

Visualisation des métadonnées des géoportails

- Premières Visualisations -

Problématiques actuelles

1. Comment mettre en avant la généalogie des données et quelle est son évolution ?
2. Comment mettre en relief les flux d'informations au sein des IDG ?
3. De quelle manière agissent les acteurs au sein des IDG et quelle est leur portée (spatiale ou thématique) ?

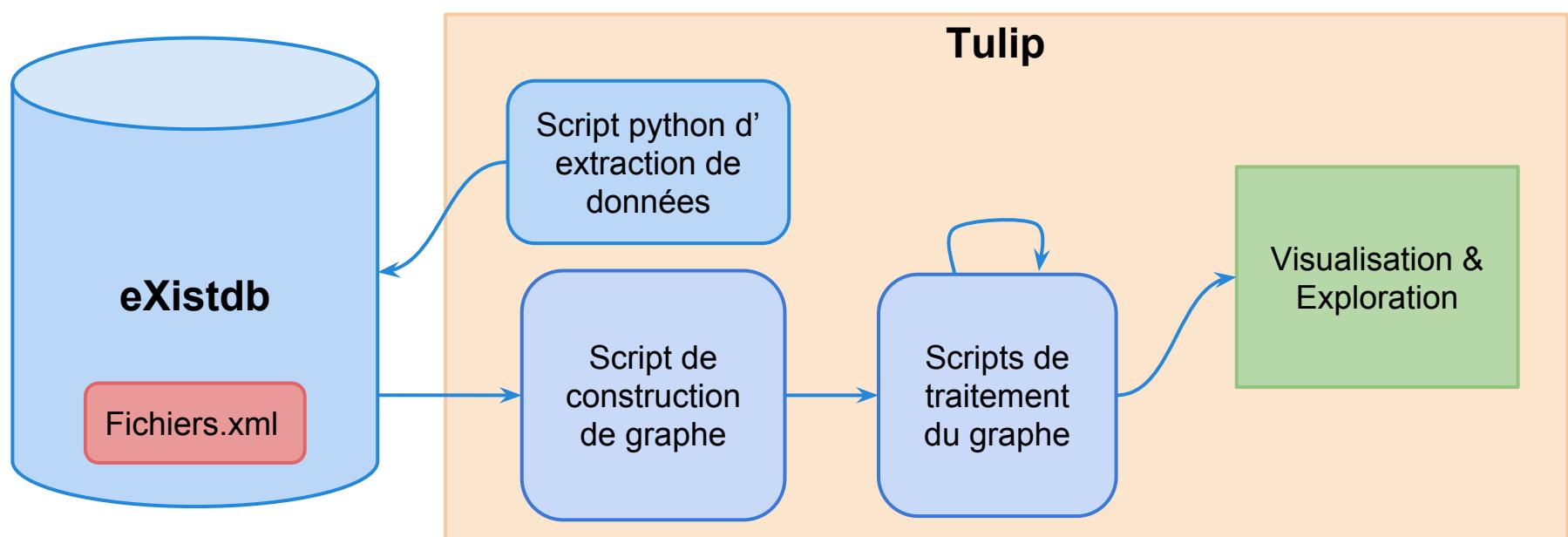
Procédure Actuelle

Prérequis :

- Données sous forme .xml
- EXistDB installé (nécessite Java JRE ≥ 7)
- Python 2.7.6 installé (incompatible > 3.x.x)
- Tulip 4.6.1 installé ou supérieur

Procédure Actuelle

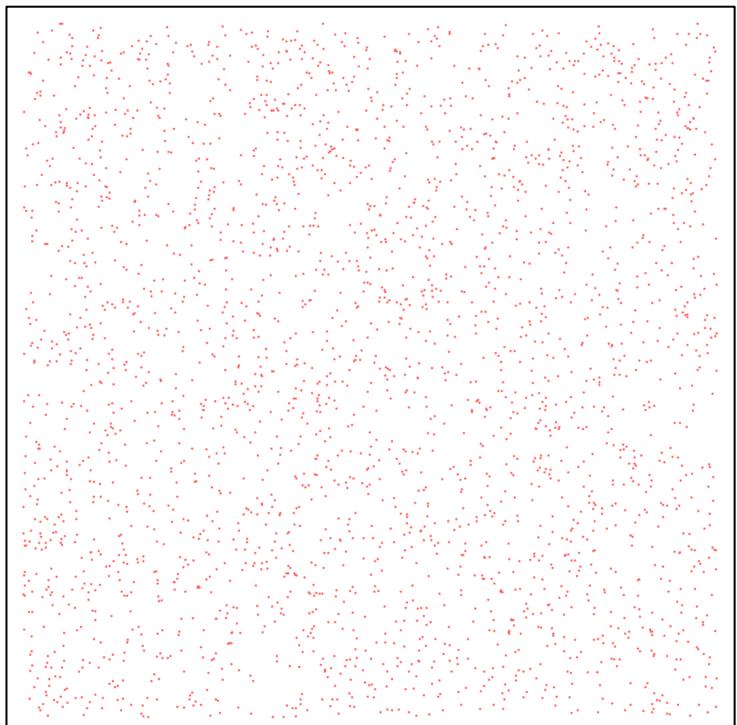
Procédure : Mettre les .xml dans eXistdb puis lancer la suite de scripts dans Tulip



Construction du graphe

Le script de construction va créer un ensemble de sommets* qui représentent chacun une fiche .xml de métadonnées.

Ici, les 2835 nodes de PIGMA :



*Ou "nodes". Ici les points rouges.

Construction du graphe

- Chaque sommet possède les informations contenues dans la fiche de métadonnées correspondante.

- Ces informations sont accessibles directement via la visualisation, la spreadsheet (ci-contre), etc.

Genealogie	Hierarchie	ID_Fiche	Lampe	OrganisationName	OrganisationRole	Role	Tags	Titre
Les données proviennent de la DGFIP qui da...	-jeu de données	-fb6da399c-58cd-4439-a078-5a7c38117828	-fre	{"VAL DE GARONNE AGGLOMERATION"}	{"owner"}		-URBANISME; PARCELLES CADASTRALES; MARMAN...	-Communauté d'Agglomération Val de Garonne
Fonds Cartographiques de la couverture GS...	-jeu de données	-96a3097c-d431-4497-a54e-8bb320678ec8	-fre	{"UNITE MIXTE DE RECHERCHE AMENAGEMENT De ..."}	{"owner"}		-ACTIVITES ECONOMIQUES; AQUITAINIE; CAMPING...	-Aquitaine : Recensement des sites web des co...
{1} Prélevements : deux sites ont été sélectio...	-jeu de données	-c2e2736d-c1df-4ef6-8ae5-4ab1c0a85c	-fre	{"UMR CNRS 5805 EPOC", "UMR CNRS 5805 EPO ..."}	{"pointOfContact", "pointOfContact", "poi ..."}		-AQUITAINIE; BRETAGNE; BASSIN D'ARCACHON; BI...	-Réponse adaptative des coques et des palour...
{1} Dates : -2004-2005 (ATPNEC Tapamor) : é...	-jeu de données	-f43783c3-9481-4175-a740-9dd2e0a0c63d	-fre	{"UMR CNRS 5805 EPOC", "UMR CNRS 5805 EPO ..."}	{"pointOfContact", "pointOfContact", "poi ..."}		-AQUITAINIE; BASSIN D'ARCACHON; ABONDANCE; ...	-Densités bactériennes associées à des bivalve...
{1} Dates : -2003-fréquence bi-hébdomadaire	-jeu de données	-2bd3983d-b3af-4c61-ad97-320a37b0a071	-fre	{"UMR CNRS 5805 EPOC", "UMR CNRS 5805 EPO ..."}	{"pointOfContact", "pointOfContact", "poi ..."}		-AQUITAINIE; BASSIN D'ARCACHON; DYNAMIQUE D...	-Suivi des abondances du pico et du nano plan...
Evrac : un lot de 16 huitres d'un état ...	-jeu de données	-f9ee240b-b3d0-488b-a2d3-22991a1b1a35	-fre	{"UMR CNRS 5805 EPOC", "UMR CNRS 5805 EPO ..."}	{"pointOfContact", "pointOfContact", "poi ..."}		-AQUITAINIE; BASSIN D'ARCACHON; BIVALVE; ...	-Analyse du comportement de mollusques bival...
{1} Dates : 31 stations échantillonées en avr...	-jeu de données	-veda28404-5d09-46c6-963-49f62af620c	-fre	{"UMR CNRS 5805 EPOC", "UMR CNRS 5805 EPO ..."}	{"pointOfContact", "pointOfContact", "poi ..."}		-AQUITAINIE; BASSIN D'ARCACHON; DIVERSITÉ; PR...	-Diversité procarionte de flores d'intérêt écolog...
{1} Dates : -1999 à 2002 et 2004: échantillon...	-jeu de données	-b472c51e-f1fa-47c4-b5b1-e1ccdb602986	-fre	{"UMR CNRS 5805 EPOC", "UMR CNRS 5805 EPOC"}	{"pointOfContact", "pointOfContact"}		-AQUITAINIE; BASSIN D'ARCACHON; FLORE; OCÉA...	-Diversité phytoplanctonique dans le Bassin d'A...
{1} Dates: déc. 2002 à janv. 2004: Échantillonna...	-jeu de données	-031f7dea-7b7b-4079-9345-d44bb4440d1	-fre	{"UMR CNRS 5805 EPOC", "UMR CNRS 5805 EPOC"}	{"pointOfContact", "pointOfContact"}		-AQUITAINIE; BASSIN D'ARCACHON; FLORE; OCÉA...	-Communautés microbiennes autoptrophes dans
{1} Dates d'échantillonnage : «A Cassy, fond...	-jeu de données	-88182e10-7f0d-48a8-97e8-f023d8b7af79	-fre	{"UMR CNRS 5805 EPOC", "UMR CNRS 5805 EPOC"}	{"pointOfContact", "pointOfContact"}		-AQUITAINIE; BASSIN D'ARCACHON; OCÉANOGRAP...	-Cycles de prélevement d'eau de 24h au milieu d...
{1} Méthodes: Nos futures recherches sur la ...	-jeu de données	-5116a497-cee6-45e6-9010-7a74130a0e6	-fre	{"UMR CNRS 5805 EPOC", "UMR CNRS 5805 EPOC"}	{"pointOfContact", "pointOfContact"}		-AQUITAINIE; BASSIN D'ARCACHON; BIÈVUE; ECOT...	-Mise en évidence d'une nouvelle pathologie ch...
Renseignements pratiques des fichiers bibliogr...	-jeu de données	-f8ed081b-75c7-479f-a2e2-2025feef1309	-fre	{"UMR CNRS 5805 EPOC"}	{"pointOfContact"}		-BIBLIOGRAPHIE; AQUITAINIE; BASSIN D'ARCACHON;	-Bibliographie UMR CNRS, Bassin d'Arcachon
Les données ont été obtenue soit par cd-ro...	-jeu de données	-52969295-e0e6-415b-b13c-0006ffrb8c3a	-fre	{"UMR CNRS 5805 EPOC"}	{"pointOfContact"}		-BASSIN D'ARCACHON; TÉLÉDETÉCTION;	-Images satellites optiques sur le Bassin d'Arc...
{1} Dates: Janvier – Juin 2005; Janvier – Jun ...	-jeu de données	-2a8c3160-1569-424a-802c-91eb61b72e8	-fre	{"UMR CNRS 5805 EPOC"}	{"pointOfContact"}		-AQUITAINIE; BASSIN D'ARCACHON; HYDROGÉOLOG...	-Flux continental d'azote et de phosphore vers
{1} Méthodes: Analyse de la fluctuation saiso...	-jeu de données	-88182d8d-96e9-48c2-8a90-61f5ea2a0644	-fre	{"UMR CNRS 5805 EPOC"}	{"pointOfContact"}		-AQUITAINIE; BASSIN D'ARCACHON; BIÈVUE; ECOT...	-Impact des contaminations métalliques chez la
{1} Échantillonnage chimique: Prélèvement ...	-jeu de données	-7c3884d8-1913-4872-a887-93994007ab82	-fre	{"UMR CNRS 5805 EPOC"}	{"pointOfContact"}		-AQUITAINIE; BASSIN D'ARCACHON; GÉOPHYSIQUE...	-Distribution de l'aide inorganique dissous dans
{1} Caractérisation physico-chimique des sé...	-jeu de données	-7616b4fb-44b0-4f09-a397-d34154c7516d	-fre	{"UMR CNRS 5805 EPOC"}	{"owner"}		-AQUITAINIE; BASSIN D'ARCACHON; INSTRUMENTA...	-Etudes in situ et ex situ des paramètres chimiq...
{1} Dates: juillet à septembre (1999 à 2003) ...	-jeu de données	-38572a0c-9b65-424f-ba8e-c33a6a516e1	-fre	{"UMR CNRS 5805 EPOC"}	{"pointOfContact"}		-AQUITAINIE; BASSIN D'ARCACHON; FAUNE; OCÉA...	-Suivi naissance : diversité zooplanctoniques dans
Photographies prises par TIGF Années: 1994...	-jeu de données	-b5ca3c2b-2a1e-41de-86c2-2179ff1ceda	-fre	{"UMR CNRS 5805 EPOC"}	{"pointOfContact"}		-AQUITAINIE; BASSIN D'ARCACHON; GÉOMORPHOL...	-Photographies séries de TIGF sur le Bassin
{1} Dates: Série en cours depuis 1997: échant...	-jeu de données	-59a7bc15-11fa-4646-867c-20f7ab6cd1	-fre	{"UMR CNRS 5805 EPOC"}	{"pointOfContact"}		-AQUITAINIE; BASSIN D'ARCACHON; FAUNE; OCÉA...	-SOAR-Préincrivre : Diversité métazoaires zoot...
{1} Dates : 2005 et 2006 A chaque sorte les ea...	-jeu de données	-a5453fc8-4485-478a-9577-ea99a306fa0	-fre	{"UMR CNRS 5805 EPOC"}	{"pointOfContact"}		-AQUITAINIE; BASSIN D'ARCACHON; OCÉANOGRAF...	-Chênes des eaux des estuaries dans la zone de G...
{1} Dates: Novembre 1998 à janvier 2000 éch...	-jeu de données	-9f101695-dc5c-4540-94b7-f582dd631ccc	-fre	{"UMR CNRS 5805 EPOC"}	{"pointOfContact"}		-AQUITAINIE; BASSIN D'ARCACHON; FAUNE; OCÉA...	-Diversité des ciliés et flagellés planctoniques da...
{1} Analyse de l'expression génétique de gén...	-jeu de données	-d6873a89-91ec-4d81-beed-199b79a77e09	-fre	{"UMR CNRS 5805 EPOC"}	{"pointOfContact"}		-AQUITAINIE; BASSIN D'ARCACHON; BIÈVUE; ECOT...	-Réponse de l'huître creuse aux variations de se...
{1} Méthodes: 3 sites d'échantillonnage: zone...	-jeu de données	-04ed482c-914b-4bb1-bbd4-89d777c31a69	-fre	{"UMR CNRS 5805 EPOC"}	{"pointOfContact"}		-AQUITAINIE; BASSIN D'ARCACHON; OCÉANOGRAF...	-Observation à long terme du zooplankton dans
{1} Dates : mars 2005, mars, mai, juillet, sept...	-jeu de données	-a3d88842-c911-4ef9-a78e-48bd1cc995	-fre	{"UMR CNRS 5805 EPOC"}	{"pointOfContact"}		-AQUITAINIE; BASSIN D'ARCACHON; OCÉANOGRAF...	-Géochimie des sédiments dans la zone de G...
287 photos	-jeu de données	-088066260-3b2a-41e0-8110-f97ff9bba1c	-fre	{"UMR CNRS 5805 EPOC"}	{"owner"}		-AQUITAINIE; BASSIN D'ARCACHON; PHOTOS; TRAI...	-Photographies aériennes du Bassin d'Arcachon
3 sites d'échantillonnage: zone interne, zone...	-jeu de données	-5566d317-608c-4a9b-88b8-1e2342332420	-fre	{"UMR CNRS 5805 EPOC"}	{"pointOfContact"}		-AQUITAINIE; BASSIN D'ARCACHON; OCÉANOGRAF...	-Observation à moyen et long terme de l'évolut...
-jeu de données	-fac43944-155c-4a0f-acb3-0731e91be4e	-	-	{"TIGF AQUITAINE"}	{"pointOfContact"}		-TRANSPORT; AQUITAINE;	-Aquitaine : Stations de Compression
-jeu de données	-8ebaa028-0003-4afe-9b8b-0cbabf899a	-	-	{"TIGF AQUITAINE"}	{"pointOfContact"}		-TRANSPORT; AQUITAINE;	-Aquitaine : Canalisations
-jeu de données	-f1c40804-31ef-4fb0-a930-79606a03e726	-	-	{"TIGF AQUITAINE"}	{"pointOfContact"}		-ENERGIE; TRANSPORT; AQUITAINE; SERVICES D'UT...	-Aquitaine : Postes de Livraison
Cette ressource provient de Gaz de France, ...	-jeu de données	-c1100c81-0081-4d4f-b701-1f7813a7c383	-	{"TIGF AQUITAINE"}	{"pointOfContact"}		-ENERGIE; TRANSPORT; AQUITAINE; SERVICES D'UT...	-Aquitaine : Postes de sectonnement et Robin...
Géolocalisation approximatif, à posteriori...	-jeu de données	-9559a8fb-0219-4f74-9bde-7e2332389a	-	{"SYSDAU"}	{"owner"}		-ENERGIE; SOURCES D'ENERGIE; AQUITAINE; LAPRA...	-Aquitaine : Tracé de l'Artère de Guyenne
Les polygones correspondent aux espaces n...	-jeu de données	-d30a3d26-b2ca-4c26-8a9b-27d595d4f4d	-	{"SYSDAU"}	{"owner"}		-OCCUPATION ET USAGE DU SOL; USAGE DES SOLS;...	-Aire Métropolitaine Bordelaise : carte de destin...
construit à partir des contours de commune...	-jeu de données	-29d27068-15d5-4b1b-84dc-b22849bdseed	-	{"SYSDAU"}	{"owner"}		-LIMITES ADMINISTRATIVES; ZONES DE GESTION; DE...	-Aire Métropolitaine Bordelaise : Périmètre du S...
orthophoto acquise pour le RIG - 2625 km2	-	-137ef98a-08af-47bc-91bf-6703b99756b8	-	{"SYSDAU"}	{"owner"}		-FONDS RASTER; ORTHO-IMAGERIE; BORDEAUX; G...	-Grande-Dordogne : Vue aérienne Estuaire-Gar...
-jeu de données	-e1c61d66-f6bf-4629-901a-4e85df9d403a	-	-	{"SYNDICAT MIXTE POUR L'AMENAGEMENT DE LA ..."}	{"owner"}		-RÉSEAUX DE TRANSPORT; LOISIRS; TOURISME; TR...	-Pays de la Vallée du Lot : Le circuit Vélo Ro...

Construction du graphe

Problème majeur :

Les métadonnées sont d'une qualité extrêmement faible :

- La champ généalogie est inexploitable à grande échelle à cause d'un champ libre saisi n'importe comment et qui peut comporter n'importe quoi.
- Les mots clés sont parfois tous énoncés dans un seul champ libre avec un séparateur choisi aléatoirement.
- Les droits d'accès (ainsi que d'autres champs) ne sont pas toujours spécifiés.
- Certains champs ont des valeurs fantaisistes : ex: "hbjghh" pour les mentions légales.
- La structure .xml peut changer pour certains IDG, etc.

Construction du graphe

Objectif :

Trouver les données les plus solides et les plus complètes afin d'établir des liens entre les fiches qui puissent répondre aux problématiques.

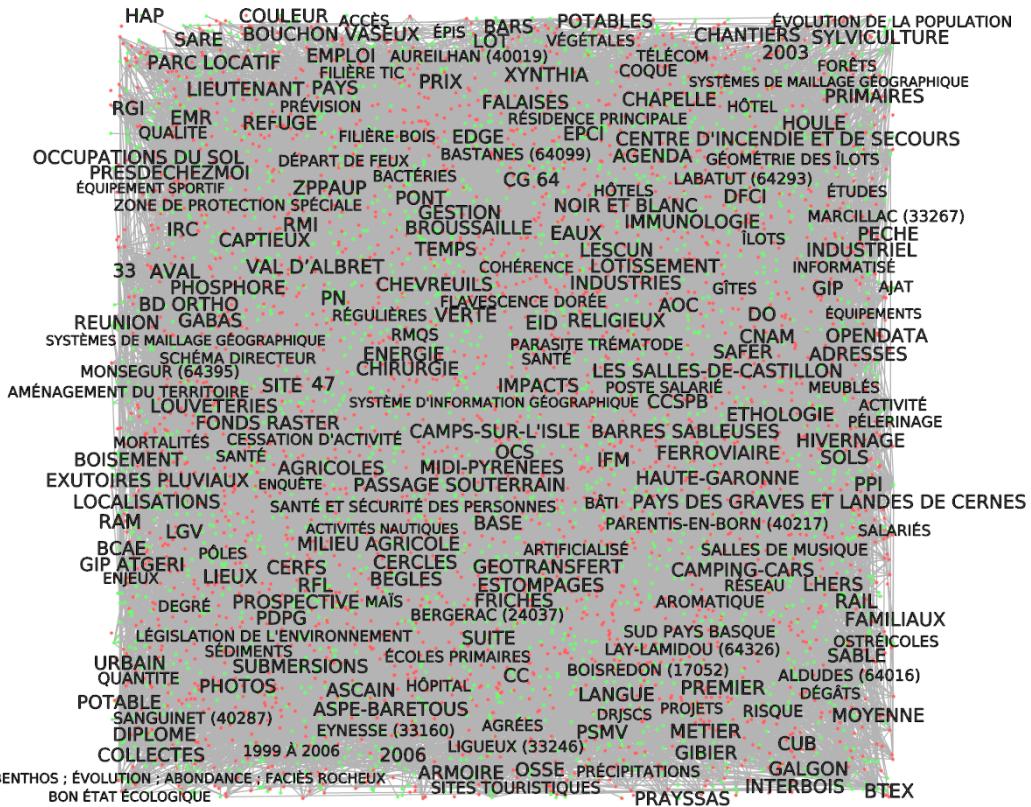
Notre choix se porte pour le moment sur les mots clés/thèmes des fiches ainsi que sur leurs acteurs.

Graphe “Keyword”

De nouveaux sommets sont créés et représentent des mots clés.

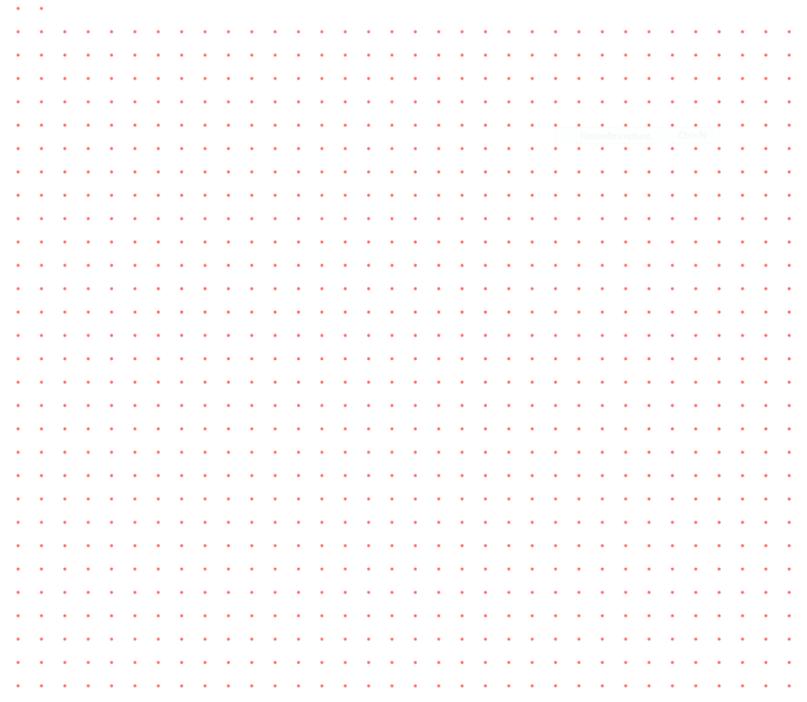
Une fiche (sommet rouge) est liée à un mot clé (sommet vert) si celui-ci est contenu dans ses métadonnées.

Le premier résultat obtenu est chaotique :



Graphe “Keyword”

En appliquant un algorithme de force (FM³ ici), on voit que de nombreuses fiches ne sont qu'à un état embryonnaire et parasitent le graphe obtenu (les sommets non connectés*) mais qu'une composante majeure semble apparaître.



S, RUMEX RUPESTRIS, FALAISES DUNAIRES, LITTORAL MÉDOCAIN

LES (MINUSCULE, ACCENTUÉ ET PLURIEL)

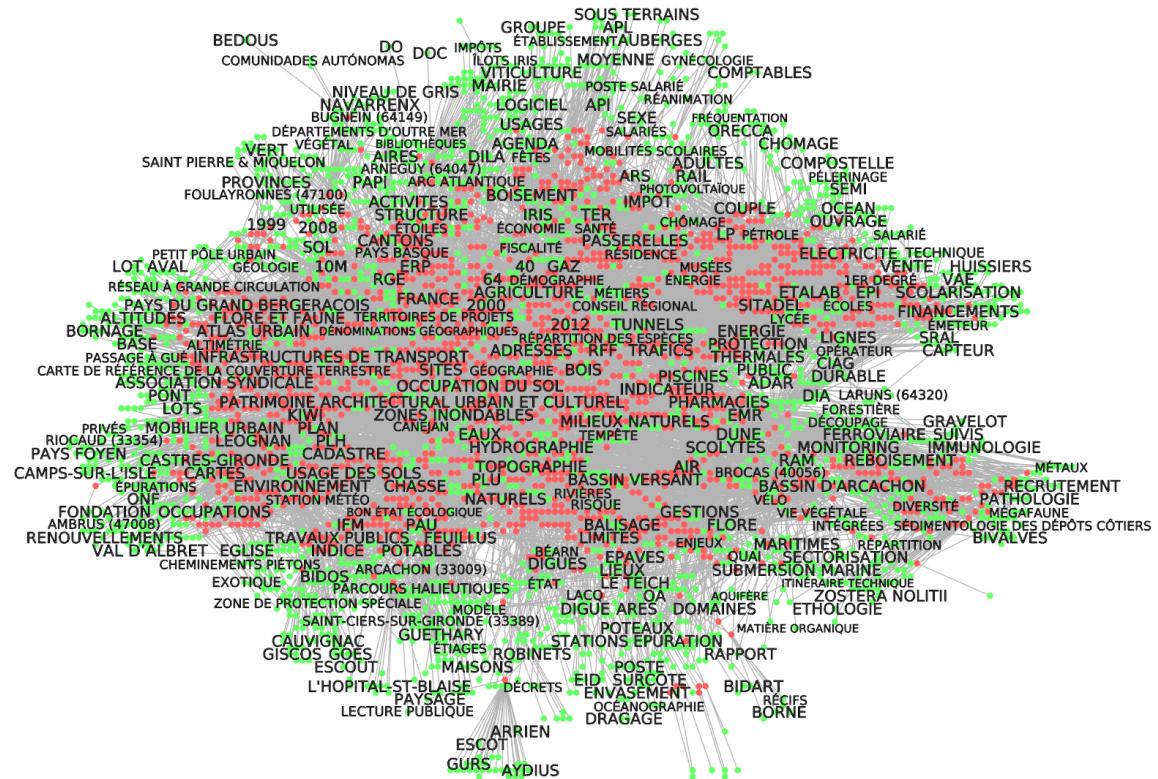
ONNÉES CARROYÉES
POTEAUX



*qui n'ont pas de liens

Graphe “Keyword”

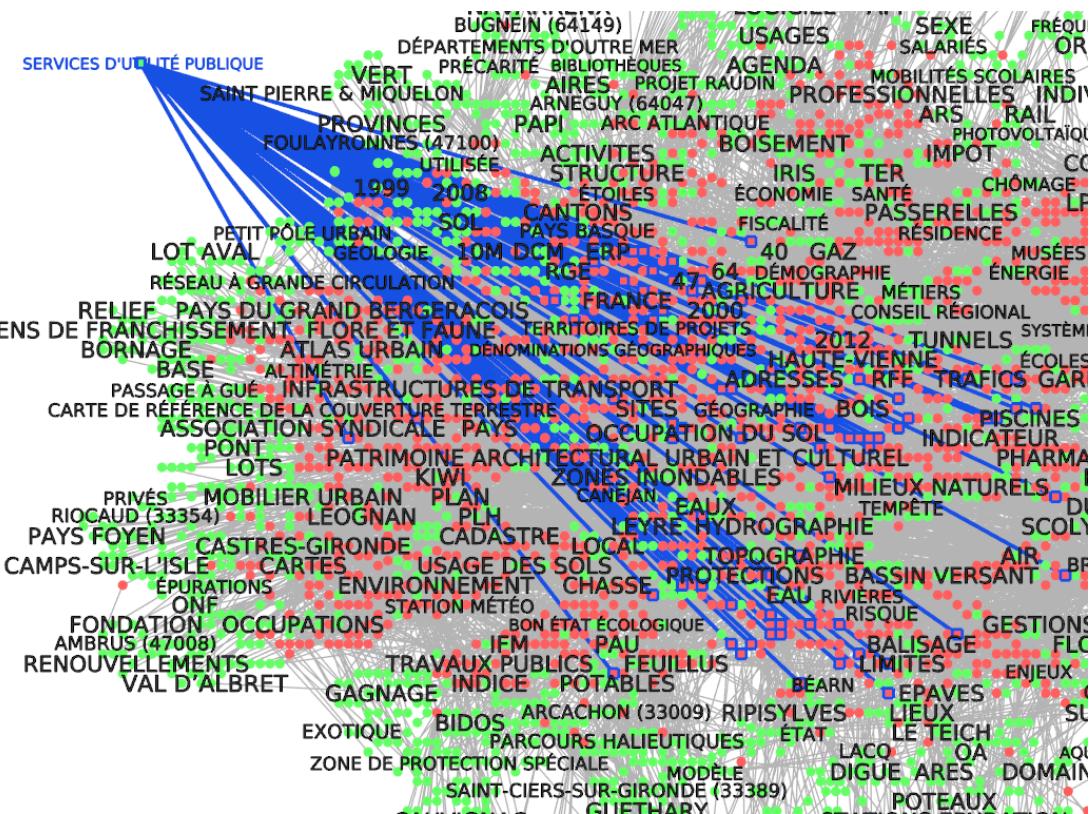
Néanmoins, si on supprime ces fiches parasites, et que l'on se concentre sur cette composante majeure, le graphe semble davantage exploitable bien que toujours difficile à lire.



Graphe “Keyword”

On peut tout de même avoir une idée de l'importance d'un thème, les fiches qui y sont liées, etc.

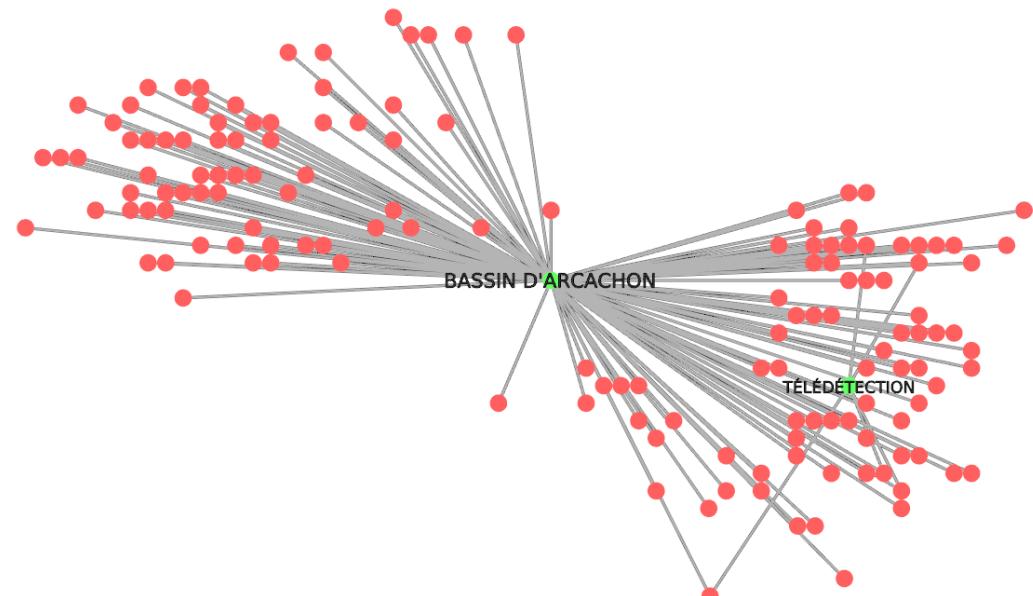
Ici, toutes les fiches relatives aux services d'utilité publique sont surlignées.



Graphe “Keyword”

De la même manière, on peut obtenir (via le module “Reachable Sub-Graph”) à partir d'une fiche choisie les fiches partageant au moins un de ses mots clés.

Toute sélection peut être isolée dans un sous-graphe pour une meilleure visualisation.



Graphe “Keyword”

Cependant, cette visualisation ne donne pas d'idée précise de communauté et n'aide pas à envisager les flux.

Graphe Similarité

Le graphe similarité :

- Le graphe est toujours basé sur les mots clés mais ne comporte que des sommets représentant des fiches '.xml'. Le graphe comporte moins d'éléments et est donc plus claire.
- Les liens sont effectués entre les fiches via un "score de similarité". Cela permet de mettre en avant des communautés de fiches par regroupement thématique.

Graphe Similarité

Un lien est établi si deux fiches présentent une similarité suffisante entre elles.

On utilise pour cela un ‘score de similarité’ :

Pour chaque paire de fiches, on calcule ce score en fonction des ‘vecteurs de similarité’ des deux fiches :

Sommets	Mots clés présents dans les fiches
1	ORTHO-IMAGERIE;; FONDS RASTER;; ORTHO;; ORTHOPHOTOGRAPHIE;; VUE AÉRIENNE;; BASSIN D'ARCACHON;; SIBA;; INSPIRE;;
2	PYRENEES-ATLANTIQUES;; FONDS RASTER;; ORTHO-IMAGERIE;; ORTHOPHOTOGRAPHIE;; ORTHO;; ASPE-BARETOUS;;
3	ORTHO-IMAGERIE;; FONDS RASTER;; BAYONNE (64102);; ORTHO;;CAMPING;; SITE WEB;;

Graphe Similarité

Paires	Mots clés communs	Score
1 2	FONDS RASTER ; ORTHO-IMAGERIE ; ORTHOPHOTOGRAPHIE ; ORTHO ;	0.57735
1 3	ORTHO-IMAGERIE;; FONDS RASTER;; ORTHO;;	0.43301
2 3	ORTHO-IMAGERIE ; FONDS RASTER ; ORTHO ;	0.54772

Exemple : Paire 1-3 :

On utilise un vecteur de 0 et de 1 dont chaque entrée correspond à un mot clé présent dans une fiche ou les deux.

- 0 : la fiche ne contient pas ce mot clé.
- 1 : la fiche contient ce mot clé.

fiche 1 : 1 1 1 1 1 1 1 1 0 0 0 → 8

fiche 3 : 1 1 1 0 0 0 0 0 1 1 1 → 6

$$\text{Score} = \frac{\sum(a^*b)}{\sqrt{(\sum a^2)} * \sqrt{(\sum b^2)}} = \frac{3}{\sqrt{8} * \sqrt{6}} = 0.43301$$

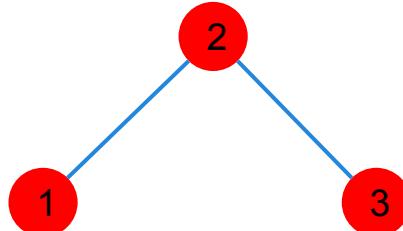
Graphe Similarité

Paires	Mots clés communs	Score
1 2	FONDS RASTER ; ORTHO-IMAGERIE ; ORTHOPHOTOGRAPHIE ; ORTHO ;	0.57735
1 3	ORTHO-IMAGERIE;; FONDS RASTER;; ORTHO;;	0.43301
2 3	ORTHO-IMAGERIE ; FONDS RASTER ; ORTHO ;	0.54772

Une fois les scores calculés, une valeur seuil est définie.

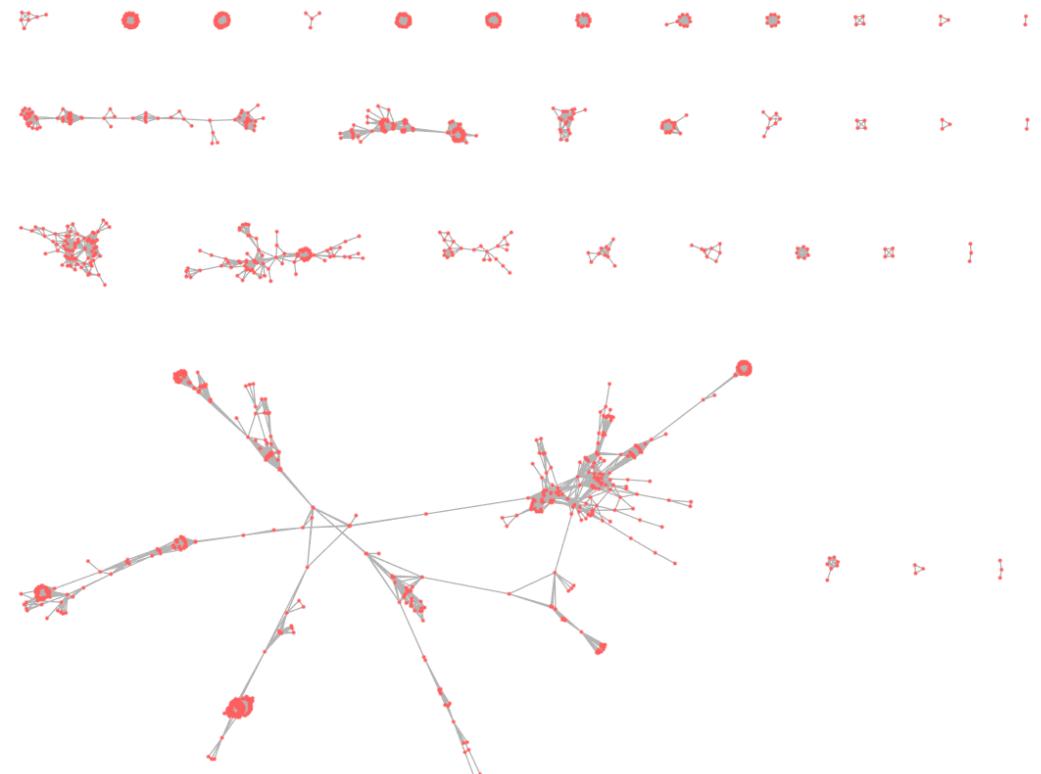
Seules les paires ayant un score supérieur au seuil auront un lien.

Ici un seuil de 0.5 :



Graphe Similarité

Des communautés se dessinent très clairement dans ce nouveau graphe



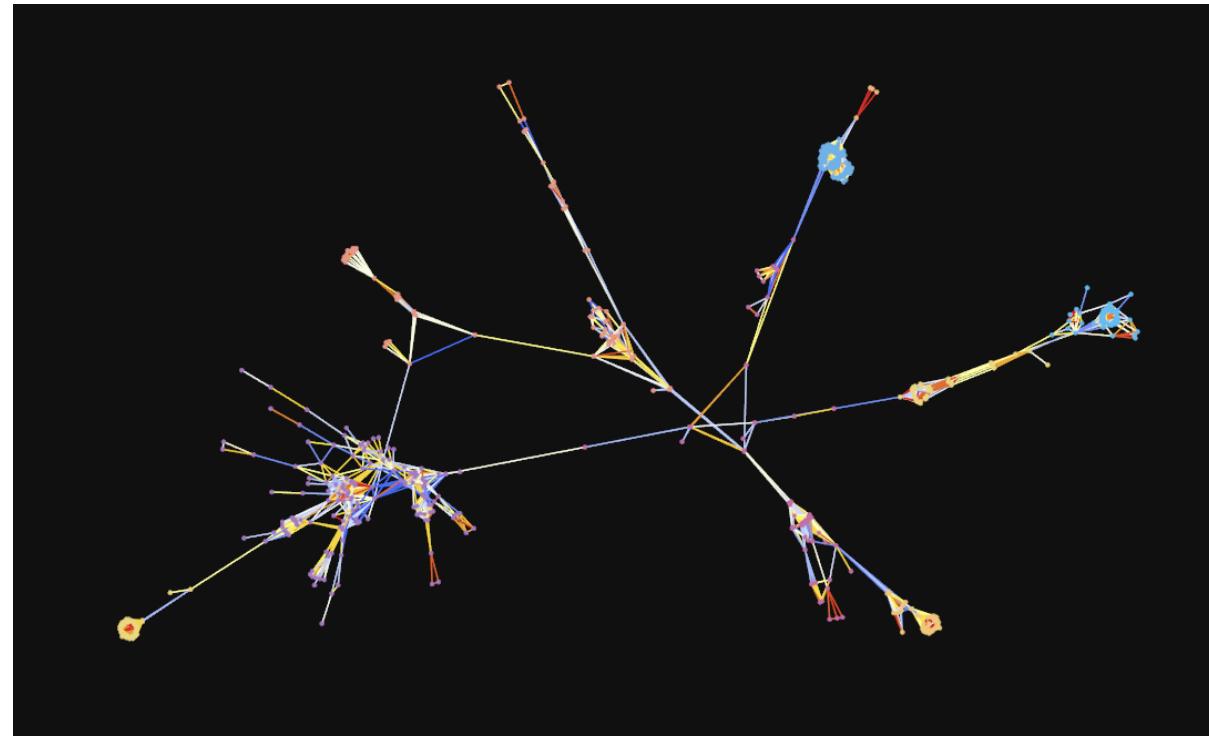
Graphe Similarité

Une nouvelle coloration est appliquée au graphe.

Plus un lien* tend vers le rouge, plus les fiches sont similaires.

Un lien qui tend vers le bleu indique donc des fiches peu similaires (très proche du seuil).

Les couleurs des sommets varient pour mettre en valeur différents communautés.

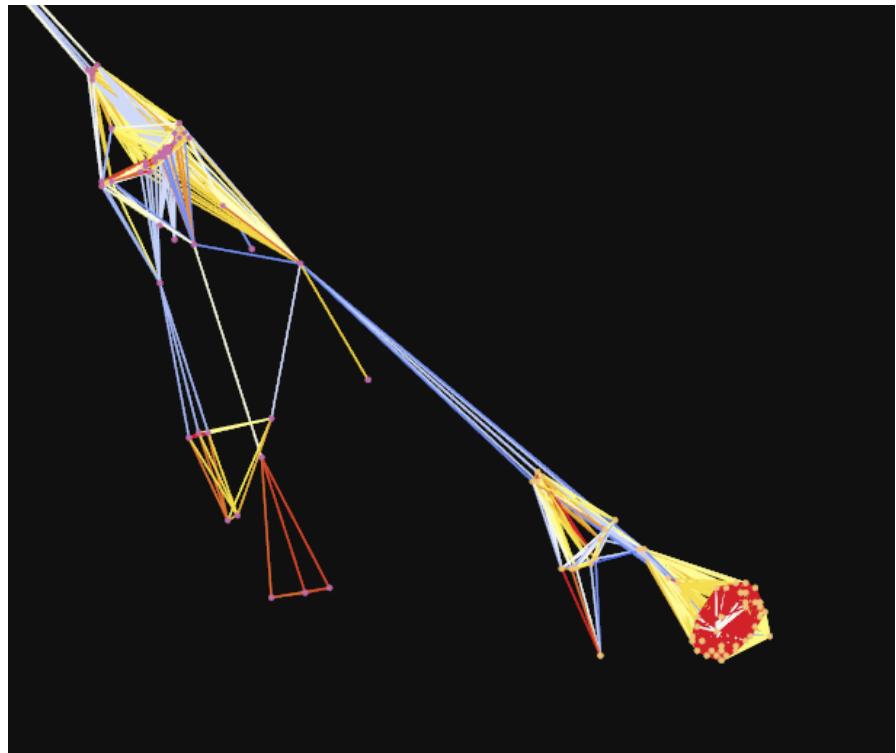


*aussi arête ou edge

Graphe Similarité

Ici on voit clairement un noyau de fiches qui correspond à des jeux de données relatifs au recensement.

La très forte similarité fait ressortir cette communauté qui est tout de même liée au reste du graphe via des liens avec des études relatives à la population et la démographie.



Graphe Similarité

Le graphe Similarité permet, contrairement au graphe “Keyword”, de mettre en valeur des liens révélateurs de communautés.

Même s'il ne s'agit pas directement de généalogie, il permet néanmoins de voir des groupements thématiques pouvant potentiellement abriter une parenté.

Graphe Acteur

Le graphe acteur est un autre graphe biparti qui isole les acteurs et montre les jeux de données qui y sont liés.



Graphe Acteur



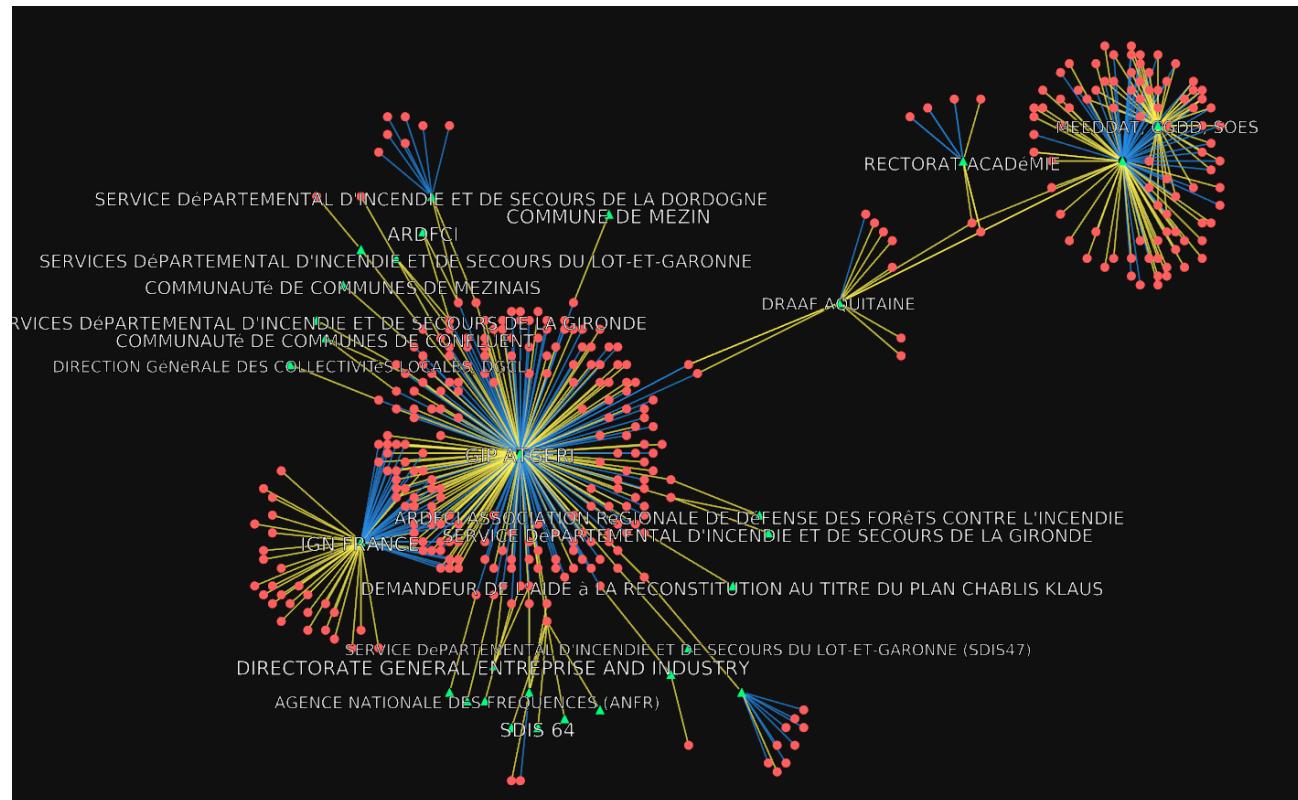
fiche de métadonnées



acteur

Lien jaune : l'acteur est le propriétaire de la fiche ("owner").

Lien bleu : l'acteur est un contact ("PointOfContact").



Graphe Acteur

On peut ainsi facilement déterminer l'importance d'un acteur et l'étendue de sa portée. Cependant ...

Problème :

On perd les communautés de fiches déterminées par les autres graphes qui pourraient permettre de savoir quelle est la spécialité des acteurs et où ils sont majoritairement intervenus.

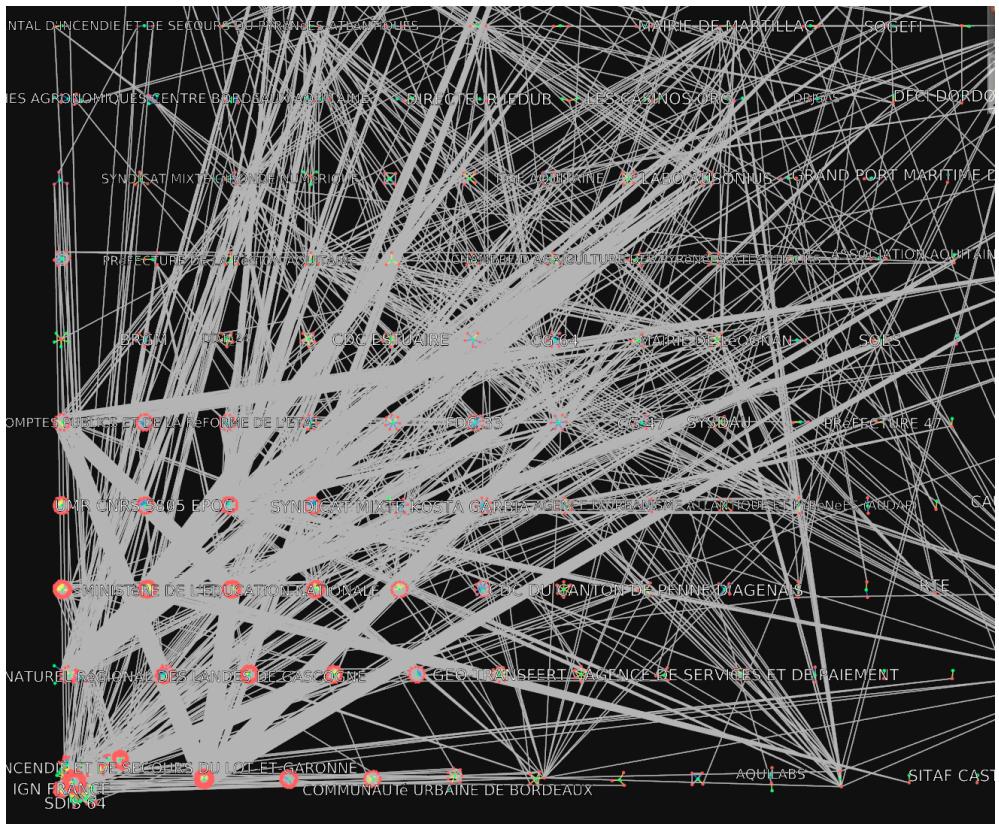
Graphe Hybride (acteur/similarité)

Solution potentielle :

Coupler les deux graphes précédents.

Résultat obtenu :

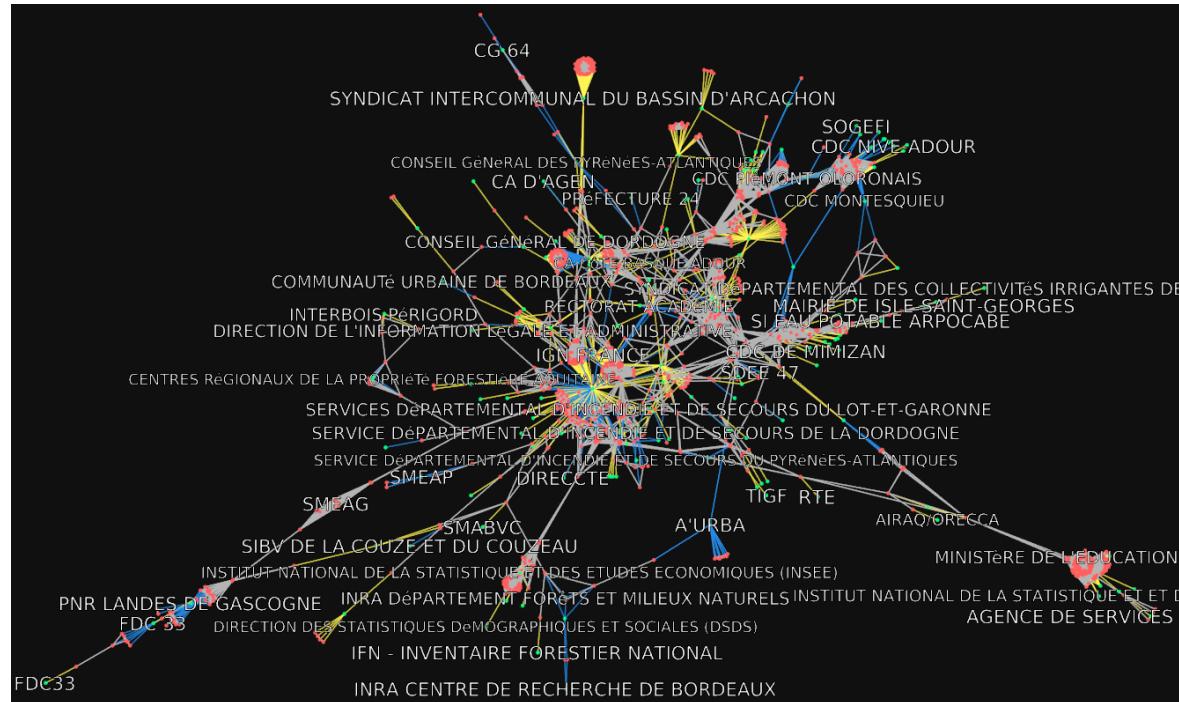
Graphe très dense et qui semble difficile à exploiter.



Graphe Hybride (acteur/similarité)

Malgré une présentation améliorée, ce graphe reste toujours difficile à lire simplement sans utiliser des méthodes d'analyses avancées.

Il est cependant possible d'obtenir des graphes pertinents en hybridant les types de donnée.

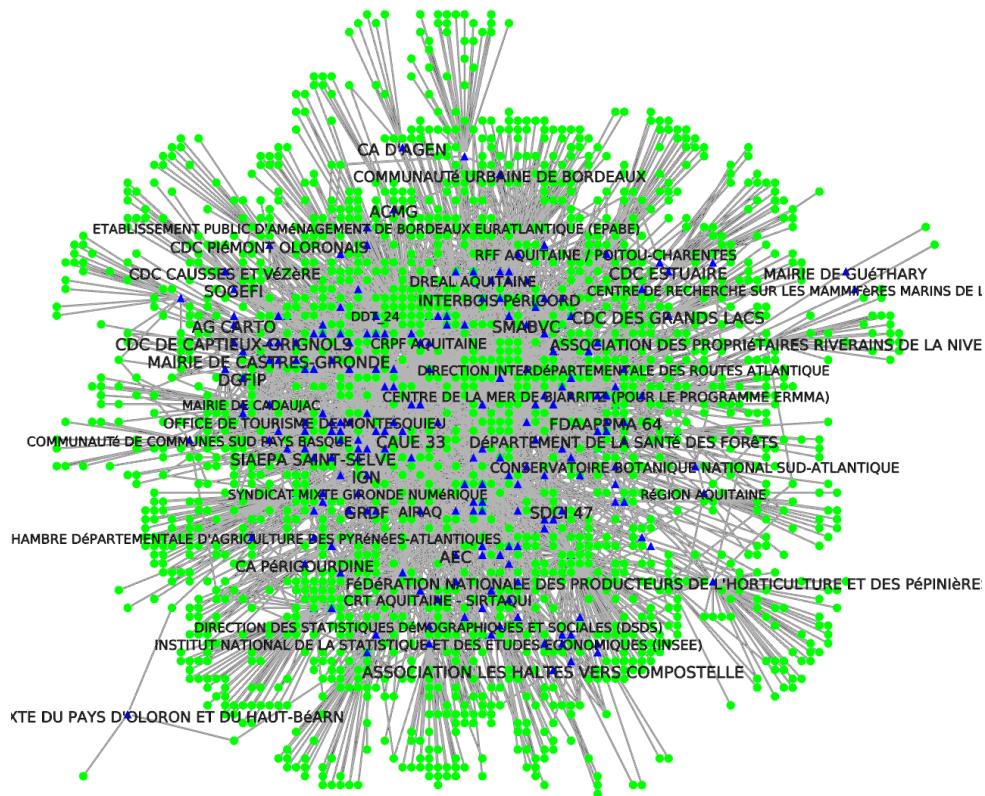


Graphes Hybrides (acteur/"keyword")

Alternatives essayées :

Ici un acteur (sommets bleus) est lié à un mot-clé (sommets verts) lorsqu'une fiche possédant le mot-clé implique cet acteur.

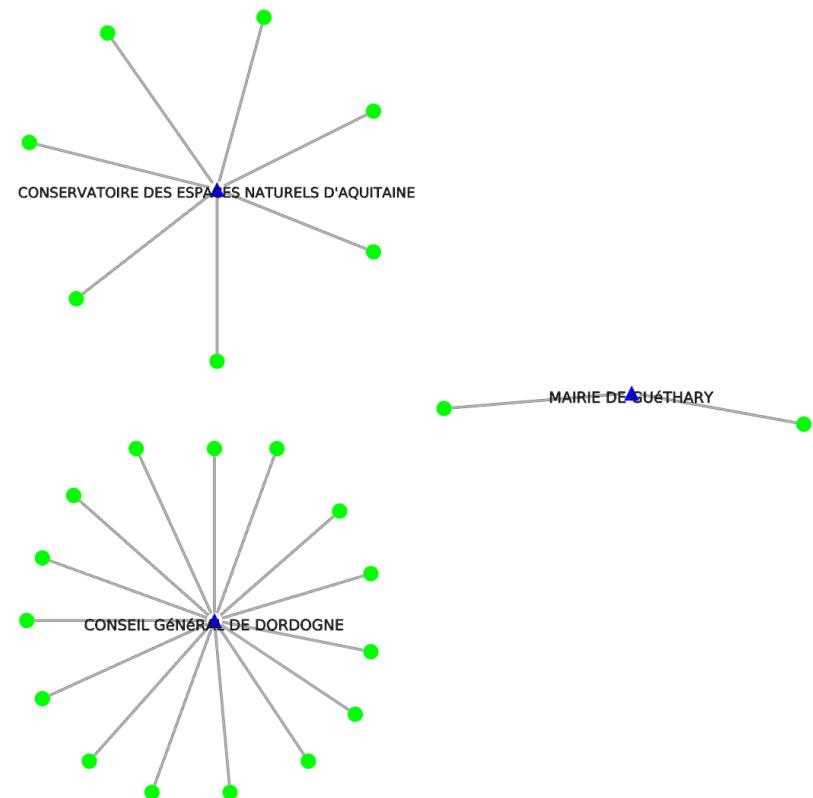
Le même phénomène est visible que pour le graphe Keyword. L'exploitation du graphe est difficile à cause du fort couplage entre les acteurs.



Graphes Hybrides (acteur/"keyword")

Alternatives essayées :

Isoler un ou plusieurs acteurs au sein du graphe peut cependant offrir une analyse intéressante des thèmes susceptibles d'être couverts par ces acteurs.

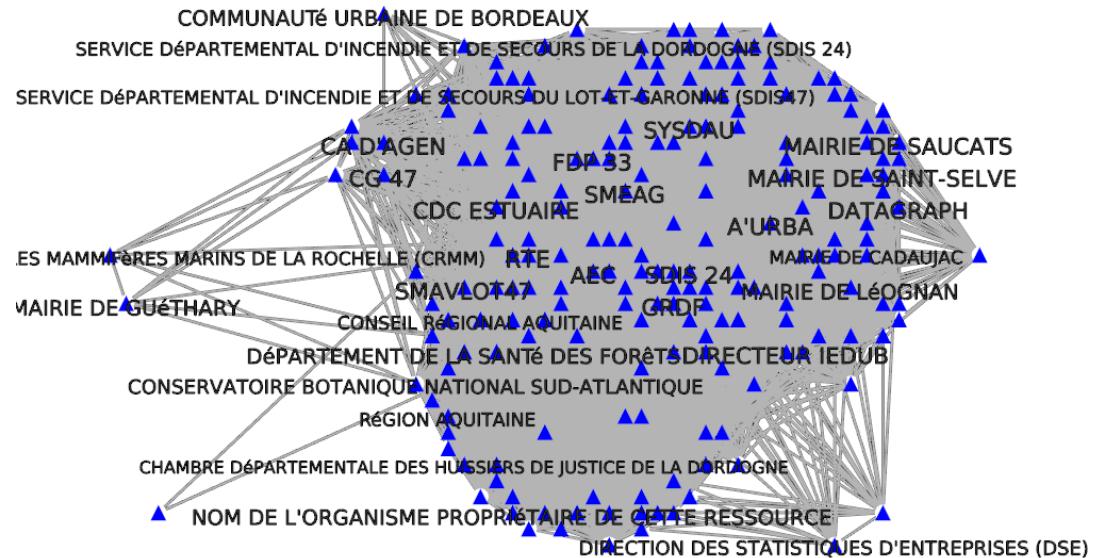


Graphes Hybrides (acteur/"keyword")

Alternatives essayées :

Ici les acteurs sont liés s'ils partagent un mot-clé dans leurs fiches respectives. C'est une version alternative du graphe précédent qui ne comportent que des sommets acteurs.

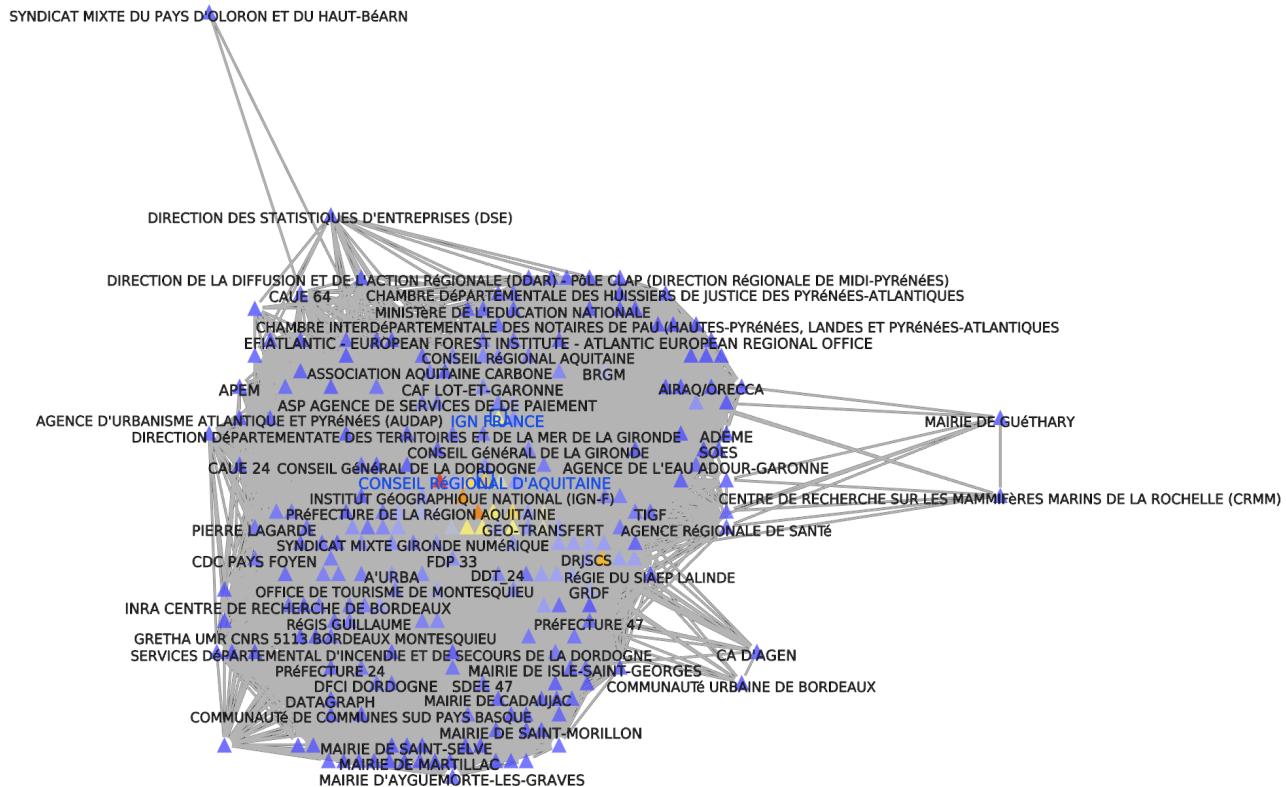
Le couplage fort rend l'analyse visuelle difficile mais des traitements restent possibles.



Graphes Hybrides (acteur/"keyword")

Alternatives essayées :

Ici un calcul de centralité permet de mettre en évidence des acteurs majeurs (couleurs chaudes) comme IGN France ou le conseil régional d'Aquitaine.



Graphes Hybrides (acteur/"keyword")

Alternatives essayées :

Cela permet de voir que ces acteurs partagent des thèmes avec presque l'intégralité des acteurs de PIGMA.



Graphes Emprise

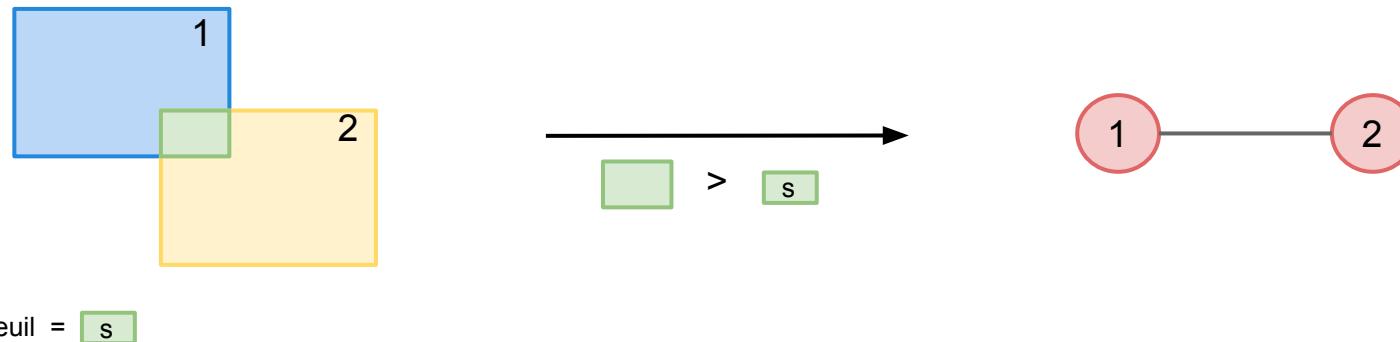
Autres pistes :

Il est aussi possible de créer plusieurs modèles de graphe basés sur les emprises spatiales afin d'avoir un modèle mathématique issu du modèle géographique.

Graphes Emprise

Modèle 1 :

Si les emprises spatiales de deux fiches se recouvrent, on calcule la surface commune entre les deux emprises. Si elle excède un certain seuil, alors on crée un lien entre ces deux fiches.



Cela permettrait de mettre en avant les
recouvrements et les zones de forte densité de
couverture.

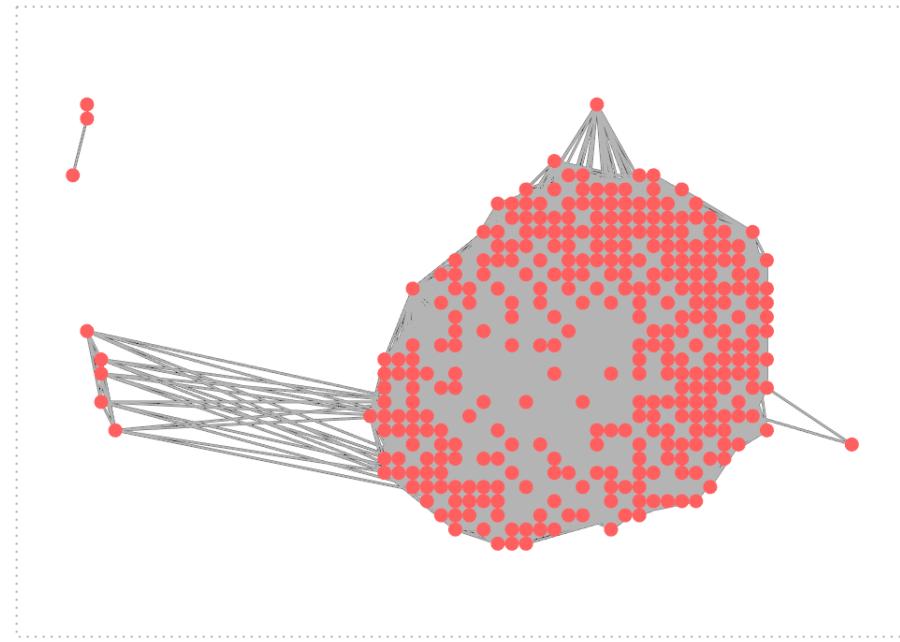
Graphes Emprise

Modèle 1 :

Problème : le recouvrement semble tellement élevé que le graphe est extrêmement dense si on n'utilise pas le seuil.

Solutions potentielles :

- Augmenter le seuil
- Appliquer un masque d'emprise



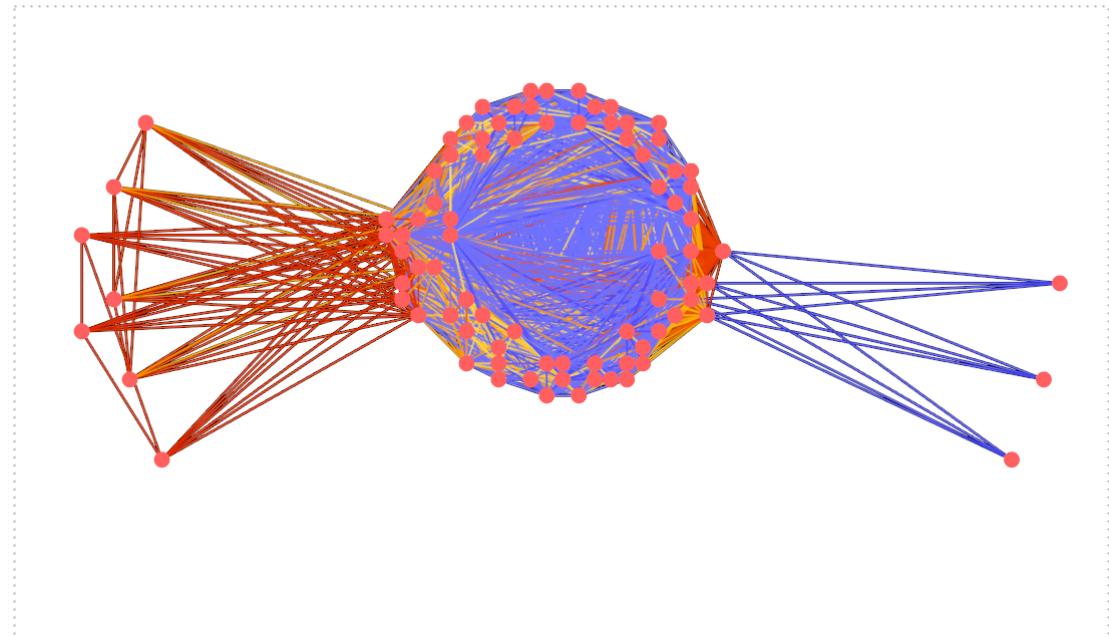
Exemple du graphe d'emprise avec un seuil nul

Graphes Emprise

Modèle 1 :

Augmenter le seuil va diminuer la taille de la composante principale et augmenter la pertinence du résultat obtenu.

Ici la couleur des liens représente la surface commune entre deux fiches. Un lien rouge signifiera que leur surface commune est élevée.

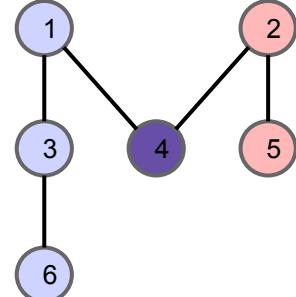


Graphes Emprise

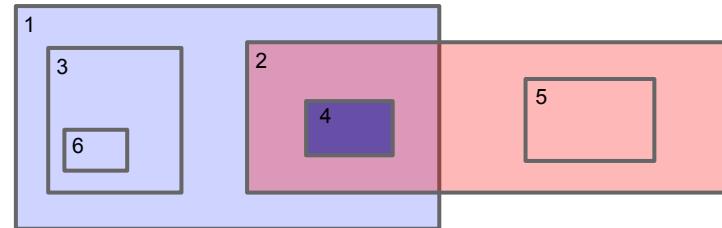
Modèle 2 : (en cours)

Ce modèle est basé sur les inclusions :

Si la surface d'une fiche B est incluse dans celle d'une fiche A alors la fiche B est la fille de la fiche A.



- 6 est inclus dans 3 qui est inclus dans 1.
- 5 est inclus dans 2.
- 4 est inclus à la fois dans 1 et 2 donc 1 et 2 se chevauchent.



Exploitation Généalogie

Le champ généalogie est le champ le plus difficilement exploitable des fiches de métadonnées :

- Il est en texte libre
- Il peut indiquer potentiellement n'importe quoi : l'échelle des données, la description des données, la méthodologie d'acquisition des données, la date des données, le type des données, etc.

Cependant, on peut essayer d'extraire les informations pertinentes en déterminant les mots importants du champ généalogie. On pourrait alors établir un graphe comme le graphe “Keyword” ou Similarité.

Exploitation Généalogie

On utilise TF-IDF (*Term Frequency-Inverse Document Frequency*) :

On calcule un score de pertinence pour chaque mot au sein d'un corpus de documents :

$$\text{Score}(\text{"mot"}) = \text{TF} * \text{IDF}$$

$$\text{TF} = \frac{\text{Nombre de fois où "mot" apparaît dans le document}}{\text{Nombre de mots dans le document}}$$

$$\text{IDF} = \log \frac{\text{Nombre de documents}}{\text{Nombre de documents où "mot" apparaît}}$$

Les mots blancs ne sont pas conservés (le, la, de, un, a, etc.).

On conserve les trois mots ayant les meilleurs scores par fiche pour opérer un traitement comme les graphes “Keyword” et Similarité avec ces mots à la place des “Keyword”.