services and feelings, societal perception...) of these systems interact over time and at each spatial scale.
After a few years of existence, ArkeoGIS aggregates 67 databases representing over 50 000 objects (sites, analyzes...). With this standardization of archaeological and paleoenvironmental information, it seemed important to test new data mining methods, to see whether "related" and complex data can be linked to these archaeological data sets. The link between aggregated-bases extracts within ArkeoGIS allowed us to set up a cross-requesting and test possibilities in a prototype developed by the consortium IndexMed. This prototype, open source, allows the establishment of links between objects from different databases.
The consortium IndexMed aims to identify and to raise the scientific challenges related to data quality and heterogeneity. The use of graphs allows us to consider data despite their disparity and without prioritization, and improve decision support using emerging data mining methods (collaborative clustering, machine learning, graphs approaches, representation knowledge). Adapting these methods in archeology allows us to go beyond the "simple" data aggregation: ArkeoGIS can therefore also be used to power such tools allowing us to mine our data and metadata.
MOTS-CLÉS. visualisation, qualification de données, graphes, système d'information décentralisé, archéologie.
KEYWORDS. visualisation, data qualification, graph, distributed information system, archeology.

Introduction / contexte

Si les archéologues utilisent l'informatique depuis plusieurs décennies, la mise en commun des données produites et leur interrogation à l'aide de méthodes plus novatrices que les désormais traditionnelles CAH et AFC restent un défi. Ces données hétérogènes sont difficilement exploitables de manière intégrée par les techniques couramment utilisées par les chercheurs en archéologie. L'analyse de cette grande quantité de données diversifiées et produites par des sources hétérogènes constitue également un vrai défi pour la science des données. Désormais, la croissance exponentielle de la quantité de données nécessite l'utilisation des outils les plus récents.

La version 4 d'ArkeoGIS (arkeogis.org) permet d'agréger des bases de données disparates au sein d'un outil libre et en ligne à l'aide d'une ontologie "bottum up" construite par les acteurs de la discipline. Forte de plusieurs décennies d'expériences plus ou moins comparables (Archaeomedes puis ArchaeDYN, Fastionline), et agrégeant des projets aussi bien archéologiques (Digital Atlas of Roman and Medieval Civilizations - DARMC, sont à l'étude Chronocarto, NOMISMA, artefacts…) que des projets environnementaux (European Pollen Database, MedMAX, Banadora ou DCCD en cours), le projet ArkeoGIS a trouvé avec Indexmed une équipe développant un outil très adapté, bien qu'initialement développé dans le cadre d'études en écologie marine (David et al 2016).

Récemment, les 3èmes journées organisées par le consortium IndexMed (https://indexmed2016.sciencesconf.org/) ont mis en évidence non seulement le potentiel des approches basées sur les graphes, mais aussi les lacunes en termes de compétences et d'expérience de la communauté des écologues et des archéologues pour adapter et utiliser ces méthodes, afin d'analyser de manière totalement intégrée leurs jeux de données multisources. Une première discussion autour de la visualisation de données hétérogènes, à laquelle ont participé les animateurs d'ArkeoGIS a notamment permis de montrer que des techniques à base de graphes peuvent être adaptées pour modéliser plus efficacement les composantes d'interactions spatiales malgré l'hétérogénéité de ce type d'information. Les participants à ces journées (STIC -Sciences des Techniques de l'Information et de la communication-, écologues et archéologues) ont sollicité l'organisation de rencontres et le développement de collaborations entre la communauté des chercheurs en science de l'écologie et de la biodiversité et celle des chercheurs en science des données et STIC.

IndexMed est un consortium pluridisciplinaire créé par l'axe *Gestion de la biodiversité et des espaces naturels* de l'IMBE (Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale). Son objectif principal est de développer la culture des bases de données et leur utilisation efficace dans le milieu de la recherche en écologie et biodiversité. Ce consortium s'est étendu à plusieurs UMR de disciplines différentes (notamment, de l'environnement pour l'expertise qualitative

de la donnée, et de l'astronomie pour l'expertise en matière de gestion des grosses masses de données). Il doit permettre de répondre à des appels à projets dans le domaine des bases de données en écologie méditerranéenne en favorisant l'interdisciplinarité et les collaborations avec d'autres entités du CNRS. Les projets qui y seront développés doivent s'appuyer sur les différentes démarches nationales et internationales et promouvoir un travail partenarial international. IndexMed doit notamment servir de relais aux réseaux et démarches en place nationalement et internationalement, et proposer une réponse aux obligations européennes (Aarhus, INSPIRE,…) auxquelles les laboratoires de recherche travaillant dans les domaines de l'environnement et des sciences humaines sont et seront de plus en plus soumis. L'objectif à court terme d'IndexMed est de mettre en place une plateforme d'indexation des données sur la biodiversité méditerranéenne et des paramètres environnementaux ayant un intérêt pour la recherche (David et al 2015). Cette indexation utilisera les outils et méthodes préconisés nationalement (SINP - Système national d'Information sur la Nature et les Paysages, MNHN - Muséum National d'Histoire Naturelle, SPN - Service du Patrimoine Naturel, RBDD Réseau Bases De Données du CNRS) ou internationalement (OBIS, GBIF, LifeWatch, GEOBON, CoL, WoRMS…) et s'appuiera sur les catalogues développés à ce niveau (IDCNP - Inventaire des Dispositifs de Collecte sur la Nature et les Paysages du SINP, Réseaux d'acteurs de la FRB - Fondation pour la Recherche sur la Biodiversité).

Les données environnementales et écologiques accessibles sont d'un grand intérêt pour contextualiser les données issues des prospections archéologiques. Les méthodes de fouille de données basées sur la fouille de graphes peuvent être transposables en archéologie, et s'appuyer sur les résultats de requêtes réalisées à l'aide d'ArkeoGIS. Cet article présente les principes de l'utilisation des graphes envisagés en lien avec ArkeoGIS, ainsi que quelques grands types de questionnements scientifiques testés grâce aux premiers exports. Il met en lumière les verrous scientifiques et techniques identifiés lors de ces premières prospections. Il dresse enfin la liste de quelques pistes pour lever ces barrières et développer ce nouveau mode de prospection en archéologie, basé sur les données hétérogènes et multi-sources.

Théorie et méthode

Principes de la représentation de l'information sous forme de graphes

Un graphe est un ensemble de points que l'on appelle des nœuds (sommets en mathématique ou cellules en informatique) reliés par des traits (segments) ou flèches nommées arêtes (ou bien liens ou arcs). L'ensemble des arêtes (edges en anglais) entre nœuds (nodes en anglais) forme une figure similaire à un réseau (Aggarwal & Wang 2010).

Afin de construire les graphes qui seront présentés, un export de données a été réalisé à partir d'ArkeoGIS. La représentation de ces données sous forme de graphes permet de relier des objets (champs de la base de donnée ou valeurs de ces champs) ayant des formats différents (quantitatif, qualitatifs ordonnés ou non ordonnés); les valeurs d'attributs d'un second champ de la base de donnée permettent de créer les liens entre ces objets. Les liens sont matérialisés par des descripteurs (une variable ayant plus d'une valeur possible, qui est aussi la valeur ou une transposition de la valeur que prend un champs de la base de donnée pour un enregistrement). Les descripteurs quantitatifs sont en général transformés en classes de valeurs.

Plusieurs descripteurs peuvent être assemblés pour former - selon la combinaison de leurs valeurs respectives - un motif, qui pour ces valeurs données sont appelés patrons (patterns en anglais). Ces patrons peuvent décrire des objets et/ou des liens et/ou des contextes (contextes qui ne participent pas à la topologie du graphe).

Les objets ayant le plus de liens en commun sont les plus proches, ceux ayant les liens les plus ténus (c'est à dire le moins de chemins possibles pour les relier à entre eux et beaucoup de nœuds

intermédiaires) sont les plus éloignés dans la représentation. La représentation sous forme de graphe permet de représenter de nouveaux objets en combinant les valeurs de différents champs. On peut ainsi traiter les champs un à un ou bien en groupe de valeurs.

D'autres champs de la base de données, nommés "contextes", sont ensuite utilisés pour colorer ou changer la forme et la grosseur des nœuds. Ils ne participent pas à la topologie du graphe. Les motifs ainsi projetés dans le graphe peuvent être (i) dispersés, auquel cas les liens qui organisent le graphe ne sont pas liés aux éléments de contexte ; ou bien (ii) regroupés dans une ou plusieurs parties du graphe, auquel cas il existe un lien entre la façon dont les nœuds sont organisés et un ou plusieurs contextes.

L'analyse des fréquences relatives de ces "motifs" et des redondances entre "plus proches voisins" par rapport à leur fréquence dans tout ou partie du graphe donne une idée de l'importance de ces corrélations. La significativité de ces motifs peut ensuite être testée par des méthodes comme le *clustering* de graphes. Le clustering (classification non supervisée en français, mais c'est le terme anglais qui est le plus usité à la place de "classification", "groupe", ou "regroupement") consiste à regrouper des éléments. Cette agrégation est un élément-clé pour l'analyse de grands graphes. Une fois les groupes obtenus, on peut ré-appliquer l'opération pour obtenir un clustering hiérarchique (basé sur une autre variable par exemple). Cette décomposition hiérarchique (ou multi-échelle) permet de modifier la complexité des algorithmes de fouille, de faciliter l'exploration des données, et de proposer une visualisation paramétrable : on parle aussi de navigation multi-échelle (Lambert et al 2013).

Dans des graphes plus complexes ou le nombre de combinaison et de liens peut croître exponentiellement, l'étude de la corrélation entre fréquence de contextes et "clusters" de nœuds peut demander de paralléliser les calculs nécessaires à une investigation des parcours possibles. Selon la question scientifique sous-jacente aux objets représentés par le graphe, certains éléments dans les liens ou les nœuds peuvent être ignorés ou simplifiés. Cet aspect prospectif dans les graphes est en cours d'élaboration avec la communauté STIC .

Le potentiel actuel d'ArkeoGIS

ArkeoGIS fonctionne comme un agrégateur requêtable de bases de données. Cela signifie que toute information présente dans ArkeoGIS peut faire l'objet de requêtes sur son emplacement, l'état de la recherche, et les périodes concernées. Chaque site peut ensuite être interrogé à l'aide de cinq filtres selon que l'information concerne les structures, le mobilier, les productions, le paysage et les analyses, ou les textes et l'iconographie. Le résultat de la requête s'affiche sous forme de carte dans l'application ; les lignes d'informations correspondantes peuvent ensuite être exportées dans un format très simple (CSV), permettant une réutilisation dans tout type de logiciel. Ce format assure une compatibilité avec des tableurs, des bases de données, des systèmes d'information géographique, des logiciels d'analyse ou de modélisation (comme R ou Netlogo p.ex.), ou encore avec le prototype de visualisation sous forme de graphe de données distantes en cours de développement dans le cadre d'IndexMed.

Initialement développé afin de mutualiser les données archéologiques et paléoenvironnementales de la vallée du Rhin, ArkeoGIS est un système d'information géographique (SIG) libre, en ligne et multilingue (allemand-anglais-espagnol-français). Actuellement dans sa quatrième version, ArkeoGIS permet de mettre en commun les données scientifiques spatialisées concernant le passé. Les bases de données sont issues de travaux de chercheurs institutionnels, d'étudiants avancés, de sociétés privées, de services d'archéologie, mais aussi de travaux de paléo-environnementalistes, d'historiens et de géographes. Tous ces travaux et bases de données sont accessibles et requêtables en ligne. L'étendue chronologique de l'outil est désormais ouverte et permet d'agréger des informations depuis la Préhistoire jusqu'à nos jours. L'emprise spatiale d'ArkeoGIS permet d'afficher des informations sur toutes les régions du monde. A ce jour, les régions les mieux renseignées sont le Rhin Supérieur, l'Ouest méditerranéen et le Proche Orient. Plusieurs dizaines de milliers de sites, objets et analyses sont d'ores et déjà accessibles. ArkeoGIS fait aussi le lien vers différents outils numériques, permettant à ses utilisateurs d'avoir connaissance de différents projets numériques en ligne.

Chaque utilisateur peut interroger en ligne tout ou partie des bases disponibles, afficher ses résultats sur plusieurs fonds de carte et exporter les résultats de sa requête vers d'autres outils. ArkeoGIS peut servir à tout travail de recherche individuel ou collectif. Il permet entre autres de gérer le *Data Management Plan* (DMP) de contrats de recherche, et constitue un outil puissant pour la préparation de recherches théoriques ou de terrain (fouilles, synthèses, thèses etc..)

Chaque auteur mettant à disposition des informations géoréférencées au format ArkeoGIS reste maître de celles-ci et peut seul décider de les modifier, un identifiant unique pérenne (DOI-HANDLE, répondant aux normes INSPIRE) qualifie chacune des bases. Le contributeur peut ainsi très facilement accéder aux informations des autres contributeurs afin d'implémenter sa base tout en citant ses sources. Un annuaire permet de mettre en contact les chercheurs, afin de développer les échanges entre pays et entre institutions.

Aspect heuristique et représentation de la connaissance

Pour l'archéologue, la spatialisation des données a une forte valeur heuristique. Elle permet de saisir instantanément l'état de la recherche ou l'avancement de la mise en commun des données, et rend possible une approche interdisciplinaire et diachronique.

Cette représentation spatiale de la connaissance est immédiatement exploitable pour les données que le chercheur maîtrise ; en revanche, pour les données issues de disciplines connexes (dans le temps, l'espace ou en provenance de travaux environnementaux ou paléo-environnementaux), d'autres outils sont nécessaires. L'exploration des jeux de données peut avantageusement compléter la représentation géographique et permet de mettre en évidence des ressemblances entre sites dans le jeu de données, sans que ceux-ci soient géographiquement proches (ce qui fait l'objet de la présente collaboration).

Enfin, les métadonnées et la façon dont nos bases sont renseignées livrent, grâce à ces représentations de la connaissance, des informations parfois inattendues sur la densité de la recherche par thématique, et se révèlent un outil précieux pour les questions de "*state of art*".

Curation de données

La mise en commun de données issues de différents chercheurs permet une curation simple, sur l'emplacement des sites par exemple, mais aussi une amélioration itérative des données. C'est-à-dire que chaque chercheur peut compléter sa base avec les informations mises à disposition par les collègues, et au passage corriger d'éventuelles erreurs. Ce travail commun implique de facto la constitution d'un vocabulaire contrôlé commun à large échelle. Lorsqu'il est validé par une communauté homogène et organisé avec des liens d'équivalence et des liens hiérarchiques, on parle alors de micro-thésaurus. Lorsque deux communautés travaillent à l'interprétation interdisciplinaire de leurs données, comme c'est ici l'ambition (archéologie / palynologie / paramètres environnementaux), une confrontation entre ces micro-thésaurus est nécessaire. Celle-ci doit prioritairement établir des correspondances sur les descripteurs et la valeur des descripteurs qui, utilisés en commun, permettent une analyse conjointe des données de disciplines différentes (cf. aussi Bernard et al 2015 pour une approche plus "traditionnelle").

Export des données pour construire les graphes

Afin de pouvoir exploiter les données exportées depuis ArkeoGIS, un certain nombre de modifications ont dû être effectuées : les champs non relevants (présence de Geonames p.ex.) ont été supprimés, les vides remplacés par "NULL" et les champs alphanumériques (Bibliographie et Remarques) ont également été ignorés. La fonction d'export de la V4 d'ArkeoGIS (figure 1) permet maintenant une intégration quasiment directe dans le prototype de visualisation d'IndexMed.

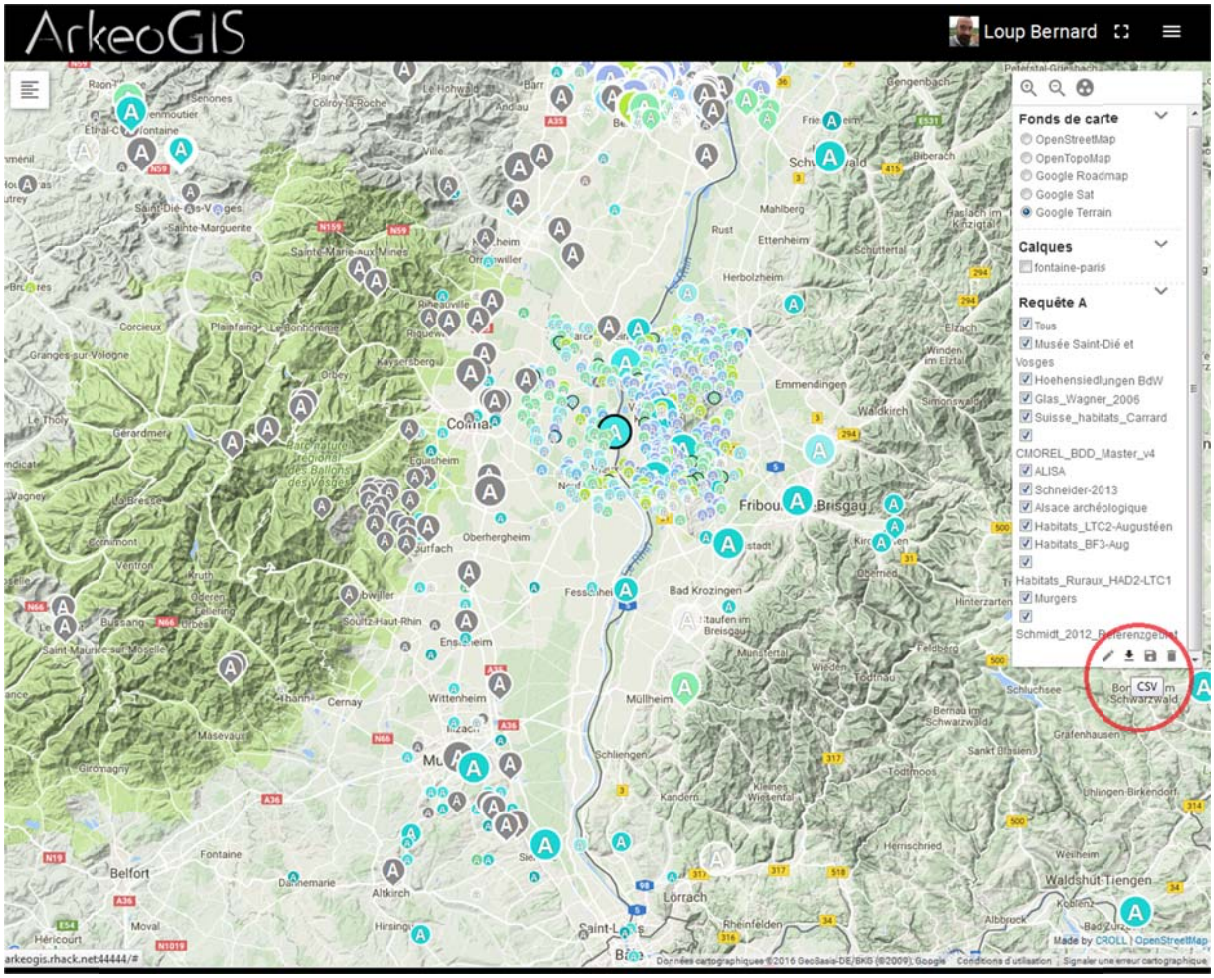

Figure 1. *Capture d'écran de la version 4 d'ArkeoGIS. Les bases ayant fourni des informations sur la région du Rhin Supérieur (extrait de requête sur le Rhin Supérieur de Bâle à Freiburg) sont listées à droite. La taille des points est indexée sur l'état de la recherche ; les sites qui n'ont pas de pointe vers le bas sont des centroïdes. Les codes couleurs correspondent à la chronologie choisie, ici "Europe continentale du Néolithique à nos jours". Les sites en gris sont de période indéterminée. Le curseur de la souris (cercle rouge) est sur l'icône qui permet d'enregistrer le résultat de cette requête au format CSV. C'est ce type de fichier qui a ensuite été utilisé avec des modifications mineures afin d'alimenter l'outil développé par le consortium IndexMed et de produire une représentation des données sous forme de graphes.*

Après cette présentation synthétique et assez formelle de l'intérêt de la mise en commun et de la représentation de données hétérogènes sous forme de graphes, voici parmi de multiples possibilités, deux premières représentations des données d'ArkeoGIS.

Prototype de visualisation et résultats préliminaires

Prototype de visualisation

L'interface du prototype utilise Neo4j <neo4j.com/>, une base de données à base de graphes mise en œuvre en java et publiée en 2010. L'édition communautaire de la base de données est sous licence GNU GPL v3. La base de données et ses modules supplémentaires (sauvegarde en ligne ou haute disponibilité) sont disponibles sous licence commerciale. Le prototype d'IndexMed permet à un opérateur néophyte d'importer des données (en CSV, XML ou JSON). Il permet d'interroger Neo4j pour produire le graphe et d'interagir avec lui à l'aide du navigateur Web. Le personnel technique d'IndexMed développe un frontend Web spécifique à l'aide du langage Ajax / Jquery. Il peut être possible de demander une base de données demandant des objets spécifiques et des relations spécifiques entre eux, sans utiliser un langage de requête technique tel que SQL ou Cypher.

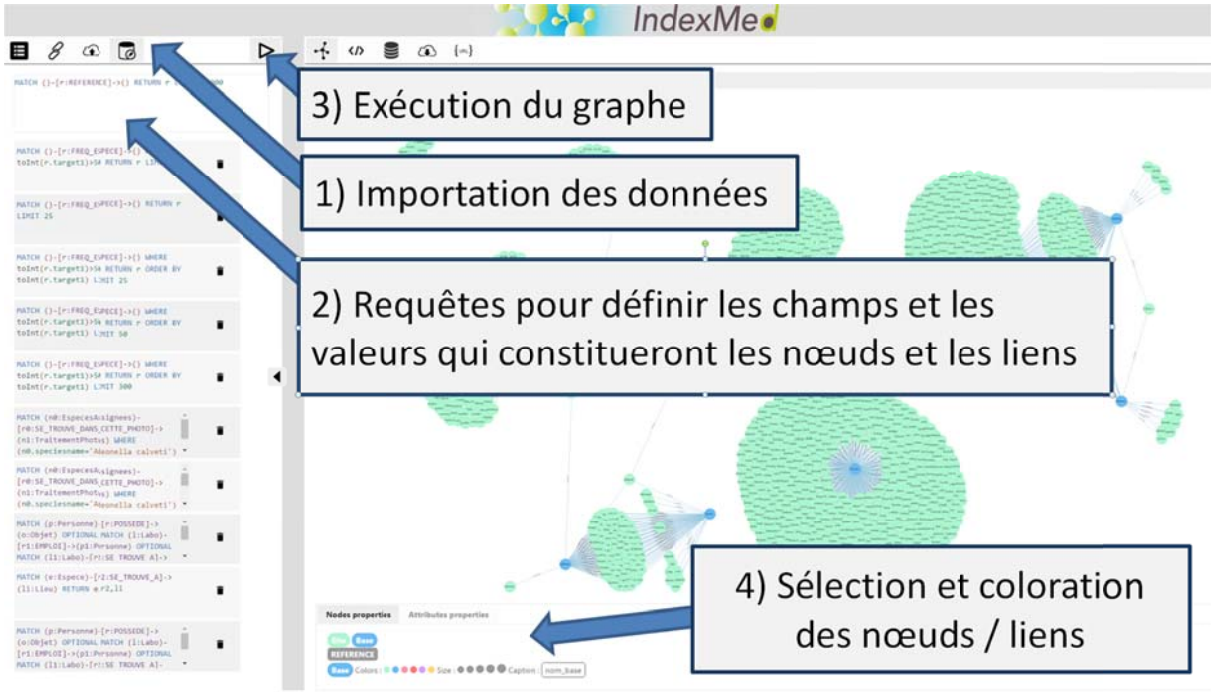

Figure 2. *Présentation générale du prototype d'IndexMed de visualisation des données représentant dans cet exemple 1492 sites d'archéologie en vert et 12 bases de données. Les informations proviennent de l'import d'ArkeoGIS mais pourraient directement être interrogées à distance (1) par le prototype sur les systèmes d'information des partenaires, (format JSON ou XML). La colonne de gauche permet d'importer, d'effectuer les requêtes (2) avec un formulaire ou le langage cypher et de les enregistrer. Un bouton (3) permet de lancer l'exécution de la nouvelle requête ou d'une requête pré-enregistrée. Le bandeau du bas (4) permet de configurer les couleurs des noeuds et des liens en fonction des valeurs de descripteurs.*

Le prototype est développé pour pouvoir être générique et permet d'intégrer n'importe quel type de données sous la forme de "valeur d'objet et d'attribut". Il suffit ensuite à l'opérateur de sélectionner la base à utiliser, les champs qui servent de noeuds, les champs qui servent de liens, et ceux qui servent à mettre en évidence des éléments de contextes. Il est aussi possible de faire ces opérations en sélectionnant certaines valeurs de champs. Ce prototype sera disponible sous forme de source ouverte pour développer, à moyen terme, l'utilisation de ces graphiques pour l'aide à la décision en matière de gestion environnementale et dans le cadre d'un projet de recherche à soumettre aux appels à projets européens (BiodivERsA, ERDF, SeasEra , H2020 ...)

Résultats préliminaires

Dans le premier graphe présenté, les bases sont des nœuds, les sites sont des nœuds reliés aux bases qui contiennent des données les concernant (figure 2), les descripteurs de base (auteurs, langues, types, descripteurs communs) permettent de colorer les nœuds ou de sélectionner une partie des bases ou des sites seulement.

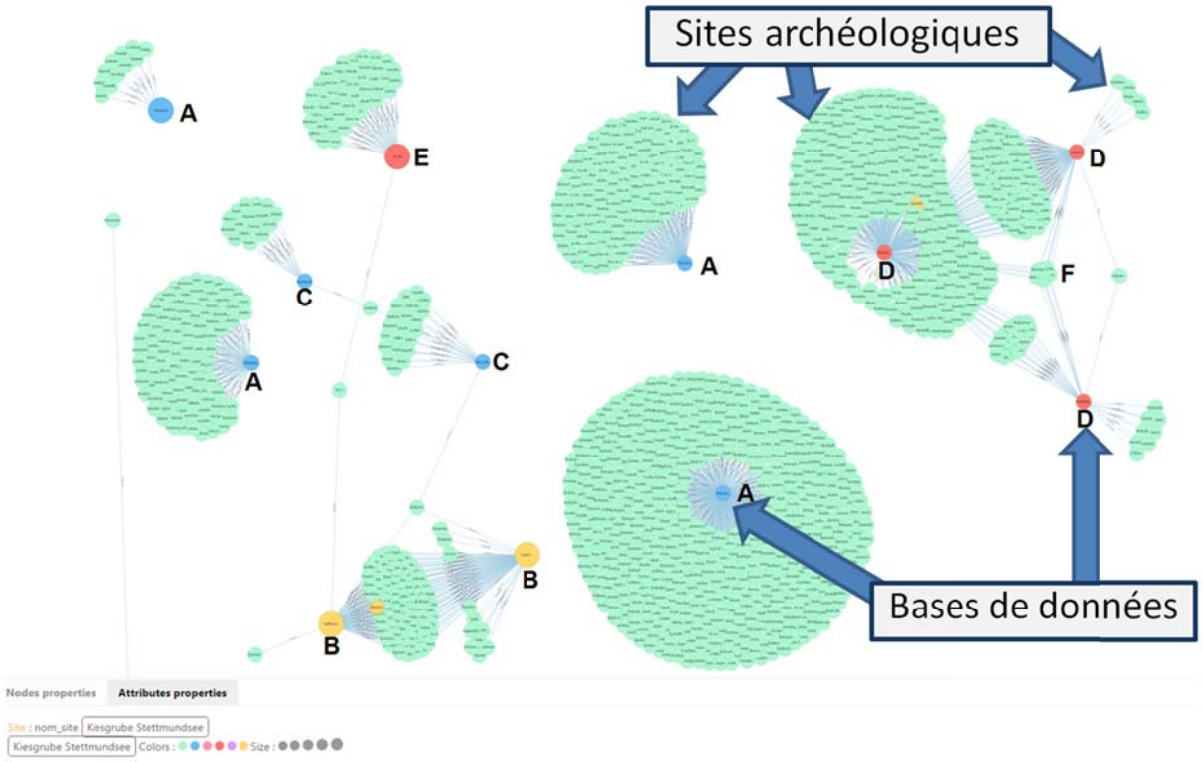

Figure 3. *Graphe avec extrait des données issus de la requête géographique présentée dans la figure 1 et reliant les bases avec les sites communs. Cette figure permet de mettre en évidence, concernant la requête réalisée, les bases qui référencent des informations sur des sites archéologiques qu'elles ne partagent avec aucune base (A), d'autres qui référencent l'essentiel des sites en commun avec une autre base (B). Dans le cas C, un seul site est commun aux autres bases. Les trois bases du cas D partagent une partie de leurs sites deux à deux, et Le cas F montre les sites communs aux trois bases à la fois. Le cas E est relié par un noeud aux cas B, par un "NULL", qui met en évidence une erreur dans le jeu de données (un enregistrement sans nom pour le site dans les deux bases de données reliées).*

Le fait de représenter les sites et les bases sur le même graphe (figure 3) permet d'étudier le système d'information en lui même. L'opérateur peut visualiser l'importance de chaque base dans la sélection géographique effectuée via Arkeogis. Il est possible aussi de sélectionner une période ou d'autres critères descripteurs des bases ou des sites eux même. Cette représentation permet aussi, comme le cas E dans la figure 3, de mettre rapidement en évidence des particularités ou des erreurs dans les jeux de données.

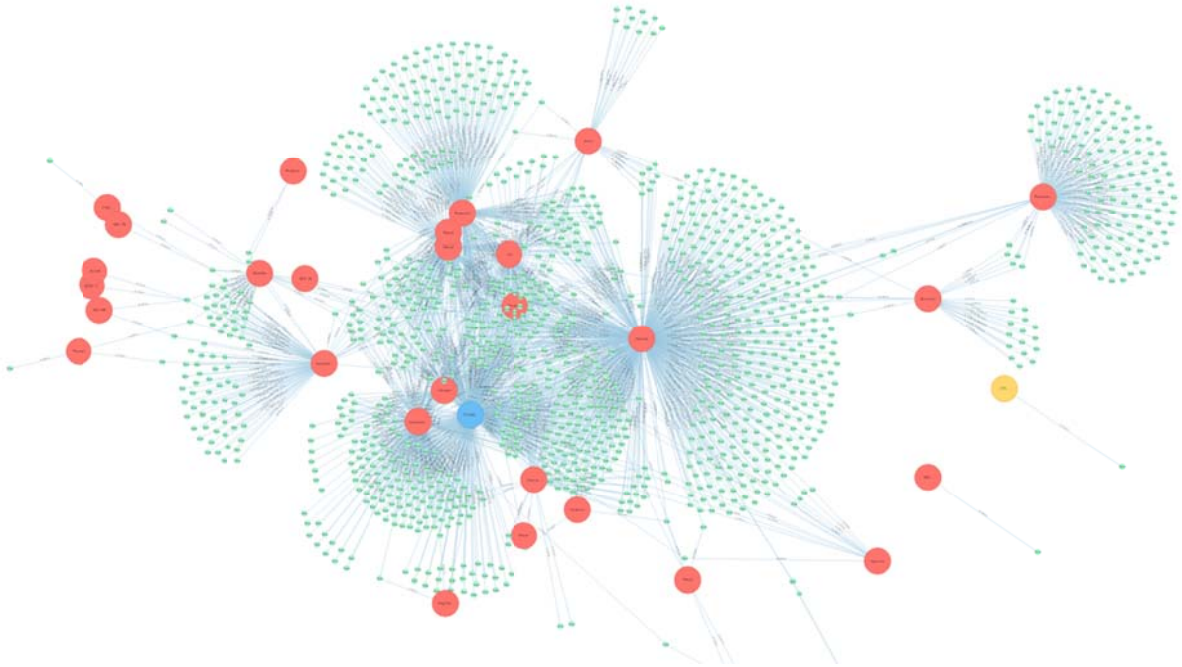

Figure 4. *Graphe issu de l'interface IndexMed, utilisant les données exportées à partir de la sélection géographique sur ArkeoGIS, représentant 1492 sites et 4950 liens matérialisant les types d'objets trouvés sur ces sites. Les sites sont représentés par les petits nœuds verts, les clusters rapprochant les sites ayant les mêmes topologies / patrons de descripteurs (ici sont utilisés uniquement les patrons d'objets de niveau 1 matérialisés par des nœuds rouges). Une sélection a été faite sur deux valeurs de descripteurs : "Bleu" pour les types d'objet de niveau 1 qualifiés de "céramique" (Le site contient donc au moins un objet de type céramique si il est lié à ce noeud), Jaune : le noeud correspond à un objet daté de la période exactement égale à "-900 -à -726".*

Le graphe suivant est un exemple de représentation du système observé (la sélection géographique de sites et les types d'objets de niveau 1). Les sites représentés au centre du graphe (figure 4) contiennent les patrons d'objets les plus communs, ceux à la périphérie des patrons plus particuliers. Certains sites déconnectés du graphe (exemple de la période sélectionnée égale à "-900 -à -726") contiennent des types d'objets (de niveau 1, les niveaux suivants étant moins systématiquement renseignés) que l'on ne trouve que sur ces sites. Des regroupements d'objets (groupes de nœuds en rouge à gauche du graphe) semblent typiques de quelques sites et sont donc rassemblés dans un cluster bien particulier, auquel correspond un contexte précis.

Lorsque des objets sont particuliers à un groupe de site (ici on a clairement un cluster d'objets en rouge, à gauche du graphe), on peut rechercher les contextes particuliers (groupes de valeurs de descripteurs significativement différents dans cette partie du graphe, par rapport aux autres clusters) grâce à des algorithmes adaptés. Le choix des algorithmes dépendent du type des objets représentés dans le graphe et de sa topologie.

L'étude de ces contextes associés à ces clusters est un champ d'investigation d'autant plus grand que le nombre de descripteurs de contextes consistant est important ; Sur 15 bases sélectionnées pour les graphes, en se limitant aux 16 champs communs de l'export utilisables en tant que descripteurs, on peut remarquer que même si 77% des champs sont renseignés de manière quasiment systématique (à plus de 90%), 20% des descripteurs sont renseignés en moyenne à moins de 60% (dont 12.1% des descripteurs sont renseignés en moyenne à moins de 10%) ce qui les rend inutilisables dans un des graphes tel que nous les proposons. Cela laisse apprécier la marge de progrès à faire si l'on souhaite élargir la liste de ces descripteurs.

Discussion et perspectives

Difficultés et propositions de résolutions

De manière générale, les verrous scientifiques à lever dans un projet de recherche interdisciplinaire futur devront être précisés par les experts du domaine des STIC, et concernent notamment i) l'augmentation des fréquences et de la densité d'acquisition des observations (développement des méthodes de reconnaissance automatique et déploiement d'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.

Le premier verrou important, surtout lorsque l'on part de bases de données "empilées", est d'avoir un dénominateur commun suffisant à chacune de ces bases pour obtenir, pour chaque objet, au moins une valeur pour les descripteurs qui servent de liens. Cette recherche de consistance des données se traduit par l'élimination systématique des descripteurs qui ne sont pas majoritairement renseignés (idéalement, il faut au moins une valeur enregistrée pour chaque objet, en comprenant bien qu'un objet n'ayant pas de lien avec un autre n'a aucun intérêt à être représenté). Plus le nombre de descripteurs intégrés dans le modèle est grand, moins il doit y avoir d'objets sans valeur pour un descripteur.

La deuxième source de difficultées liée à cette approche est l'utilisation involontaire de descripteurs équivalents, ou au moins partiellement redondants : ceux-ci peuvent entraîner une "déformation" du graphe, et les rapports de distances entre les objets représentés être mal interprétés. Des techniques - à explorer - de comparaisons de topologies de graphes, faites avec les descripteurs dont on soupçonne la redondance, sont prévues dans le cadre du développement de cette recherche.

Pour intégrer ces formats différents de variables, il faut donc travailler sur le dénominateur commun pour chaque type de variable descriptive ou de contexte. Cela peut à minima être fait sous la forme de booléens (présence/absence), mais demande de faire accepter à tous les acteurs un processus permettant un consensus sur les descripteurs, leur définition, leurs nombres de valeurs possibles et les possibilités de dupliquer les descripteurs d'une même qualité avec des niveaux de précision différents et les équivalences entre ces niveaux.

Les perspectives interdisciplinaires de cette approche de fouille de données basée sur les graphes sont applicables à la majorité des sciences humaines et sociales. Elles requièrent la recherche et l'acceptation de termes/valeurs structurants dans les jeux de données, basé sur les standards dans chaque discipline (et donc une mise en cohérence des standards entre disciplines, ce qui est parfois un peu compliqué). L'internationalisation des standards, les nuances entre langues qui impactent la compréhension exacte des termes en anglais et l'évolution en propre des standards sont aussi des facteurs à prendre en compte.

Cette cohérence est une condition *sine qua* non de l'interdisciplinarité. Ce travail d'homogénéisation peut commencer par un relevé des problèmes de polysémie de descripteurs et de termes/valeurs structurants. Il doit éviter les jargons ou les usages trop locaux et/ou non entérinés par les communautés. L'objectif doit être d'élaborer une proposition d'amélioration des standards par recherche de consensus entre les communautés, et ceci de manière itérative.

La taxonomie utilisée pour désigner des descripteurs et leur donner des valeurs a une importance primordiale pour construire un graphe qui puisse répondre à un questionnement scientifique. Elle agit principalement sur sa topologie. Sans entrer dans les détails, par exemple, son pouvoir descripteur / discriminant dépend du nombre de valeur pour chaque descripteur et de la répartition des valeurs possibles du descripteur sur l'ensemble des enregistrements.

Les archéologues sont majoritairement des littéraires, leurs outils sont traditionnellement bibliographiques, et depuis quelques décennies maintenant statistiques. Il nous semble essentiel d'accompagner le transfert vers de nouveaux outils, par exemple en incluant de la bibliographie dans le code, ou en codant des articles ou des hypothèses.

Afin que les participants au projet ne se sentent pas tributaires d'une "*black box*" (i.e. le fait que les utilisateurs non-avertis ne comprennent pas l'intégralité des calculs effectués), le code sera fourni en open source (avec la bibliographie correspondante). Cela devrait permettre à l'opérateur d'identifier et / ou de modifier le code.

Concernant enfin la manipulation de variables dans les graphes, afin de lever les suspicions de boîte noire, une bonne méthode nous semble de tester des hypothèses simples et avérées. Pour une approche archéologique, cela pourrait correspondre à des faciès ou des groupes culturels connus, par exemple, afin de vérifier que ces entités apparaissent regroupées sur le graphe aussi clairement que sur le SIG. Les graphes pourraient permettre ici de tester des marqueurs annexes qui "ressortent" alors qu'ils n'avaient pas été pris en compte lors de l'hypothèse originale. Une démarche secondaire et complémentaire pourrait être de tester une hypothèse absurde afin de mettre en avant le fait que l'outil ne fonctionne pas s'il est mal utilisé.

Perspectives

Les premiers tests de ces méthodes sont néanmoins encourageants car ils permettent d'appréhender tout ou partie du système observé (un ensemble de sites) en intégrant des données de contextes de format différents. Ils permettent aussi d'étudier le système d'observation ainsi que les efforts de prospection, ou d'avoir une approche visuelle de la répartition des compétences en archéologie. Les aspects chronologiques et les requêtes spatiales n'ont pas été mises en oeuvre de manière plus poussée, car un travail d'homogénéisation de ces descripteurs est encore nécessaire (nombre de catégories et valeurs partagées par toutes les bases de données utilisées). Dans un second temps, ils feront l'objet d'analyses plus complexes basés sur des graphes issus de requêtes plus spécifiques. Chacun de ces trois aspects pourra être développé dans le cadre de futures recherches interdisciplinaires, dans lesquelles les questionnements scientifiques en archéologie côtoient les questions scientifiques en sciences et techniques de l'information et de la communication.

La mise en place d'une dynamique d'échange entre des experts en écologie / biodiversité / archéologie et des experts du domaine des STIC est la prochaine étape prévue par le consortium IndexMed. Proposée à différents financeurs sous la forme d'une action, celle-ci regroupe des représentants des deux champs disciplinaires et permettra de formaliser des besoins en terme d'analyses de données hétérogènes de la part de la communauté écologie / biodiversité / archéologie et de stimuler la recherche en STIC afin de proposer des solutions plus adéquates pour l'analyse et la gestion des données écologiques dans le contexte du Big Data (prise en compte de la dimension temporelle et spatiale et des données multi-échelles et hétérogènes).

En ce qui concerne les techniques et approches STIC étudiées et développées, la recherche sera dirigée vers des techniques de gestion et d'analyse de graphes qui puissent prendre en compte la complexité des données hétérogènes et notamment passer à l'échelle sur des jeux des données volumineux sans détériorer la qualité des résultats obtenus. Sous l'égide d'IndexMed, il est prévu de réaliser une première carte des compétences de laboratoires en informatique qui pourront apporter des outils méthodologiques ou des techniques algorithmiques adéquates pour l'analyse des données issue de l'écologie et des Sciences humaines et sociales.

Conclusion

Les approches proposées dans le cadre de la collaboration entre ArkeoGIS et IndexMed seront testées sur d'autres correspondances de patrons utilisant des objets hétérogènes dans les domaines des sciences environnementales liées à l'archéologie. L'ambition est de développer des modèles d'études libres et ouverts sur le long terme aux experts de l'analyse de données (STIC), ainsi que des processus de test et de choix de nouveaux algorithmes à l'échelle globale pour les utilisateurs potentiels (environnementalistes comme archéologues dans le cas décrit ici).

Cet élargissement de la collaboration sera l'occasion de préciser les besoins, intérêts et attentes de chaque communauté, en termes de recherche autant que de formation. Une plateforme ouverte aux collaborateurs désirant investiguer ces nouvelles méthodes, pourrait prendre par exemple la forme d'une « forge » permettant aux deux communautés de faire évoluer leurs recherches sur le long terme, mettant ainsi en place un vrai *Linked Open Data* utilisable par les chercheurs travaillant sur des sujets interdisciplinaires à l'aide de nos outils. Les acteurs de la recherche intéressés peuvent prendre contact avec les deux communautés via arkeogis.org et indexmed.eu

Remerciements

La construction du premier prototype du consortium IndexMed a été est financé par le défi CNRS « VIGI- GEEK1 » et le PEPS Blanc CNRS INEE avec le projet "Charliee2".

Nous remercions tous les membres actifs du consortium IndexMed pour leurs contributions et les GDR MaDICS et EcoStat pour leurs labellisations et soutiens. Les auteurs tiennent évidemment à remercier leurs communautés respectives, concernant ArkeoGIS plus particulièrement les auteurs des bases utilisés : G. Hoffmann, M. McCormick, C. Morrissey, C. Morel, M. Trautmann, N. Schneider, H. Wagner, C. Jeunesse, M. Roth-Zehner, D. Schwartz et C. Schmid-Merkl, et pour la relecture effectuée par Dino Ienco concernant les termes propres aux STIC.

Références

AGGARWAL, C & WANG, H. 2010. C. Aggarwal, H. Wang, *Managing and Mining Graph Data*, Springer, 1st Edition., 2010, XXII, 600 p.

BERNARD et al. 2015. Bernard L., Ertlen D., Schwartz D., "ArkeoGIS, Merging Geographical and Archaeological Datas Online",*in* Giligny F., Djindjian, F., Costa, L., MoscatiI, P., Robert, S. (éds.) *Concepts, methods and tools. Proceedings of the 42nd Annual Conference on Computer Applications and Quantitative Methods in Archaeology* Paris, Archaeopress 2015 : 401-406.

DAVID et al 2015. David R., 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.

DAVID et al 2016. David R., 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, A.E. (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

LAMBERT et al 2013. Lambert A., R. Bourqui, D. Auber, "Graph Visualization for Geography. Methods for Multilevel Analysis and Visualisation of Geographical Networks", *Springer*, : 81-102, 2013. <hal-00841188>

