TELECOM ParisTech

Institut Mines-Télécom

SemBib

Dépôt local d'articles scientifiques sémantiquement décrits

Retour d'expérience

Télécom ParisTech - IDS

Jean-Claude Moissinac

TextMine - Janvier 2018

Objectifs de SemBib

- **Donner des outils pour observer et naviguer dans la production scientifique de Telecom ParisTech**
  - Thématiques et chercheurs associés
    - Notamment pour observer des thèmes transverses
  - Tendances
  - Donner des repères
  - Rendre visible des faits implicites
- **Explorer les possibilités offertes par la combinaison de méthodes NLP classiques avec des méthodes 'sémantiques'**
- **Disposer d'un jeu de données à but pédagogique**

Contexte: les publications scientifiques

- **Tendances**
  - Ouverture de leur accès
  - Outils d'indexation et de recherche à grande échelle
  - Augmentation rapide du nombre de publications

Nombre de publications

- **Tendances**
  - Ouverture de leur accès
  - Outils d'indexation et de recherche globaux
  - Augmentation du nombre
    - Doublement tous les 9 ans
- **Les initiatives se multiplient pour exploiter cette masse de données**
- **Notre approche**
  - Interconnecter des sources de données locales grâce aux représentations sémantiques

Groupes, projets…

Philosophie du web

S'appuyer sur des solutions décentralisées et interconnectées

Questions de droits

Autorisation de certains éditeurs pour publier sur le site d'une institution

=> outiller des dépots locaux pour faciliter leur mise en œuvre et faciliter leur interconnexion

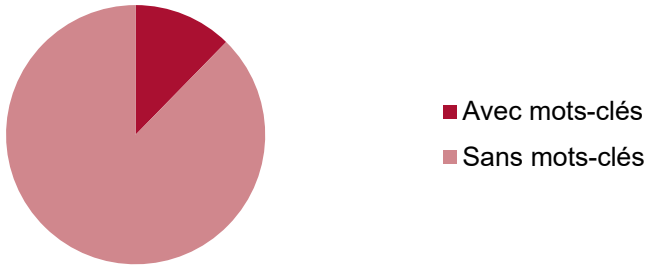
Cf atelier OpenMinTeD

Pourquoi un dépot local

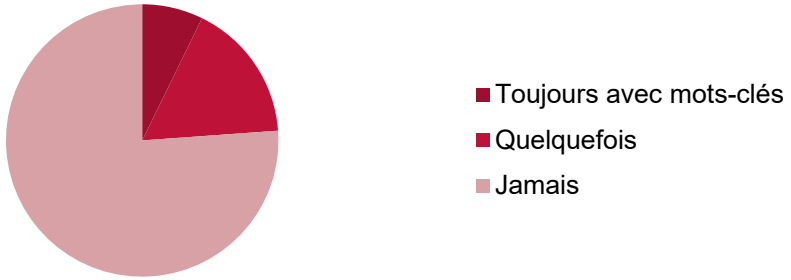
Sources de données internes

- **Serveur Bibliographique**
  - http://biblio.telecom-paristech.fr/
- De 1970 à 2017: 11311 publications référencées
  - De 7262 auteurs (avec les co-auteurs extérieurs ou étudiants)
- 1048 avec un lien (supposé) vers le document
- => besoin de récolter les documents (sur le Web)

Références



Auteurs



Sources nous représentant (mal?)

Source	Nombre d'entités associées à Télécom ParisTech
Sembib	11311 (20/11/2017)
HAL	3106
ISTEX	350
DBLP	321
SemanticScholar	124
ABES	81
Arxiv	1
CrossRef, OpenCitations Wikidata	0

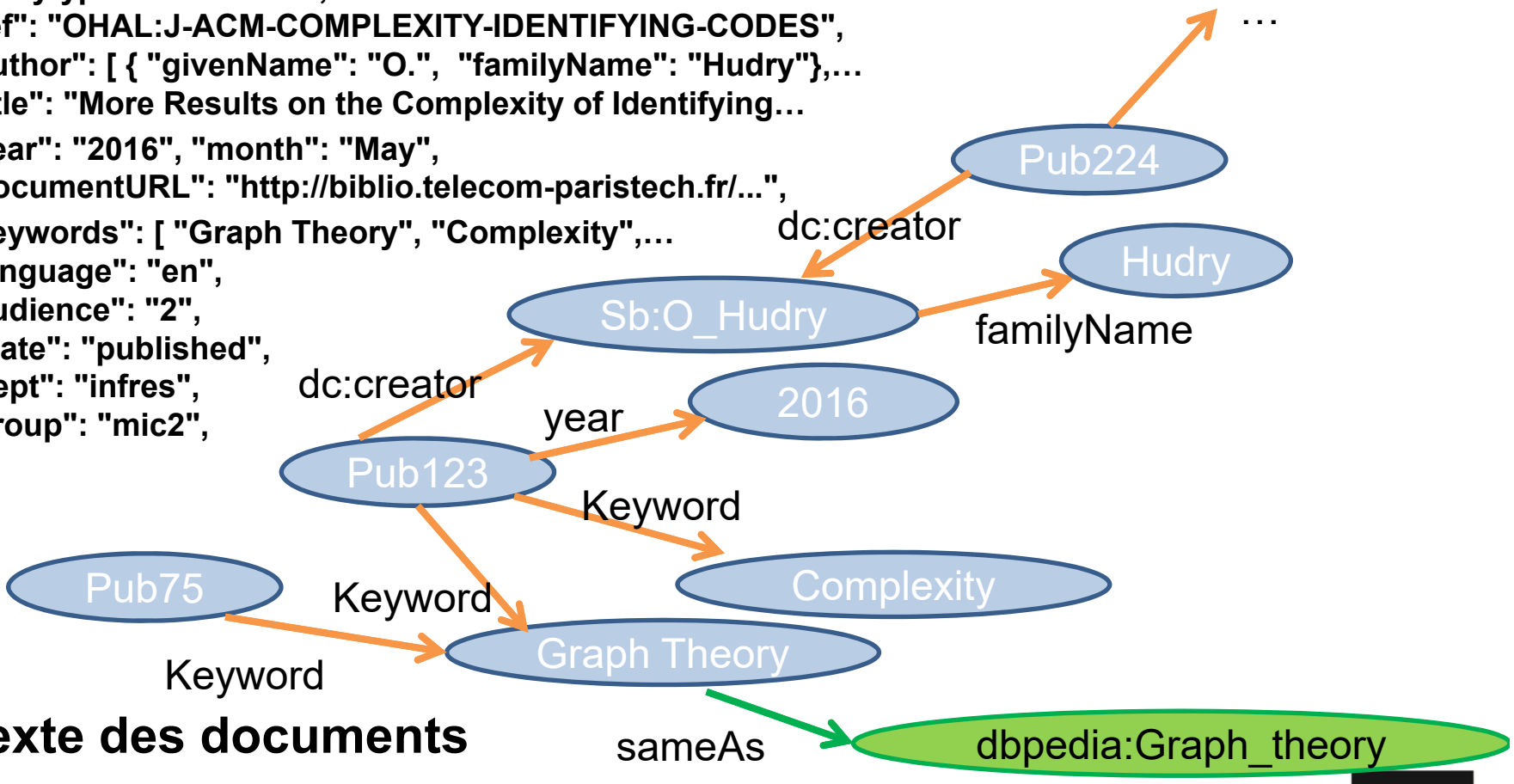
Et nombreuses autres sources: Sudoc, Microsoft Academic Graph, CrossRef…

Difficultés avec les données

- **Pour les récolter**
- **Pour les analyser**
  - Différences dans les formats de citations
  - Différences de structures
- **Incomplétude**
- **Inconsistance**
  - Des noms de personnes et d'institution
  - Des champs multiples
  - Variantes typographiques et abréviations
  - Références: DOI
- **Temporalité des données**
  - Affiliation des chercheurs change au cours du temps
  - Statuts des publications change

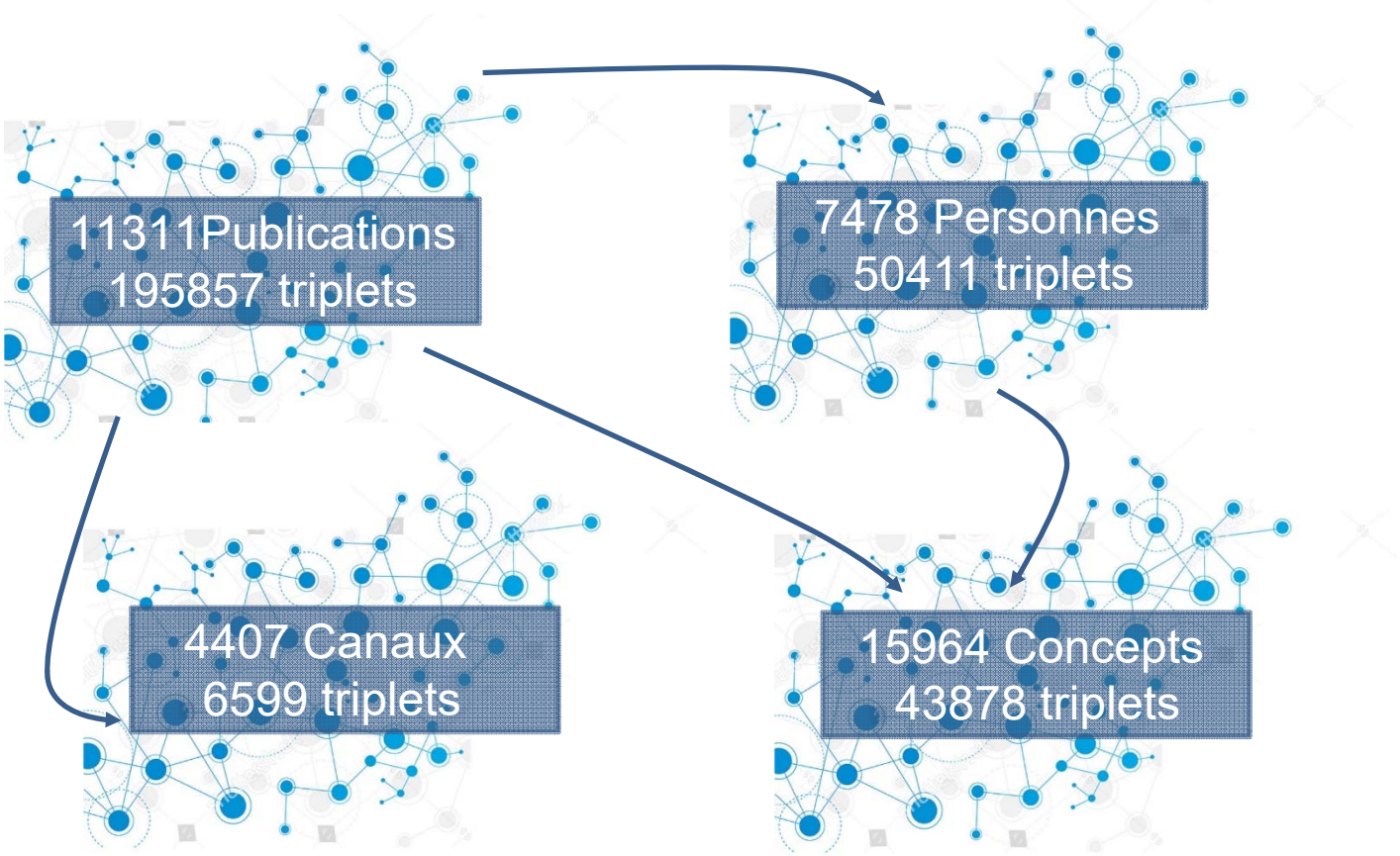
Vocabulaires: tirés des données

- **Structure de la base bibliographique d'origine**
  - **"entrytype": "ARTICLE",**
    "ref": "OHAL:J-ACM-COMPLEXITY-IDENTIFYING-CODES",
    "author": [{ "givenName": "O.", "familyName": "Hudry"},…
    "title": "More Results on the Complexity of Identifying…
    "year": "2016", "month": "May",
    "documentURL": "http://biblio.telecom-paristech.fr/...",
    "keywords": ["Graph Theory", "Complexity",…
    "language": "en",
    "audience": "2",
    "state": "published",
    "dept": "infres",
    "group": "mic2",



- **Texte des documents**

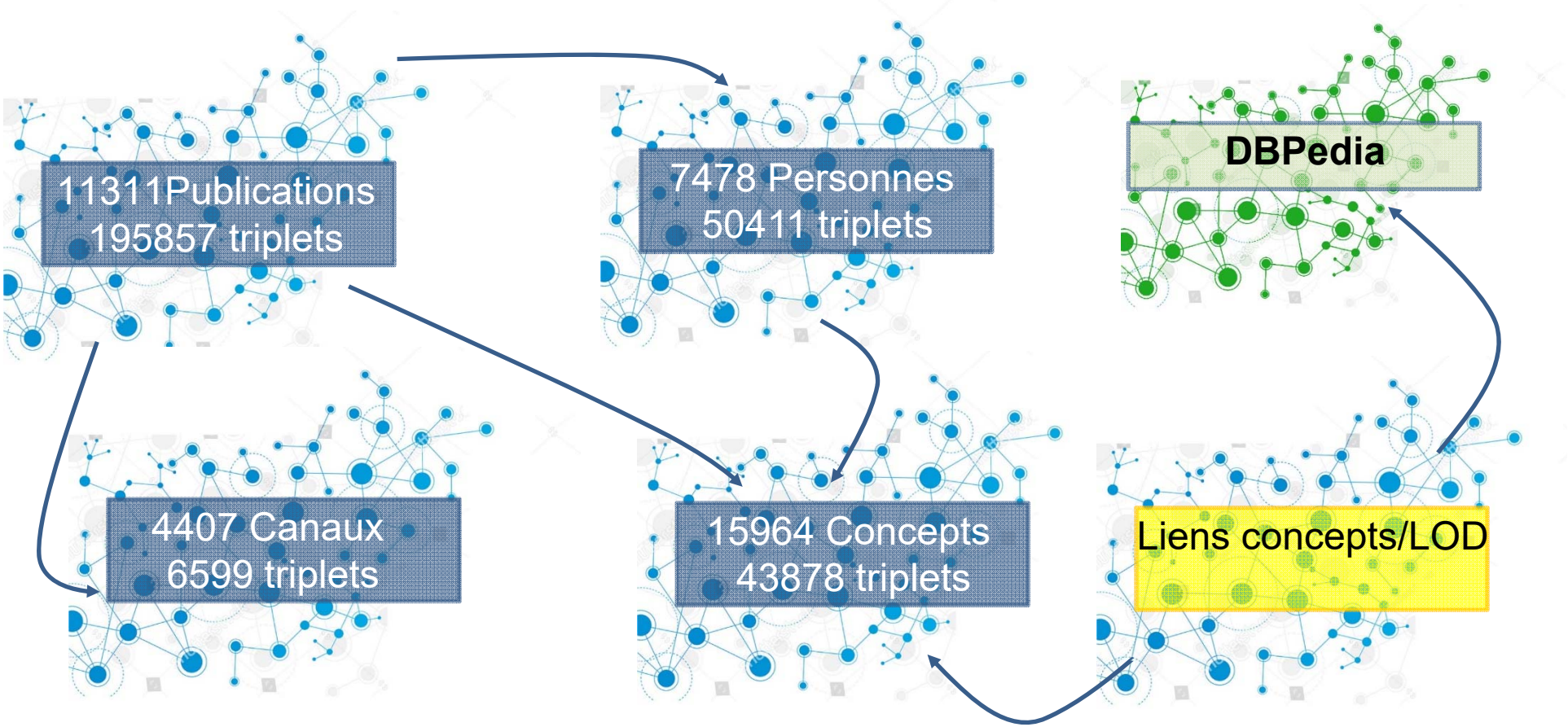
Quatre graphes principaux



RDF Stores : 3 expérimentations

- Apache Jena/Fuseki, ARC2, Virtuoso

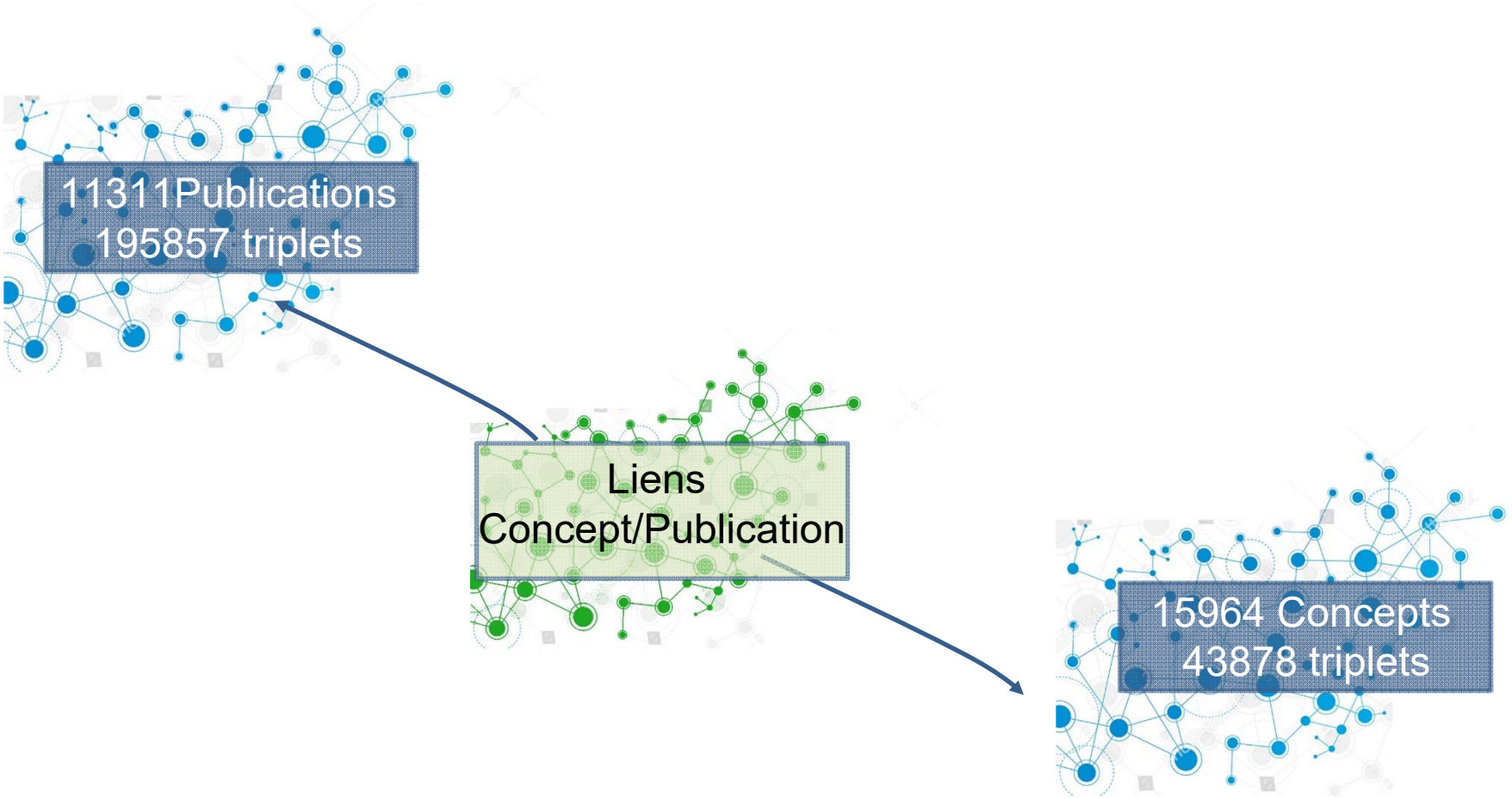
Une répartition en graphes séparés



RDF Stores : 2 expérimentations

- Apache Jena/Fuseki, ARC2

Une répartition en graphes séparés



La plus simple, mais offrant le moins de granularité

Sélection des graphes utilisés avec FROM

Requêtes sur les graphes séparés

FROM NAMED http://example.org/alice

Les graphes sont désignés dans la requête

GRAPH <http://example.org/alice> { …

But: Visibilité des données

Pages web enrichies en RDFa
http://givingsense.eu/sembib/onto/persons/tpta
http://givingsense.eu/sembib/onto/persons/David_Bertrand

Portail SemBib

http://givingsense.eu/sembib/onto/persons/David_Bertrand.json
Accès SPARQL
http://givingsense.eu/sembib/sparql
Actuellement sur une base ARC2, en cours Jean/Fuseki
ARC2 n'implémente pas complètement SPARQL 1.1

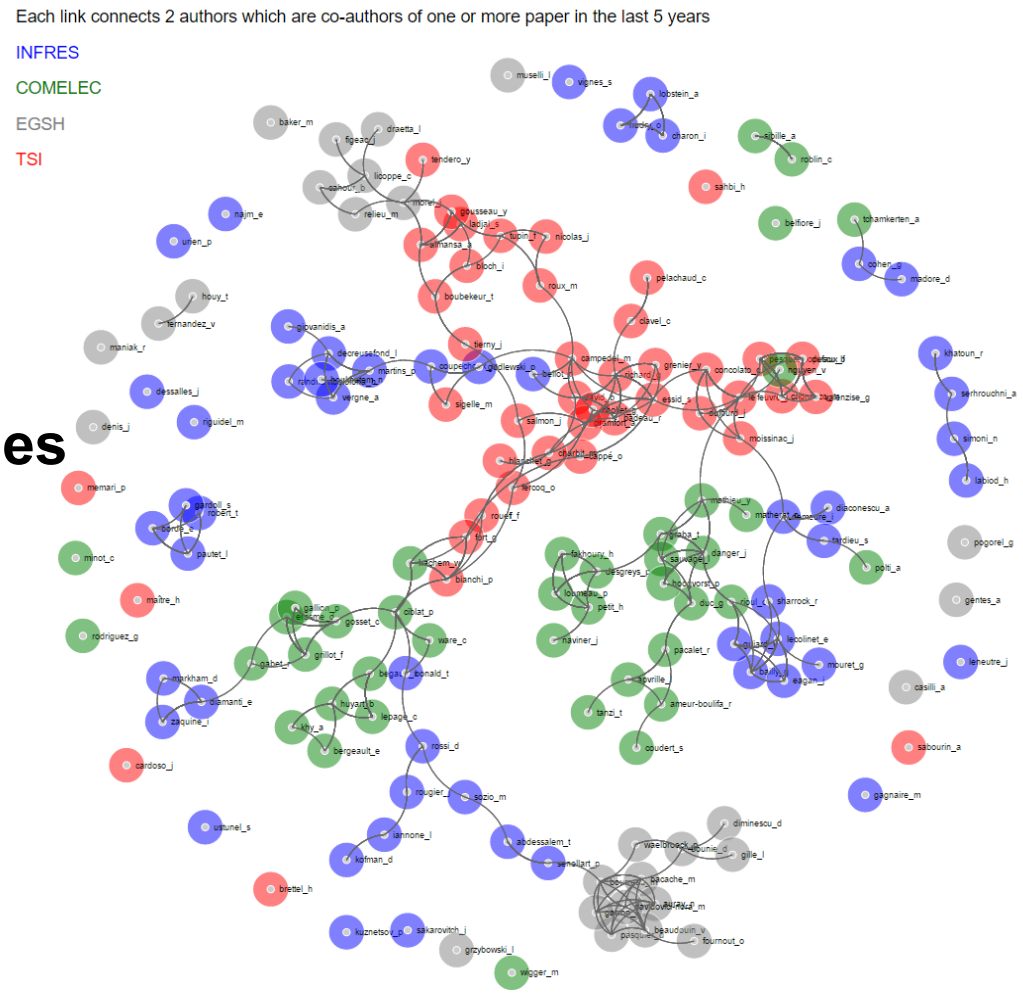
Publication des graphes

https://github.com/moissinac/sembib-graphs

Rendre visibles des faits implicites

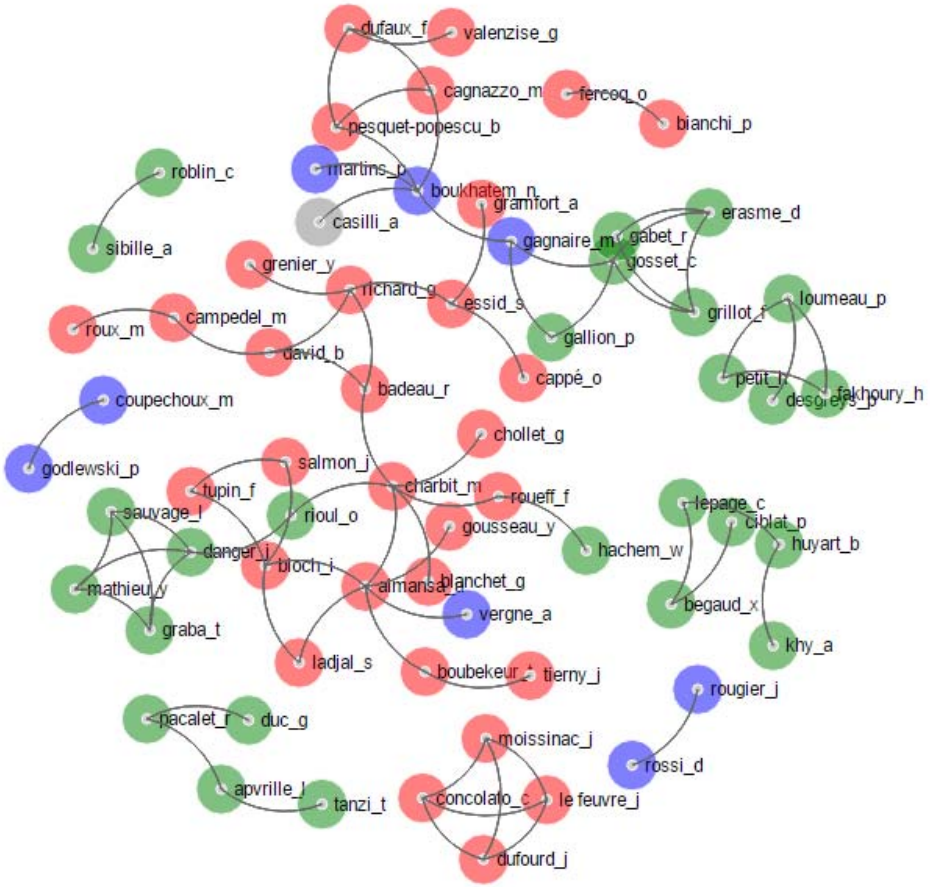
Possibilités

- Réseaux de co-publication
- Réseaux de citation
- Tendances thématiques
- Collaborations
- Cartels de citations
- …



Co-publication entre permanents

Auteurs partageants plusieurs mots-clés



51 différentes pour la série de conférences ICASSP

Collecte des articles

[redacted]00 collectés sur les 11311 connus

Extraction du texte et de sastructure (GROBID)

Vérifications (auteurs, affiliations…)

4937 documents conservés (au 23/1/2018)

Recherche de mots-clés

Travail sur la qualité des données

SemBib: Suivre l'avancement

- SemBib
  - Une construction progressive facilitée par la représentation RDF
  - Des données ouvertes organisées et traitées localement…
  - Reliées au Web des Données
- **https://onsem.wp.mines-telecom.fr**
  - https://onsem.wp.mines-telecom.fr/2016/06/02/extraire-le-texte-de-pdf-avec-python/
  - https://onsem.wp.mines-telecom.fr/2016/06/03/utiliser-nltk-sur-heroku-avec-python/
  - https://onsem.wp.mines-telecom.fr/2016/07/18/une-instance-de-fuseki-sur-openshift/

Lien avec des graphes externes (ex. DBPedia)

Conclusion et perspectives

Exploitation des connaissances externes

Dans les parcours de graphes
Dans les techniques d'apprentissage
Ex: vecteur de mots->vecteur de concepts->vecteur de classes

Représentation des textes dans les graphes

TELECOM ParisTech

Institut Mines-Télécom

Merci de votre attention

Eléments bibliographiques

Constantin,A.,S.Peroni,S.Pettifer,D.Shotton,etF.Vitali(2016). The document components ontology (**DoCO**). Semantic Web 7

Duan, S., A. Kementsietsidis, K. Srinivas, et O. Udrea (2011). Apples and oranges : a comparison of **RDF benchmarks** and real RDF datasets. In SIGMOD Conference, pp. 145–156. ACM. Larsen, P. O. et M. von Ins (2010). The rate of growth in scientific publication and the decline in coverage provided by science citation index.

Lopez, P. (2009). **Grobid** : Combining automatic bibliographic data recognition and term extraction for scholarship publications. In Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries,ECDL'09,Berlin,Heidelberg,pp.

Luggen, M., A. Gschwend, B. Anrig, et P. Cudré-Mauroux (2015). **Uduvudu** : a graph-aware and adaptive ui engine for linked data. In LDOW@WWW.

Mirizzi, R., T. D. Noia, E. D. Sciascio, et A. Ragone (2012). Using **DBpedia** for searching related terms in the IT domain. Technical report,PolitecnicodiBari,ViaOrabona,4,70125 Bari, Italy.

Moissinac, J.-C. (2017). Pour une fédération de dépôts locaux d'articles scientifiques sémantiquement reliés. In ToTh.

Rizzo, G., Tomassetti Federico, A. Vetrò, L. Ardito, M. Torchiano, Morisio Maurizio, et R. Troncy (2015). Semantic enrichment for recommendation of primary studies in a systematic literature review. Digital Scholarship in the Humanities, Oxford University Press, 13 August 2015.

Sateli,B.,F.Löffler,B.König-Ries,etR.Witte(2016). Semantic user profiles:Learning scholars' competences by analyzing their publications. In Semantics, Analytics, Visualisation : Enhancing Scholarly Data (SAVE-SD 2016). Springer : Springer.

Tang,J.,J.Zhang,L.Yao,J.Li,L.Zhang,etZ.Su(2008). **Arnetminer** : extractionandmining of academic social networks. ACM Knowledge Discovery and Data Mining, 990–998. Vincent,G.,J.-C.

Moissinac,etA.Luc(2014). Automated generation of **a "lossless semantic" eBook.** In 17ème Colloque International sur le Document Numérique (CIDE 17), le livre post-numérique : historique, mutations et perspectives., Fès, Morocco.

Vocabulaires externes

- **Taxonomies et concepts**
  - ACM (https://www.acm.org/publications/class-2012)
  - DBPedia
- **Ontologies -> recherche avec LOV**
  - SPAR (http://www.sparontologies.net/)
  - ensemble d'ontologies pour les publications scientifiques, dont
    - fabio: description d'un document (basé sur frbr et rdfs)
    - biro: représentation de citation
    - cito: typage de citations