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. Comment agissent les acteurs au sein des IDG et de quelle manière ?

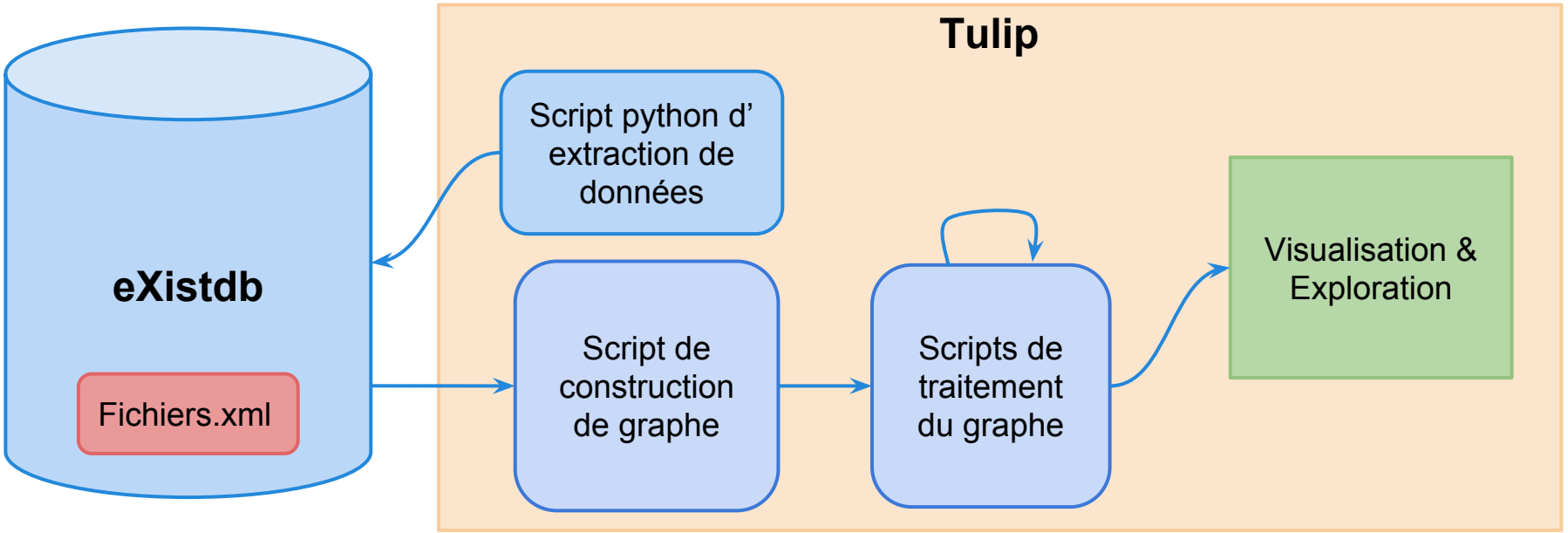
Procédure Actuelle

Prérequis :

- Données sous forme .xml
- EXistDB installé (nécessite Java JRE >= 7)
- Python 2.7.6 installé
- Tulip 4.6.1 installé

Procédure Actuelle

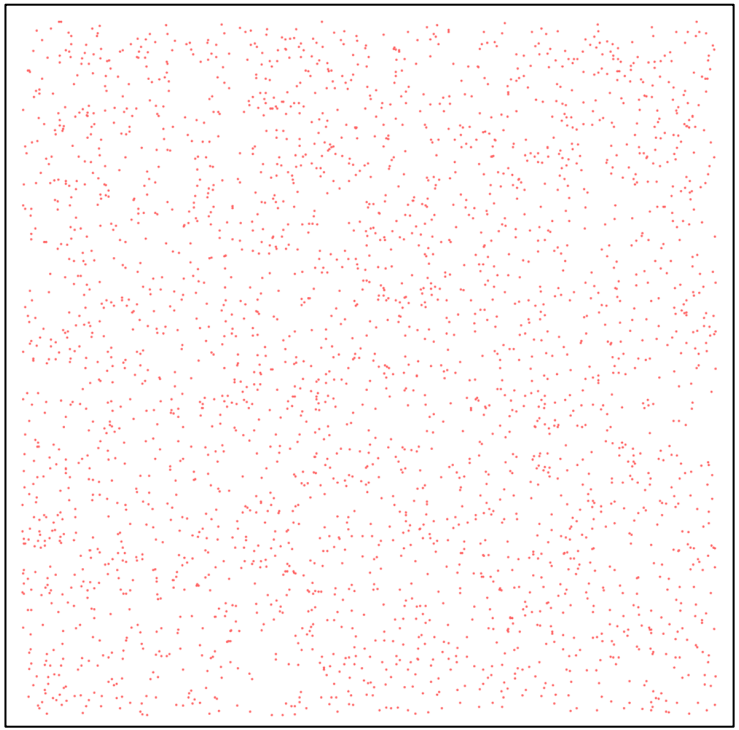
Procédure : Mettre les .xml dans eXistdb puis lancer la suite de scripts dans Tulip



Construction du graphe

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

Ici, les 2835 nodes de PIGMA :

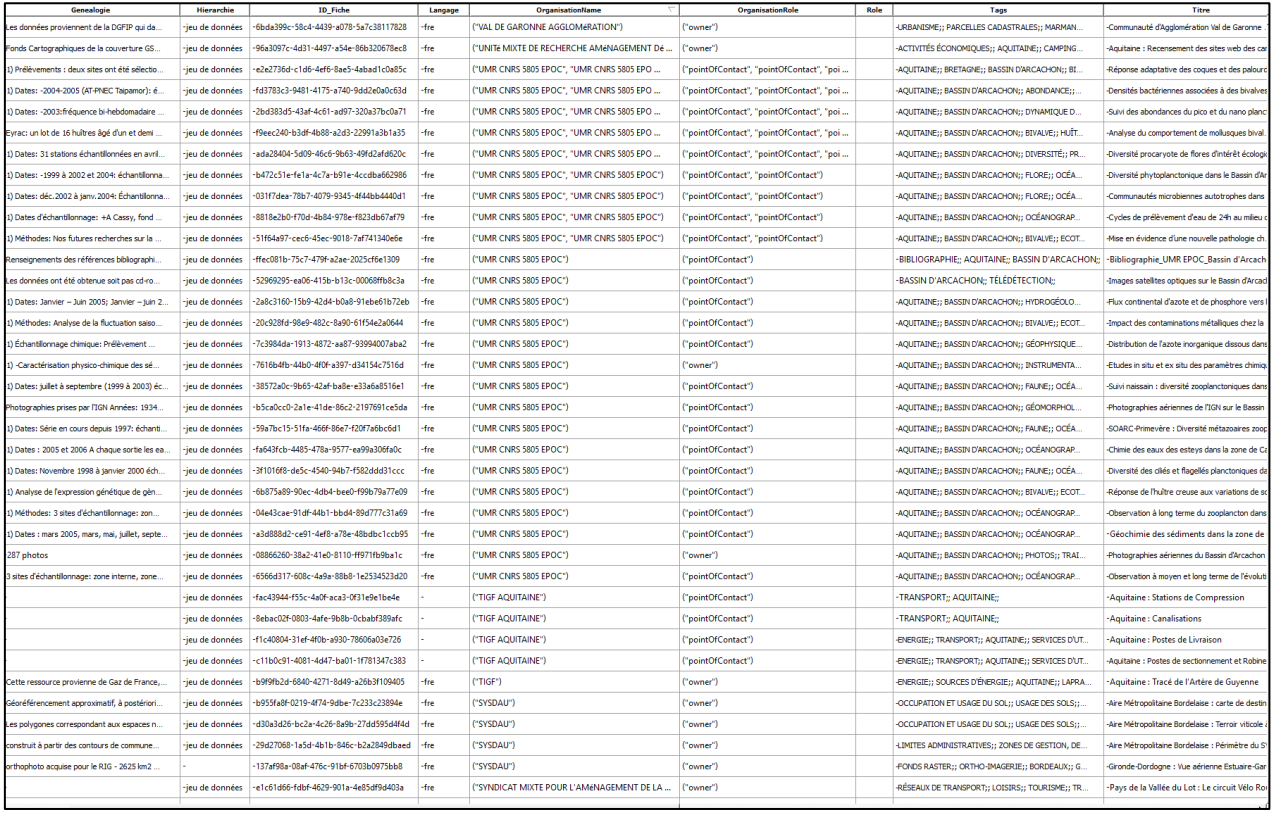


*Ou sommets. Ici les points rouges.

Construction du graphe

- Chaque node possède les informations contenues dans la fiche de métadonnée correspondante.

- Ces informations sont accessibles directement via la visualisation, la spreadsheet, etc.



Genealogie	Hierarchie	ID_Fiche	Langage	OrganisationName	OrganisationRole	Role	Tags	Titre
Les données proviennent de la DGFiP qui da...	jeu de données	6bda399c-58c4-4439-a078-5a7c38117828	fre	("VAL DE GARONNE AGGLOMéRATION")	("owner")		URBANISME;; PARCELLES CADASTRALES;; MARMAN...	Communauté d'Agglomération Val de Garonne
Fonds Cartographiques de la couverture GS...	jeu de données	96a3097c-4d31-4497-a54e-86b320678ec8	fre	("UNITé MIXTE DE RECHERCHE AMéNAGEMENT De ...	("owner")		ACTIVITÉS ÉCONOMIQUES;; AQUITAINE;; CAMPING...	Aquitaine : Recensement des sites web des ca
1) Prélèvements : deux sites ont été sélectio...	jeu de données	e2e2736d-c1d6-4ef6-8ae5-4abad1c0a85c	fre	("UMR CNRS 5805 EPOC", "UMR CNRS 5805 EPO ...	("pointOfContact", "pointOfContact", "poi ...		AQUITAINE;; BRETAGNE;; BASSIN D'ARCACHON;; BI...	Réponse adaptative des coques et des palourc
1) Dates: -2004-2005 (AT-PNEC Taipamor): é...	jeu de données	fd3783c3-9481-4175-a740-9dd2e0a0c63d	fre	("UMR CNRS 5805 EPOC", "UMR CNRS 5805 EPO ...	("pointOfContact", "pointOfContact", "poi ...		AQUITAINE;; BASSIN D'ARCACHON;; ABONDANCE;;...	Densités bactériennes associées à des bivalves
1) Dates: -2003:fréquence bi-hebdomadaire ...	jeu de données	2bd383d5-43af-4c61-ad97-320a37bc0a71	fre	("UMR CNRS 5805 EPOC", "UMR CNRS 5805 EPO ...	("pointOfContact", "pointOfContact", "poi ...		AQUITAINE;; BASSIN D'ARCACHON;; DYNAMIQUE D...	Suivi des abondances du pico et du nano planc
Eyrac: un lot de 16 huîtres âgé d'un et demi ...	jeu de données	f9eec240-b3df-4b88-a2d3-22991a3b1a35	fre	("UMR CNRS 5805 EPOC", "UMR CNRS 5805 EPO ...	("pointOfContact", "pointOfContact", "poi ...		AQUITAINE;; BASSIN D'ARCACHON;; BIVALVE;; HUÎT...	Analyse du comportement de mollusques bival
1) Dates: 31 stations échantillonnées en avril...	jeu de données	ada28404-5d09-46c6-9b63-49fd2afd620c	fre	("UMR CNRS 5805 EPOC", "UMR CNRS 5805 EPO ...	("pointOfContact", "pointOfContact", "poi ...		AQUITAINE;; BASSIN D'ARCACHON;; DIVERSITÉ;; PR...	Diversité procaryote de flores d'intérêt écologi
1) Dates: -1999 à 2002 et 2004: échantillonna...	jeu de données	b472c51e-fe1a-4c7a-b91e-4ccdba662986	fre	("UMR CNRS 5805 EPOC", "UMR CNRS 5805 EPOC")	("pointOfContact", "pointOfContact")		AQUITAINE;; BASSIN D'ARCACHON;; FLORE;; OCÉA...	Diversité phytoplanctonique dans le Bassin d'Ar
1) Dates: déc.2002 à janv.2004: échantillonna...	jeu de données	031f7dea-78b7-4079-9345-4f44bb4440d1	fre	("UMR CNRS 5805 EPOC", "UMR CNRS 5805 EPOC")	("pointOfContact", "pointOfContact")		AQUITAINE;; BASSIN D'ARCACHON;; FLORE;; OCÉA...	Communautés microbiennes autotrophes dans
1) Dates d'échantillonnage: +A Cassy, fond ...	jeu de données	8818e2b0-f70d-4b84-978e-f823db67af79	fre	("UMR CNRS 5805 EPOC", "UMR CNRS 5805 EPOC")	("pointOfContact", "pointOfContact")		AQUITAINE;; BASSIN D'ARCACHON;; OCÉANOGRAP...	Cycles de prélèvement d'eau de 24h au milieu d
1) Méthodes: Nos futures recherches sur la ...	jeu de données	51f64a97-cec6-45ec-9018-7af741340e6e	fre	("UMR CNRS 5805 EPOC", "UMR CNRS 5805 EPOC")	("pointOfContact", "pointOfContact")		AQUITAINE;; BASSIN D'ARCACHON;; BIVALVE;; ECOT...	Mise en évidence d'une nouvelle pathologie ch
Renseignements des références bibliographi...	jeu de données	ffec081b-75c7-479f-a2ae-2025cf6e1309	fre	("UMR CNRS 5805 EPOC")	("pointOfContact")		BIBLIOGRAPHIE;; AQUITAINE;; BASSIN D'ARCACHON;;	Bibliographie_UMR EPOC_Bassin d'Arcach
Les données ont été obtenue soit pas cd-ro...	jeu de données	52969295-ea06-415b-b13c-00068ffb8c3a	fre	("UMR CNRS 5805 EPOC")	("pointOfContact")		BASSIN D'ARCACHON;; TÉLÉDÉTECTION;;	Images satellites optiques sur le Bassin d'Arcad
1) Dates: Janvier – Juin 2005; Janvier – Juin 2...	jeu de données	2a8c3160-15b9-42d4-b0a8-91ebe61b72eb	fre	("UMR CNRS 5805 EPOC")	("pointOfContact")		AQUITAINE;; BASSIN D'ARCACHON;; HYDROGÉOLO...	Flux continental d'azote et de phosphore vers l
1) Méthodes: Analyse de la fluctuation saiso...	jeu de données	20c928fd-98e9-482c-8a90-61f54e2a0644	fre	("UMR CNRS 5805 EPOC")	("pointOfContact")		AQUITAINE;; BASSIN D'ARCACHON;; BIVALVE;; ECOT...	Impact des contaminations métalliques chez la
1) Échantillonnage chimique: Prélèvement ...	jeu de données	7c3984da-1913-4872-aa87-93994007aba2	fre	("UMR CNRS 5805 EPOC")	("pointOfContact")		AQUITAINE;; BASSIN D'ARCACHON;; GÉOPHYSIQUE...	Distribution de l'azote inorganique dissous dans
1) -Caractérisation physico-chimique des sé...	jeu de données	7616b4fb-44b0-4f0f-a397-d34154c7516d	fre	("UMR CNRS 5805 EPOC")	("owner")		AQUITAINE;; BASSIN D'ARCACHON;; INSTRUMENTA...	Etudes in situ et ex situ des paramètres chimiq
1) Dates: juillet à septembre (1999 à 2003) éc...	jeu de données	38572a0c-9b65-42af-ba8e-e33a6a8516e1	fre	("UMR CNRS 5805 EPOC")	("pointOfContact")		AQUITAINE;; BASSIN D'ARCACHON;; FAUNE;; OCÉA...	Suivi naissain : diversité zooplanctoniques dans
Photographies prises par l'IGN Années: 1934...	jeu de données	b5ca0cc0-2a1e-41de-86c2-2197691ce5da	fre	("UMR CNRS 5805 EPOC")	("pointOfContact")		AQUITAINE;; BASSIN D'ARCACHON;; GÉOMORPHOL...	Photographies aériennes de l'IGN sur le Bassin
1) Dates: Série en cours depuis 1997: échanti...	jeu de données	59a7bc15-51fa-466f-86e7-f20f7a6bc6d1	fre	("UMR CNRS 5805 EPOC")	("pointOfContact")		AQUITAINE;; BASSIN D'ARCACHON;; FAUNE;; OCÉA...	SOARC-Primevère : Diversité métazoaires zoop
1) Dates : 2005 et 2006 A chaque sortie les ea...	jeu de données	fa643fcb-4485-478a-9577-ea99a306fa0c	fre	("UMR CNRS 5805 EPOC")	("pointOfContact")		AQUITAINE;; BASSIN D'ARCACHON;; OCÉANOGRAP...	Chimie des eaux des esteys dans la zone de Cz
1) Dates: Novembre 1998 à janvier 2000 éch...	jeu de données	3f1016f8-de5c-4540-94b7-f582ddd31ccc	fre	("UMR CNRS 5805 EPOC")	("pointOfContact")		AQUITAINE;; BASSIN D'ARCACHON;; FAUNE;; OCÉA...	Diversité des ciliés et flagellés planctoniques de
1) Analyse de l'expression génétique de gèn...	jeu de données	6b875a89-90ec-4db4-bee0-f99b79a77e09	fre	("UMR CNRS 5805 EPOC")	("pointOfContact")		AQUITAINE;; BASSIN D'ARCACHON;; BIVALVE;; ECOT...	Réponse de l'huître creuse aux variations de sc
1) Méthodes: 3 sites d'échantillonnage: zon...	jeu de données	04e43cae-91df-44b1-bbd4-89d777c31a69	fre	("UMR CNRS 5805 EPOC")	("pointOfContact")		AQUITAINE;; BASSIN D'ARCACHON;; OCÉANOGRAP...	Observation à long terme du zooplancton dans
1) Dates : mars 2005, mars, mai, juillet, septe...	jeu de données	a3d888d2-ce91-4ef8-a78e-48bdbc1ccb95	fre	("UMR CNRS 5805 EPOC")	("pointOfContact")		AQUITAINE;; BASSIN D'ARCACHON;; OCÉANOGRAP...	Géochimie des sédiments dans la zone de
287 photos	jeu de données	08866260-38a2-41e0-8110-ff971fb9ba1c	fre	("UMR CNRS 5805 EPOC")	("owner")		AQUITAINE;; BASSIN D'ARCACHON;; PHOTOS;; TRAI...	Photographies aériennes du Bassin d'Arcachon
3 sites d'échantillonnage: zone interne, zone...	jeu de données	6566d317-608c-4a9a-88b8-1e2534523d20	fre	("UMR CNRS 5805 EPOC")	("pointOfContact")		AQUITAINE;; BASSIN D'ARCACHON;; OCÉANOGRAP...	Observation à moyen et long terme de l'évolut
	jeu de données	fac43944-f55c-4a0f-aca3-0f31e9e1be4e	-	("TIGF AQUITAINE")	("pointOfContact")		TRANSPORT;; AQUITAINE;;	Aquitaine : Stations de Compression
	jeu de données	8ebac02f-0803-4afe-9b8b-0cbabf389afc	-	("TIGF AQUITAINE")	("pointOfContact")		TRANSPORT;; AQUITAINE;;	Aquitaine : Canalisations
	jeu de données	f1c40804-31ef-4f0b-a930-78606a03e726	-	("TIGF AQUITAINE")	("pointOfContact")		ENERGIE;; TRANSPORT;; AQUITAINE;; SERVICES D'UT...	Aquitaine : Postes de Livraison
	jeu de données	c11b0c91-4081-4d47-ba01-1f781347c383	-	("TIGF AQUITAINE")	("pointOfContact")		ENERGIE;; TRANSPORT;; AQUITAINE;; SERVICES D'UT...	Aquitaine : Postes de sectionnement et Robine
Cette ressource provienne de Gaz de France,...	jeu de données	b9f9fb2d-6840-4271-8d49-a26b3f109405	fre	("TIGF")	("owner")		ENERGIE;; SOURCES D'ÉNERGIE;; AQUITAINE;; LARRA...	Aquitaine : Tracé de l'Artère de Guyenne
Géoréférencement approximatif, à postériori...	jeu de données	b955fa8f-0219-4f74-9dbe-7c233c23894e	fre	("SYSDAU")	("owner")		OCCUPATION ET USAGE DU SOL;; USAGE DES SOLS;;...	Aire Métropolitaine Bordelaise : carte de destin
Les polygones correspondant aux espaces n...	jeu de données	d30a3d26-bc2a-4c26-8a9b-27dd595d4f4d	fre	("SYSDAU")	("owner")		OCCUPATION ET USAGE DU SOL;; USAGE DES SOLS;;...	Aire Métropolitaine Bordelaise : Terroir viticole
construit à partir des contours de commune...	jeu de données	29d27068-1a5d-4b1b-846c-b2a2849dbaed	fre	("SYSDAU")	("owner")		LIMITES ADMINISTRATIVES;; ZONES DE GESTION, DE...	Aire Métropolitaine Bordelaise : Périmètre du S
orthophoto acquise pour le RIG - 2625 km2 ...	-	137af98a-08af-476c-91bf-6703b0975bb8	fre	("SYSDAU")	("owner")		FONDS RASTER;; ORTHO-IMAGERIE;; BORDEAUX;; G...	Gironde-Dordogne : Vue aérienne Estuaire-Gar
	jeu de données	e1c61d66-fdbf-4629-901a-4e85df9d403a	fre	("SYNDICAT MIXTE POUR L'AMéNAGEMENT 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 données 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 comporte n’importe quoi.
- Les mots clés sont parfois tous énoncés dans un seul champ libre avec un séparateur choisi aléatoirement.
- Les droit d'accès ne sont pas toujours spécifiés.
- 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 portent pour le moment sur les mots clés/thèmes des fiches ainsi que sur leurs acteurs.

Graphe Keyword

De nouveaux nodes sont crées et représentent des mots clés.

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

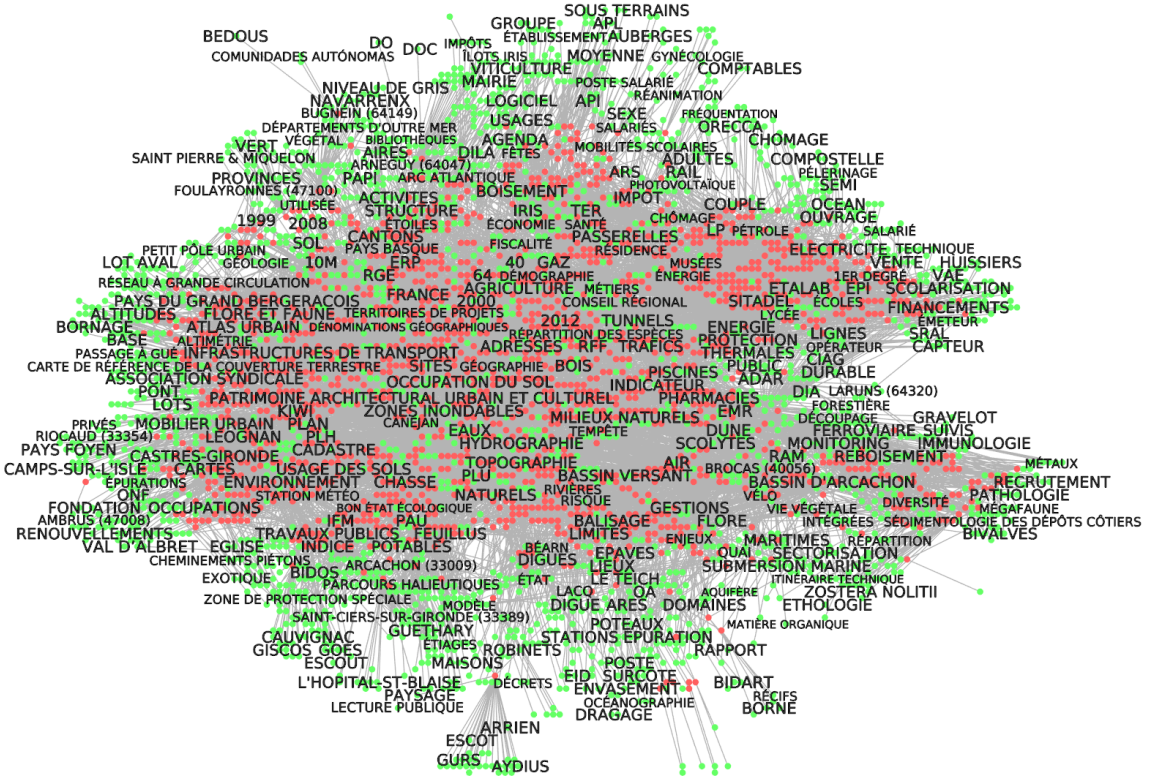
Le premier résultat obtenu est chaotique :

Graphe Keyword

En appliquant un algorithme de force (FM^3 ici), on voit que de nombreuses fiches ne sont qu'à un état embryonnaire et parasitent le graphe obtenu (les nodes non connectés) mais qu'une composante majeure semble apparaître.

Graphe Keyword

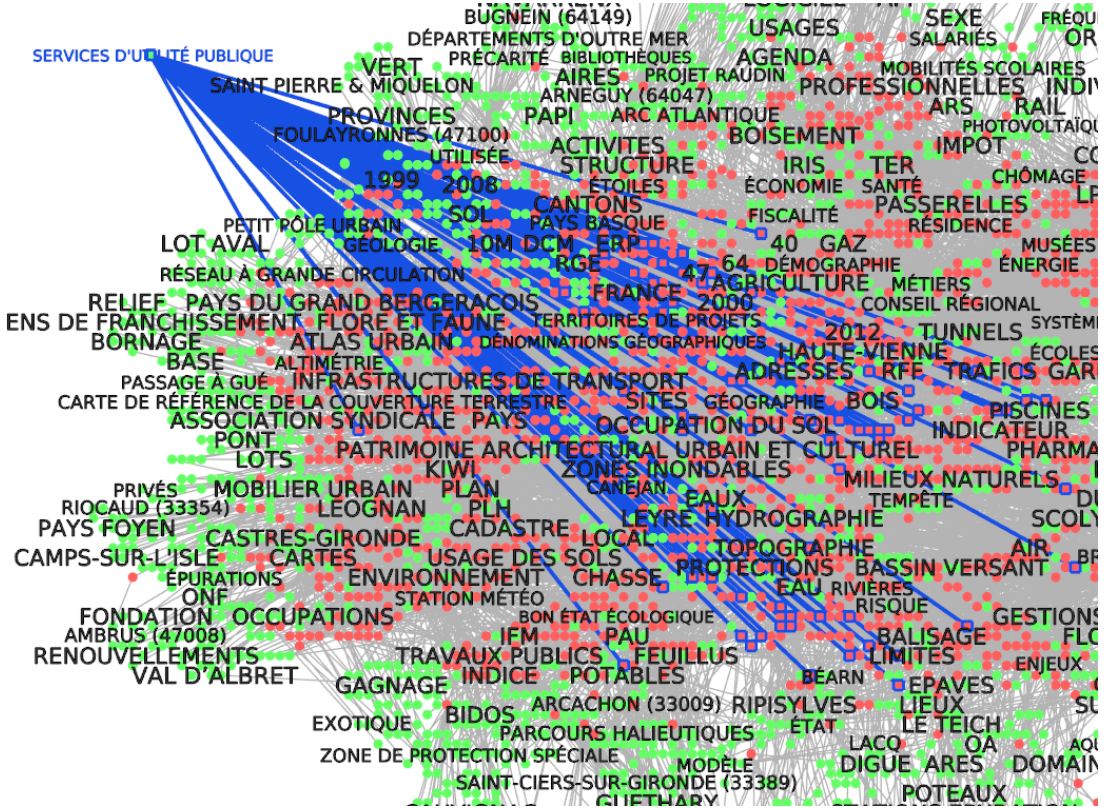
Néanmoins, si on détruit ces fiches parasites, et que l'on se concentre sur cette composante majeure, le graphe semble davantage exploitable bien que toujours difficile à lire.



Graphe Keyword

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

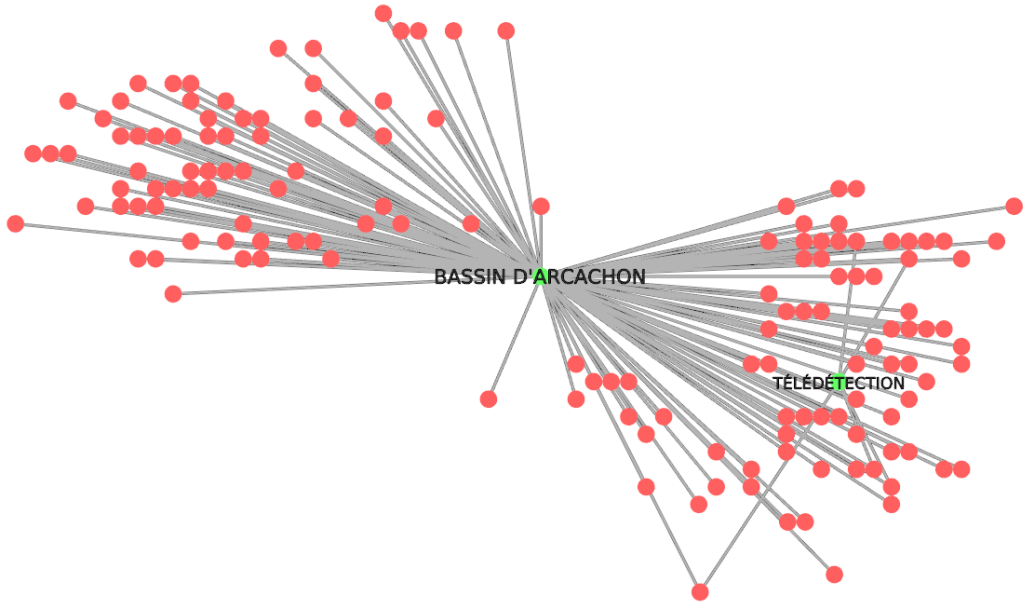
Ici, l'utilisation du module "Reachable Sub-Graph" permet de voir toutes les fiches relatives aux service d'utilité publique.



Graphe Keyword

De la même manière, on peut obtenir 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 meilleur visualisation.



Graphe Keyword

Cependant, cette visualisation ne donne pas d’idée 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 nodes 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 établit 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 :

Nodes	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

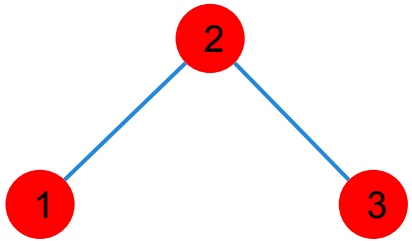
$$\text{Score} = \frac{\sum(a*b)}{\sqrt{(\sum a^2)}*\sqrt{(\sum b^2)}} = \frac{3}{\sqrt{(8)}*\sqrt{(6)}} = 0.43301$$

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

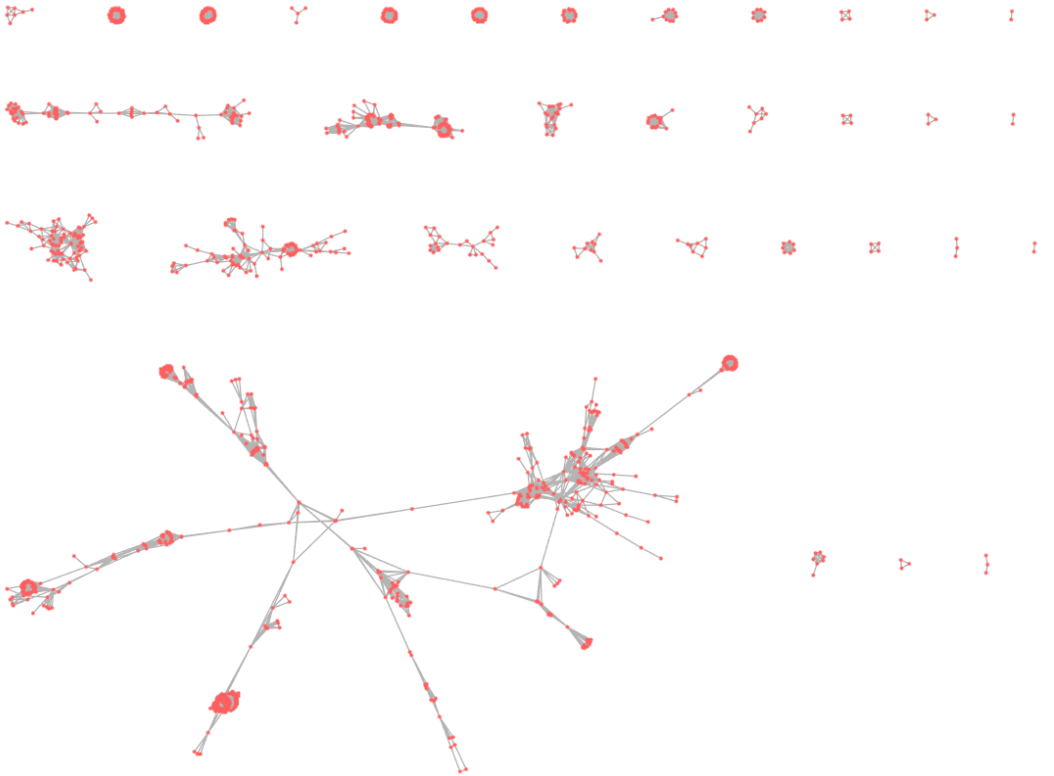
Une fois les scores calculés, une valeur seuil est définie.
Seules les paires ayant un score supérieur au seuil auront un lien.

Ici un seuil de 0.5 :



Graphe Similarité

Des communautés se dessinent très clairement dans ce nouveau graphe

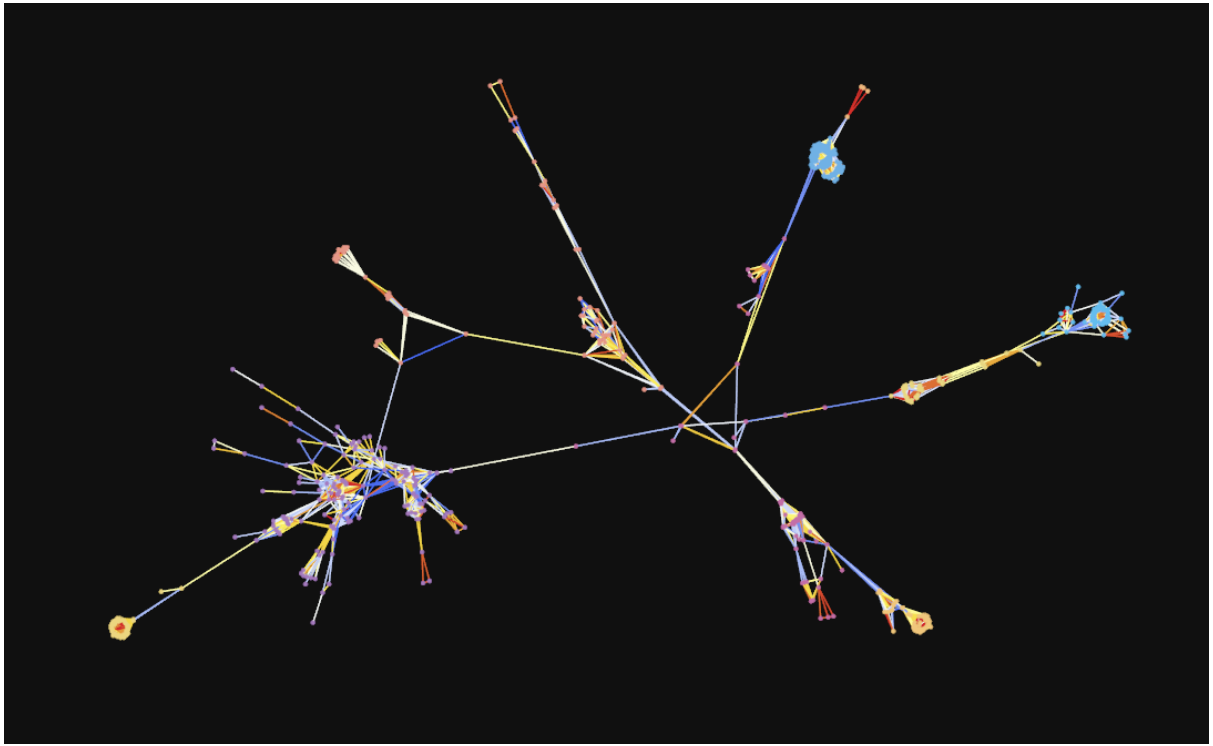


Graphe Similarité

Une nouvelle coloration est appliquée au graphe.

Plus un edge (un lien) tend vers le rouge, plus les fiches sont similaires. Un lien qui tend vers le bleu indiquent donc des fiches peu similaires.

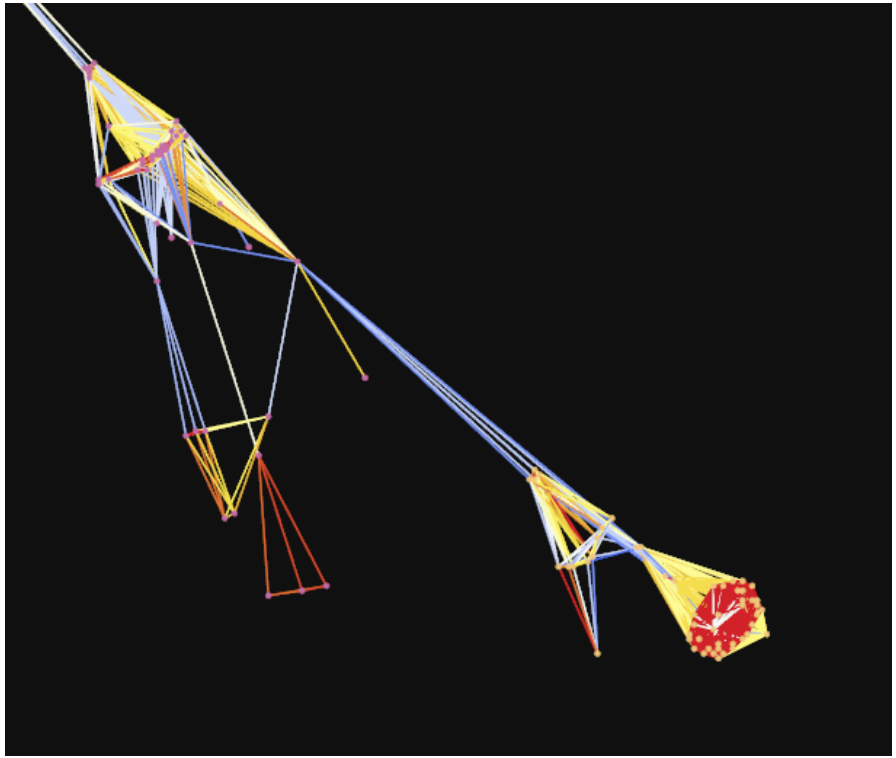
Les couleurs des nodes varient pour mettre en valeur différents communautés.



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 toute de même liée au reste du graphe via des liens avec des études relatives à la population et la démographie.



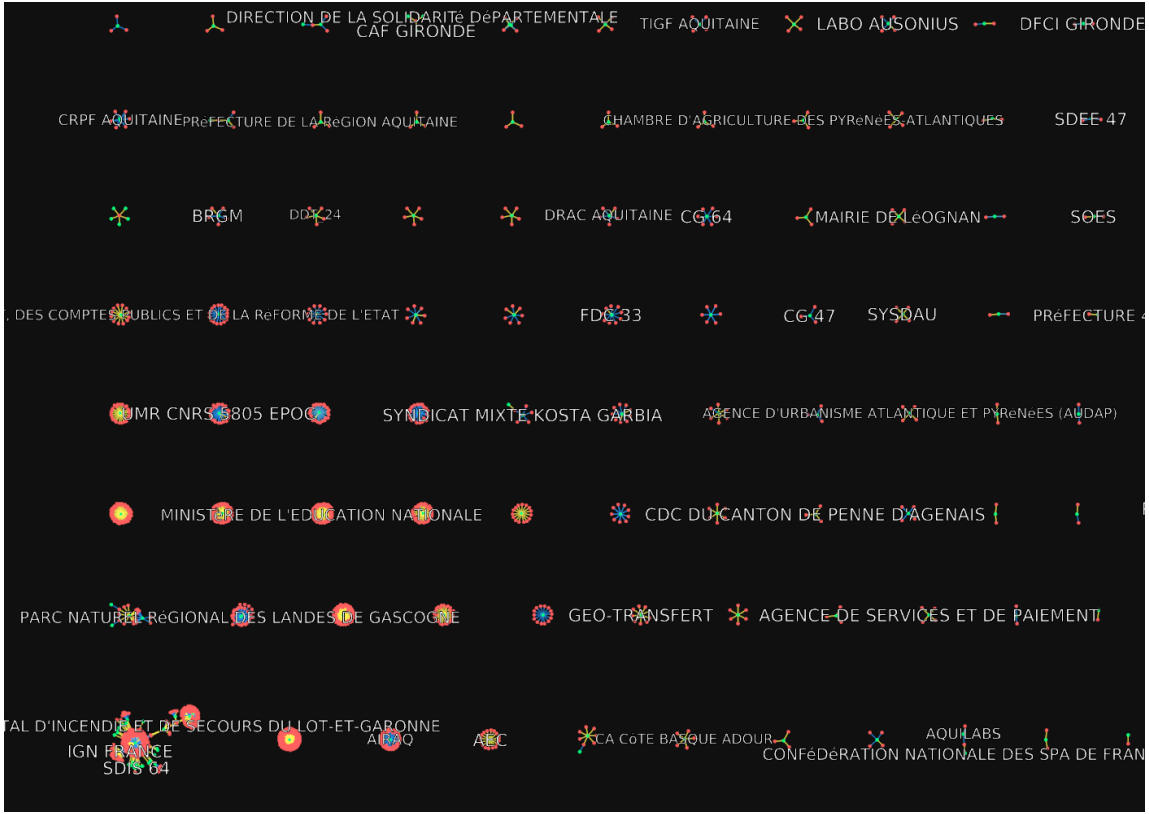
Graphe Similarité

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

Même si il ne s’agit pas directement de généalogie, il permet néanmoins de voir des groupements thématiques pouvant potentiellement abrités 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.



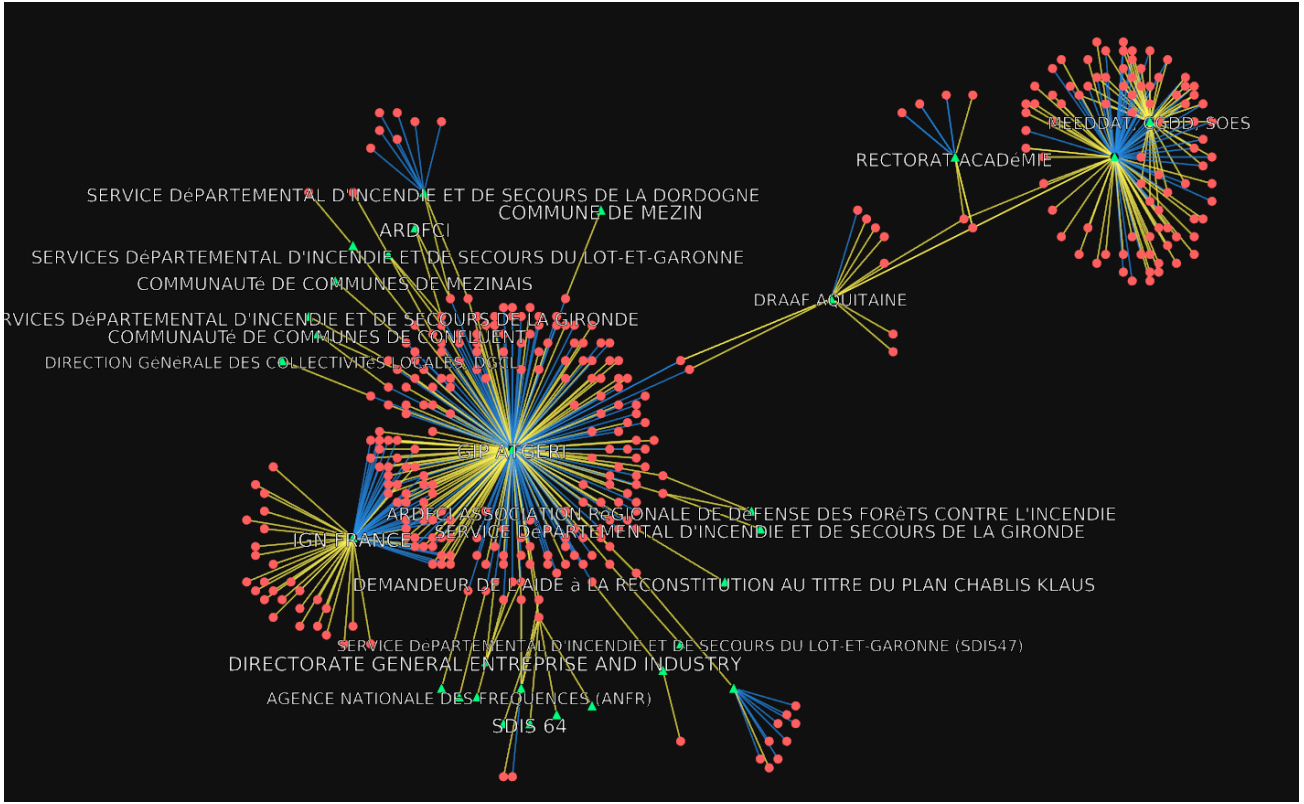
Graphe Acteur

fiche de métadonnées

acteur

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

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



Graphe Acteur

On peut ainsi facilement déterminer l'importance d'un acteur et l'étendu 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.

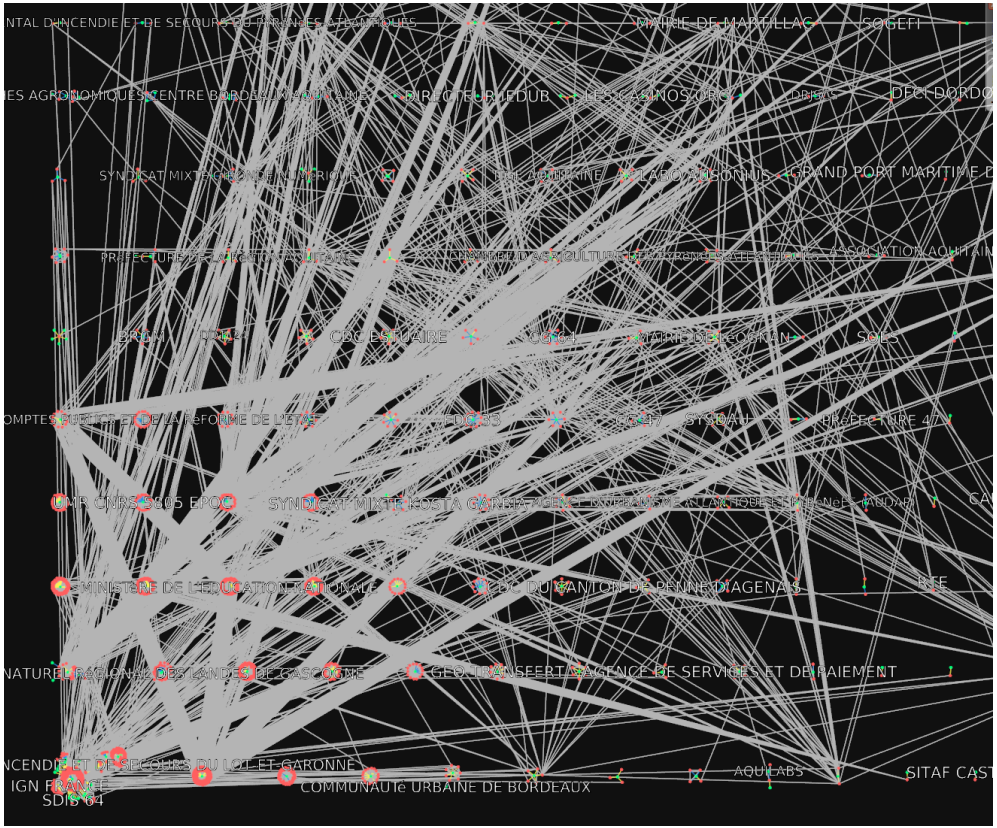
Graphe Hybride (acteur/similarité)

Solution potentielle :

Coupler les deux graphes précédents.

Résultat obtenu :

Graphe très dense et qui semble difficile à exploiter.

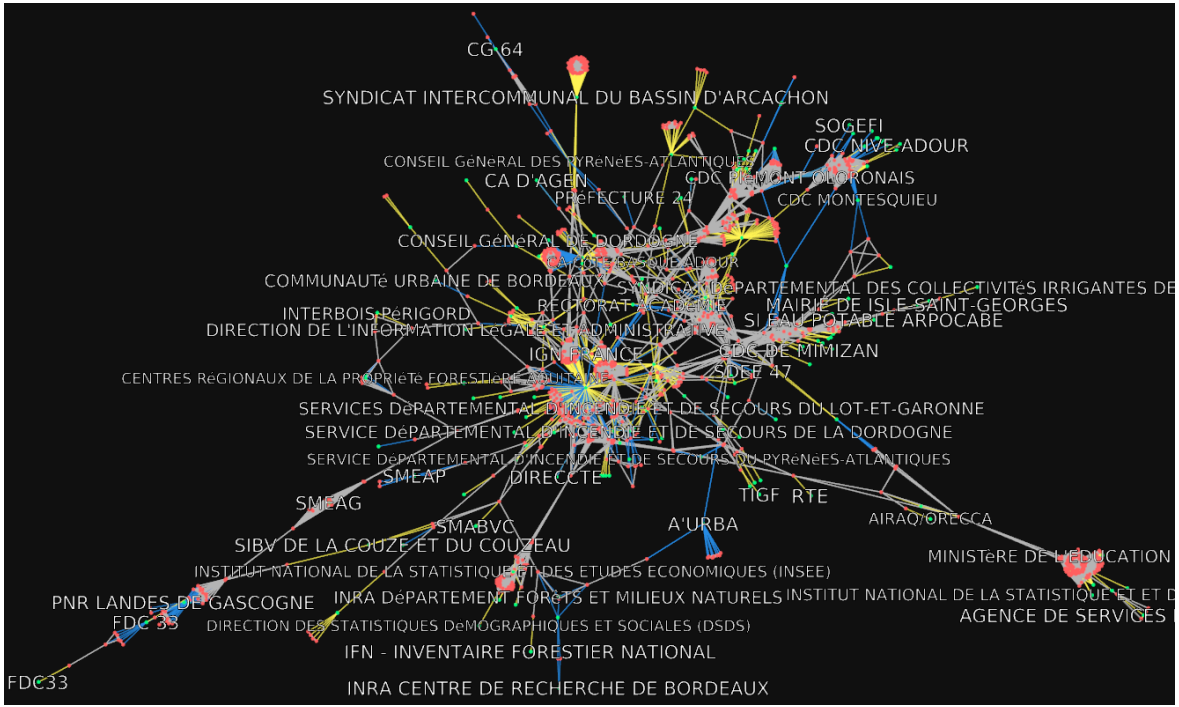


Graphe Hybride (acteur/similarité)

Travaux nécessaires :

Essayer d'établir une mise en forme plus facile à lire.

Trouver d'autres liens moins nombreux (augmenter le seuil de similarité ? parser la généalogie ? etc.)



Autres horizons

- Graphe basé sur la couverture spatiale
- Creuser d'autres méthodes de parsing de la généalogie (résultats pertinents peu probables)
- Utiliser d'autres paramètres pour les liens et essayer différents couplages/traitements
- Porter les solutions les plus pertinentes sur D3 pour répondre à la demande "accès par navigateur"