Séminaire ISEA : informatique

Les bases de données graphes et une application à la fouille sur les données de HAL

romuald.thion@unc.nc

Vendredi 23 avril 2021 13h – amphi Guy Agniel

Introduction

Figure – Mots-clefs des travaux de recherche

Cette présentation

Objectifs

1. Introduire les bases de graphes *RDF*
2. Illustrer sur la base d'archives ouvertes HAL
3. Défricher le projet de recherche
4. Susciter des échanges dans le laboratoire

Plan

Les principaux paradigmes de gestion de données

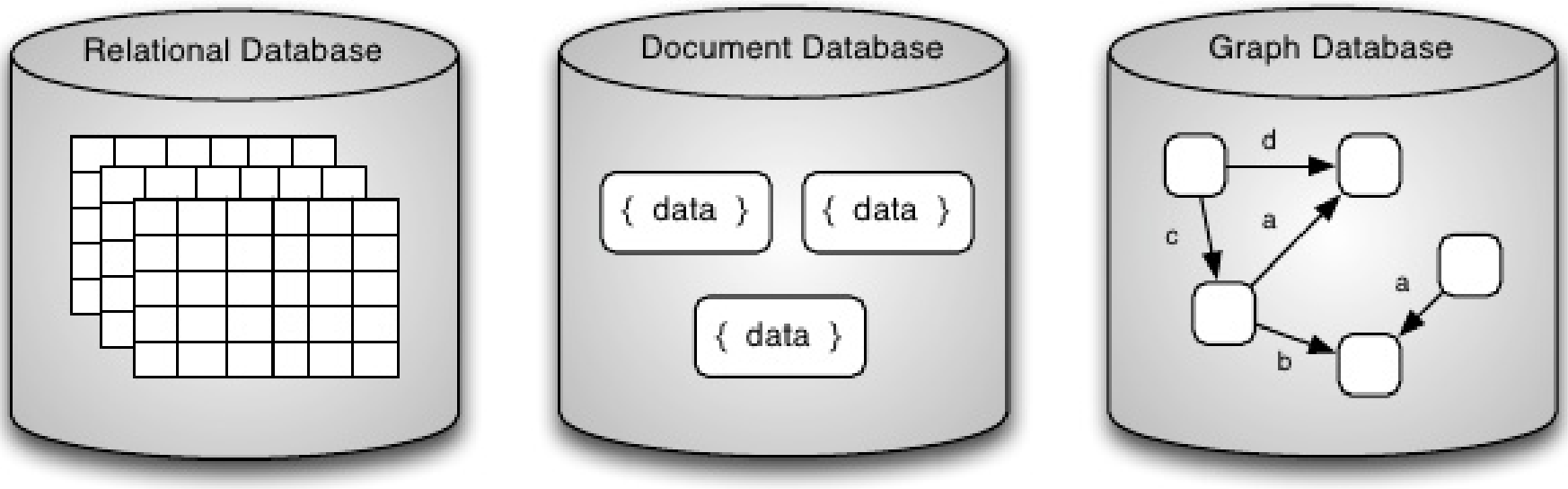
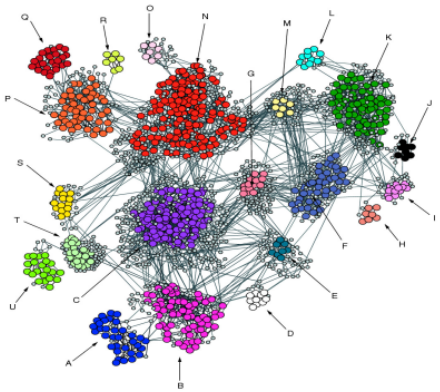
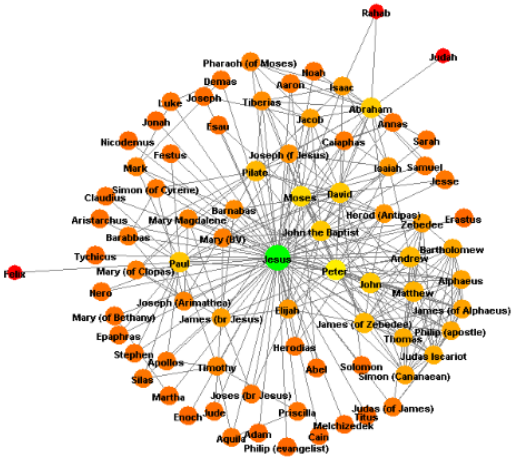


Figure – BD relationnelles, documents et graphes (CC Marko Rodriguez)



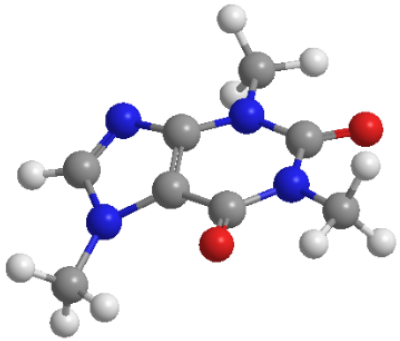
Magwene *et al. Genome Biology* 2004 **5**:R100

Co-expression Network

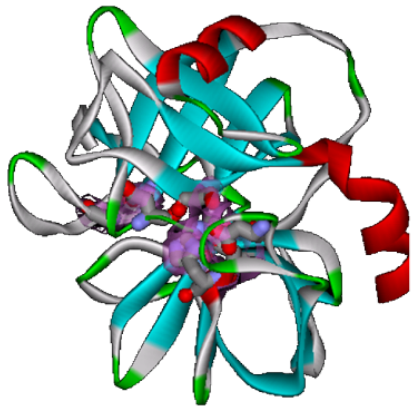


Social Network

Program Flow



Chemical Compound



Protein Structure

Figure – L'omniprésence des graphes

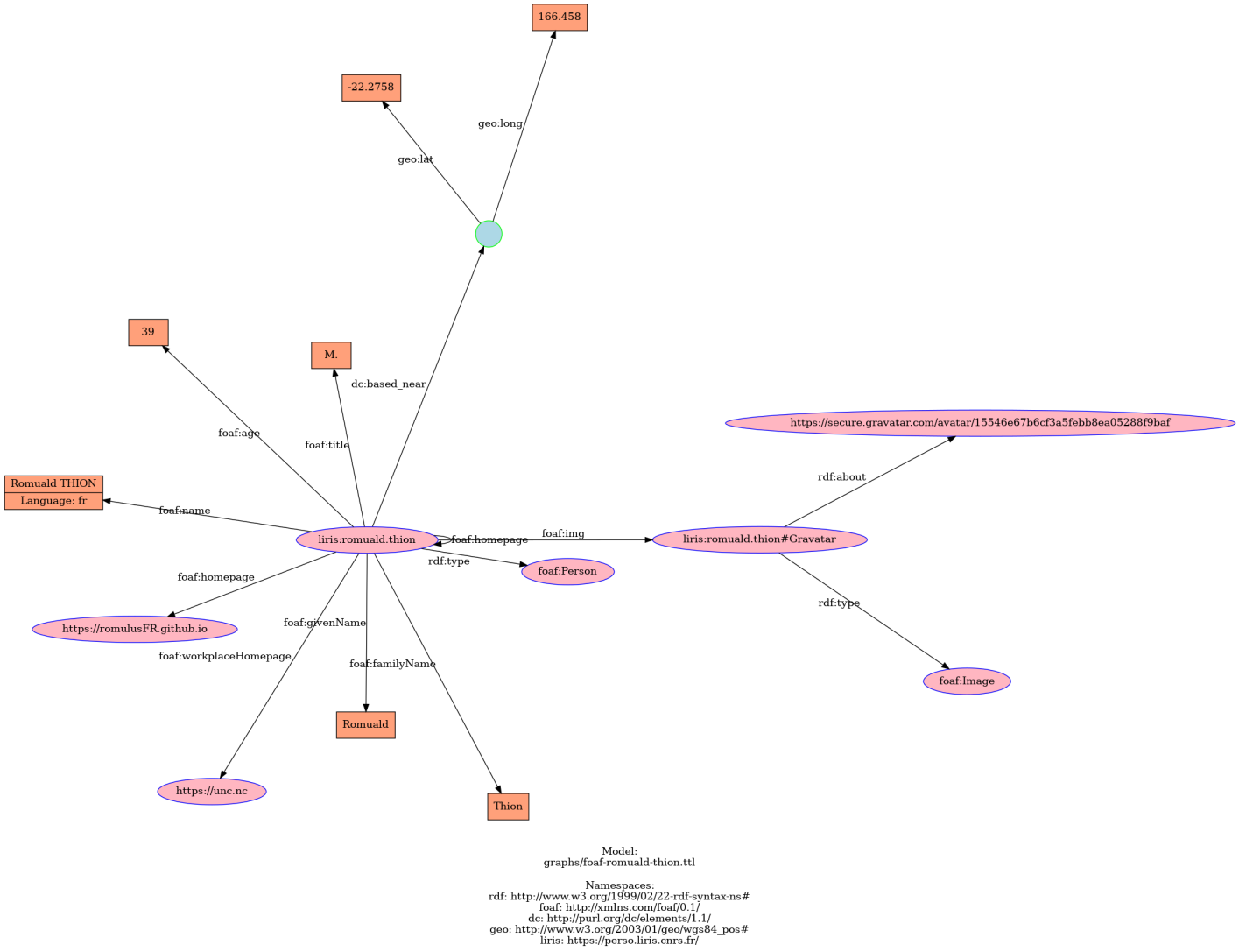


Figure – carte de visite en graphe RDF (HD)

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>.

<https://perso.liris.cnrs.fr/romuald.thion#Card>
    a foaf:Person ;
    foaf:img <https://perso.liris.cnrs.fr/romuald.thion#Gravatar> ;
    foaf:familyName "Thion" ;
    foaf:givenName "Romuald" ;
    foaf:workplaceHomepage <https://unc.nc> ;
    foaf:homepage <https://romulusFR.github.io> ;
    foaf:homepage liris:romuald.thion ;
    foaf:name "Romuald THION"@fr ;
    foaf:age "39" ;
    foaf:title "M." ;
    dc:based_near [geo:lat "-22.2758"; geo:long "166.458"].

<https://perso.liris.cnrs.fr/romuald.thion#Gravatar>
    a foaf:Image ;
    rdf:about <https://secure.gravatar.com/avatar/15546e67b6cf3a5febb8ea05288f9baf>
```

Figure – carte de visite en RDF/Turtle

Resource Description Framework (RDF)
Linked Open Data (LOD)

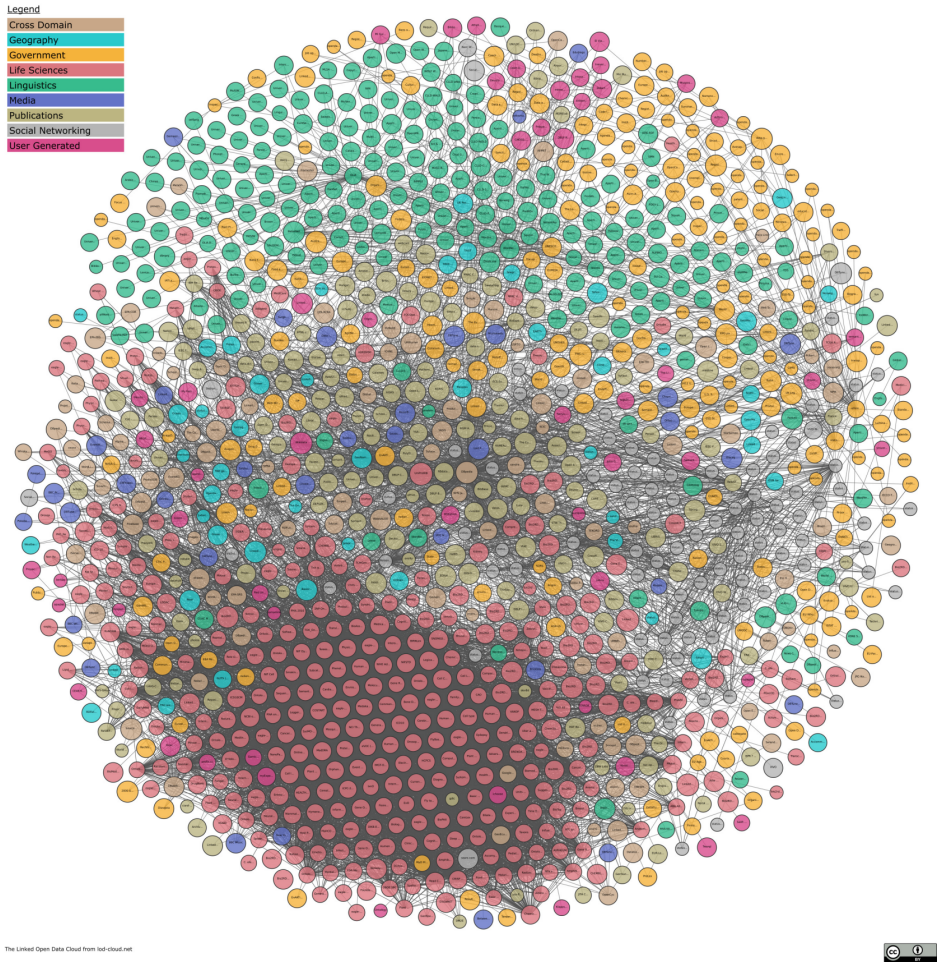


Figure – https://lod-cloud.net/

RDF et LOD

Quelques vocabulaires (ontologies) ouverts

- `foaf`, "*Friend Of A Friend*"
  `http://www.foaf-project.org/` (visualisation)
- `ore`, "*Object Reuse and Exchange*"
  `http://www.openarchives.org/ore/1.0/primer` (visualisation)
- `skos`, "*Simple Knowledge Organization System*"
  `https://www.w3.org/TR/skos-reference/` (visualisation)
- `dc`, "*Dublin Core Metadata Initiative*"
  `https://dublincore.org/schemas/rdfs/` (visualisation)
- `fabio`, "*FRBR-aligned Bibliographic Ontology*"
  `https://sparontologies.github.io/` (visualisation)

http://data.archives-ouvertes.fr

https://data.archives-ouvertes.fr/doc/schema
206 653 071 triplets *(sujet predicat objet)*

La base HAL au 2021-04-23

- types : 59 types et 26 490 225 informations de typage (`rdfs:type`)
- organisations : 326 234
  https://data.archives-ouvertes.fr/structure/529607
- personnes : 4 164 795
  https://data.archives-ouvertes.fr/author/romuald-thion
  https://data.archives-ouvertes.fr/author/818689
- publications : 691 637 dans actes, 1 339 822 articles, 129 088 thèses
  https://data.archives-ouvertes.fr/document/hal-01896276v1
- disciplines : 714 dont 102 inutilisées et 310 nommées « `domain_xxx` »
  https://data.archives-ouvertes.fr/subject/info.info-db

L'ontologie de HAL

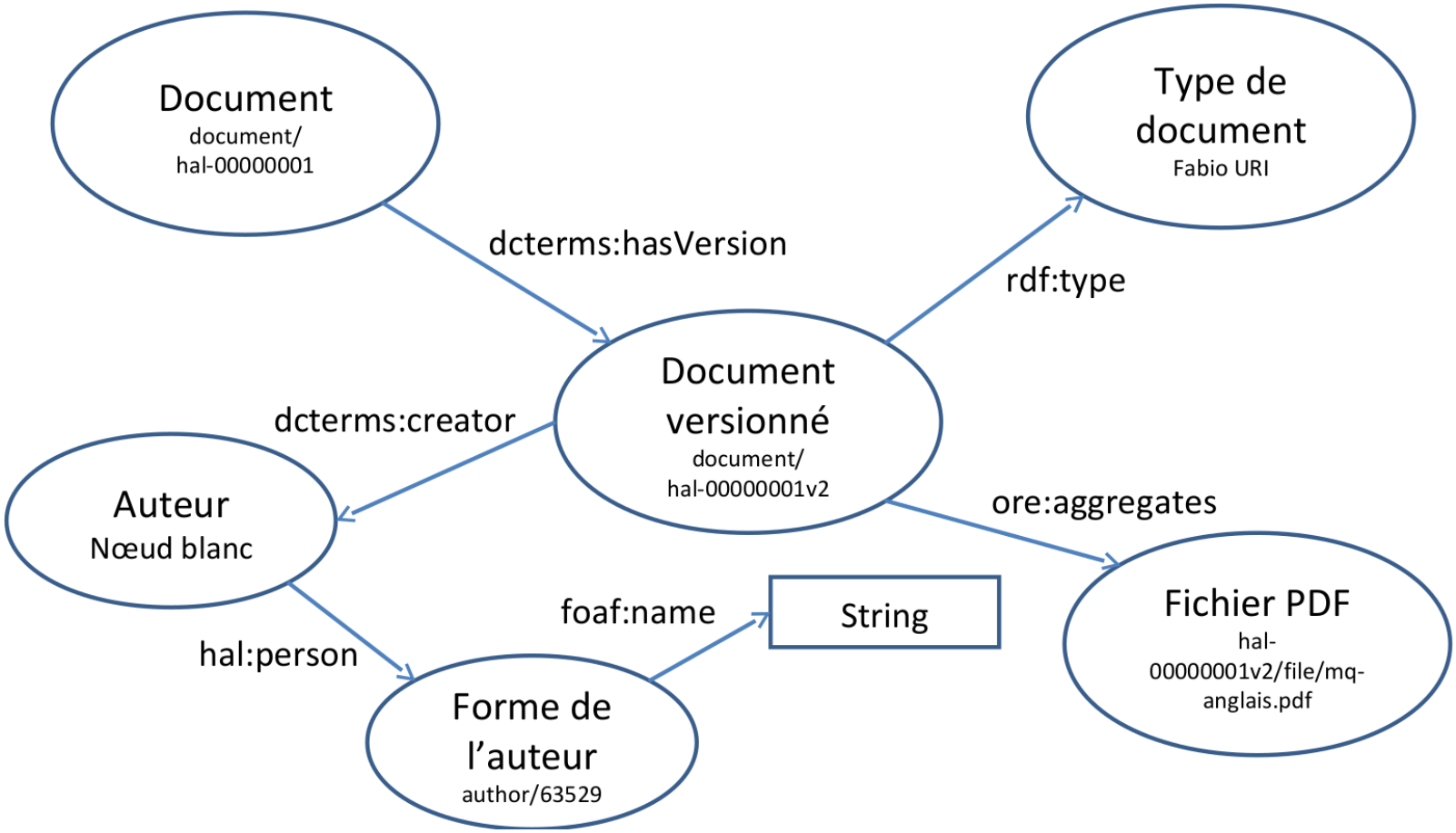


Figure – Vocabulaire principal de HAL (CC G. Poupeau)

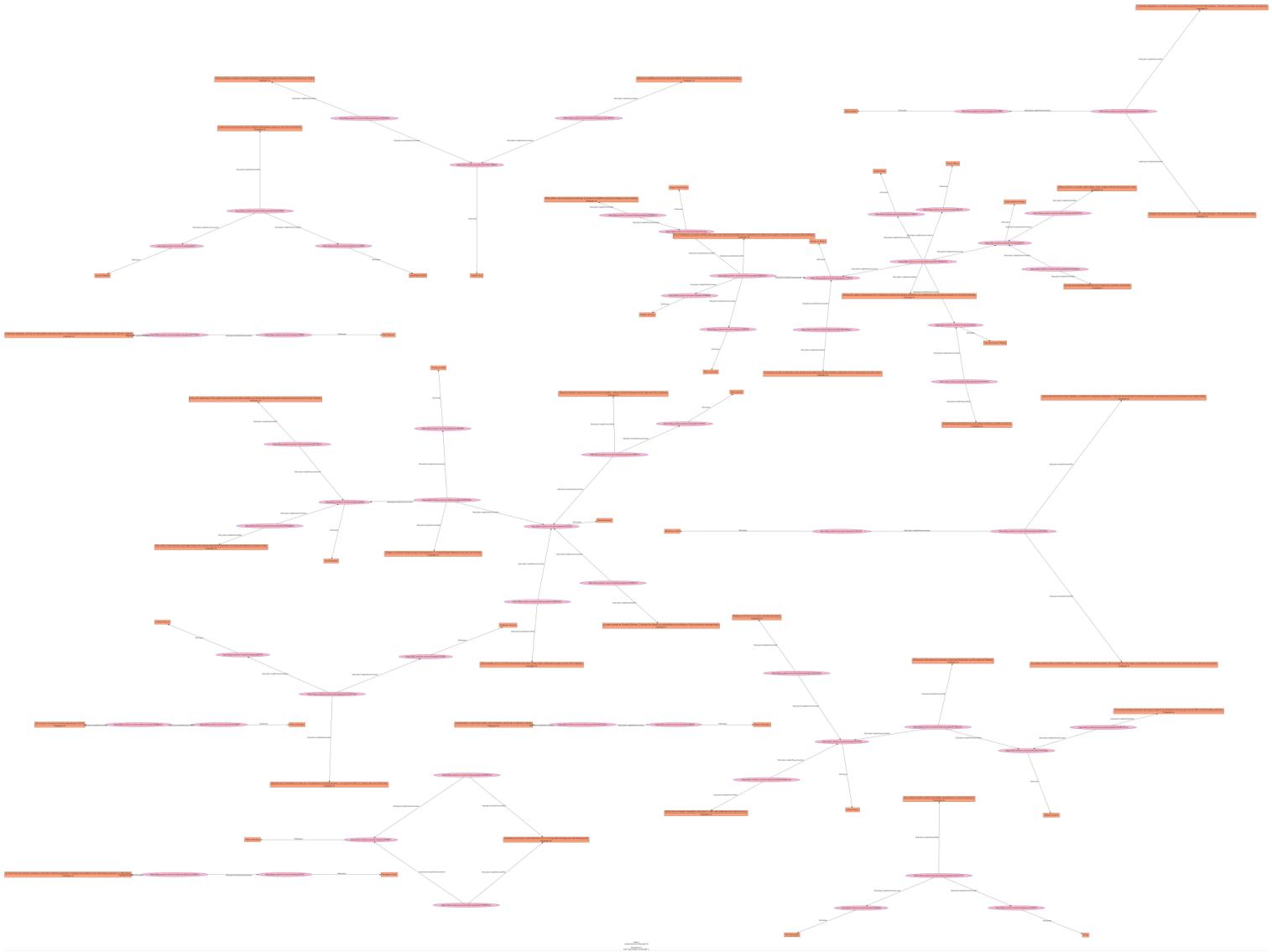


Figure – Sous-graphe des membres de l'ISEA (HD)

Interroger les bases graphes : SPARQL

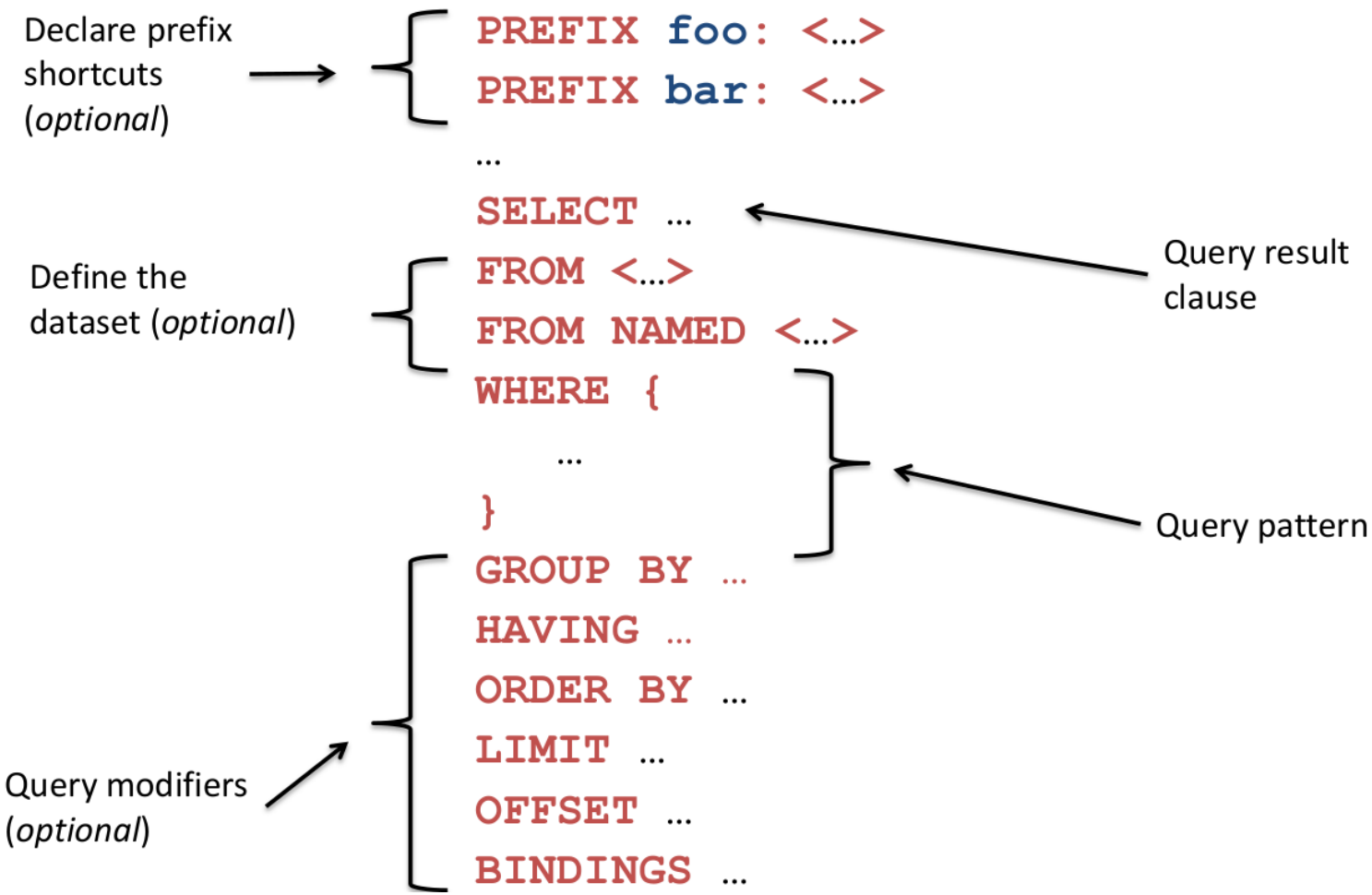


Figure – Anatomie d'une requête SPARQL (CC Lee Feigenbaum)

Interroger les bases graphes : SPARQL

Les sujets les plus présents chez les membres de l'ISEA

```
SELECT ?topic, count(distinct ?pers) as ?nb
WHERE {
  ?sub schema:structure struct:529607;
       a schema:Author;
       schema:person ?pers.
  ?pers foaf:topic_interest ?topic.
  FILTER (LANG(?topic)= "en")
}
GROUP BY ?topic
ORDER BY DESC(?nb)
LIMIT 10
```

Voir résultat en live

Interroger les bases graphes : SPARQL

Les membres de l'ISEA et leurs disciplines

```
SELECT DISTINCT
  STRAFTER(str(?p1),str(author:)) as ?idhal,
  ?n1 as ?name,
  group_concat(?ishort, '|') as ?interests
WHERE {
  ?pub dcterms:creator ?c1.
  ?c1 hal:structure struct:529607; hal:person ?p1.
  ?p1 foaf:name ?n1; foaf:familyName ?fn; foaf:interest ?i.
  BIND(STRAFTER(str(?i), str(subject:)) as ?ishort)
}
GROUP BY ?p1 ?n1 ?fn
ORDER BY ?fn
```

Voir résultat en live

Des graphes RDF aux graphes usuels

Le graphe des coauteurs de l'ISEA : requête CONSTRUCT

```
CONSTRUCT {
  ?p1 foaf:name ?n1; foaf:knows ?p2.
  ?p2 foaf:name ?n2; foaf:knows ?p1.
}
WHERE {
  ?pub dcterms:creator ?c1, ?c2.
  ?c1 hal:person ?p1;
      hal:structure struct:529607.
  ?p1 foaf:name ?n1.
  ?c2 hal:person ?p2;
      hal:structure struct:529607.
  ?p2 foaf:name ?n2.
  FILTER(?p1 < ?p2)
}
```

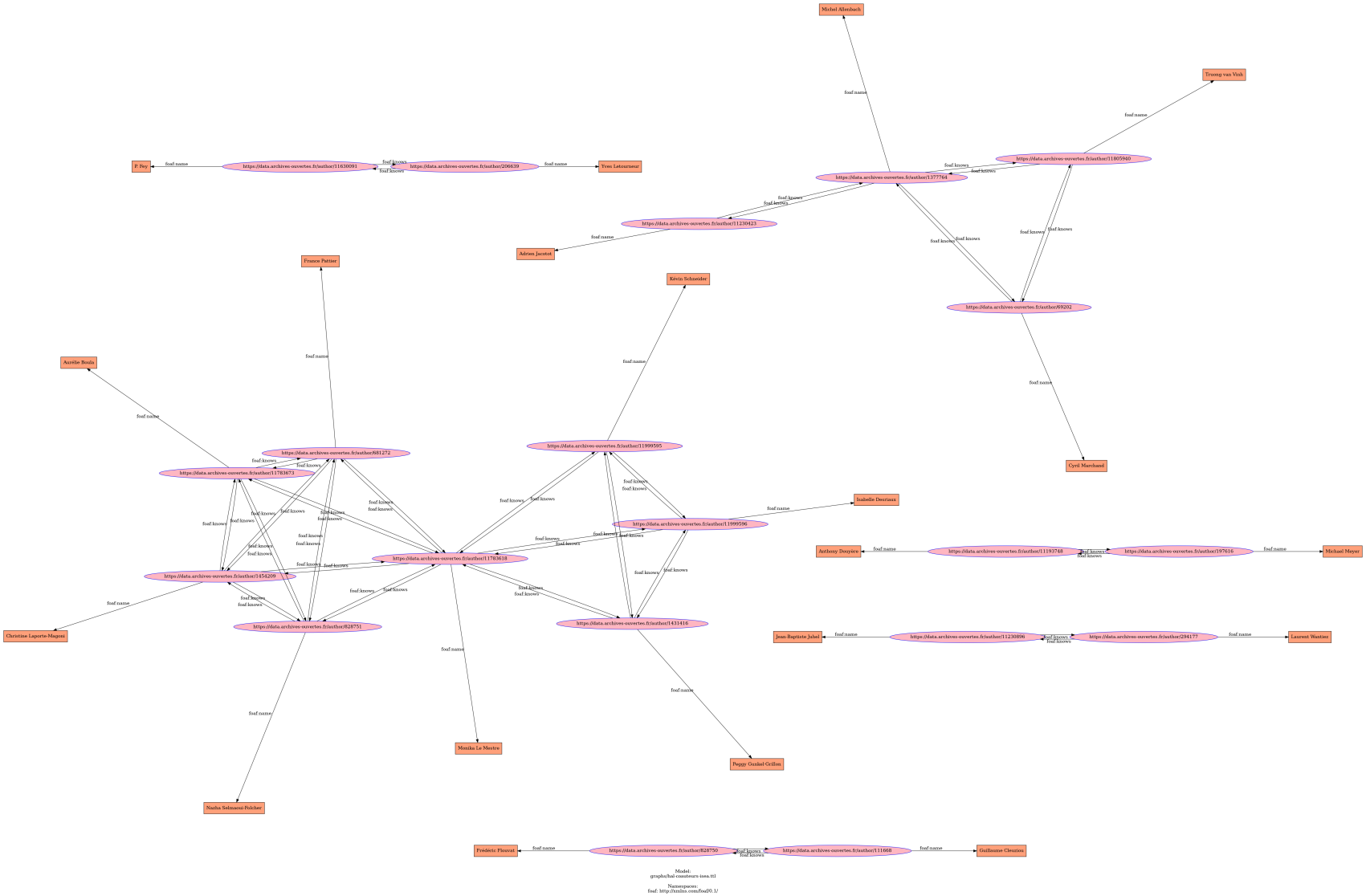


Figure – Membres de l'ISEA et leurs publis (HD)

Des graphes RDF aux graphes usuels

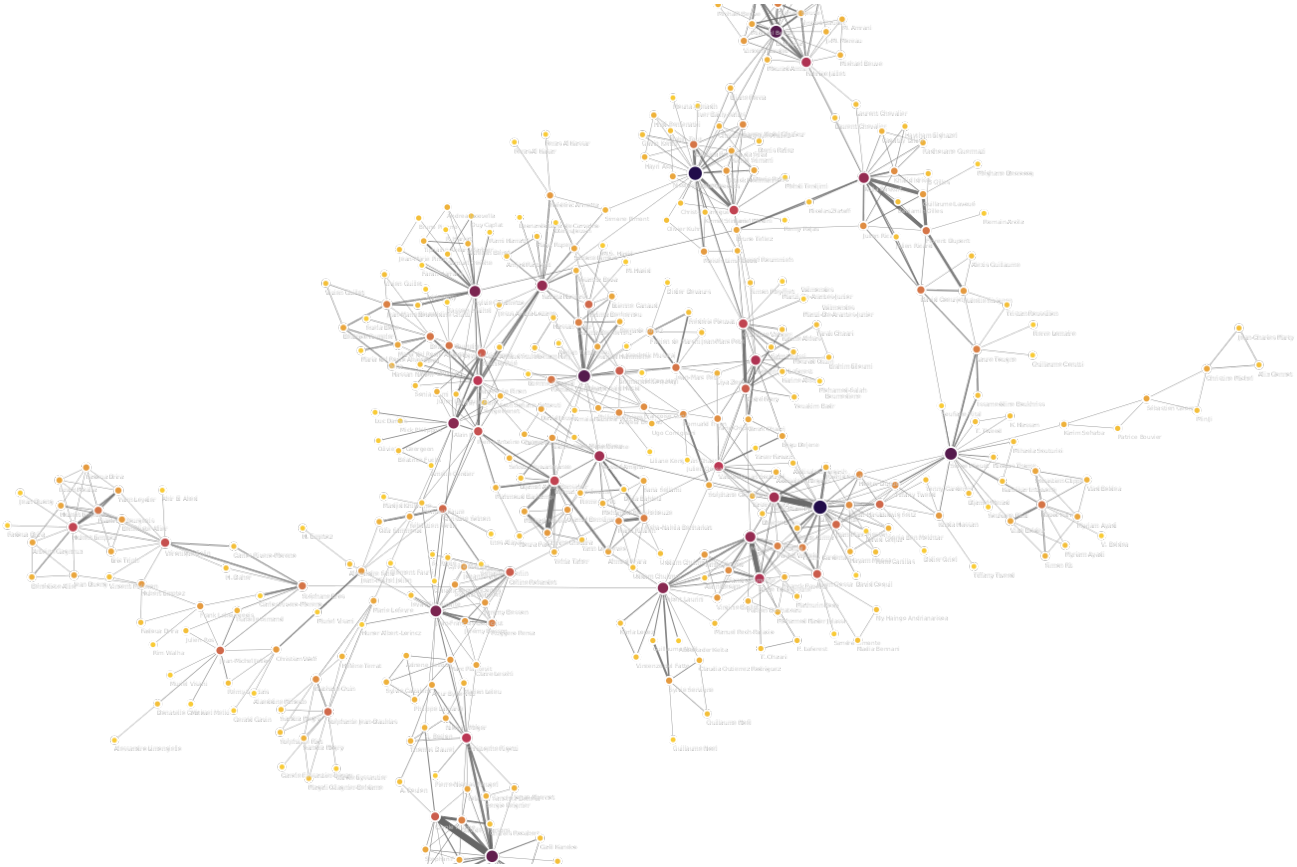


Figure – Graphe non-dirigé $G = (V, E \subseteq V \times V, w : E \to \mathbb{N})$ de la relation *co-publier*, étiquettée avec le nombre de publis en commun pour le laboratoire LIRIS, généré avec https://cytoscape.org/.

Des graphes RDF aux graphes usuels

Traitement Python avec https://networkx.org/ sur le graphes des co-auteurs de l'ISEA.

```
Type: Graph
Number of nodes: 20
Number of edges: 24
Average degree:   2.4000
Density: 0.126
Average clustering: 0.488
Assortativity: 0.430
Connected components: 6
Size of largest component: 8 (40%)
Cliques: 8
```

Clustering des co-occurences de disciplines

Extractions de graphes « usuels » depuis HAL

- La taxonomie H : la relation « être une sous-discipline de », c'est un arbre, un cas très particulier de graphe, représentatif des ontologies de domaines
- Le graphe $G = (V, E, w)$: les co-occurences entre disciplines (V), avec les poids $w(u, v) = |\{d \in Doc \mid d \text{ est taggé avec } u \text{ et avec } v\}|$

Informations sur H et G

- H : $|V| = 715$, $|E| = 714$, $\bar{d} = 1$, clust. $= 0\%$, den. $= 0.1\%$
- G : $|V| = 613$, $|E| = 34529$, $\bar{d} = 112$, clust. $= 64\%$, den. $= 18\%$
- G′ : $|V| = 399$, $|E| = 7706$, $\bar{d} = 39$, clust. $= 57\%$, den. $= 10\%$ après filtrage $w(u, v) < 32$ et disciplines mal renseignées

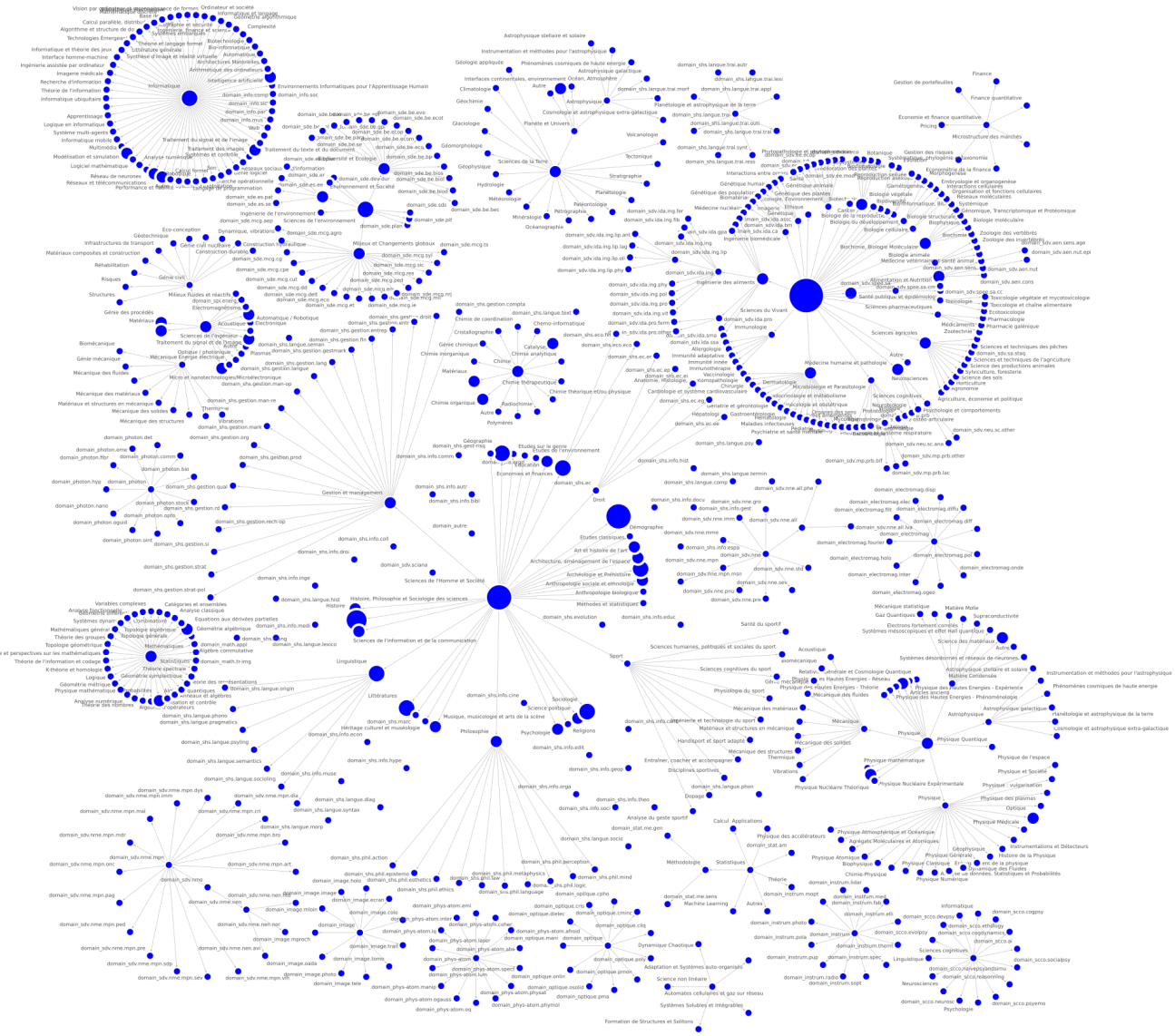


Figure – Graphe H taxonomie des disciplines (HD)

Taxonomie des disciplines H

Les disciplines avec le plus de documents

('Sciences du Vivant', 250 983),
("Sciences de l'Homme et Société", 172 496),
('Droit', 137 373),
('Histoire', 105 694),
("Sciences de l'environnement", 90 579),
('Archéologie et Préhistoire', 86 727),
('Littératures', 78 315),
('Sociologie', 74 663),
('Informatique', 74 019),
('Economies et finances', 70 463)

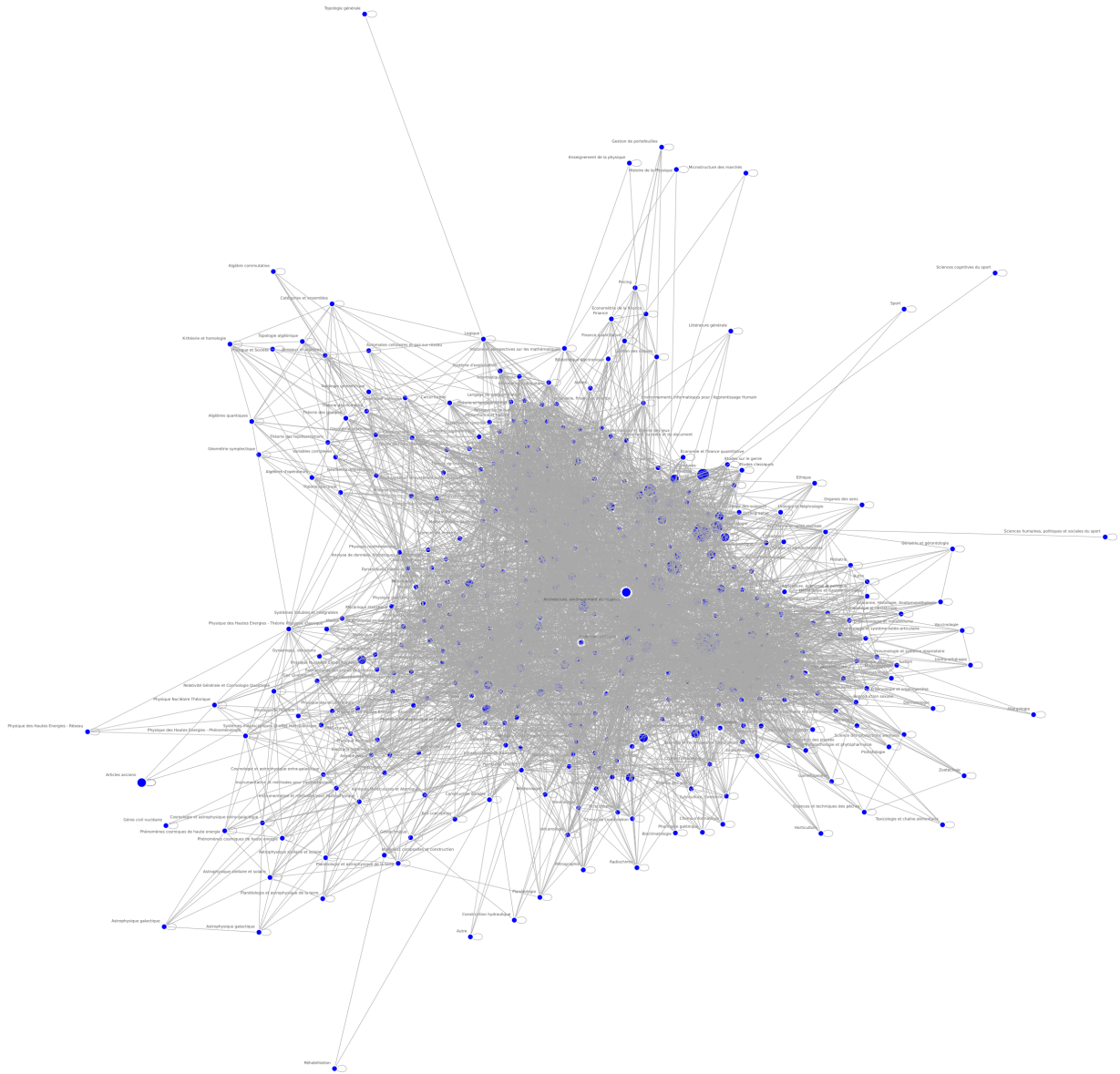


Figure – Graphe G des co-occurences de disciplines (HD)

Co-occurrences entre disciplines G

Les disciplines de plus haut degré

```
('Modélisation et simulation', 193),
('Sciences du Vivant', 162),
("Traitement du signal et de l'image", 147),
("Traitement du signal et de l'image", 132),
('Informatique', 121),
('Environnement et Société', 121),
('Biodiversité et Ecologie', 115),
('Milieux et Changements globaux', 115),
('Intelligence artificielle', 114),
("Sciences de l'environnement", 111)
```

Co-occurrences entre disciplines G

Les co-occurrences les plus fortes

```
('Sciences du Vivant', "Sciences de l'environnement", 23 641),
("Sciences de l'Homme et Société",
 'Archéologie et Préhistoire', 22 458),
('Sciences du Vivant', 'Informatique', 18 243),
("Traitement du signal et de l'image",
 "Traitement du signal et de l'image", 16 058),
("Sciences de l'Homme et Société", 'Histoire', 14 812),
('Science politique', 'Sociologie', 14 669),
("Sciences de l'Homme et Société",
 'Sciences du Vivant', 12 866),
('Archéologie et Préhistoire', 'Histoire', 10 053),
("Sciences de l'Homme et Société", 'Littératures', 9 514),
('Histoire', 'Littératures', 8 970)
```

Fouille de données avec les graphes RDF

Modéliser *l'intérêt subjectif* dans la fouille

- Les méthodes de fouilles peuvent donner *beaucoup de résultats*
- On veut garder pour l'utilisateur les *intéressants*, c-à-d les plus :
  - fréquents ?
  - gros ?
  - rares ?

Intuition caractériser l'intérêt des résultats avec l'ontologie

On va partionner les disciplines scientifiques de HAL en *cluster* et rechercher les plus intéressants, au sens qu'on ne les attendait pas *a priori* en connaissant la taxonomie des disciplines.

Exploiter la taxonomie dans le clustering

Idée 1 : évaluer l'intérêt *a posteriori* en comparant avec H

- le plus petit ancètre commun d'un *cluster* est `top` : le *cluster* fait intervenir des disciplines différentes
- il y a une faible proportion de cousins dans un *cluster* : des sous-disciplines cousines sont plus proches que leurs frères
- il y a équilibre entre les représentants des fratries dans le *cluster* : une discipline émergeante

Idée 2 : modifier les données *a priori*

On modifie les co-occurences de disciplines en utilisant en ajoutant des arcs $(u, v, k \times |doc_u|)$ à G pour chaque $(u, v) \in$ H, avec k un facteur de poids :

- un *cluster* qui résiste à l'ajout de H a des liens internes plus forts que le lien « *être sous-discipline* »

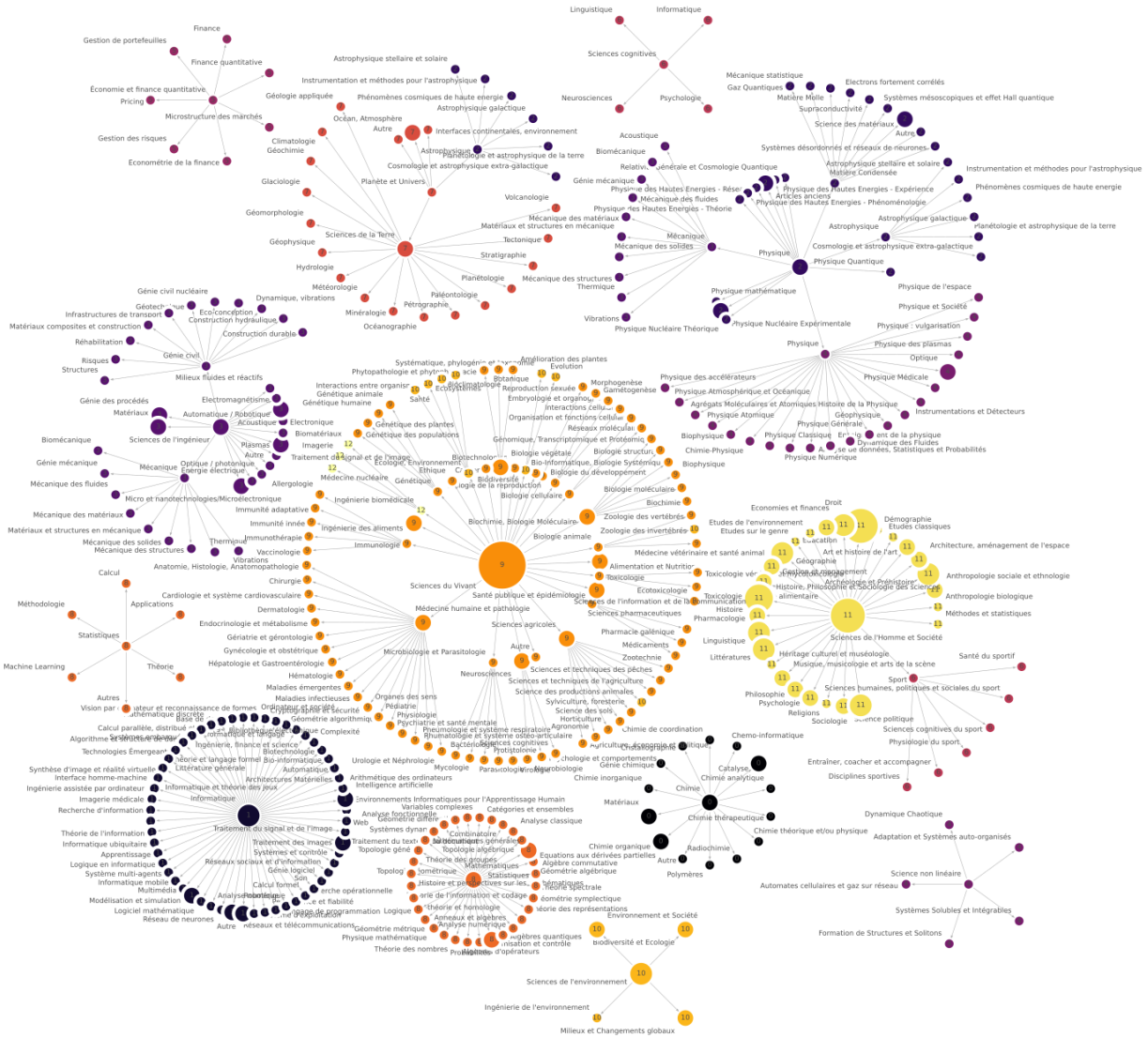


Figure – *Sanity check* : clustering de G avec H pour $k = 1$ (HD)

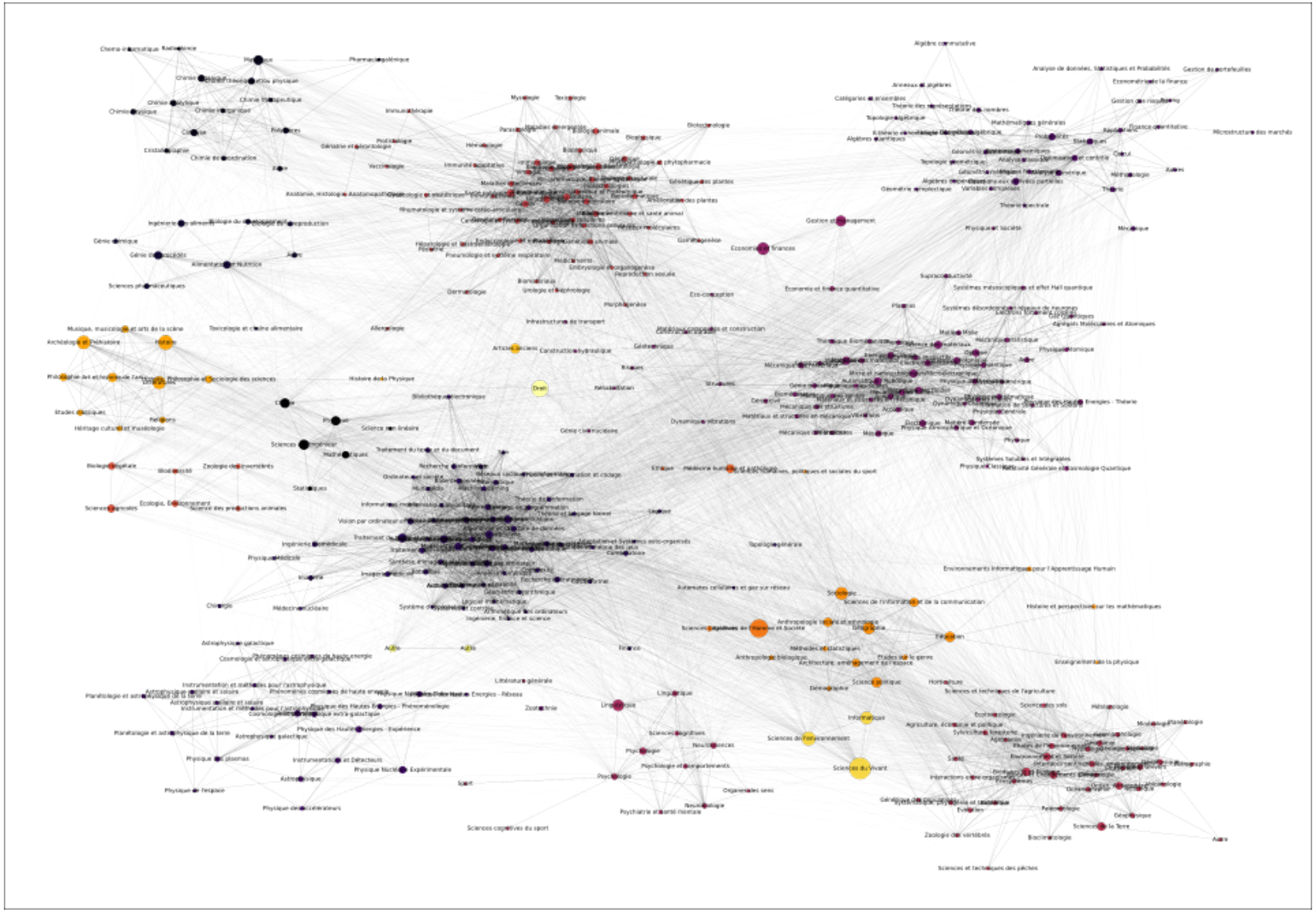


Figure – Clustering de G avec (python-louvain HD)

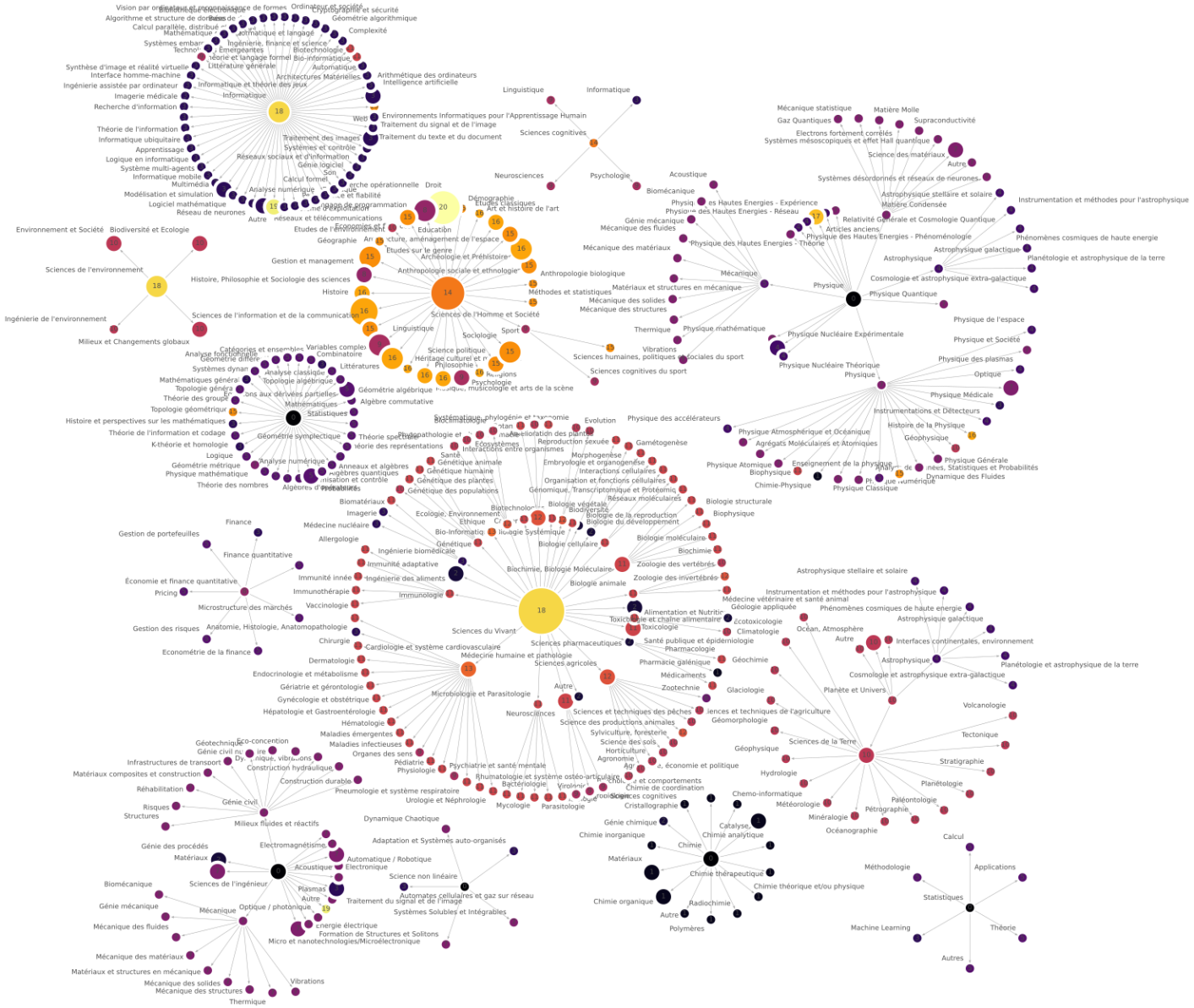


Figure – Le même que précédent, mais vu sur H (HD)

Quelques résultats empiriques (idée #1)

Contre-exemples : des *clusters* attendus en informatique

- Arithmétique des ordinateurs (info.info-ao)
- Architectures Matérielles (info.info-ar)
- Systèmes embarqués (info.info-es)
- Système d'exploitation (info.info-os)

Choix (erreur) de modélisation, les *statistiques*

- Probabilités (math.math-pr)
- Statistiques (math.math-st)
- *Applications (stat.ap)*
- *Calcul (stat.co)*
- *Théorie (stat.th)*

Quelques résultats empiriques (idée #1)

Les homononymes *science pour l'ingénieur* et *physique*

- Thermique (phys.meca.ther) (spi.meca.ther)
- Vibrations (phys.meca.vibr) (spi.meca.vibr)
- Acoustique (phys.meca.acou) (spi.acou)

Les homononymes *science de l'univers* et *physique*

- Phénomènes cosmiques de haute energie (phys.astr.he) (sdu.astr.he)
- Instrum. et méthodes pour l'astrophysique (phys.astr.im) (sdu.astr.im)
- Géophysique (phys.phys.phys-geo-ph) (sdu.stu.gp)

Des résultats plutôt dans le sens la fouille *pour* l'ontologie que l'inverse.

Quelques résultats empiriques (idée #1)

Les cousins en SHS

- Histoire et perspectives sur les mathématiques (math.math-ho)
- Enseignement de la physique (phys.phys.phys-ed-ph)
- Histoire de la Physique (phys.phys.phys-hist-ph)
- Environnements Informatiques pour l'Apprentissage Humain (info.eiah)
- Agriculture, économie et politique (sdv.sa.aep)
- Ethique (sdv.eth)

Apparaissent[a] dans des clusters exclusivement SHS

a. Selon les paramètres et les *coin flips* pour les algortihmes randomisés

Une cousine de la chimie

- Pharmacie galénique (sdv.sp.pg)

Quelques résultats empiriques (idée #1)

Un cousin des sciences pour l'ingénieur

- Ingénierie assistée par ordinateur (info.info-ia)

Un cousin de l'informatiqe

- Complexité (info.info-cc)
- Géométrie algorithmique (info.info-cg)
- Mathématique discrète (info.info-dm)
- Algorithme et structure de données (info.info-ds)
- *Combinatoire (math.math-co)*

Ces cousins sont des outliers dans un groupe majoritaire d'une autre discipline : ils apparaissent comme des disciplines frontières.

Quelques résultats empiriques (idée #1)

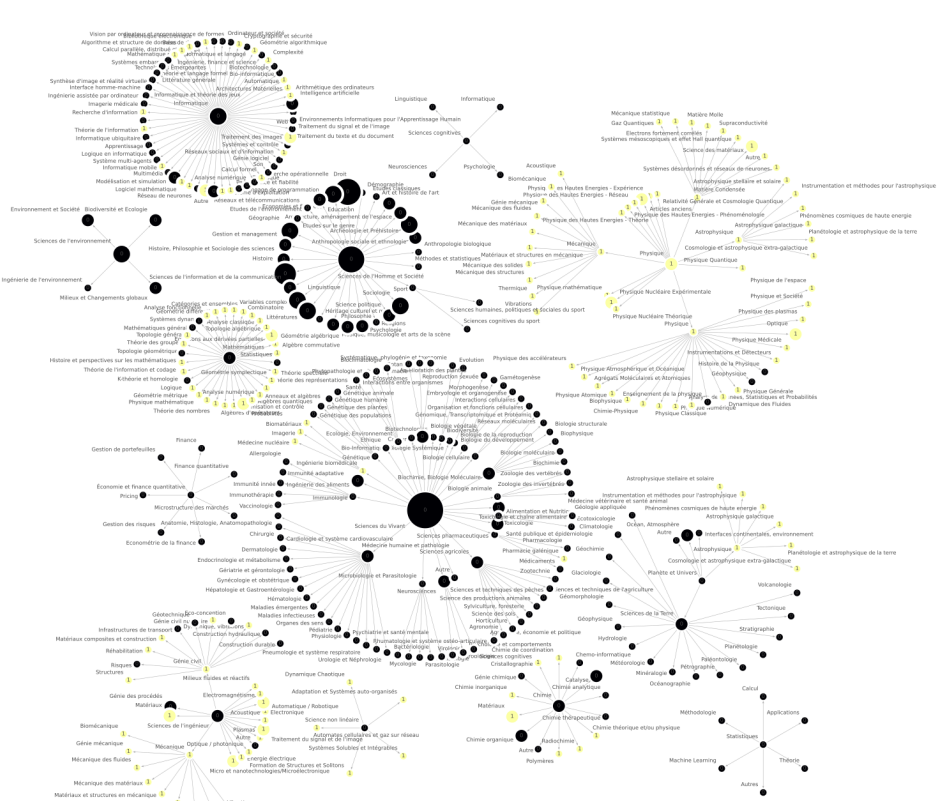
Un cluster pluri-disciplinaire : la robotique ?

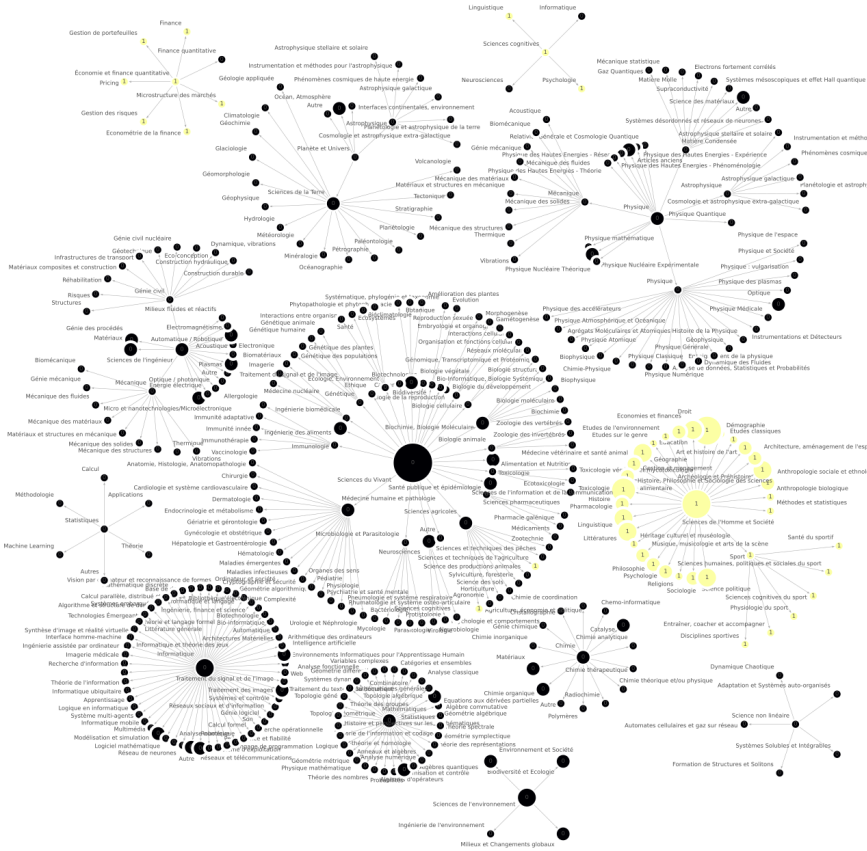
- Automatique (info.info-au)
- Systèmes et contrôle (info.info-sy)
- Automatique / Robotique (spi.auto)
- Energie électrique (spi.nrj)

Un cluster pluri-disciplinaire : physique/math ?

- Physique mathématique (math.math-mp)
- Algèbres quantiques (math.math-qa)
- Mécanique statistique (phys.cond.cm-sm)
- Physique mathématique (phys.mphy)
- Analyse de données, Stats. et Probas. (phys.phys.phys-data-an)

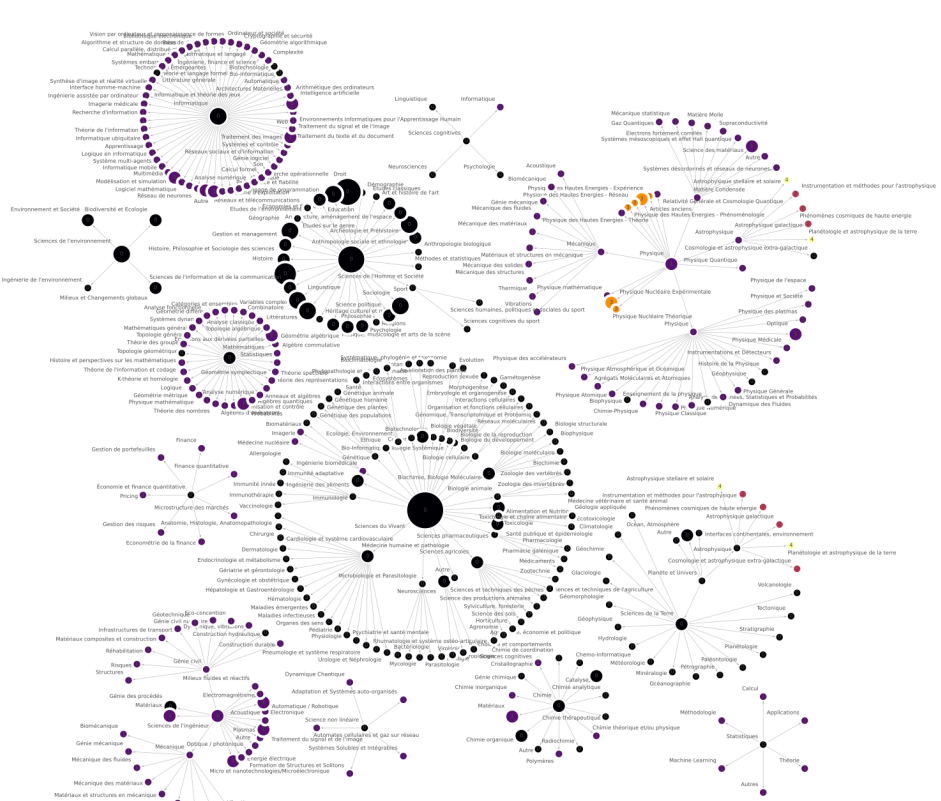
Quelques résultats empiriques (idée #2)

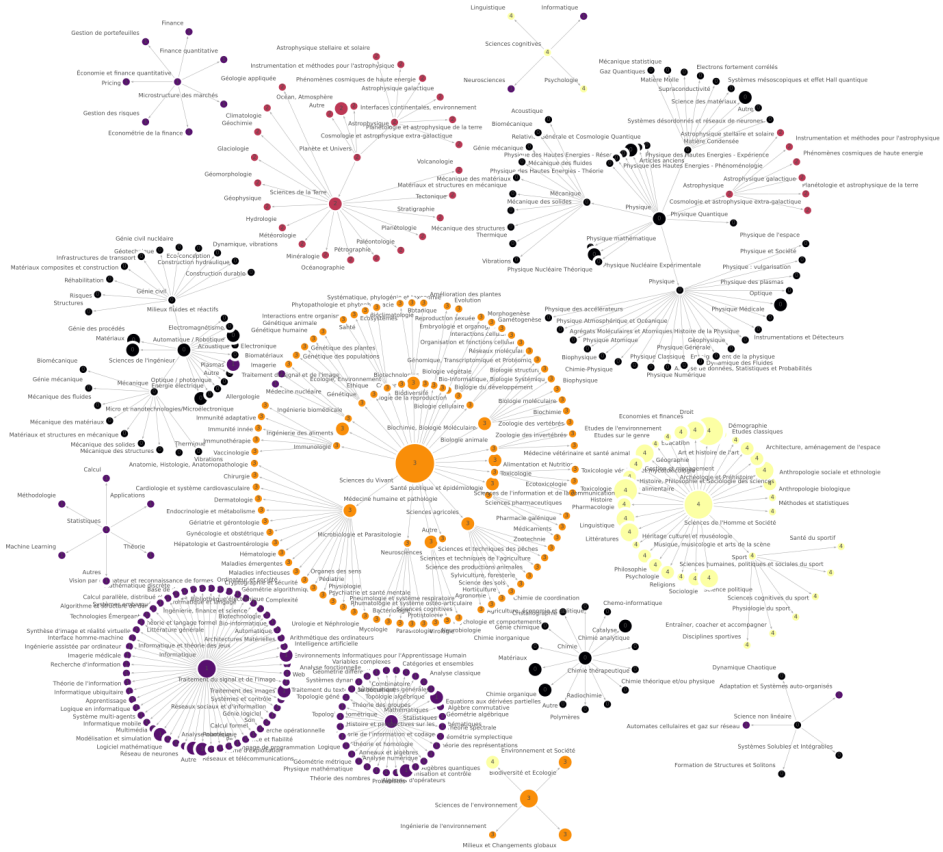




Par *spectral clustering* on demande 2 *clusters* avec $k = 0.0$ (gauche) et $k = 1.0$ (droite) (HD, HD)

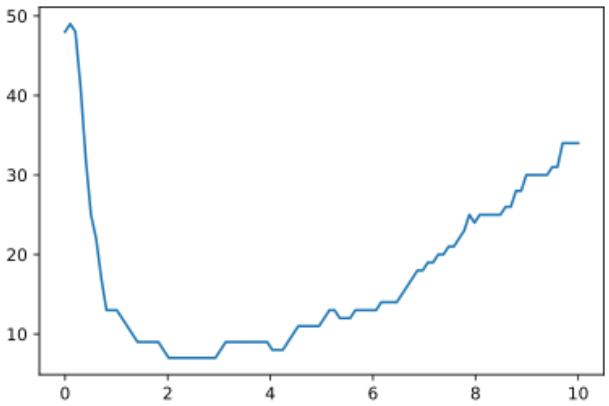
Quelques résultats empiriques (idée #2)

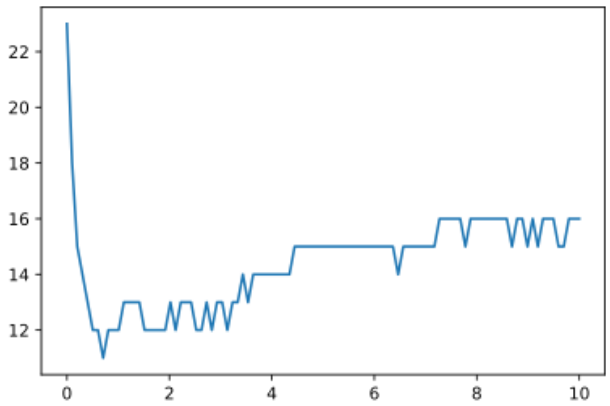


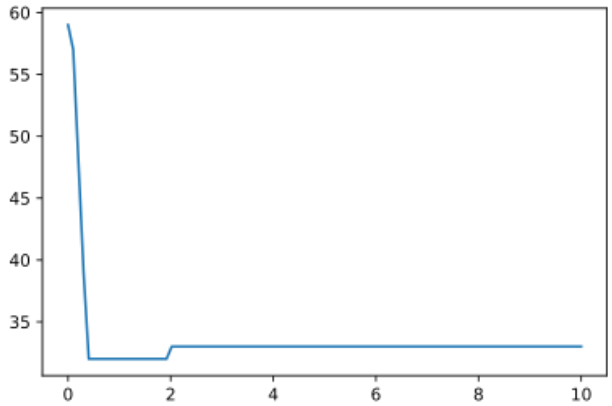


Rigidification des *clusters* avec H pour $k = 5$: une variation qualitative ?
(HD, HD)

Quelques résultats empiriques (idée #2)







Nombre de *clusters* en fonction du paramètre k pour `DBSCAN` (gauche) et `louvain` (milieu) et `AffinityPropagation` (droite) : une forme d'optimum ?

Les ontologies en science : PubCHEM

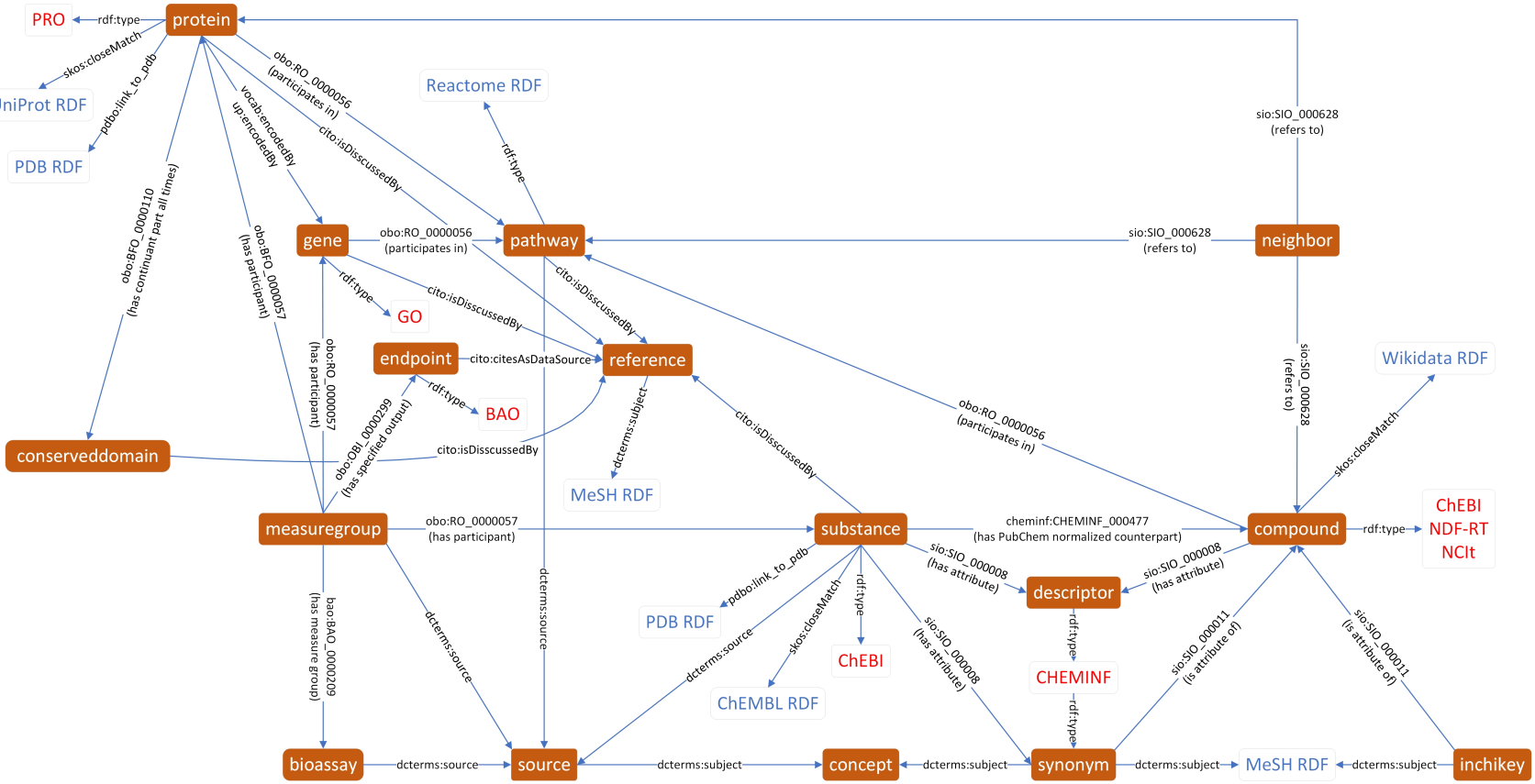


Figure – Ontologie PubChemRDF

Les ontologies en science : PubCHEM

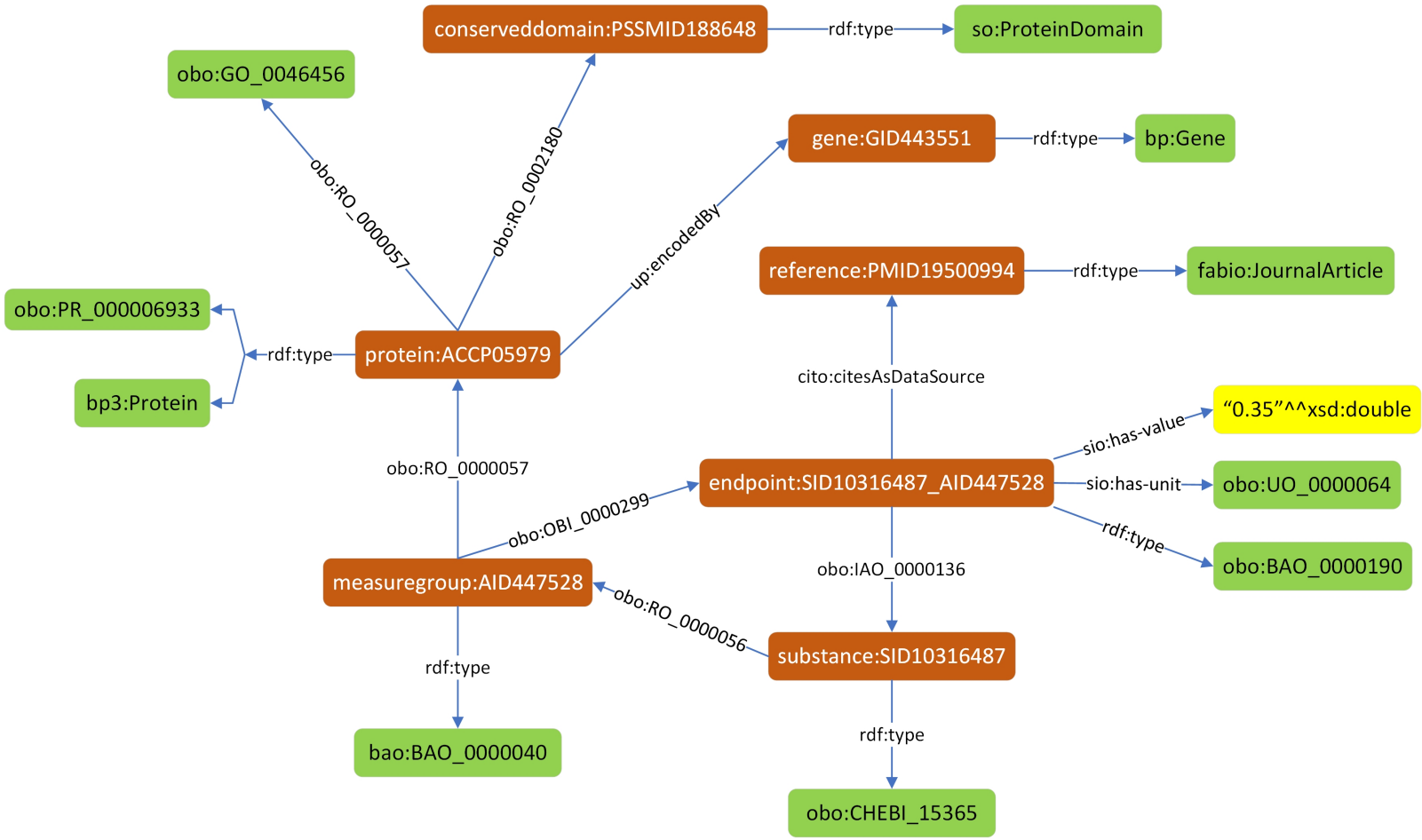


Figure – Faits sur PubChemRDF

Les ontologies en science : ChEBI

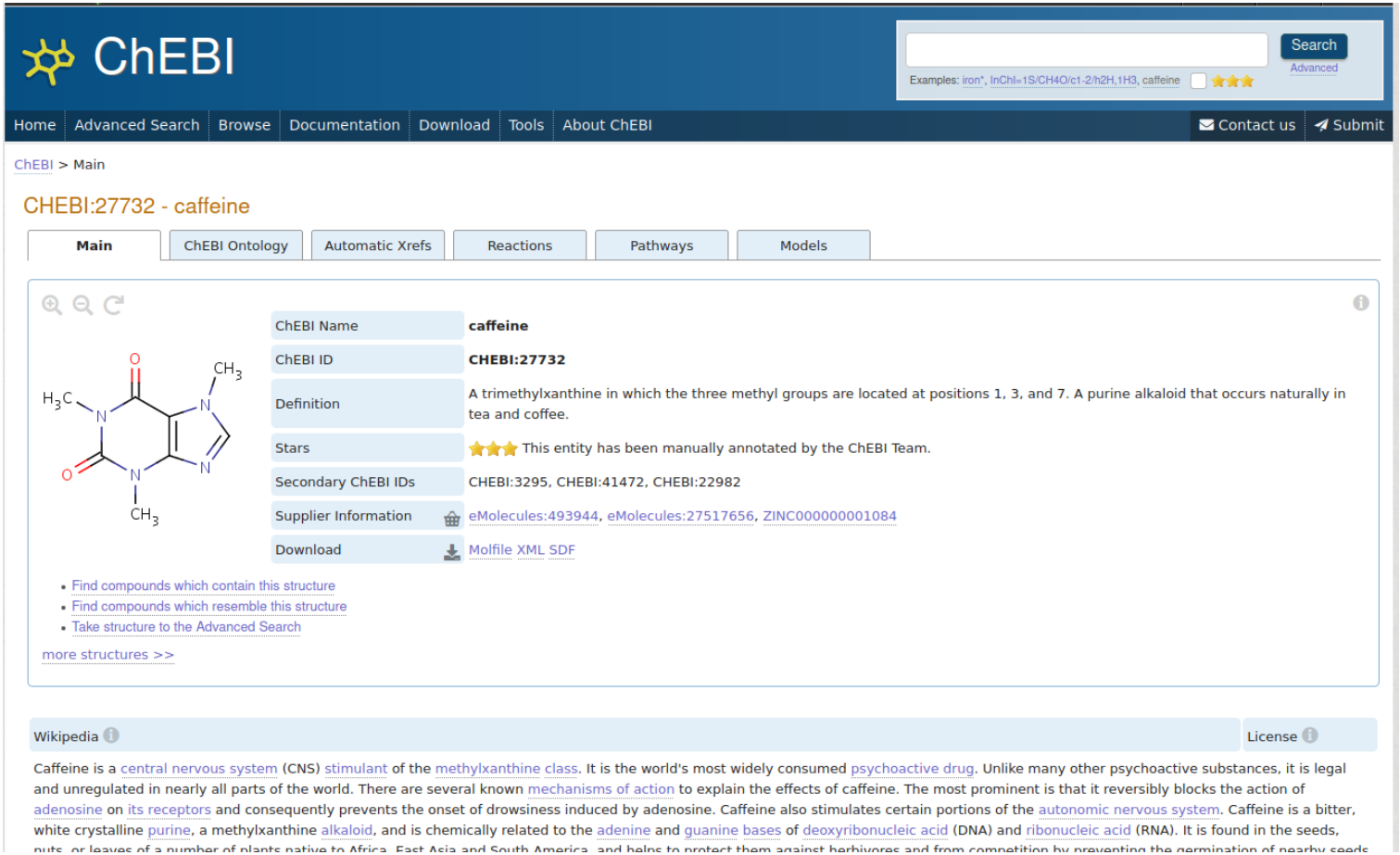


Figure – la caffeine sur ChEBI (Chemical Entities of Biological Interest)

En résumé

Un séminaire introductif

- à la gestion de données sous forme de graphes RDF
- aux ontologies
- aux méthodes de *clustering*

Campagne expérimentale sur les données de HAL

L'illustration sur les données de HAL montre :

- une interprétabilité des *clusters*,
- des effets originaux sur la qualité des *clusters*

Travaux futurs

Perspectives de recherche

Une ouverture vers le domaine du *semantic data mining*.

- Définir et mesurer l'intérêt subjectif avec les ontologies
  - modéliser plus précisement la mesure de l'intérêt
  - exploiter d'autres relations que `rdfs:subClassOf`
- Diriger le *clustering* avec les ontologies

TODO : exploiter des données réelles avec des ontologies existantes

Séminaire ISEA : informatique

Les bases de données graphes et une application à la fouille sur les données de HAL

romuald.thion@unc.nc

Vendredi 23 avril 2021 13h – amphi Guy Agniel