API Hour!

Des outils à consommer sans modération pour augmenter le taux de full-texts dans HAL

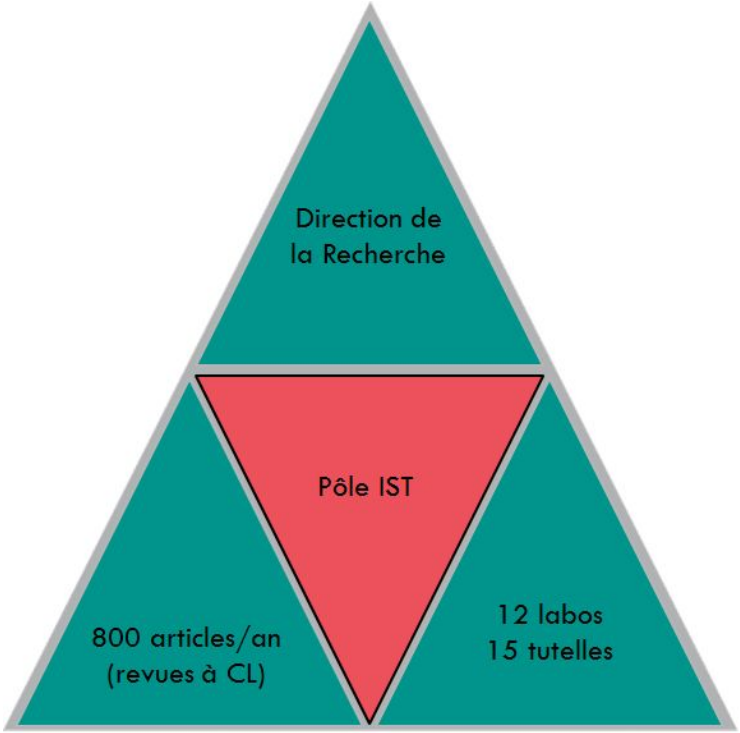
Romain Boistel - Frédérique Bordignon

Journées CasuHAL - 18 juin 2019

École des Ponts
ParisTech

casuHAL
club utilisateur

Contexte général



Contexte politique

- Mandat de l'École des Ponts (2017) pour le partage de la science
- Proposition d'un bonus Open Science au Conseil des labos (mars 2018)
- Décision finale : repérage des chercheurs mauvais élèves et mailing
- Soutien de la Direction de la Recherche
- Contexte (HCERES) propice ~~au dépôt~~ à la création de notices
- Nouvelle campagne en 2019

Des outils et des APIs

Sources des données

unpaywall

DOAJ

HAL archives-ouvertes.fr

Crossref

OCdHAL

Outils de traitement

Excel

Google Sheets

OU

Script en ligne de Philippe Gambette

Outils pour le mailing

YAMM

École des Ponts ParisTech

+ableau

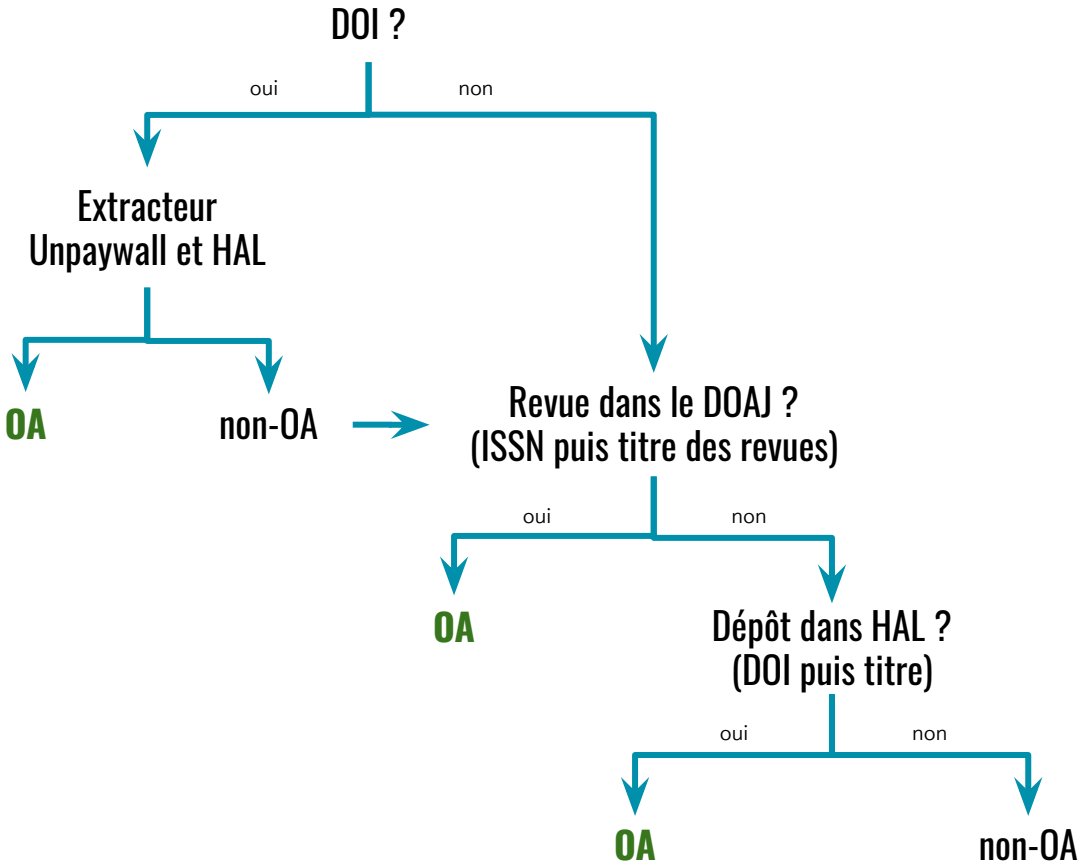
Google Fusion Tables

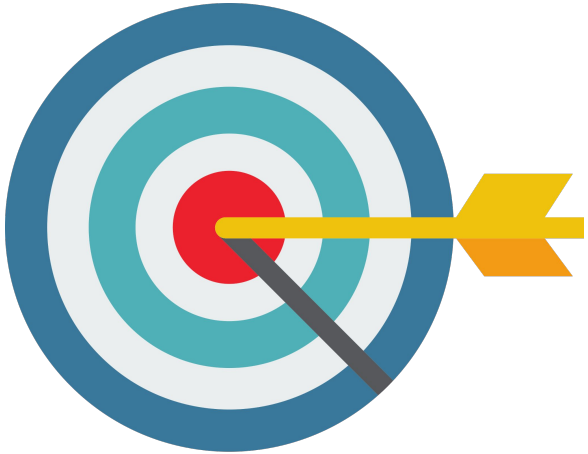
Gmail

Étape 1
Identifier les full-texts qu'on veut récupérer

Identifier ce qui est déjà en OA dans un corpus

Ceinture-bretelles...





Filtrage, nettoyage

Sont considérés en OA tous les articles pour lesquels :

- sur la base du DOI :
  - Unpaywall renvoie une réponse positive sur l'article, ou sur la revue dans le DOAJ
  - HAL renvoie une réponse positive sur la présence d'un full-text ou si *LinkExtId* renvoie une réponse sauf si c'est Istex
- sur la base de la revue :

  on trouve le titre ou l'ISSN de la revue dans l'export csv du DOAJ
- sur la base du titre de l'article :

  on trouve s'ils sont en OA dans l'export de la collection ENPC (*fileMain* ou *linkExtId*)

Ultimes vérifications manuelles pour détecter certaines barrières mobiles, certains articles dont le titre existe en 2 langues, les articles qui n'ont pas de DOI etc...

Fonctionnement d'une API

Plusieurs types d'API. Ici on utilise des API REST avec la méthode GET :

- Requête depuis la barre d'adresse du navigateur : une URL et des paramètres
- Uniquement pour récupérer des données, pas pour écrire dans une base de données
- Réponse du serveur dans de multiples formats possibles : csv, xml, json, html...
- Une API vient toujours avec une documentation : HAL, Unpaywall...

http://site.com/?requete=valeur1&champs_demandes=valeur2&format_reponse=valeur3

&authentification=email_ou_cle

Racine de l'URL Paramètres

Limites fréquentes : nombre de requêtes / heure ; nombre de réponses / requête...

Focus sur l'API d'Unpaywall

Documentation : https://unpaywall.org/products/api

- Réponse en JSON
- Champs utilisés par l'extracteur :
  is_oa ; journal_is_oa ; journal_is_in_doaj ; best_oa_location/evidence
- Autres infos disponibles : URL du PDF, couleur OA, informations bibliographiques...

Extension JSONView

Non-OA selon Unpaywall :

https://api.unpaywall.org/v2/10.1007/s11401-012-0755-7?email=mail@mail.com

OA selon Unpaywall :

https://api.unpaywall.org/v2/10.1007/s40314-013-0076-9?email=mail@mail.com

Focus sur l'API de HAL

https://api.archives-ouvertes.fr/search/?wt=csv&fq=docType_s:("ART"OR"UNDEFINED")&fq=collCode_s:"ENPC"&rows=10000&fq=producedDateY_i:[2015 TO 2018]&fl=halId_s,authLastNameFirstName_s,producedDateY_i,title_s,journalIssn_s,journalEissn_s,doiId_s,journalTitle_s,docType_s,fileMain_s,labStructName_s,citationFull_s,linkExtId_s,linkExtUrl_s

Version JSON Version XML

https://api.archives-ouvertes.fr/search/	URL de l'API de HAL
?wt=**csv**	Format de retour csv
&fq=docType_s:("**ART**"OR"**UNDEFINED**") &fq=collCode_s:"**ENPC**" &fq=producedDateY_i:[**2015** TO **2018**]	Critères de requêtes **articles** ou **preprints**, collection **ENPC**, **2015** à **2018**
&rows=10000	Limite de 30 résultats retournés par défaut
&fl=**halId_s**,**fileMain_s**,**linkExtId_s**,**linkExtUrl_s**,**citationFull_s**,authLastNameFirstName_s,producedDateY_i,title_s,journalIssn_s,journalEissn_s,doiId_s,journalTitle_s,docType_s,labStructName_s	Valeurs retournées **identifiant HAL**, **full-text sur HAL**, **OA externe**, **citation bibliographique**

Pratique

API HAL &
API Unpaywall

Rendez-vous ici

pour les ressources nécessaires :

http://bit.ly/tutocasuhal19

(lien non pérenne, voir les pages de CasuHAL 2019
pour accéder au document associé)

Focus sur l'extracteur Unpaywall-HAL

- Script ImportJSON pour collecter proprement les données en JSON dans un GoogleSheet
- API d'Unpaywall (oaDOI) pour :
  - détecter s'il existe une version OA de l'article
  - détecter si la revue est dans le DOAJ
- API de HAL pour :
  - détecter les full-texts déjà présents dans HAL
  - collecter à nouveau les résultats d'Unpaywall mais via HAL
  - récupérer les *halId*

Quota de Google !

Démo de l'extracteur Unpaywall - HAL

http://bit.ly/ExtracteurUnpaywallHAL

Faites une copie !

Passez à la pratique !

Alternative

Extracteur expérimental de P Gambette

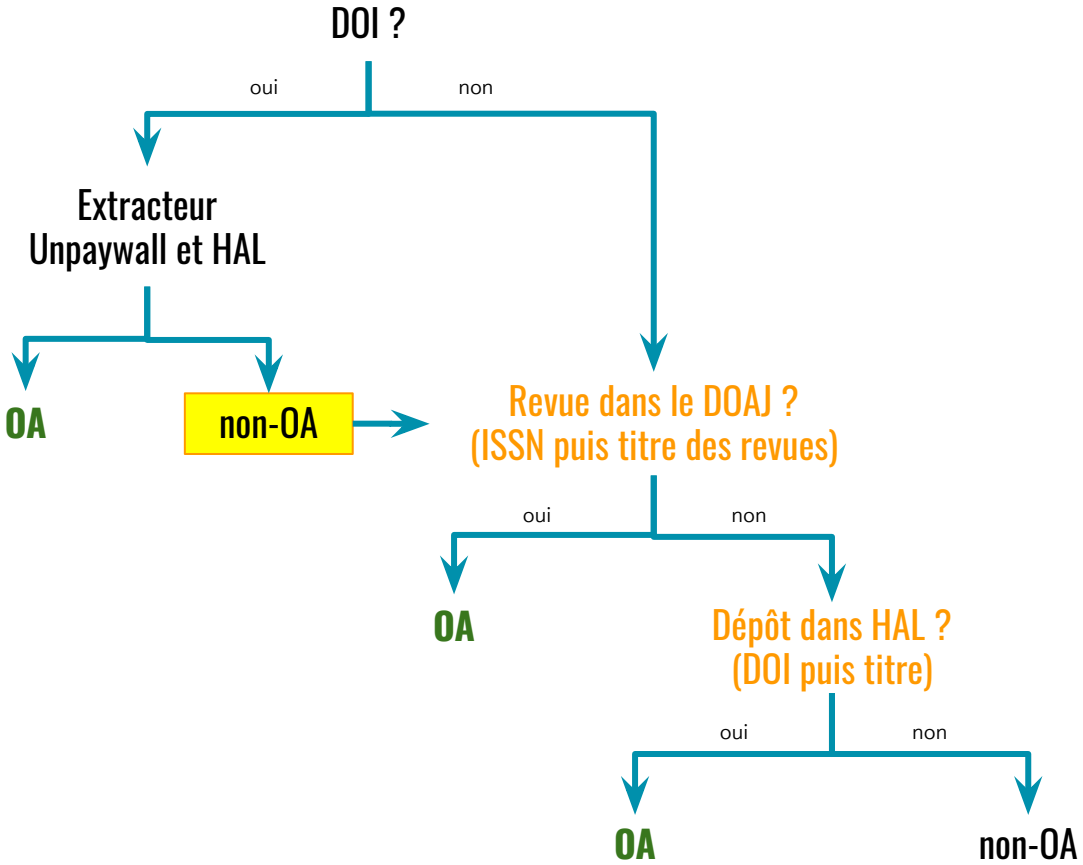
http://igm.univ-mlv.fr/~gambette/ExtractionHAL/ExtracteurUnpaywall/

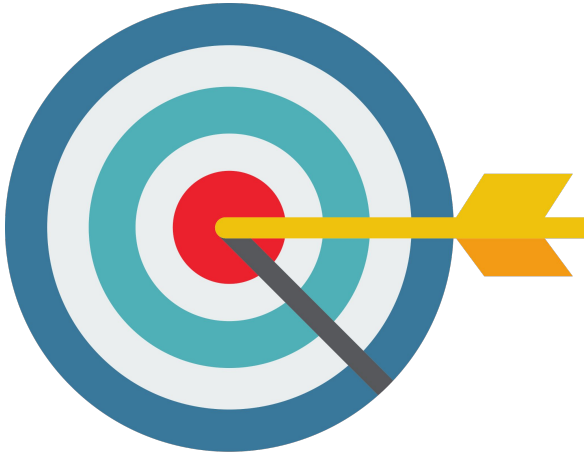
Passez à la pratique !

Pratique

Identifier ce qui est déjà en OA dans un corpus

Ceinture-bretelles...





Interroger le DOAJ et sonder le corpus exporté de HAL

- Vérifier si la revue est dans le DOAJ (ISSN puis Titre)
  `RechercheV` dans Excel

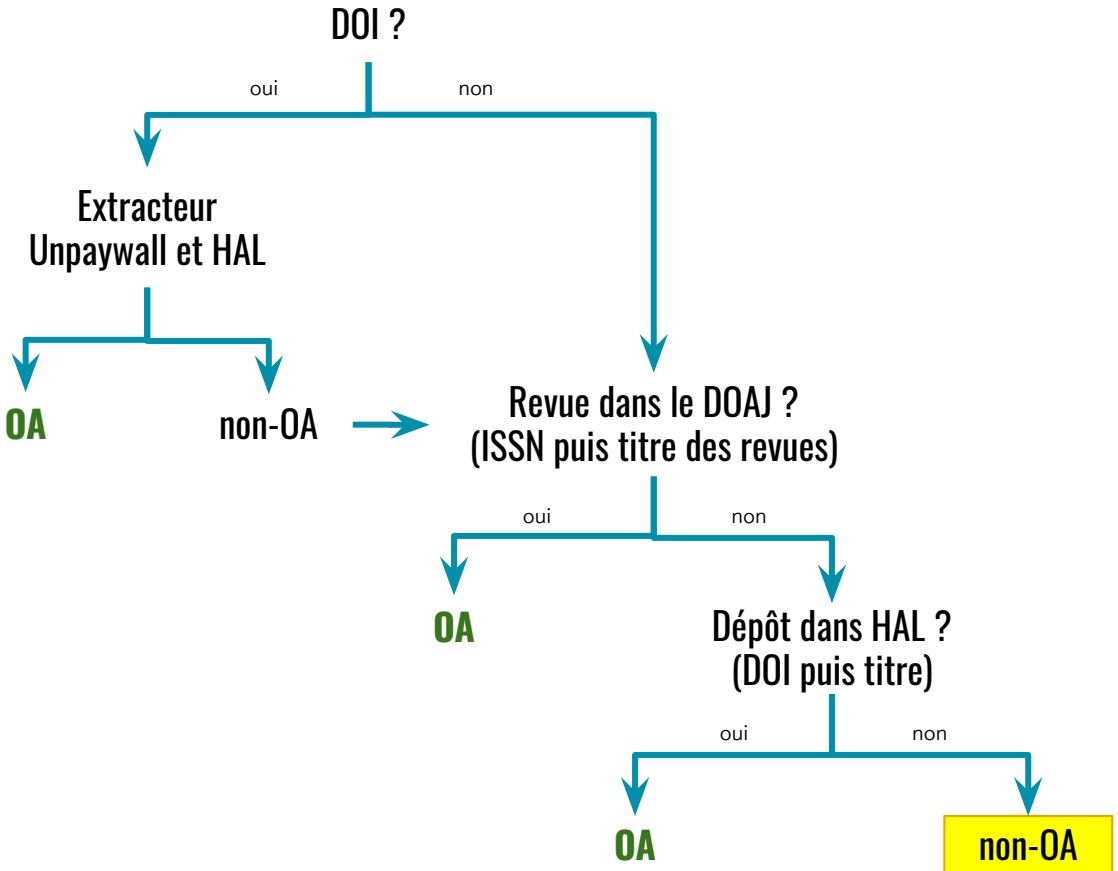
- Vérifier s'il existe un dépôt dans l'export de HAL (DOI puis Titre)
  `RechercheV` dans Excel
  Bonus : Macro de "nettoyage" (retrait de la ponctuation)

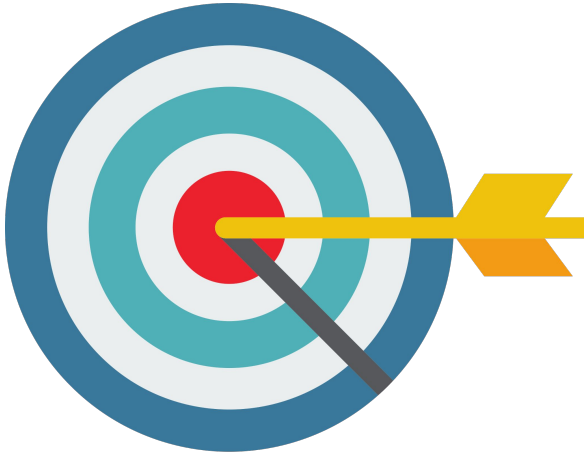
Warnings !

- Disparités entre Unpaywall pur et Unpaywall via HAL
- Disparités entre DOAJ et DOAJ selon Unpaywall
- DOIs qui contiennent des () et problèmes de casse
- Doublons dans HAL
- Dépôts hors collection exportée (+ mauvaise ou pas d'affiliation)

Identifier ce qui est déjà en OA dans un corpus

Ceinture-bretelles…





Étape 2
Identifier les auteurs et campagne de mailing

Retrouver ses auteurs

Les auteurs ne sont pas toujours dans l'annuaire de l'établissement → vérification avec OCdHAL

IDArt	Auteurs
Art1	Moisan, Lionel; Moulon, Pierre; Monasse, Pascal

Corpus de travail

Nom	Prénom	Affiliation
Moulon	Pierre	LIGM
Monasse	Pascal	LIGM
Vandamme	Matthieu	NAVIER
Montel	Nathalie	LATTS

OCdHAL

Retrouver ses auteurs

Les auteurs ne sont pas toujours dans l'annuaire de l'établissement → vérification avec OCdHAL

1. Split auteurs

IDArt	Auteurs
Art1	Moisan, Lionel; Moulon, Pierre; Monasse, Pascal

IDArt	IDAut	Auteurs
Art1	Aut1	Moisan, Lionel
Art1	Aut2	Moulon, Pierre
Art1	Aut3	Monasse, Pascal

Corpus de travail →

Nom	Prénom	Affiliation
Moulon	Pierre	LIGM
Monasse	Pascal	LIGM
Vandamme	Matthieu	NAVIER
Montel	Nathalie	LATTS

OCdHAL

Retrouver ses auteurs

Les auteurs ne sont pas toujours dans l'annuaire de l'établissement → vérification avec OCdHAL

IDArt	Auteurs
Art1	Moisan, Lionel; Moulon, Pierre; Monasse, Pascal

1. Split auteurs

IDArt	IDAut	Auteurs
Art1	Aut1	Moisan, Lionel
Art1	Aut2	Moulon, Pierre
Art1	Aut3	Monasse, Pascal

2. Nettoyage auteurs

IDArt	IDAut	Auteurs	Auteurs clean
Art1	Aut1	Moisan, Lionel	MoisanL
Art1	Aut2	Moulon, Pierre	MoulonP
Art1	Aut3	Monasse, Pascal	MonasseP

Corpus de travail →

Nom	Prénom	Affiliation
Moulon	Pierre	LIGM
Monasse	Pascal	LIGM
Vandamme	Matthieu	NAVIER
Montel	Nathalie	LATTS

1. Nettoyage auteurs

Nom	Prénom	Auteurs clean	Affiliation
Moulon	Pierre	MoulonP	LIGM
Monasse	Pascal	MonasseP	LIGM
Vandamme	Matthieu	VandammeM	NAVIER
Montel	Nathalie	MontelN	LATTS

OCdHAL →

Retrouver ses auteurs

Les auteurs ne sont pas toujours dans l'annuaire de l'établissement → vérification avec OCdHAL

IDArt	Auteurs
Art1	Moisan, Lionel; Moulon, Pierre; Monasse, Pascal

1. Split auteurs

IDArt	IDAut	Auteurs
Art1	Aut1	Moisan, Lionel
Art1	Aut2	Moulon, Pierre
Art1	Aut3	Monasse, Pascal

2. Nettoyage auteurs

IDArt	IDAut	Auteurs	Auteurs clean
Art1	Aut1	Moisan, Lionel	MoisanL
Art1	Aut2	Moulon, Pierre	MoulonP
Art1	Aut3	Monasse, Pascal	MonasseP

3. Correspondance d'affiliations

IDArt	IDAut	Auteurs	Auteurs clean	Affiliation
Art1	Aut1	Moisan, Lionel	MoisanL	
Art1	Aut2	Moulon, Pierre	MoulonP	LIGM
Art1	Aut3	Monasse, Pascal	MonasseP	LIGM

Corpus de travail

Nom	Prénom	Affiliation
Moulon	Pierre	LIGM
Monasse	Pascal	LIGM
Vandamme	Matthieu	NAVIER
Montel	Nathalie	LATTS

OCdHAL

1. Nettoyage auteurs

Nom	Prénom	Auteurs clean	Affiliation
Moulon	Pierre	MoulonP	LIGM
Monasse	Pascal	MonasseP	LIGM
Vandamme	Matthieu	VandammeM	NAVIER
Montel	Nathalie	MontelN	LATTS

Trouver les adresses mails

IDArt	IDAut	Auteurs	Auteurs clean	Affiliation
Art1	Aut1	Moisan, Lionel	MoisanL	
Art1	Aut2	Moulon, Pierre	MoulonP	LIGM
Art1	Aut3	Monasse, Pascal	MonasseP	LIGM

Corpus de travail

Nom	Prénom	Email
Argoul	Pierre	mail1@enpc.fr
Monasse	Pascal	mail2@enpc.fr
Caré	Sabine	mail3@enpc.fr
Château	Camille	mail4@enpc.fr

Annuaire

Trouver les adresses mails

IDArt	IDAut	Auteurs	Auteurs clean	Affiliation
Art1	Aut1	Moisan, Lionel	MoisanL	
Art1	Aut2	Moulon, Pierre	MoulonP	LIGM
Art1	Aut3	Monasse, Pascal	MonasseP	LIGM

Corpus de travail

1. Nettoyage auteurs

Nom	Prénom	Email
Argoul	Pierre	mail1@enpc.fr
Monasse	Pascal	mail2@enpc.fr
Caré	Sabine	mail3@enpc.fr
Château	Camille	mail4@enpc.fr

Nom	Prénom	Auteurs clean	Email
Argoul	Pierre	ArgoulP	mail1@enpc.fr
Monasse	Pascal	MonasseP	mail2@enpc.fr
Caré	Sabine	CareS	mail3@enpc.fr
Château	Camille	ChateauC	mail4@enpc.fr

Annuaire →

Trouver les adresses mails

IDArt	IDAut	Auteurs	Auteurs clean	Affiliation
Art1	Aut1	Moisan, Lionel	MoisanL	
Art1	Aut2	Moulon, Pierre	MoulonP	LIGM
Art1	Aut3	Monasse, Pascal	MonasseP	LIGM

Corpus de travail

1. Correspondance d'emails

IDArt	IDAut	Auteurs	Auteurs clean	Affiliation	Email
Art1	Aut1	Moisan, Lionel	MoisanL		
Art1	Aut2	Moulon, Pierre	MoulonP	LIGM	
Art1	Aut3	Monasse, Pascal	MonasseP	LIGM	mail2@enpc.fr

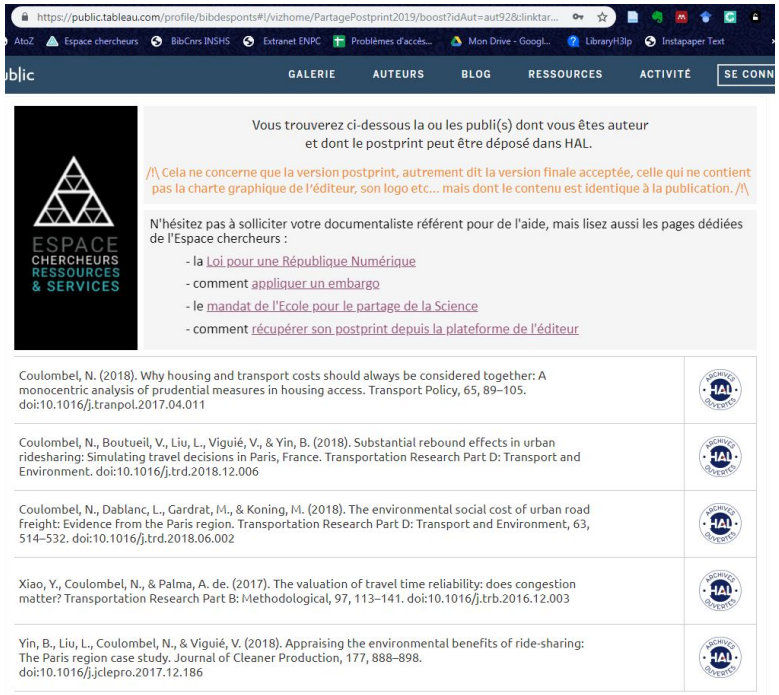
1. Nettoyage auteurs

Nom	Prénom	Email
Argoul	Pierre	mail1@enpc.fr
Monasse	Pascal	mail2@enpc.fr
Caré	Sabine	mail3@enpc.fr
Château	Camille	mail4@enpc.fr

Annuaire

Nom	Prénom	Auteurs clean	Email
Argoul	Pierre	ArgoulP	mail1@enpc.fr
Monasse	Pascal	MonasseP	mail2@enpc.fr
Caré	Sabine	CareS	mail3@enpc.fr
Château	Camille	ChateauC	mail4@enpc.fr

Campagne de mailing



Bonjour,

Suite à mon message du 10 mai dernier, je vous rappelle que vous pouvez partager le texte intégral pour 5 publication(s) dont vous êtes l'auteur ou le co-auteur. J'ai en effet repéré pour vous celle ou celles qui n'étaient pas encore en Open Access, cliquez ici pour procéder au dépôt dans HAL : https://public.tableau.com/profile/bibdesponts#!/vizhome/PartagePostprint2019/boost?idAut=aut92&:linktarget=_blank

/!\ Cela ne concerne que la version postprint, autrement dit la version finale acceptée, celle qui ne contient pas la charte graphique de l'éditeur, son logo etc... mais dont le contenu est identique à la publication. /!

Pour information :

En coordination avec la Direction de la Recherche et conformément au Mandat pour le partage de la science de l'École des Ponts, nous menons une campagne auprès des chercheurs afin de leur demander de partager le texte intégral pour un maximum de leurs publications. Comme l'année dernière, nous espérons que de nombreux dépôts seront faits sur HAL pour augmenter la visibilité de vos travaux et celle de votre laboratoire.

Depuis octobre 2016, la Loi pour une République Numérique vous protège et vous est très favorable, vous pouvez en effet réduire la durée de l'embargo à 6 mois pour les Sciences, Techniques et Médecine et à 12 mois pour les Sciences Humaines et Sociales.

Si vous rencontrez des difficultés, n'hésitez pas à me solliciter.

Récupérer les citations complètes :

- API CrossRef
  https://api.crossref.org/works/10.5555/12345678/transform/text/x-bibliography
- API HAL, valeur de citationFull_s

Une seule URL pour toutes les refs :

- Tableau Software
- ou Google Fusion

Envoi des mails avec YAMM

Après le mailing... le déluge !

- Suivi des dépôts dans HAL
- Démaquillage
- Assistance, voire dépôts

Bilan de cette *petite* opération

Bilan chiffré 1 mois après le 1er mail

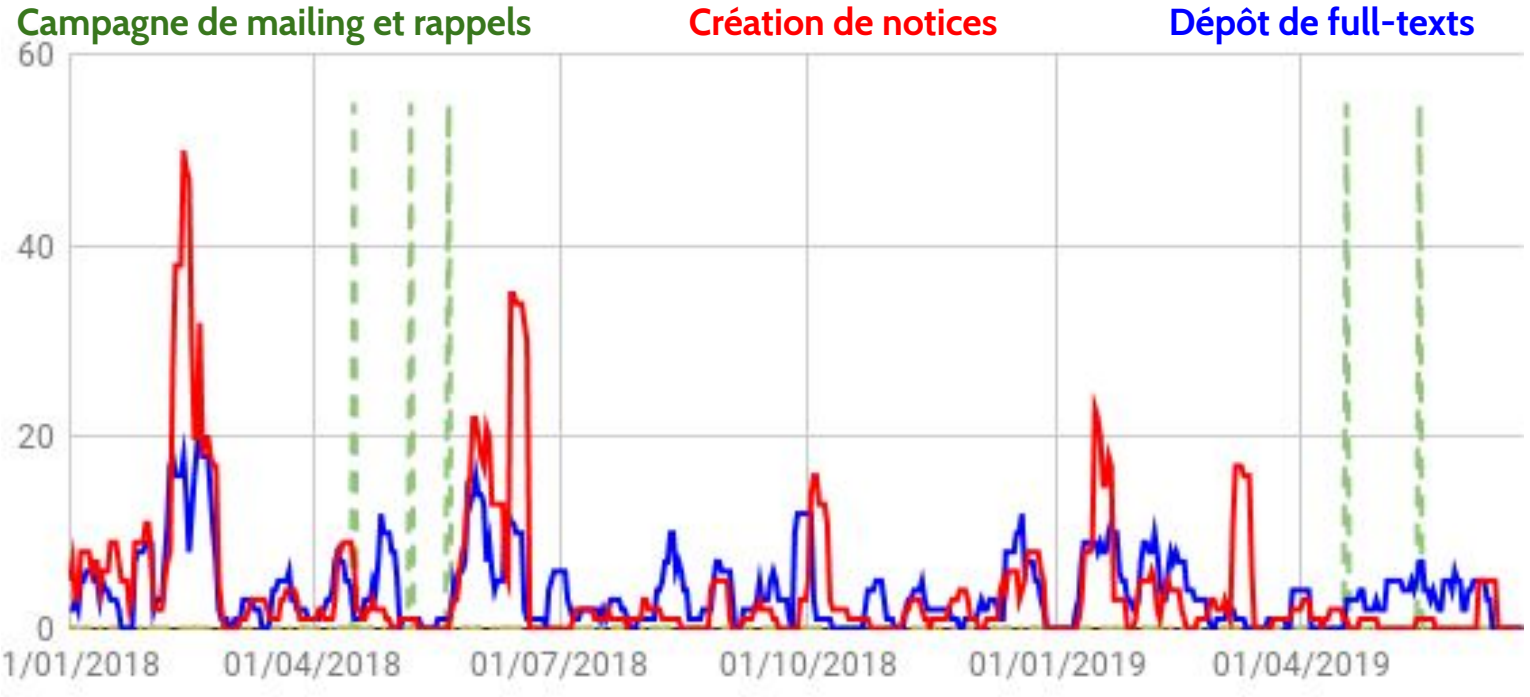
+79 dépôts

Taux d'OA = 65%

à comparer aux 41% pour la production française, avec une méthodologie différente

Mesure de l'effet de la campagne

[Macro Google Sheet de Philippe Gambette]



Bilan non-chiffré

- Montée en compétences sur de nombreux outils
- Contact avec les chercheurs (sensibilisation, pédagogie)

- ...et tout le plaisir d'avoir partagé ça avec vous !

Boîte à outils complète...

- Extracteur Unpaywall - HAL (avec GSheets)
- Extracteur Unpaywall - HAL (en ligne)
- OcdHAL

- YAMM (outil de mailing)
- 3 macros utilitaires
  - Splitter sur plusieurs lignes l'information d'une cellule qui contient plusieurs valeurs
  - Retirer les caractères de ponctuation des titres
  - Concaténer des noms d'auteurs

Merci !

Romain Boistel

romain.boistel@enpc.fr

@RomBoistel

Frédérique Bordignon

frederique.bordignon@enpc.fr

@freddie2310

https://carnetist.hypotheses.org