Institut
Mines-Télécom

SemBib
Sémantique et Publications

Télécom ParisTech - SID

Jean-Claude Moissinac – SebWeb Pro

Novembre 2016

Objectif

- **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’**

Contexte: les publications scientifiques

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

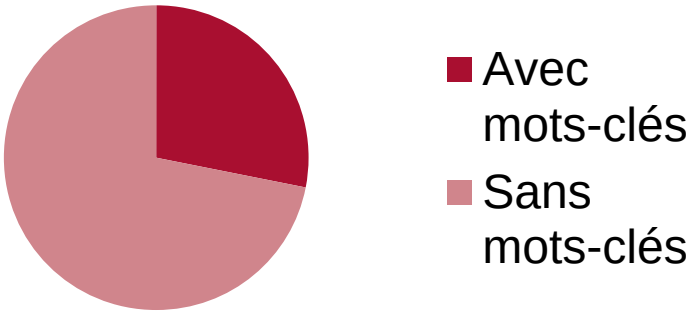
Nombre de publications

- **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 avec la sémantique

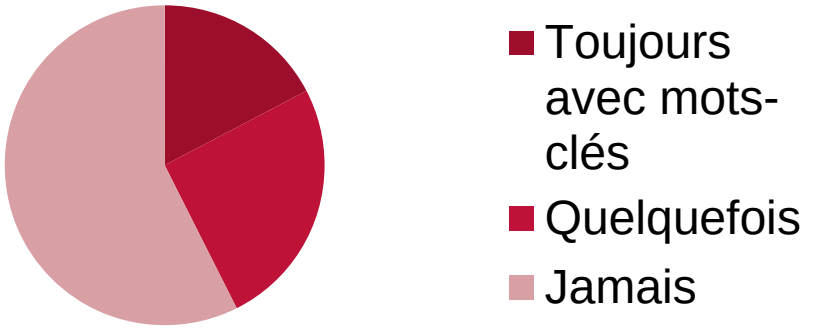
Sources de données internes

- **Serveur Bibliographique**
  - http://biblio.telecom-paristech.fr/
  - De 2011 à 2015: 3797 publications référencées
    - De 3093 auteurs (avec les co-auteurs extérieurs ou étudiants)
  - 1313 avec un lien (supposé) vers le document

Références



Auteurs



- **Rapports d'activités**
  - Mots-clés, tendances thématiques…

Sources de données externes

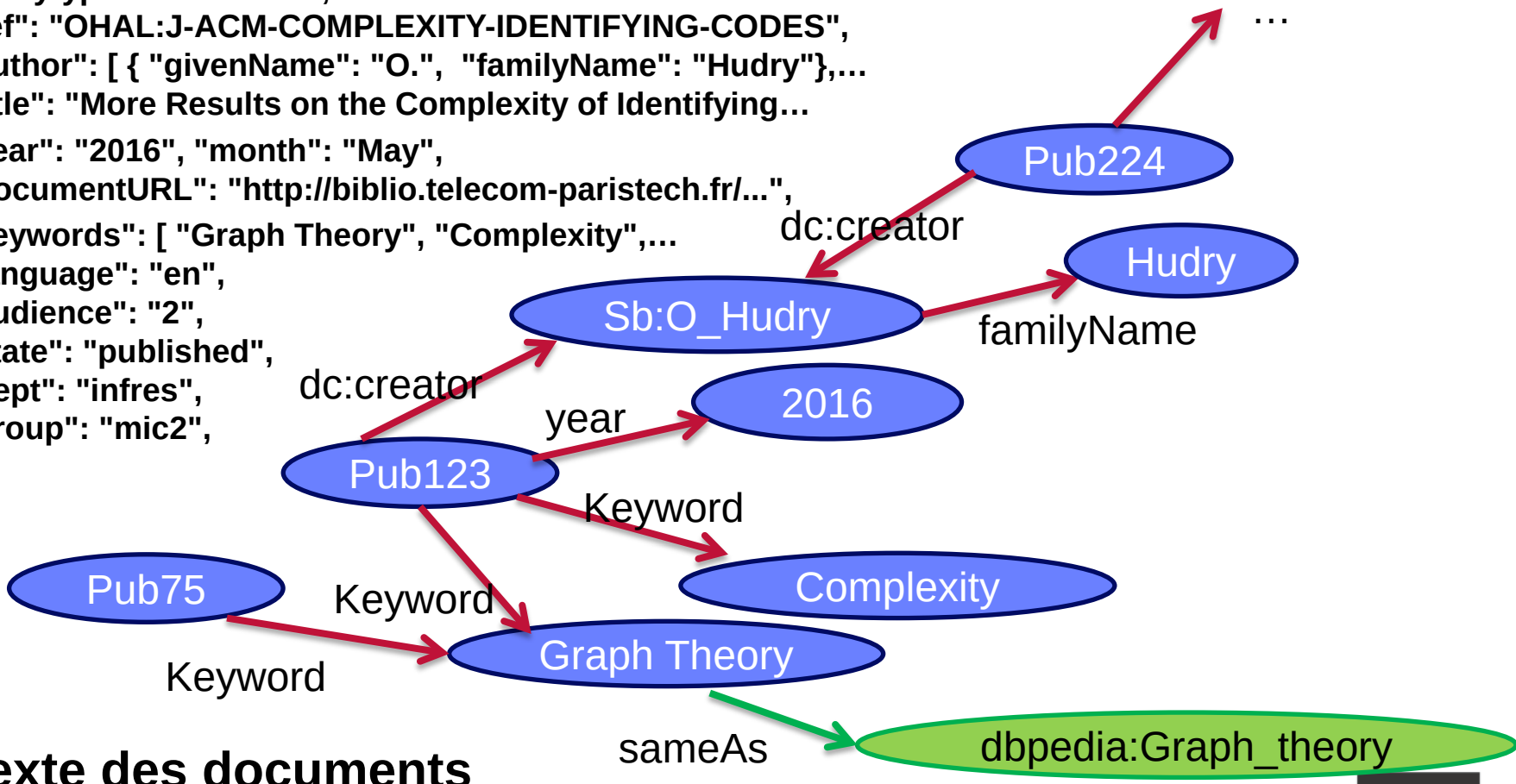
- **Springer (accès sparql)**
- **Google Scholar**
  - pas d'API ou de données publiques
- **SemanticScholar**
- **Academic.microsoft.com**
- **DBLP**
- **CrossRef**
- **OpenCitations**
- **ArXiv**
  - Limitations d'accès
- **Sudoc (base des thèses)**
- **Dbpedia (citations and references challenge)**

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 bibliographie**
  - **"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**

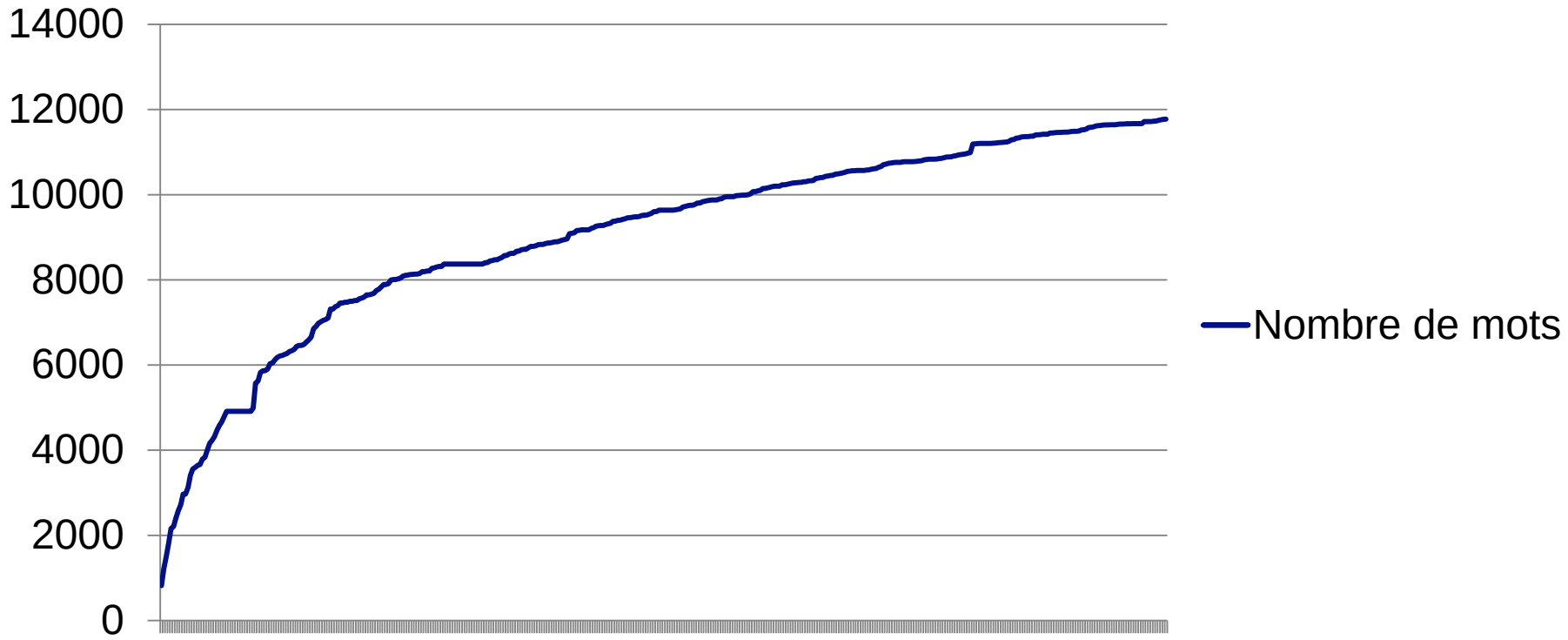
Mots-clés de la base bibliographique

Mot-clé	Usage	Auteurs
security	162	40
access control	90	12
audio source separation	90	14
exposure index	90	28
transmission techniques	76	26
beam combining	74	11
nonnegative matrix factorization	69	15
particle filter	60	11
optical packet switching	60	14
cloud	54	17
adaptive optics	50	14
lisp	50	14
sysml	48	7
visualization	46	23
complexity	44	11
propagation	44	14
laser	42	20
interaction techniques	40	19
graph theory	40	5
cognitive radio	39	13
music information retrieval	39	7
electroencephalography (eeg)	38	11
magnetoencephalography (meg)	38	11

Seulement 39 mots-clés présents plus de 5 fois dans la base

Vocabulaire sélectionné dans les textes

- **Augmentation de la taille du vocabulaire lors de l'ajout d'un article (avec tfIdf sur 418 articles)**



Les mots les plus significatifs (TfIdf) sont ajoutés comme « mots-clés »

Vocabulaires externes

- **Taxonomies**
  - Ex: ACM (https://www.acm.org/publications/class-2012)
- **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

Processus

Sources Web

Crawler4J

PdfMiner

Bag of Words

Fuseki/Sparql

Fuseki

NLTK

LSI

Latent Semantic Indexing

TF-IDF

Term x within document y

RDF Store

- **Fuseki**
- 68 686 379 triplets (au 20/11/2016)
- 3526 publications décrites
- 14083 mots liés à des publications
  - Liage de ces mots avec DBPedia en cours

Rendre visibles des faits implicites

- Réseaux de co-publication
- Réseaux de citation
- Tendances thématiques
- Dynamique des collaborations
- Cartels de citations

Co-publication entre permanents

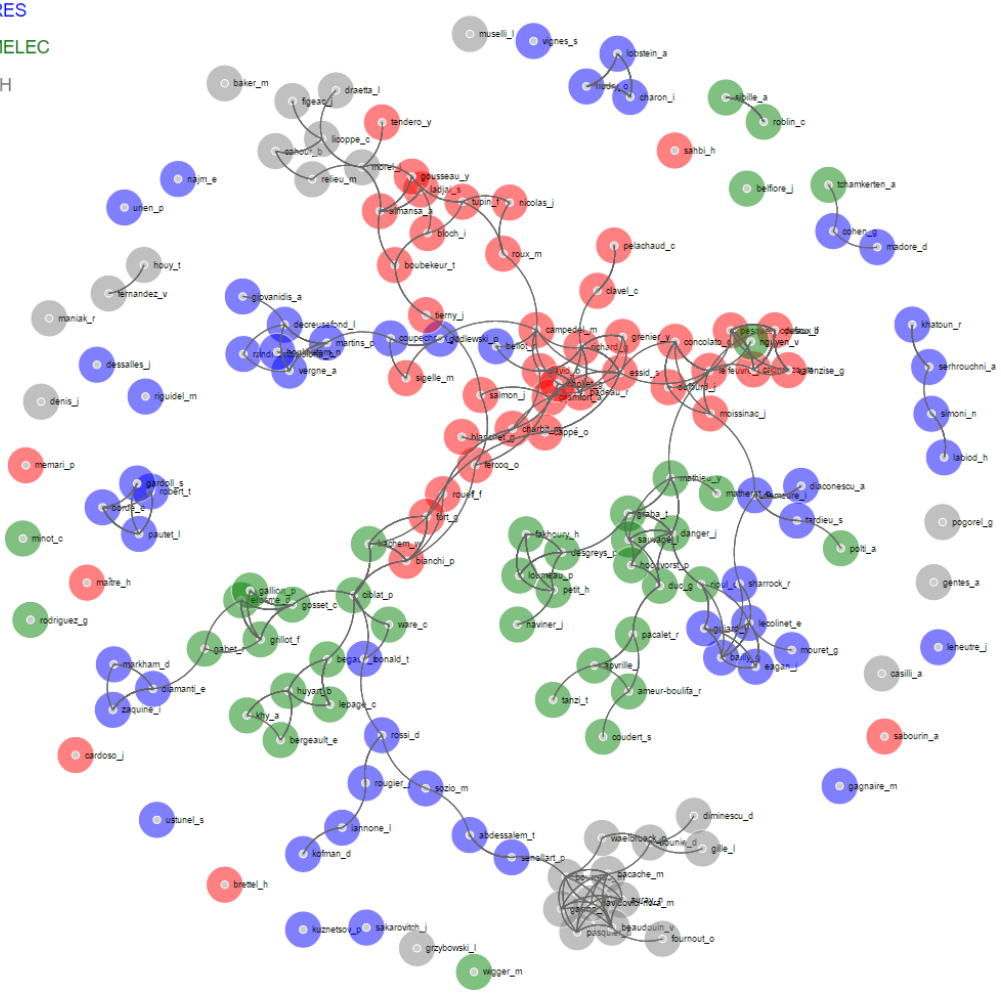
Each link connects 2 authors which are co-authors of one or more paper in the last 5 years

INFRES

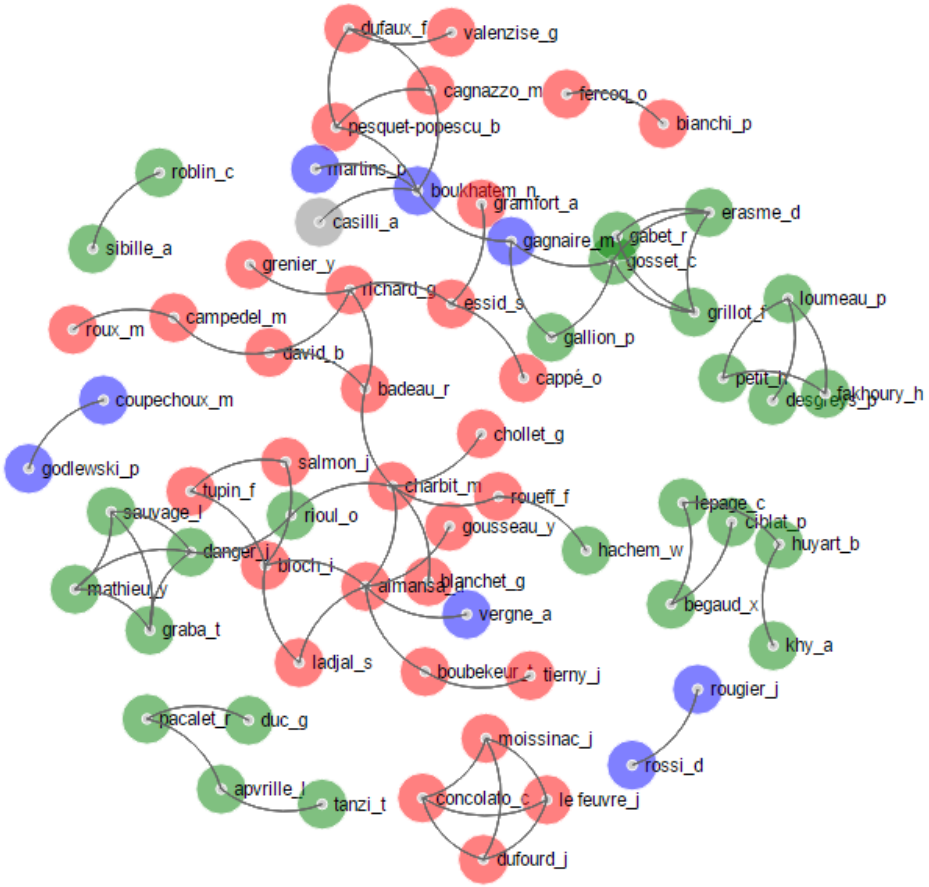
COMELEC

EGSH

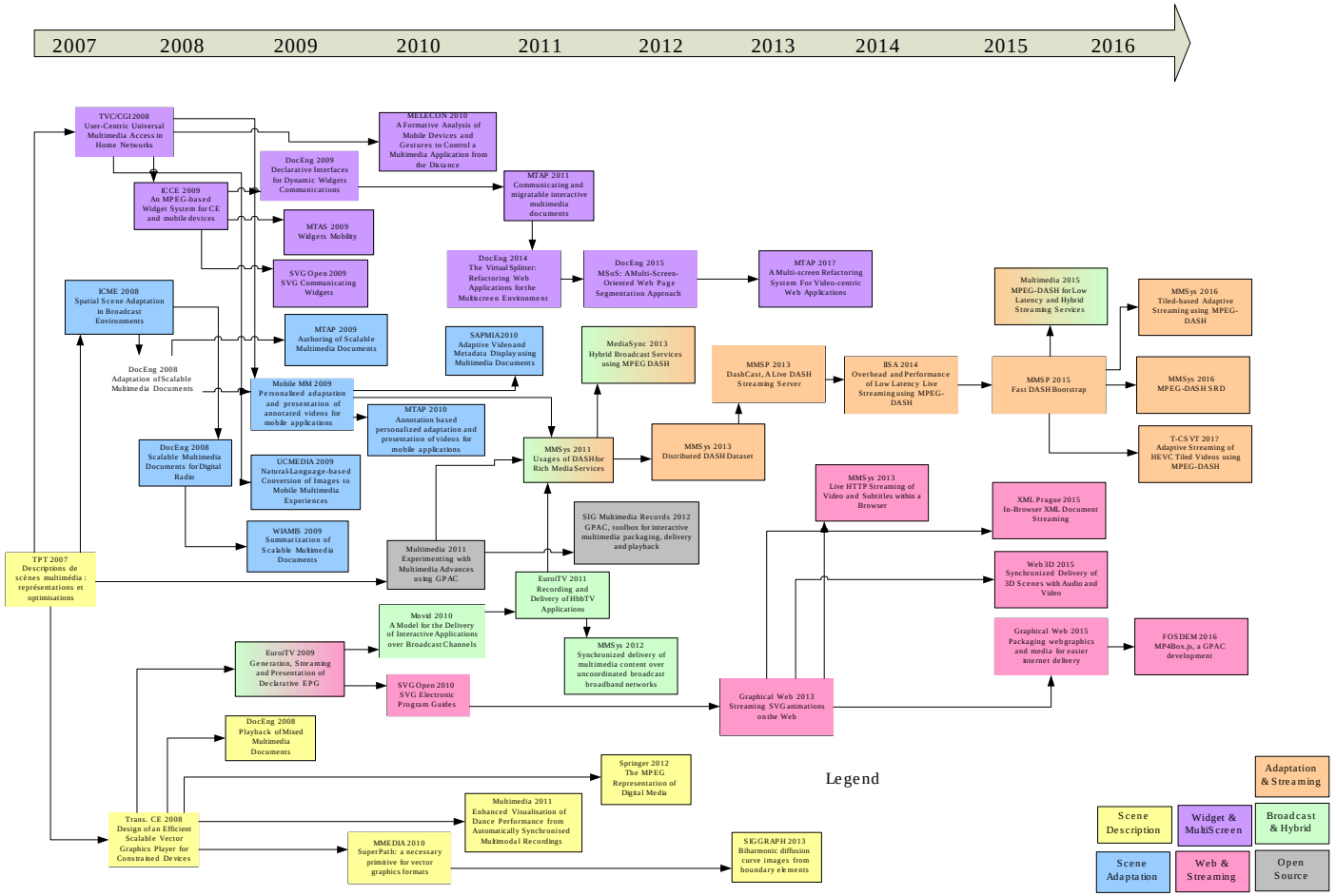
TSI

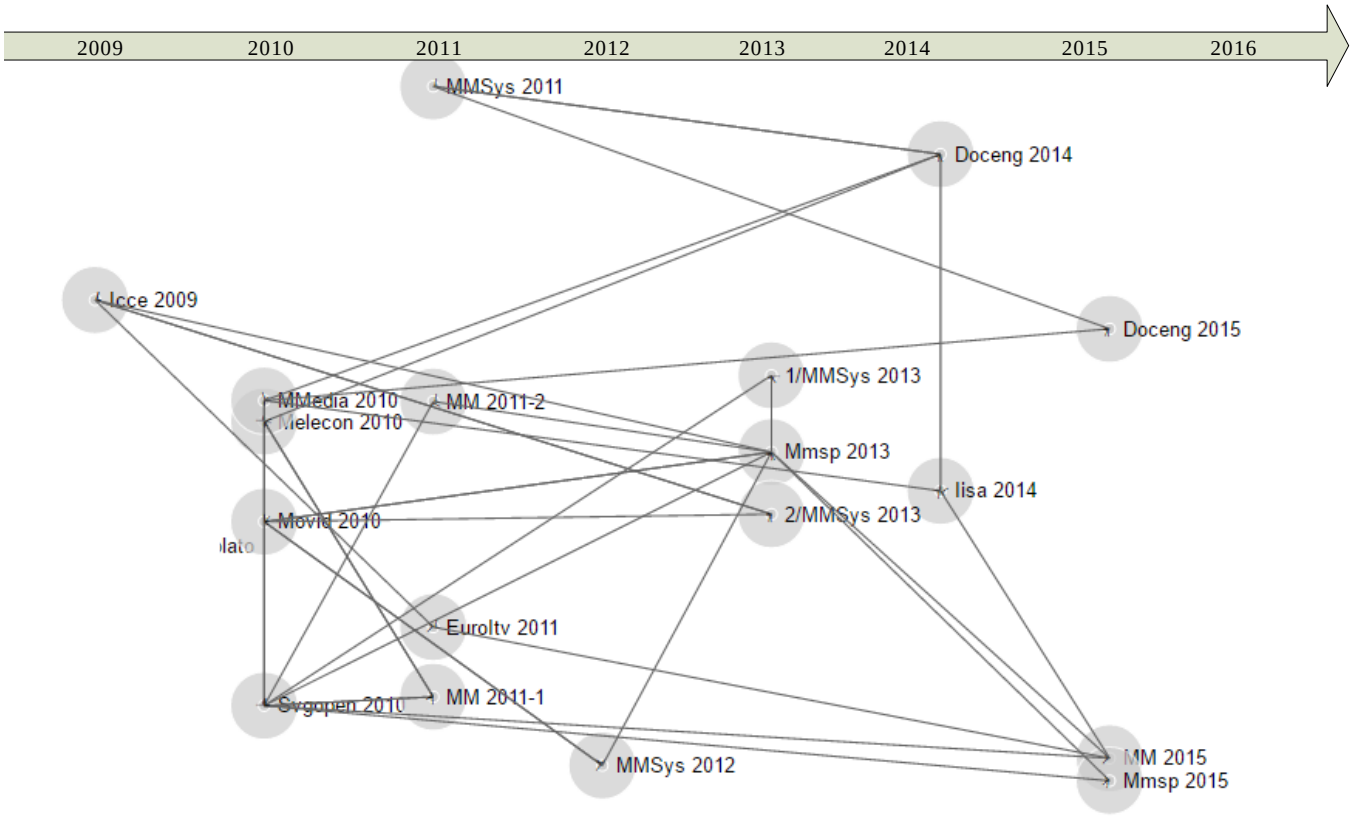


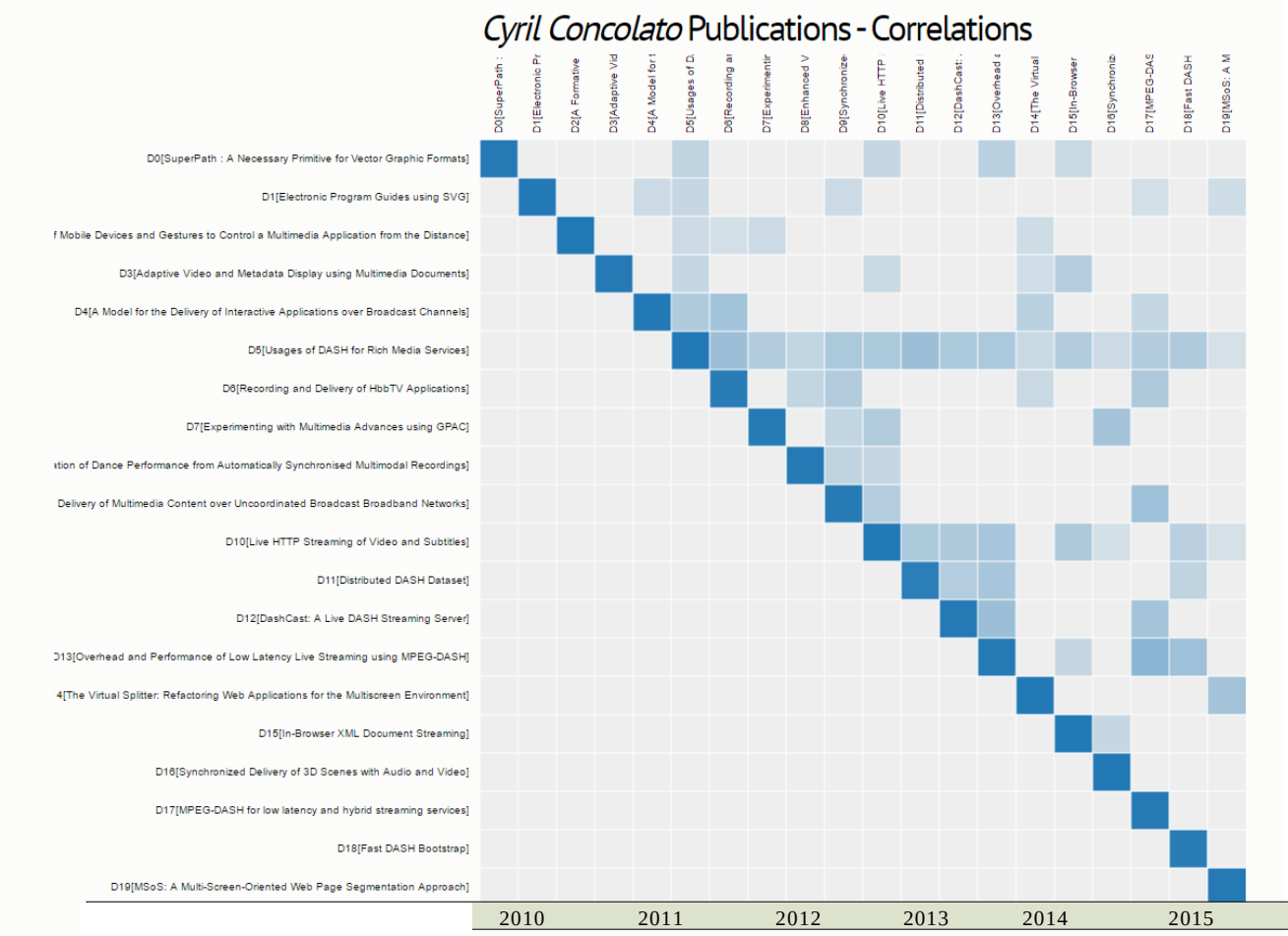
Auteurs partageants plusieurs mots-clés



Timeline des publications d'un chercheur




2009
2010
2011
2012
2013
2014
2015
2016
MMSys 2011
Doceng 2014
Icce 2009
Doceng 2015
1/MMSys 2013
MMedia 2010
MM 2011-2
Mmsp 2013
Iisa 2014
2/MMSys 2013
Euroltv 2011
MM 2011-1
MMSys 2012
MM 2015
Mmsp 2015


Cyril Concolato Publications - Correlations
D0[SuperPath : A Necessary Primitive for Vector Graphic Formats]
D1[Electronic Program Guides using SVG]
f Mobile Devices and Gestures to Control a Multimedia Application from the Distance]
D3[Adaptive Video and Metadata Display using Multimedia Documents]
D4[A Model for the Delivery of Interactive Applications over Broadcast Channels]
D5[Usages of DASH for Rich Media Services]
D6[Recording and Delivery of HbbTV Applications]
D7[Experimenting with Multimedia Advances using GPAC]
stion of Dance Performance from Automatically Synchronised Multimodal Recordings]
Delivery of Multimedia Content over Uncoordinated Broadcast Broadband Networks]
D10[Live HTTP Streaming of Video and Subtitles]
D11[Distributed DASH Dataset]
D12[DashCast: A Live DASH Streaming Server]
D13[Overhead and Performance of Low Latency Live Streaming using MPEG-DASH]
4[The Virtual Splitter: Refactoring Web Applications for the Multiscreen Environment]
D15[In-Browser XML Document Streaming]
D16[Synchronized Delivery of 3D Scenes with Audio and Video]
D17[MPEG-DASH for low latency and hybrid streaming services]
D18[Fast DASH Bootstrap]
D19[MSoS: A Multi-Screen-Oriented Web Page Segmentation Approach]
2010
2011
2012
2013
2014
2015
2016

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/

Institut
Mines-Télécom

Merci de votre attention