Comment booster le taux de full-texts dans HAL

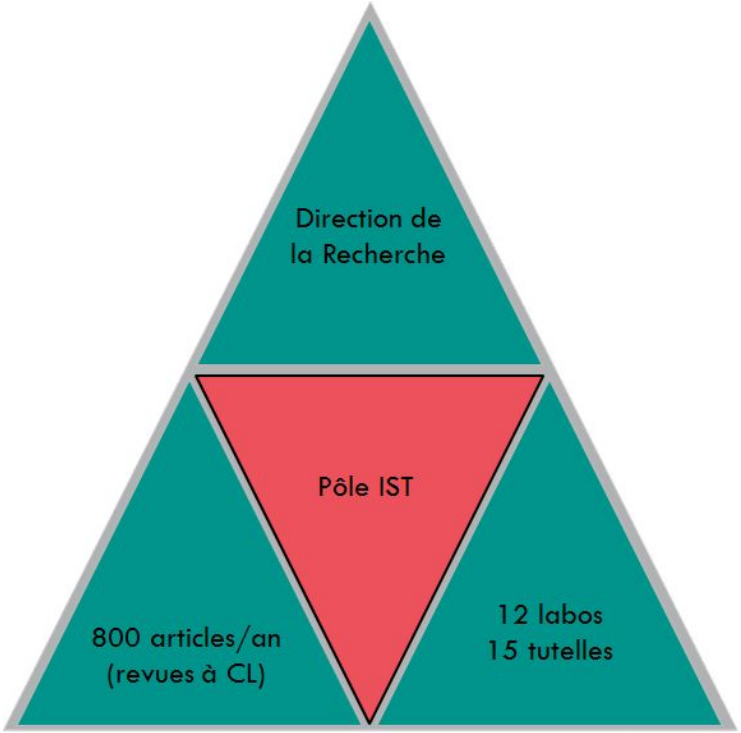
Romain Boistel - Frédérique Bordignon

Journées CasuHAL - 1er juin 2018

École des Ponts
ParisTech

casuHAL
club utilisateur

Contexte général



Contexte politique

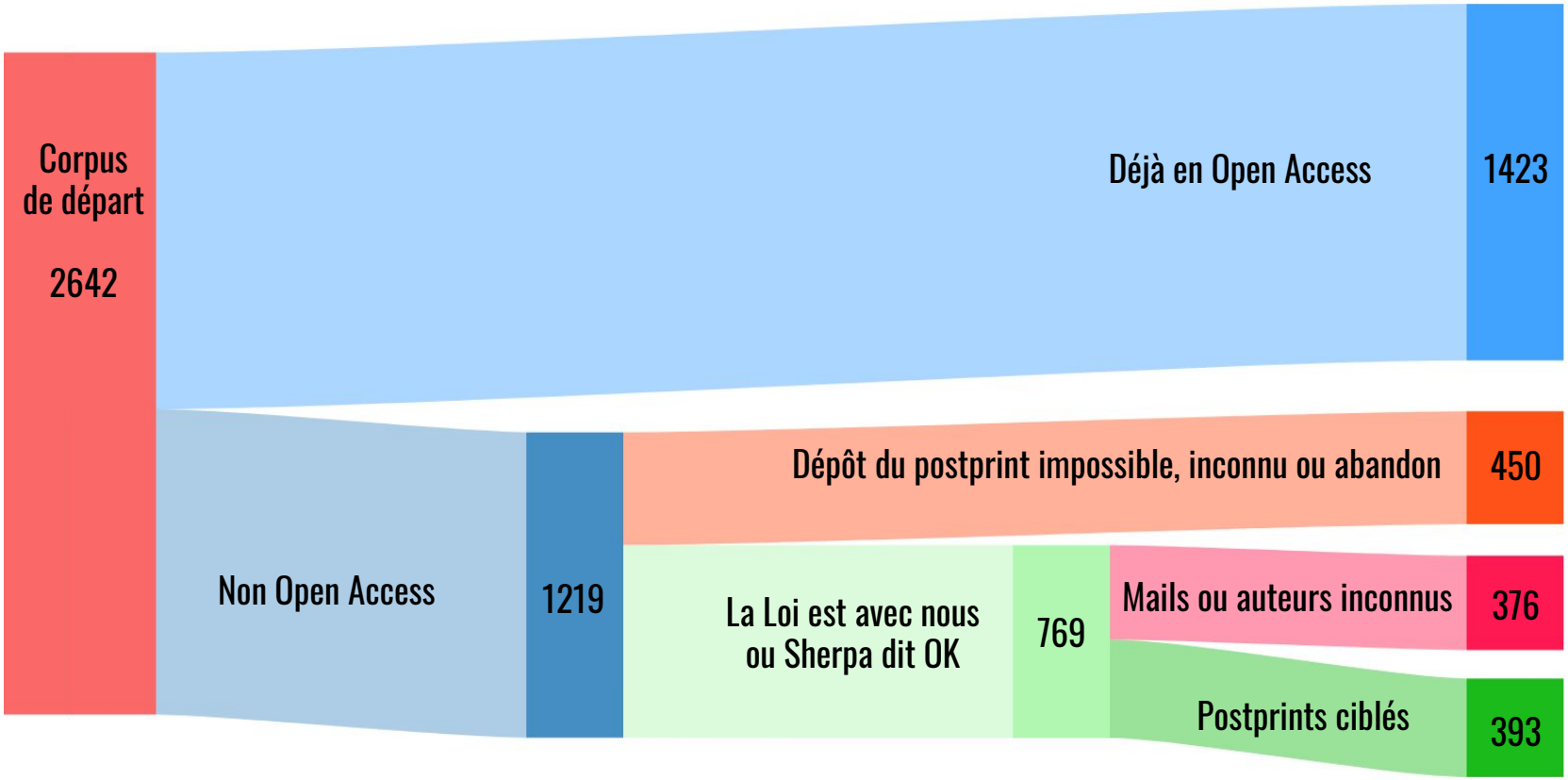
- Mandat de l'École des Ponts pour le partage de la science
- Proposition d'un bonus Open Science au Conseil des labos du 17 mars
- Décision finale : repérage des chercheurs mauvais élèves et mailing
- Soutien de la Direction de la Recherche
- Contexte propice au dépôt (HCERES)

Démarche : vue d'ensemble

Sur la base d'un corpus d'articles :

1. identifier les full-texts qu'on **veut** récupérer
2. identifier les full-texts qu'on **peut** déposer
3. identifier les **auteurs** à qui on veut envoyer un mail

Démarche : plus de détails



Des outils et des APIs (1)

Sources des données

unpaywall

DOAJ

HAL archives-ouvertes.fr

Crossref

SHERPA

OCdHAL

Outils de traitement

Excel

Google Sheets

Outils pour le mailing

YAMM

École des Ponts ParisTech

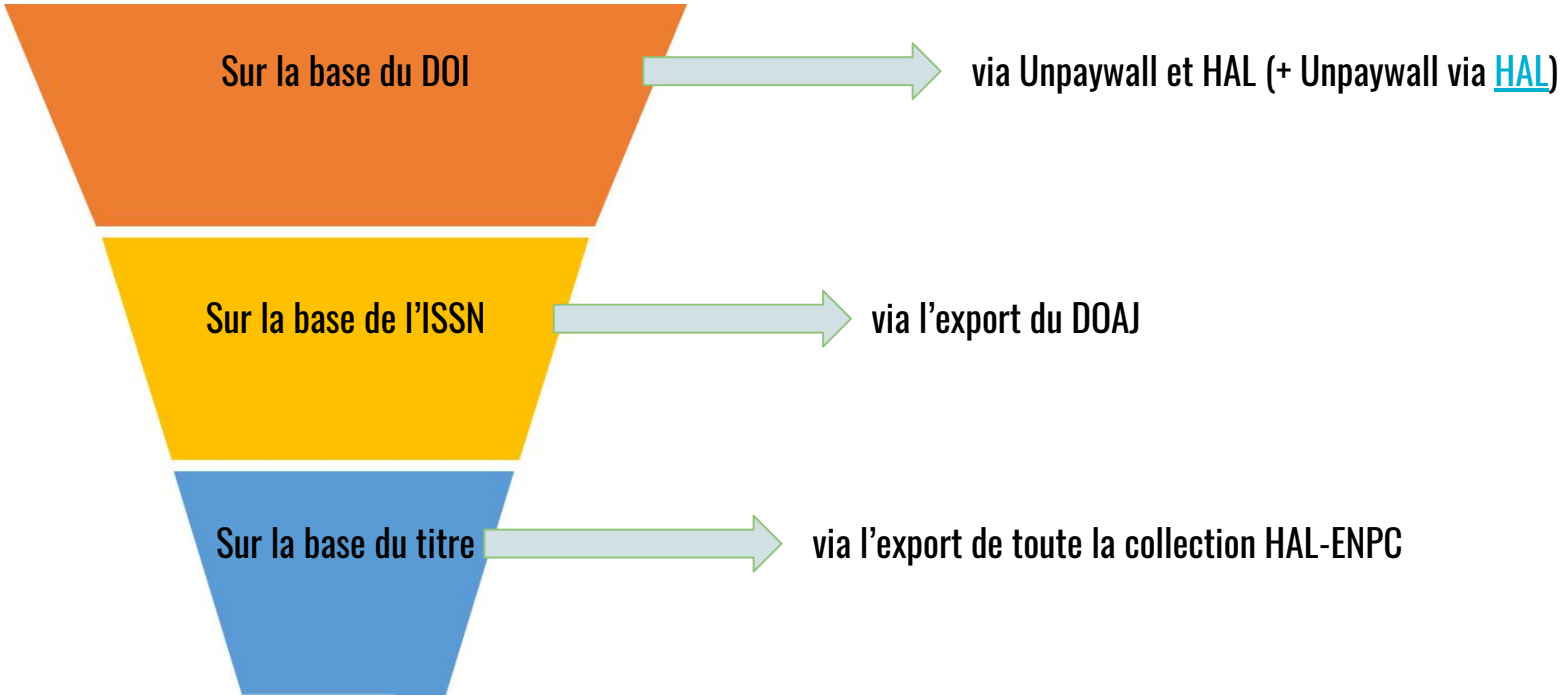
+ableau

Gmail

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

Identifier ce qui est déjà en OA

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 :

  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...

Focus sur la requête dans HAL via l'API

https://api.archives-ouvertes.fr/search/?wt=csv&fq=docType_s:("ART"OR"UNDEFINED")&fq=collCode_s:"ENPC"&rows=10000&fq=producedDateY_i:[2012 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:[**2012** TO **2018**]	Critères de requêtes **articles** ou **preprints**, collection **ENPC**, **2012** 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**

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 (exemple (mettez votre email à la fin))
  - 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*

Démo de l'extracteur Unpaywall - HAL

http://bit.ly/ExtracteurUnpaywallHAL

Faites une copie !

Warnings !

- Quota de Google !
- 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

Étape 2
Identifier les full-texts qu'on peut déposer

Pour quels articles peut-on déposer le postprint ?

Pour les articles de 2017 → La Loi pour une République Numérique prévaut.

Pour les articles de 2013 à 2016 → extracteur de données Sherpa

- Dépôt du postprint autorisé ou pas
- Durée d'embargo le cas échéant

Les mails sont envoyés avec une phrase contextuelle sur l'origine de l'autorisation :

- 2017 : *La loi est avec vous.*
- 2013-2016 + embargo passé : *On a vérifié pour vous, le dépôt est possible.*
- 2013-2016 + embargo en cours : *Vous pourrez déposer avec un embargo de...*

Focus sur l'extracteur de conditions Sherpa RoMEO

Sherpa renvoie les infos au format XML (version web) → à exploiter dans Excel

- Une métadonnée pour le dépôt du postprint :

  <postarchiving>**can**</postarchiving>

- Un ensemble de conditions parmi lesquelles une concerne la durée d'embargo du postprint :

  <condition>**Publisher copyright and source must be acknowledged**</condition>

  <condition>**Author's post-print after 12 months after publication**</condition>

On écarte les postprints interdits.

On concatène toutes les conditions.

On y recherche toutes les chaînes de caractères que Sherpa utilise pour les conditions d'embargo de postprint (*Große Funktion*).

→ Quand il y a un match = on a la durée d'embargo !

Démo de l'extracteur Sherpa RoMEO

https://doi.org/10.5281/zenodo.1255919

Étape 3
Identifier les auteurs à qui on veut envoyer un mail

Trouver les adresses mails

- Export d'OCdHAL (onglet Auteurs) pour repérer les auteurs locaux
- Export de l'annuaire de l'École des Ponts
- Split de notre corpus de full-texts ciblés pour n'avoir qu'un auteur par ligne

- Nettoyage, simplification, concaténation pour lier un maximum de mails à nos auteurs locaux (macro + à l'ancienne)

- Attribution d'un ID par auteur (pour Tableau)

Étape 4
Mailing

Récupérer la citation propre : 2 voies

Pour la coller dans le mail

- API CrossRef
  https://api.crossref.org/works/10.5555/12345678/transform/text/x-bibliography

- API HAL, valeur de citationFull_s (déjà exportée au début...), à nettoyer

Mail

Envoyé avec YAMM

Destinataires

Partager vos publications ! Dernier rappel...

Bonjour,

Vous pouvez partager le texte intégral pour {{nbre_publis}} publication(s) dont vous êtes l'auteur ou le co-auteur. Cliquez ici pour voir lesquelles et procéder au dépôt dans HAL :
{{lien_tableau}}

/!\ 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.

Si vous n'êtes pas co-propriétaire de ce dépôt, demandez la propriété à la personne qui a fait le dépôt : après vous être authentifié sur HAL, il suffit de cliquer sur le bouton « Demander la propriété », en bas de la page présentant le dépôt.

Pour information :
Il a été décidé en **Conseil des labos le 17 mars** de mener une campagne auprès des chercheurs afin de leur demander de partager le texte intégral pour un maximum de leurs publications.

Au-delà d'augmenter la visibilité de vos travaux et de celle de votre laboratoire, la publication en Open Access ou le dépôt en archive ouverte est une obligation définie dans le Mandat pour le partage de la science de l'Ecole des Ponts.

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 SHS.

Pour les publications plus anciennes, j'ai vérifié pour vous s'il était possible de déposer le postprint (la version finale acceptée).

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

Cordialement,

Tableau avec Tableau

- Une URL par auteur
- La référence complète
- Action possible sur le postprint
- Lien vers HAL (direct vers la notice si elle existe)

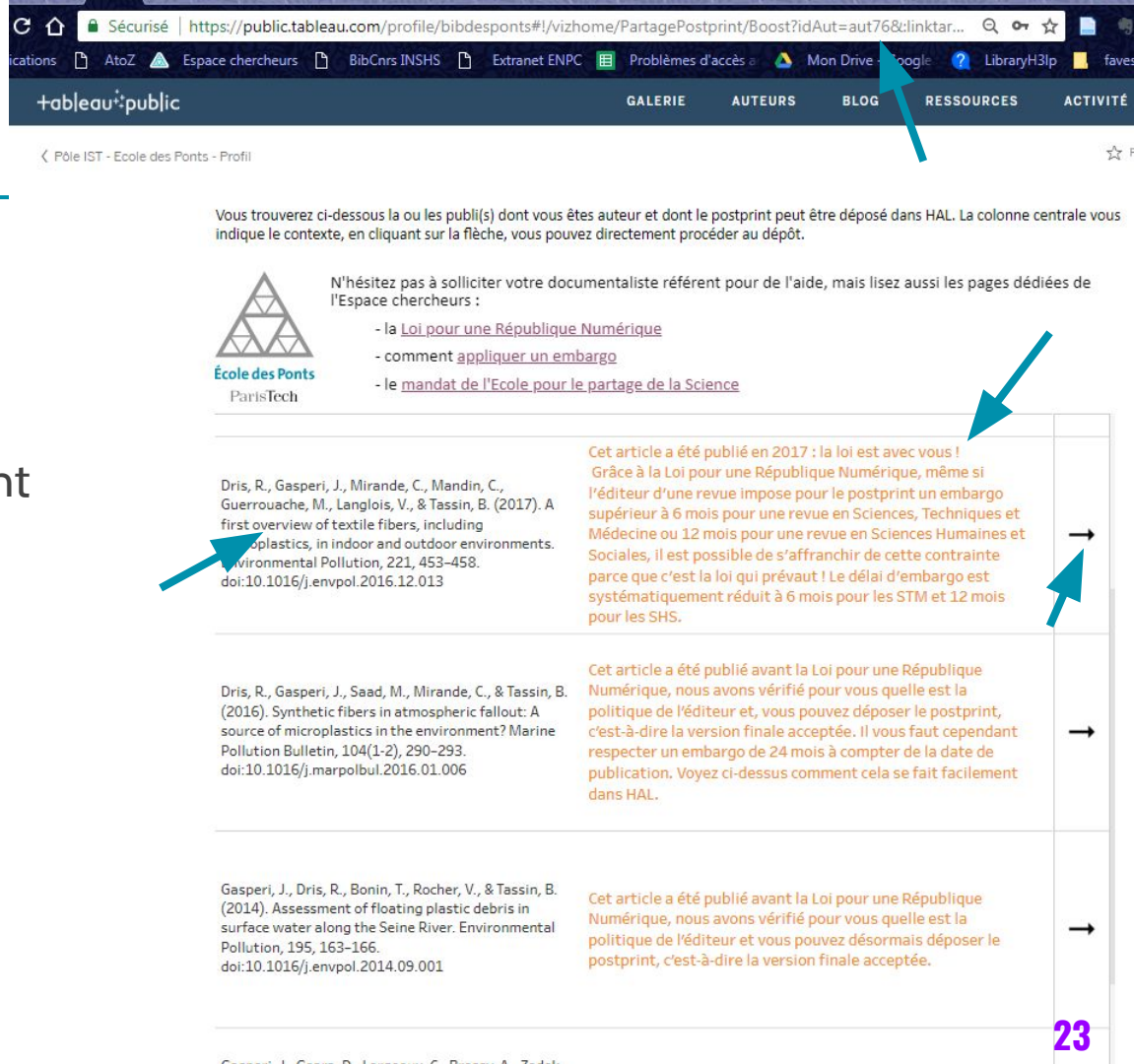
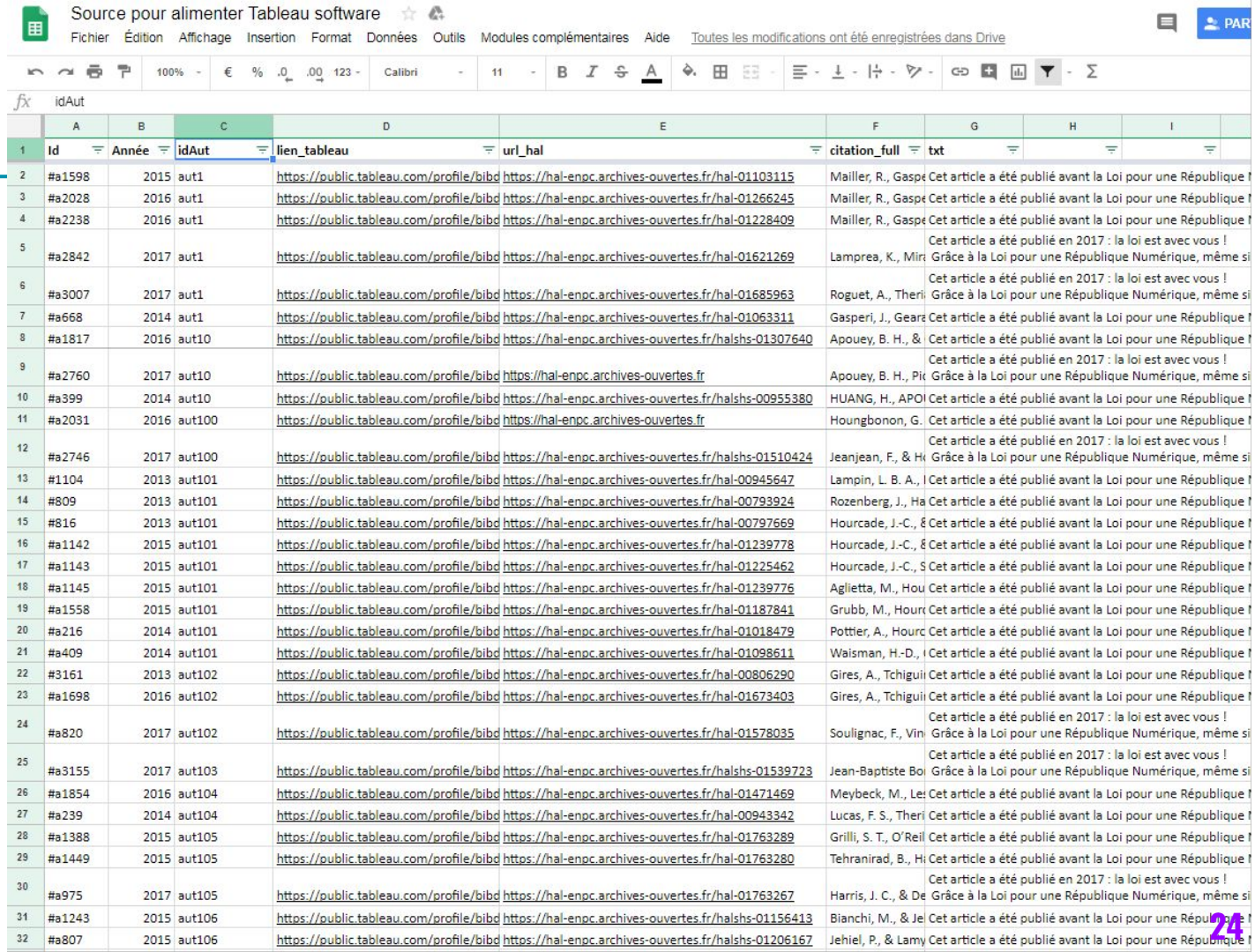


Tableau source pour Tableau



Source pour alimenter Tableau software

Fichier Édition Affichage Insertion Format Données Outils Modules complémentaires Aide Toutes les modifications ont été enregistrées dans Drive

idAut

	A	B	C	D	E	F	G	H	I
1	Id	Année	idAut	lien_tableau	url_hal	citation_full	txt		
2	#a1598	2015	aut1	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-01103115	Mailler, R., Gaspe	Cet article a été publié avant la Loi pour une République N		
3	#a2028	2016	aut1	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-01266245	Mailler, R., Gaspe	Cet article a été publié avant la Loi pour une République N		
4	#a2238	2016	aut1	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-01228409	Mailler, R., Gaspe	Cet article a été publié avant la Loi pour une République N		
5	#a2842	2017	aut1	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-01621269	Lamprea, K., Mira	Cet article a été publié en 2017 : la loi est avec vous ! Grâce à la Loi pour une République Numérique, même si		
6	#a3007	2017	aut1	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-01685963	Roguet, A., Theri	Cet article a été publié en 2017 : la loi est avec vous ! Grâce à la Loi pour une République Numérique, même si		
7	#a668	2014	aut1	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-01063311	Gasperi, J., Geara	Cet article a été publié avant la Loi pour une République N		
8	#a1817	2016	aut10	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/halshs-01307640	Apouey, B. H., &	Cet article a été publié avant la Loi pour une République N		
9	#a2760	2017	aut10	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr	Apouey, B. H., Pie	Cet article a été publié en 2017 : la loi est avec vous ! Grâce à la Loi pour une République Numérique, même si		
10	#a399	2014	aut10	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/halshs-00955380	HUANG, H., APOI	Cet article a été publié avant la Loi pour une République N		
11	#a2031	2016	aut100	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr	Houngbonon, G.	Cet article a été publié avant la Loi pour une République N		
12	#a2746	2017	aut100	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/halshs-01510424	Jeanjean, F., & He	Cet article a été publié en 2017 : la loi est avec vous ! Grâce à la Loi pour une République Numérique, même si		
13	#1104	2013	aut101	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-00945647	Lampin, L. B. A., I	Cet article a été publié avant la Loi pour une République N		
14	#809	2013	aut101	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-00793924	Rozenberg, J., Ha	Cet article a été publié avant la Loi pour une République N		
15	#816	2013	aut101	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-00797669	Hourcade, J.-C., &	Cet article a été publié avant la Loi pour une République N		
16	#a1142	2015	aut101	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-01239778	Hourcade, J.-C., &	Cet article a été publié avant la Loi pour une République N		
17	#a1143	2015	aut101	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-01225462	Hourcade, J.-C., S	Cet article a été publié avant la Loi pour une République N		
18	#a1145	2015	aut101	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-01239776	Aglietta, M., Hou	Cet article a été publié avant la Loi pour une République N		
19	#a1558	2015	aut101	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-01187841	Grubb, M., Hourc	Cet article a été publié avant la Loi pour une République N		
20	#a216	2014	aut101	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-01018479	Pottier, A., Hourc	Cet article a été publié avant la Loi pour une République N		
21	#a409	2014	aut101	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-01098611	Waisman, H.-D., I	Cet article a été publié avant la Loi pour une République N		
22	#3161	2013	aut102	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-00806290	Gires, A., Tchigui	Cet article a été publié avant la Loi pour une République N		
23	#a1698	2016	aut102	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-01673403	Gires, A., Tchigui	Cet article a été publié avant la Loi pour une République N		
24	#a820	2017	aut102	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-01578035	Soulignac, F., Vin	Cet article a été publié en 2017 : la loi est avec vous ! Grâce à la Loi pour une République Numérique, même si		
25	#a3155	2017	aut103	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/halshs-01539723	Jean-Baptiste Boi	Cet article a été publié en 2017 : la loi est avec vous ! Grâce à la Loi pour une République Numérique, même si		
26	#a1854	2016	aut104	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-01471469	Meybeck, M., Le	Cet article a été publié avant la Loi pour une République N		
27	#a239	2014	aut104	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-00943342	Lucas, F. S., Theri	Cet article a été publié avant la Loi pour une République N		
28	#a1388	2015	aut105	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-01763289	Grilli, S. T., O'Reil	Cet article a été publié avant la Loi pour une République N		
29	#a1449	2015	aut105	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-01763280	Tehranirad, B., Hi	Cet article a été publié avant la Loi pour une République N		
30	#a975	2017	aut105	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/hal-01763267	Harris, J. C., & De	Cet article a été publié en 2017 : la loi est avec vous ! Grâce à la Loi pour une République Numérique, même si		
31	#a1243	2015	aut106	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/halshs-01156413	Bianchi, M., & Je	Cet article a été publié avant la Loi pour une Répu		
32	#a807	2015	aut106	https://public.tableau.com/profile/bibd	https://hal-enpc.archives-ouvertes.fr/halshs-01206167	Jehiel, P., & Lamy	Cet article a été publié avant la Loi pour une Républiqu		

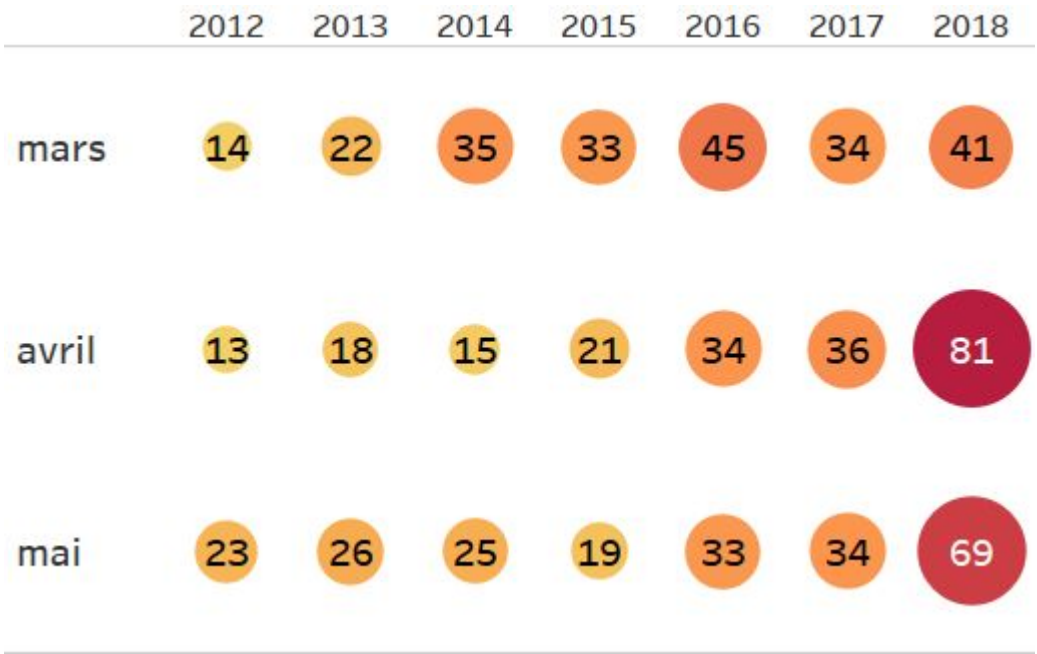
Après le mailing... le déluge !

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

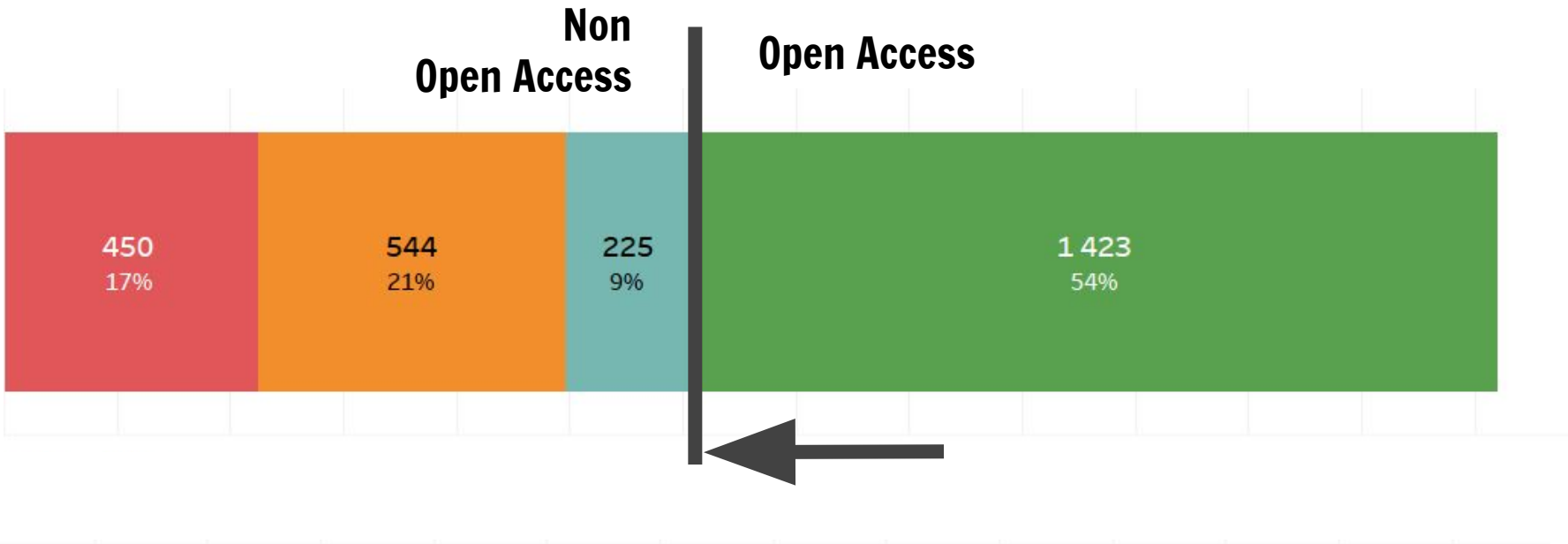
Bilan de cette *petite* opération

Teasing...

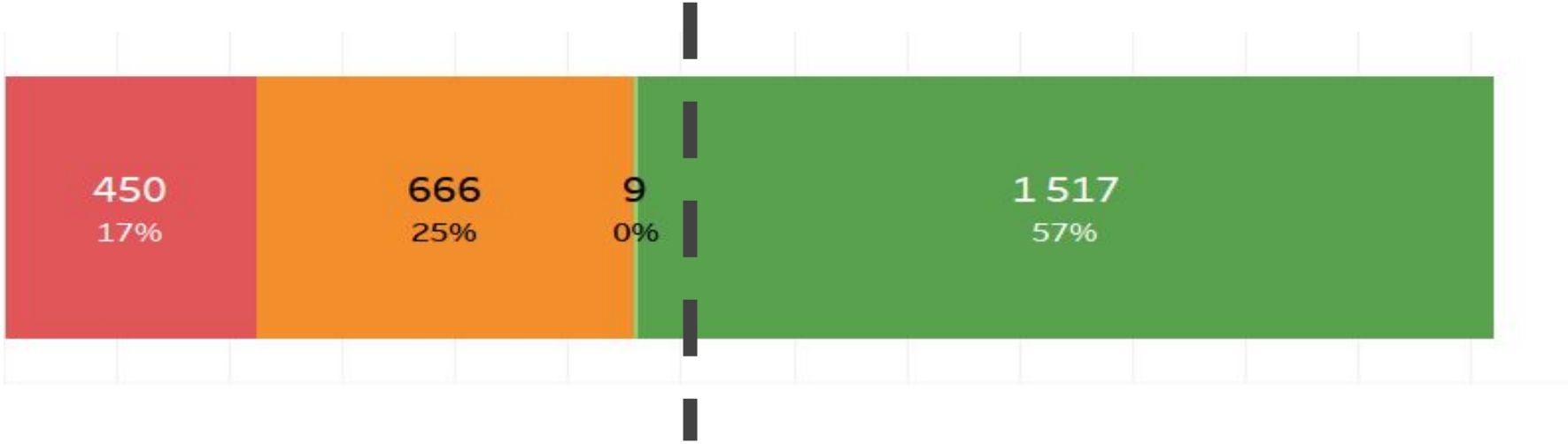
un frémissement dans le nombre de dépôts mensuels



Bilan chiffré : avant



Bilan chiffré : après



+94 dépôts (+9 à venir)
+3 points du taux d'Open Access

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...

- Extracteur Unpaywall - HAL
- Extracteur Sherpa RoMEO
- 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

Code parrainage Quitoque : ROMX35J

Frédérique Bordignon

frederique.bordignon@enpc.fr

@freddie2310

https://carnetist.hypotheses.org

Code parrainage Quitoque : FREMLTD