CNRS - Toulouse INP - UT - UT Capitole - UT2

Institut de Recherche en Informatique de Toulouse

Enjeux et outils de dépollution de la littérature scientifique

Formation URFIST - Toulouse - 2 Avril 2026

Tristan Salord - Phd. Post-doctorat IRIT-Université de Toulouse

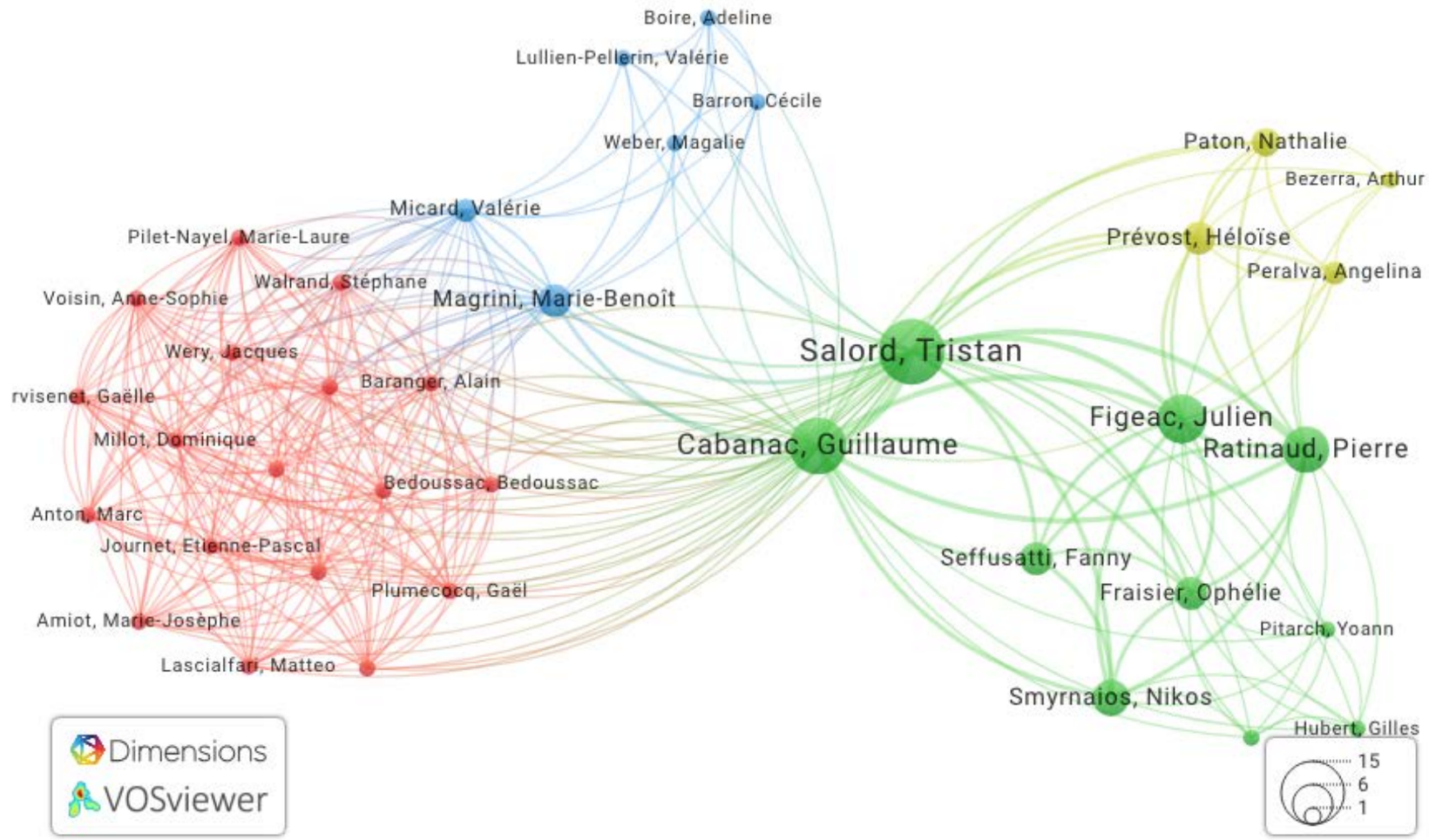
Tristan Salord

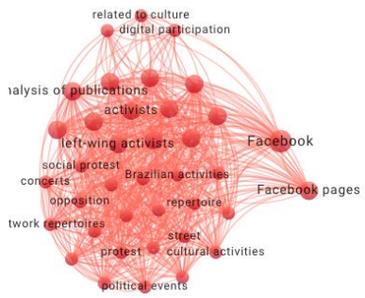
37
Documents

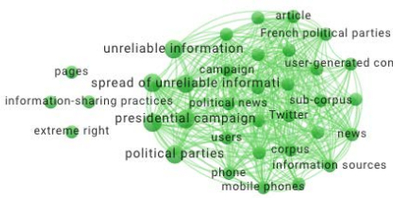
Présentation

D'une formation initiale **pluridisciplinaire** en sciences humaines et sociales (sociologies pragmatiques & critiques, géographie humaine, ethnologie contemporaine), mon parcours s'est progressivement ancré dans le champ des **humanités digitales** par le développement de compétences ciblées dans le **traitement des données textuelles** (analyses lexicométriques, Traitement Automatisé du Langage Naturel) et l'**analyse de la production scientifique** grâce aux nombreux programmes de recherche auxquels j'ai participé.

Dans un monde scientifique de plus en plus confronté à un afflux massif et constant de données, la révolution scientifique du "Big Data" conduit à être de plus en plus vigilant quant aux fondamentaux de communicabilité et transparence, reproductibilité et falsifiabilité qui structurent le champ scientifique. J'attache à cet effet une importance toute particulière dans mes travaux à ouvrir les boîtes noires de la production scientifique.





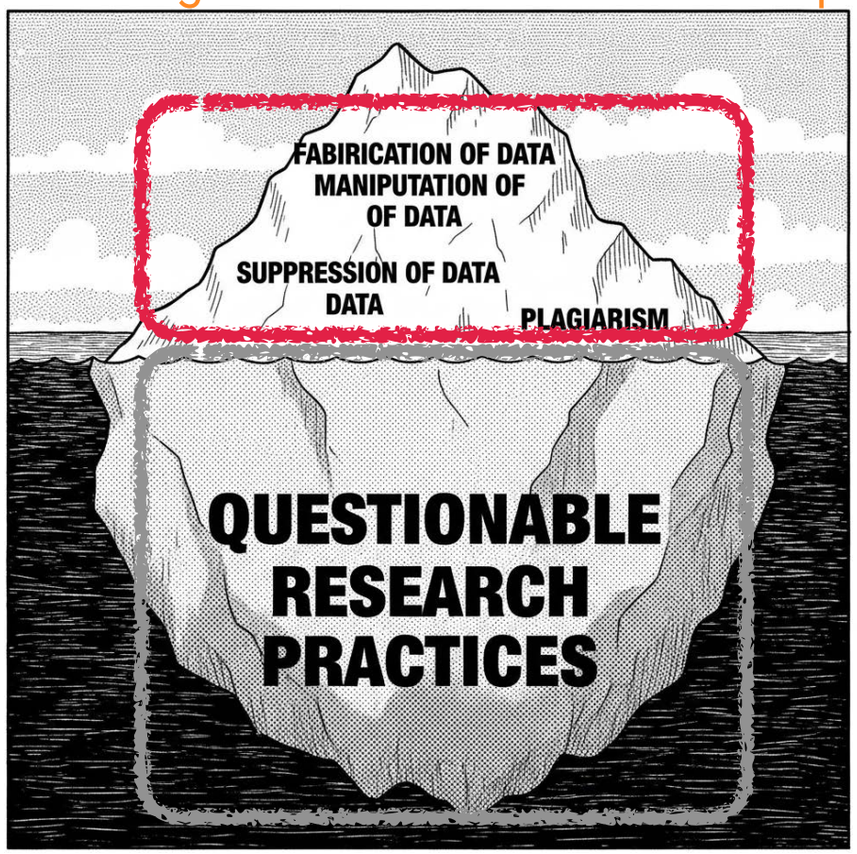


Plan de la présentation

1. Tendances épidémiologiques des méconduites scientifiques
2. Facteurs explicatifs des déviances scientifiques
3. Outiller l'intégrité scientifique

Questions de vocabulaires… et de posture

L'iceberg des méconduites scientifiques



Méconnaissance méthodologique	Pratiques discutables en recherche	Fraude
◗ Méthodes « faibles » ◗ Méthodes inappropriées ◗ Échantillon trop faible ◗ Erreurs statistiques ◗ Pas de recherche documentaire avant le travail ◗ Non-respect des réglementations	◗ « Torture » ou « massage » des données ◗ Changement du critère de jugement ◗ Choix sélectif/omission de données ◗ Références erronées ◗ Changements de tests statistiques *P-hacking* ou *P-HARKing** ◗ Manipulations d'images ◗ Paternité des articles : conflit d'auteurs ! ◗ Études animales trompeuses ◗ Non-publication de recherches financées ◗ Résumés, communiqués de presse embellis…	Fabrication Falsification Plagiat

Non intentionnel — Intentionnel

*Hypothezing After Results are Known

FIGURE 39 – *Le continuum proposé initialement par S.Lock (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2544685/?page=1) et adapté depuis par plusieurs auteurs (La revue du praticien, vol.66, Décembre 2016)*

The intense competition for making original scientific contributions—the "race for priority"—and the peer recognition that comes with it created pressures for deviant behavior ("undue" secrecy, "cutting corners," even, as we have seen, manufacturing fake evidence). But that same intense competition as a system-property also focuses the attention of scientists on particular problems, intensifies their critical review of others' work in the field, and encourages efforts to check important new truth claims through replication. This should increase the chances that any deviant behavior which does occur will be identified.

Harriet Zuckerman, *Deviant Behavior and Social Control in Science*, 1977, p.131

Tendances épidémiologiques du spectre des déviances scientifiques:FFP

➡**Une multiplication des « affaires »?**:

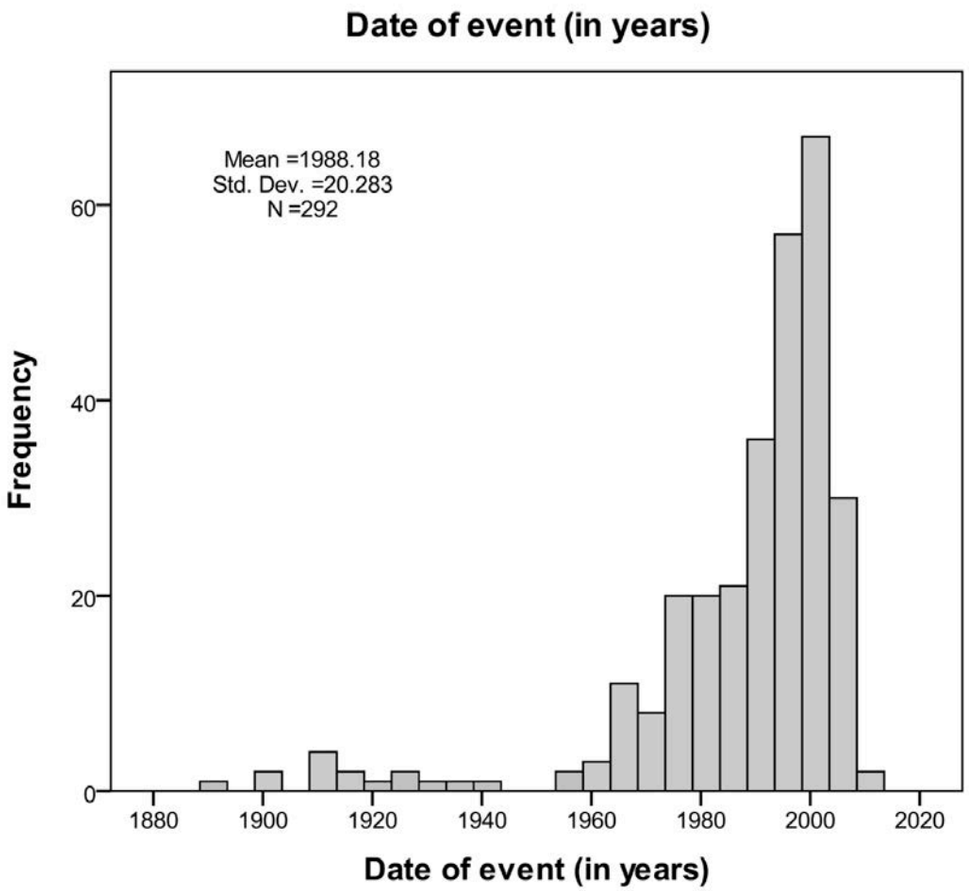
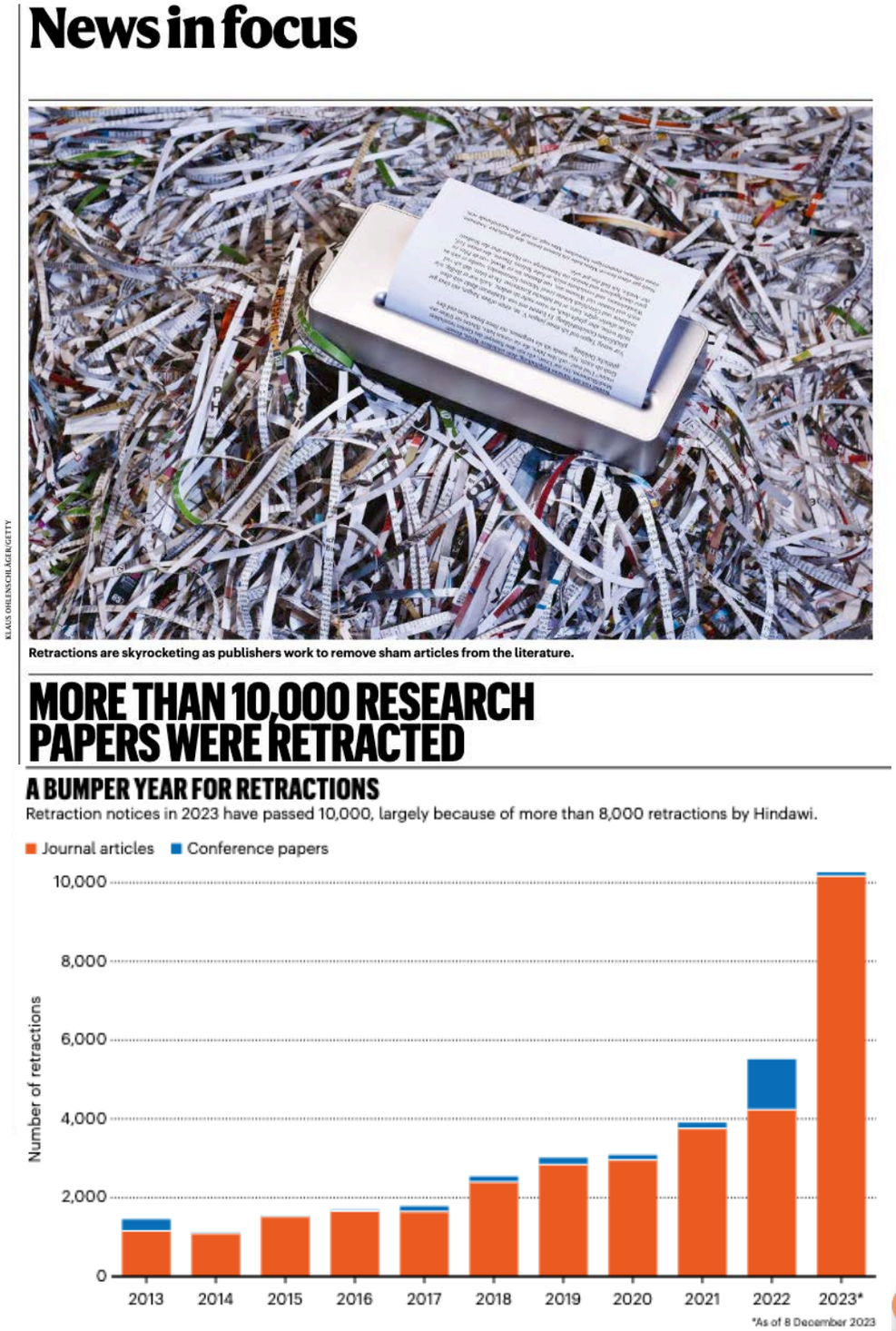


Fig. 2.1. Distribution of date of fraud in research event

Ben Yehuda et al., *Fraud and Misconduct in Research*, 2017, p.28

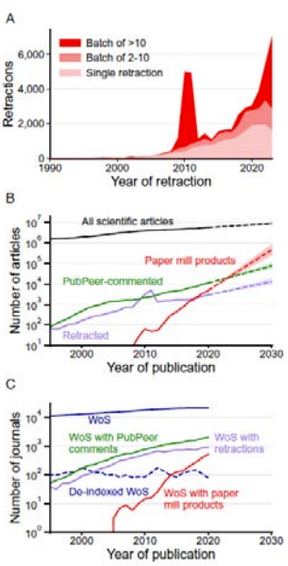
➡**Une forte croissance des rétractations**:



Tendances épidémiologiques du spectre des déviances scientifiques:FFP

➡**<u>Le symptôme des Paper Mills</u>**:

« organisations lucratives à géométrie variable, oeuvrant selon le modèle des « shadow markets », et offrant, contre finances, une diversité de services (« ghostwritting », place de co-auteur, augmentation du taux de citation, peer-review…) permettant à des chercheurs en besoin de publication et/ou de reconnaissance de contourner les règles du marché de la publication et de la compétition scientifique. »



- Les publications de « Paper Mills », une croissances **10x >** à la croissance de la littérature scientifique mondiale

*« (…) we note that the **number of retracted articles and PubPeercommented** articles has been **doubling every 3.3 y and every 3.6 y**, respectively, while the total **number of publications** has been d**oubling every 15.0 y** (80, 81). However, suspected **<u>paper mill products have been doubling every 1.5 y</u>**(…). Notably, suspected paper mill products now outnumber annually retracted articles and are projected to soon outnumber the number of PubPeer-commented articles. »*

Richardson et al., *The entities enabling scientific fraud at scale are large, resilient,* and growing rapidly, 2025, p.7

- **9-10%** des papiers issus de la recherche sur le cancer seraient issus de « Paper Mills »

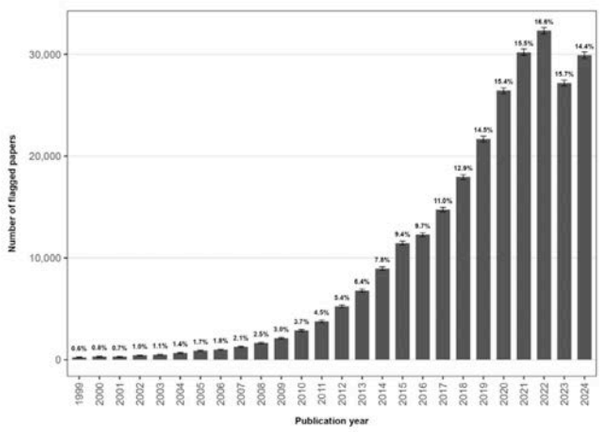


Figure 3: Number of papers per year in the cancer research corpus, flagged because their titles and abstracts were similar to those of retracted paper mill papers. The percentage of flagged papers among all cancer research papers published each year is shown above the bars. Error bars are 95% confidence intervals estimated using bootstrap resampling.

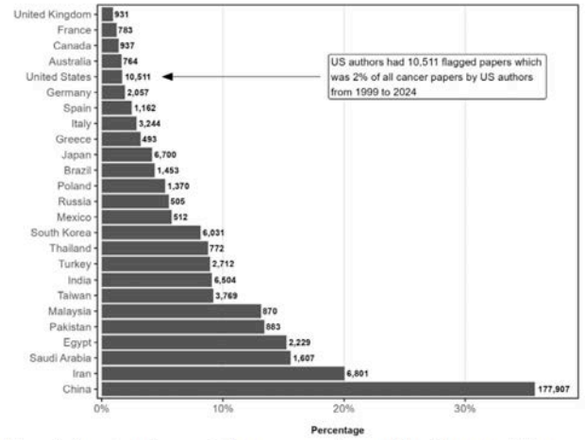
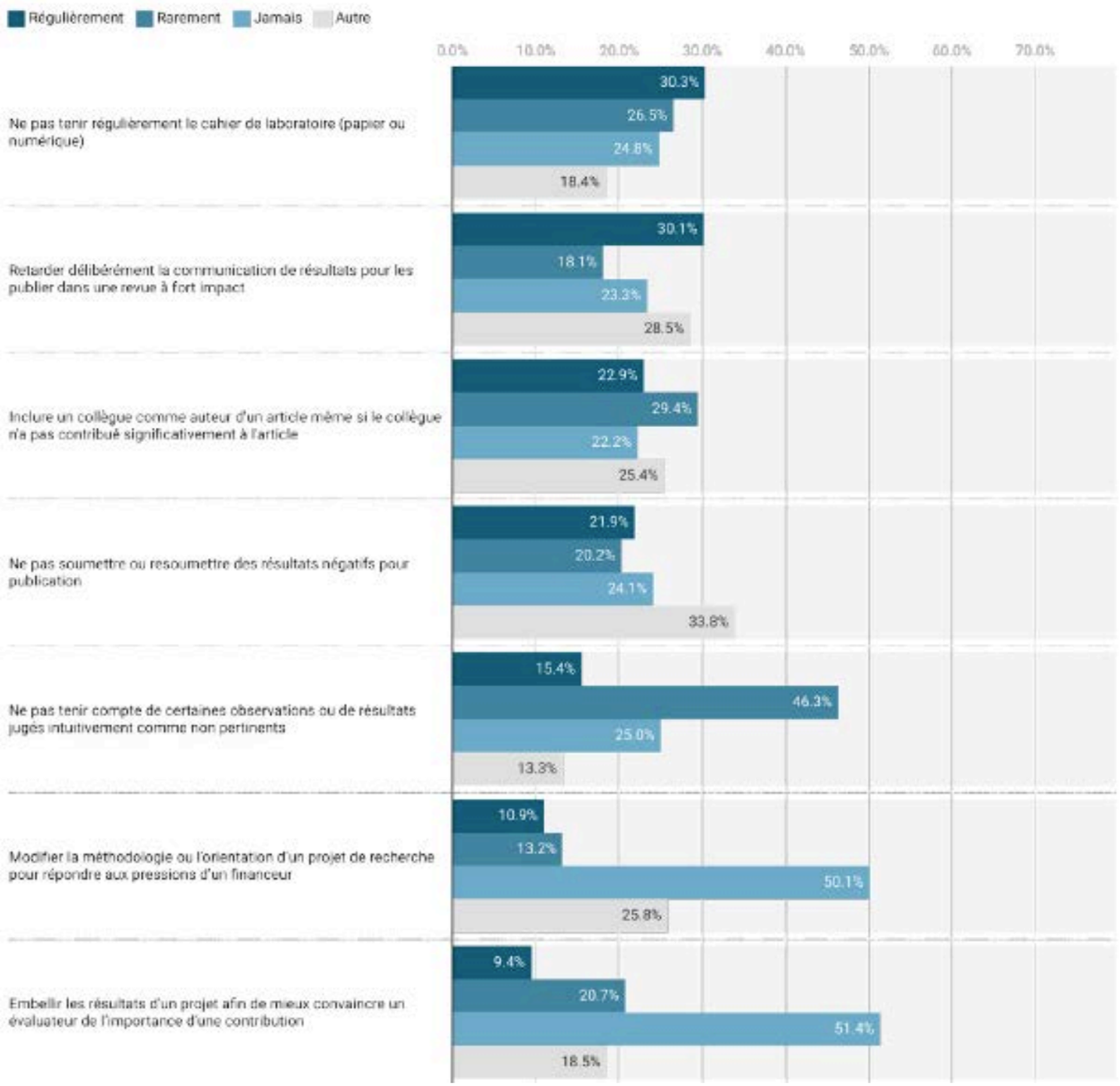


Figure 4: Percentage of papers in the cancer research corpus, flagged by our model because their titles and abstracts were similar to those of retracted paper mill papers, across the 25 countries most frequently flagged based on the first author's affiliation country. The numbers of flagged cancer research papers per country are given next to each bar.

Scancar et al., *Revealing the Paper Mill Iceberg: AI-Based Screening of Cancer Research Publications*, 2025

Tendances épidémiologiques du spectre des déviances scientifiques:QRPs

France, INSERM, 2024



Michel Dubois and Catherine Guaspare, 2024 - *Scientific Integrity and Research Ethics 2024: A Questionnaire Survey of Inserm Staff*, p. 72.

Meta-analysis, USA, 2009 :

- Fabrication/Falsification des données : 0,3% <4,9% = projeté **1,97%.**
- Témoignages de Fab./Fals. des données : 5,2% à 33,3% = projeté **14,2%.**
- QRPs : **33,7**%.
- Témoignages de QRPs: 6,2% < 72% = estimé **28,53%**.

Fanelli, 2009, *How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data*, Plos One, pp.3 à 6

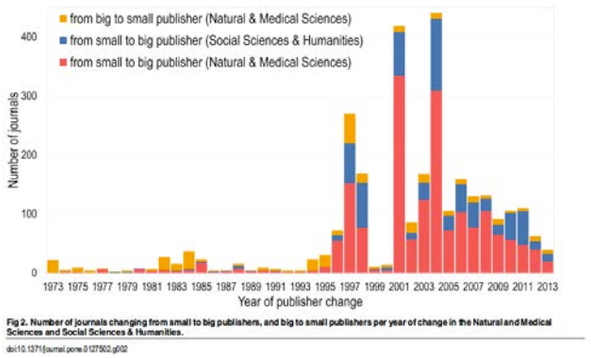
Pays-Bas 2022 :

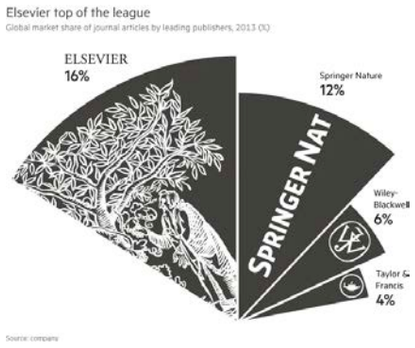
- **>50 %** des répondants témoignent avoir eut recours à plusieurs reprise à des QRPs
- **4,3 %** des répondants admettent avoir fabriqués des données,
- **4,2 %** des répondants admettent avoir falsifié des données de recherche ou des résultats
- **8,3 %** prévalence de n'importe quel type de Fal. ou Fab.

Gopalakrishna et al. - 2022 - *Prevalence of questionable research practices, research misconduct and their potential explanatory factors*, Plos One, p.12

Facteurs explicatifs: dérégulation des marchés éditoriaux

- Dépendance à un marché de la publication **fortement concentré** (oligopolistique) et digitalisé





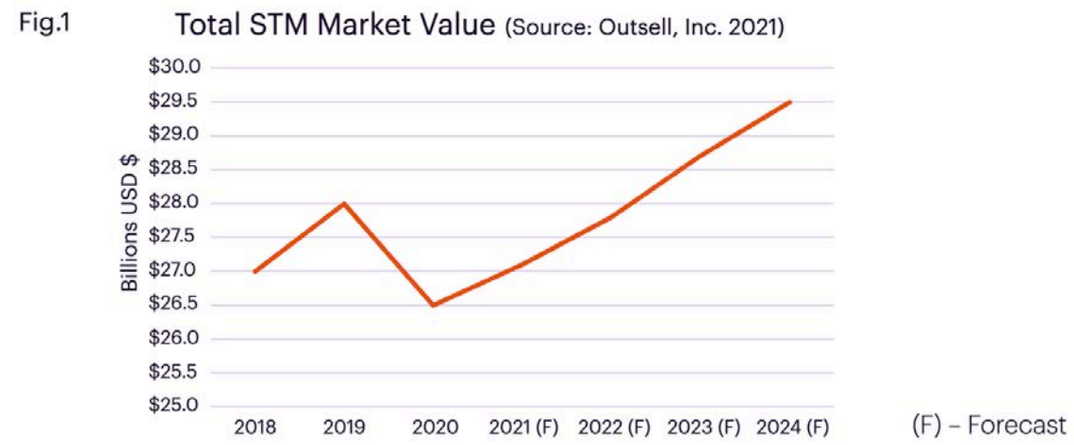
Larivière et al., The oligopoly of academic publishers in the Digital Era, p. 5

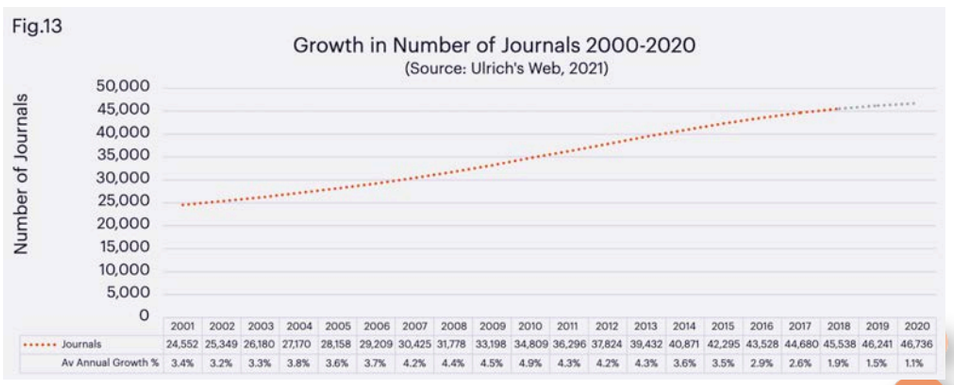
« *Unfortunately, researchers are still dependent on one essentially* ***symbolic function of publishers****, which is to* ***allocate academic capital****, thereby explaining why the scientific community is so dependent on 'The Most Profitable Obsolete Technology in History'.* »

Larivière et al., The oligopoly of academic publishers in the Digital Era, p. 13

- Capitalisme oligopolistique ou capitalisme de plateforme?

Capitalisme de plateforme= économie digitale reposant sur la valorisation de données recueillies massivement par l'intermédiaire de plateformes monopolistiques contraignant/ dirigeant en retour les activités humaines dans un but de profitabilité: « ***The imposition of the platform capitalism model upon commercially funded scientific research has become the next phase of the evolution of modern science***. » (Mirowski, *The evolution of Platform Science*, 2023,p.730)





STM Global Brief 2021 – Economics & Market Size - Fig 1, p.6, Fig 13, p. 15

Facteurs explicatifs: dérégulation des marchés éditoriaux

- Digitalisation de la publication scientifique = insistance sur le volume = désincitation à la vigilance éditoriale = transfert de responsabilité et d'intégrité sur les auteurs

> « *(…)The integrity of scholarly publishing, and its role as a trusted provider of knowledge, rests on authors carrying out research responsibly and writing it up accurately, and on editors and reviewers scrutinising the words, data and images with care. […]The move to digital publishing has removed that constraint and changed publishers' (if not researchers') need for vigilance […]* » (Crosetto et al., The Drain of Scientific Publishing, 2025, p.6)

- En transférant la responsabilité de l'intégrité sur les auteurs, la digitalisation du marché éditorial affaiblit les mécanismes de contrôle social externalisés de la science = niche pour le développement d'un marché de la fraude

Table 1.1 Estimated valuation of the combined global academic fraud industry

Industry	Estimated valuation (US dollars)	Sources
Contract cheating (e.g., homework help sites, commercial file-sharing, thesis-writing services, essay mills)	$15 billion	Eaton (2021a, 2022); Lancaster and Prakhar (2022)
Fake and fraudulent degree industry	$7 billion	Ezell (2022)
Scientific and scholarly paper mills	Unknown	No information available as of 2022.
Total	$21 billion	

Sarah Elaine Eaton, Jamie J. Carmichael - *Fake Degrees and Fraudulent Credentials in Higher Education* - 2023

Paper Mills >> 10M$/an, Matt Hodgkinson, Head of Editorial DOAJ

Facteurs explicatifs: le rôle-clef des politiques publiques

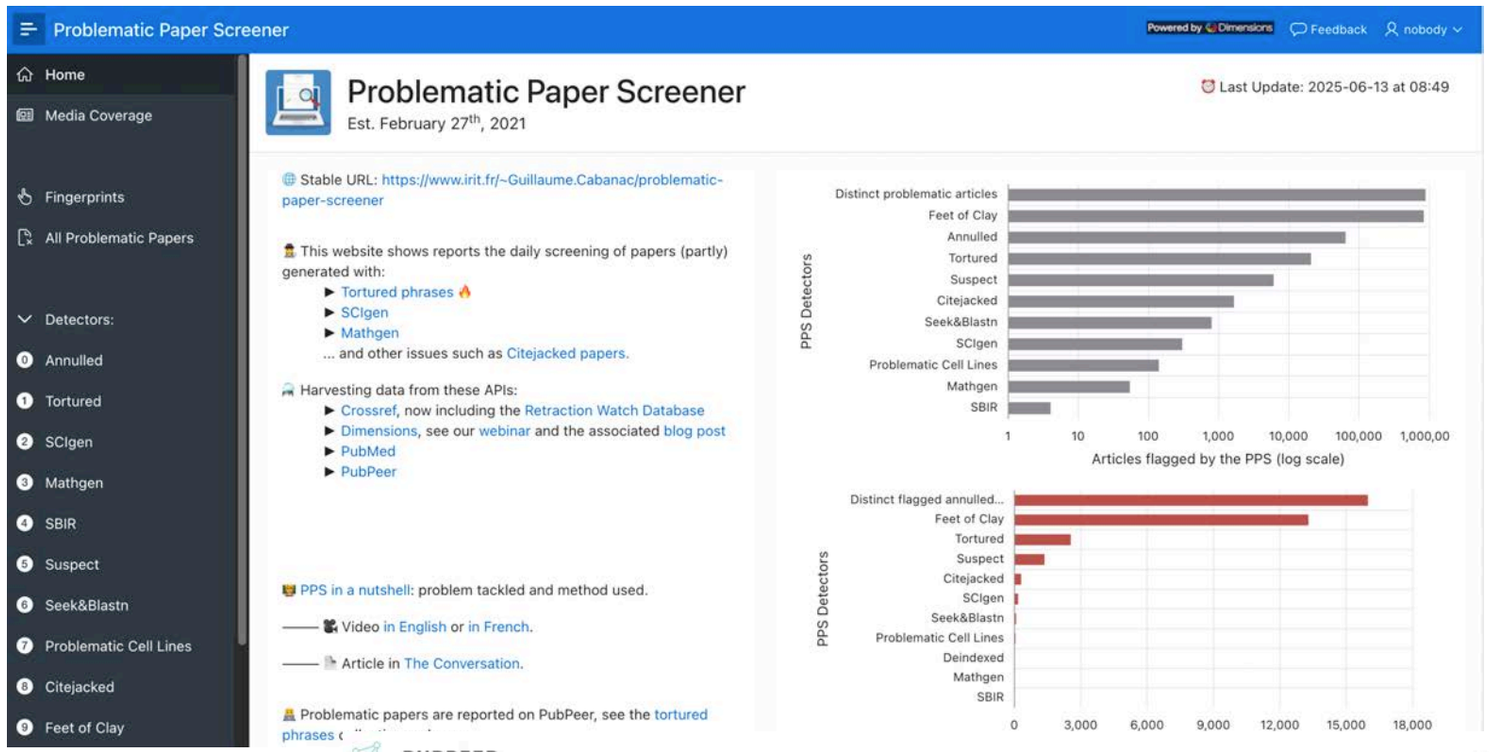
- **Contre-Intuitif** : la culture du « publish or perish » n'est pas statistiquement corrélée à plus de fraude (Fanelli, Costas & Larivière, 2015; Lacetera & Zirulia, 2011)

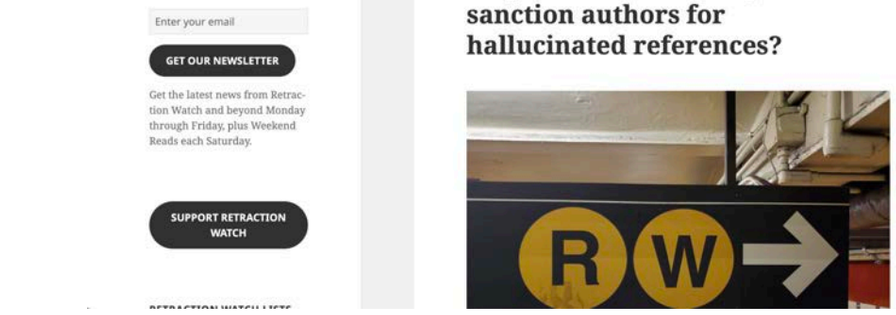
> *« (…)our findings support previous suggestions that national policies, socio-cultural conditions, research environment (including number of co-authors) and situational factors (i.e. career stage) are significant determinants of responsible and irresponsible practices. »*
> *Fanelli et al., 2015, Misconduct Policies, Academic Culture and Career Stage, Not Gender or Pressures to Publish, Affect Scientific Integrity, p.11*

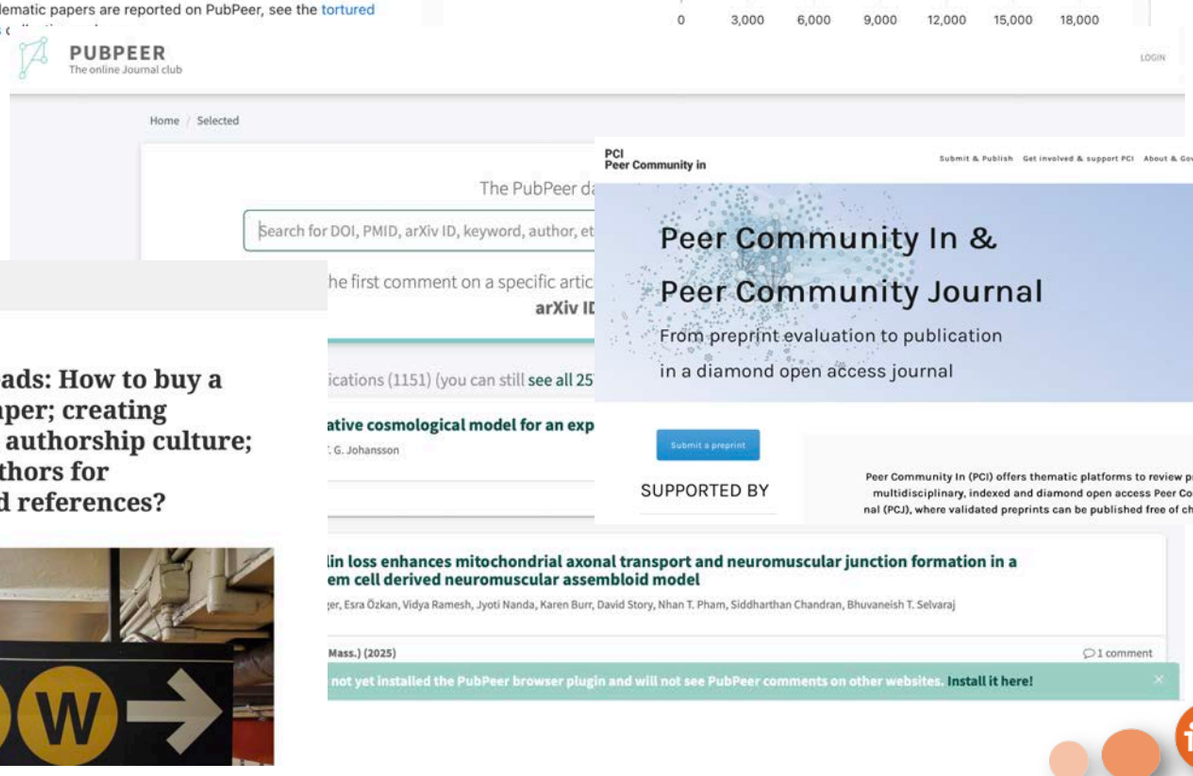
- Par contre, les **contextes organisationnels** et les **modalités pratiques d'incitation à l'excellence** sont incitatifs ou dissuasifs :
  - Absence de **politiques nationales d'intégrité scientifique**
  - **Primes financières directes** à la publication
  - Absence de **culture de la critique** entre pairs
  - **Stade précoce de carrière** – vulnérabilité spécifique

- Une tradition d'analyse des fraudes scientifiques en économie standard (Wibble) et économie behavioriste (Necker, Lacetera, …):
  - La fraude comme choix rationnel sous **contrainte informationnelle asymétrique**,
  - Une forte pression à publier peut paradoxalement *réduire* la fraude en stimulant la vigilance mutuelle,
  - Les *star scientists* frauderaient plus mais seraient moins détectés (pare-feu réputationnel),
  - …

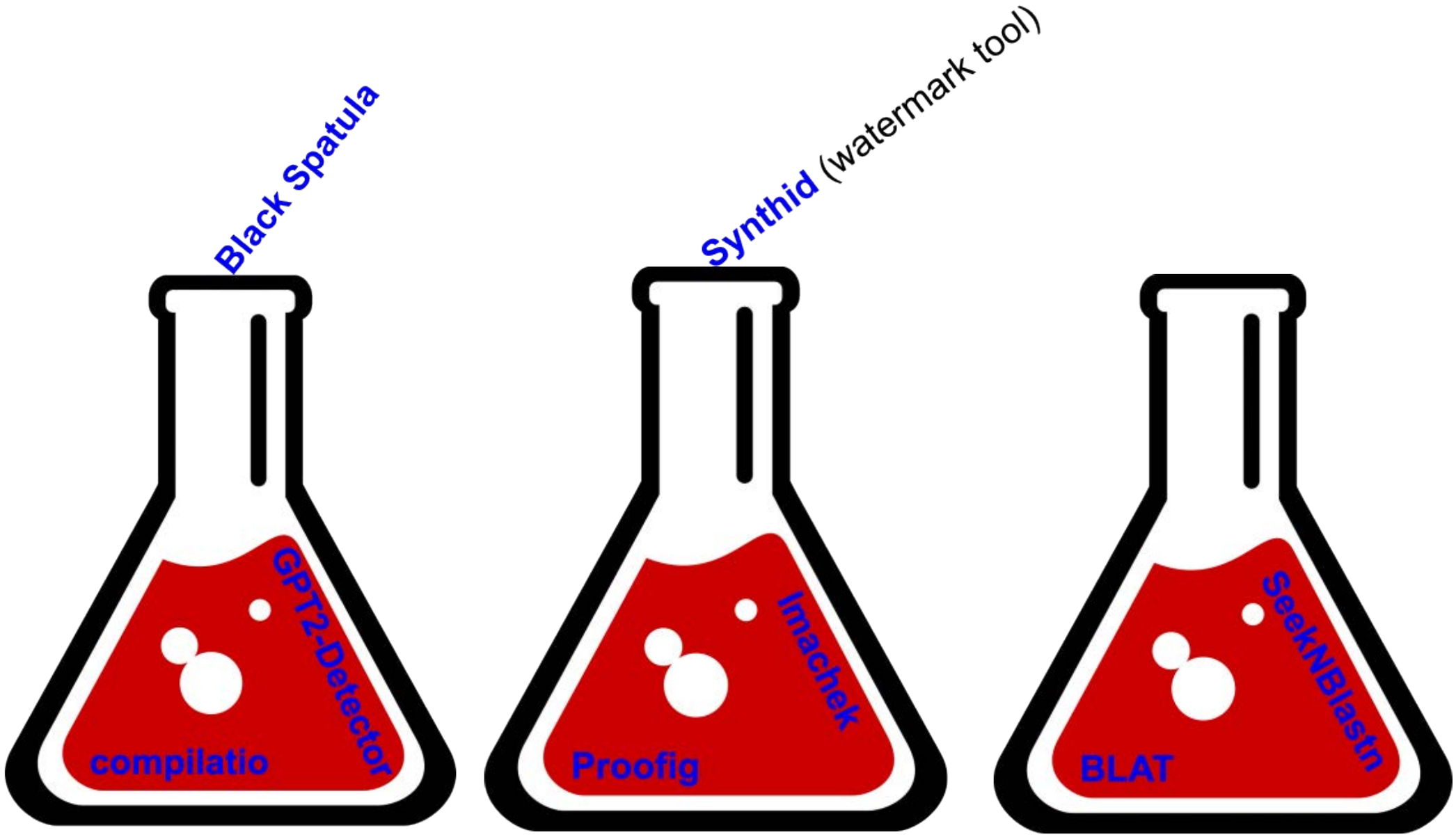
Outiller l’intégrité scientifique







Une offre restreinte et spécialisée d'outils dédiés à l'intégrité scientifique



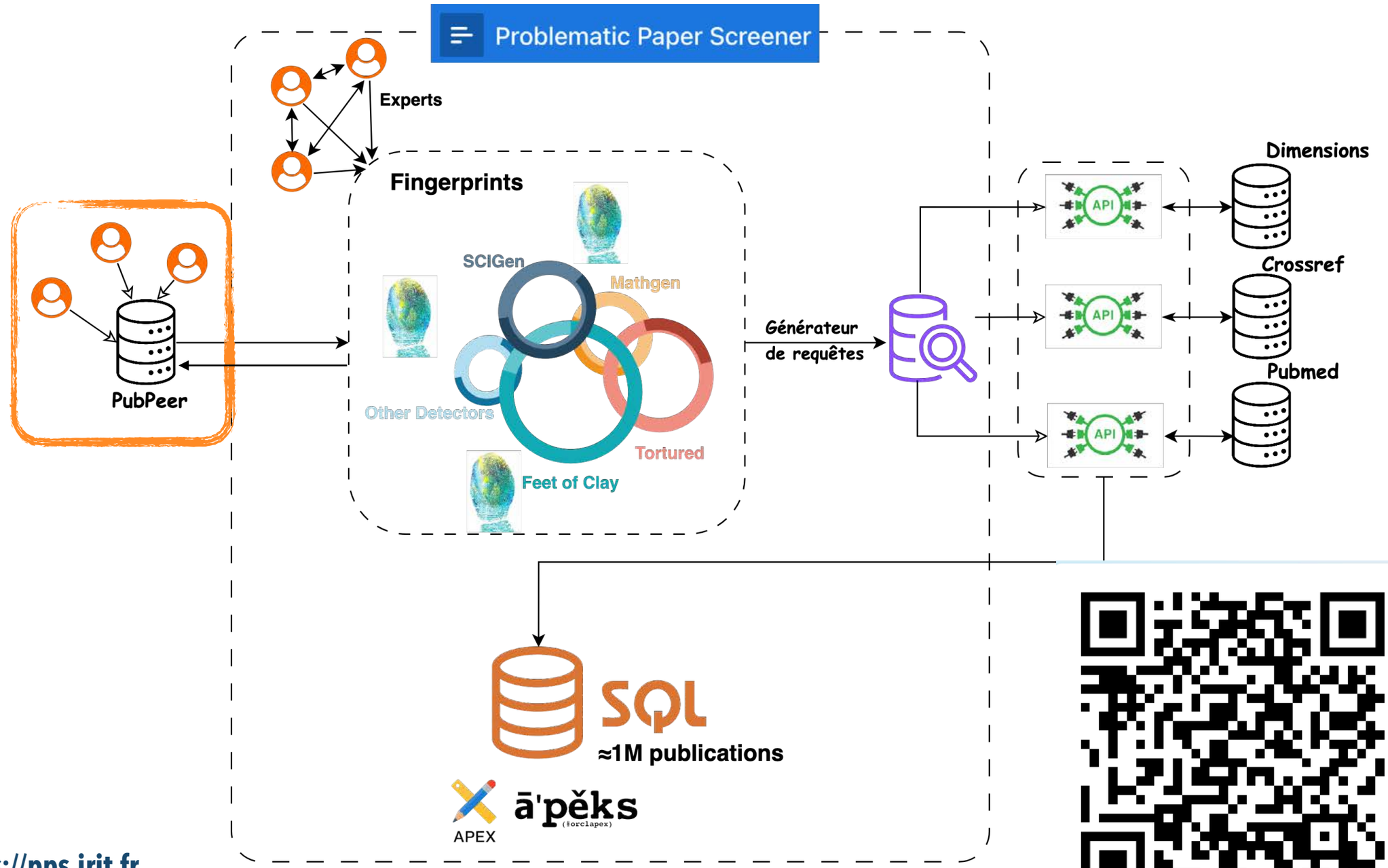
Détection de textes générés par IA

Détection d'images dupliquées/trafiquées

Détection de composés chimériques

Le Problematic Paper Screener : un système d'information unifié dédié à l'intégrité scientifique

💡 Créé en 2022 par **Guillaume Cabanac, Cyril Labbé et Alexander Magazinov** (https://doi.org/10.48550/arXiv.2210.04895)

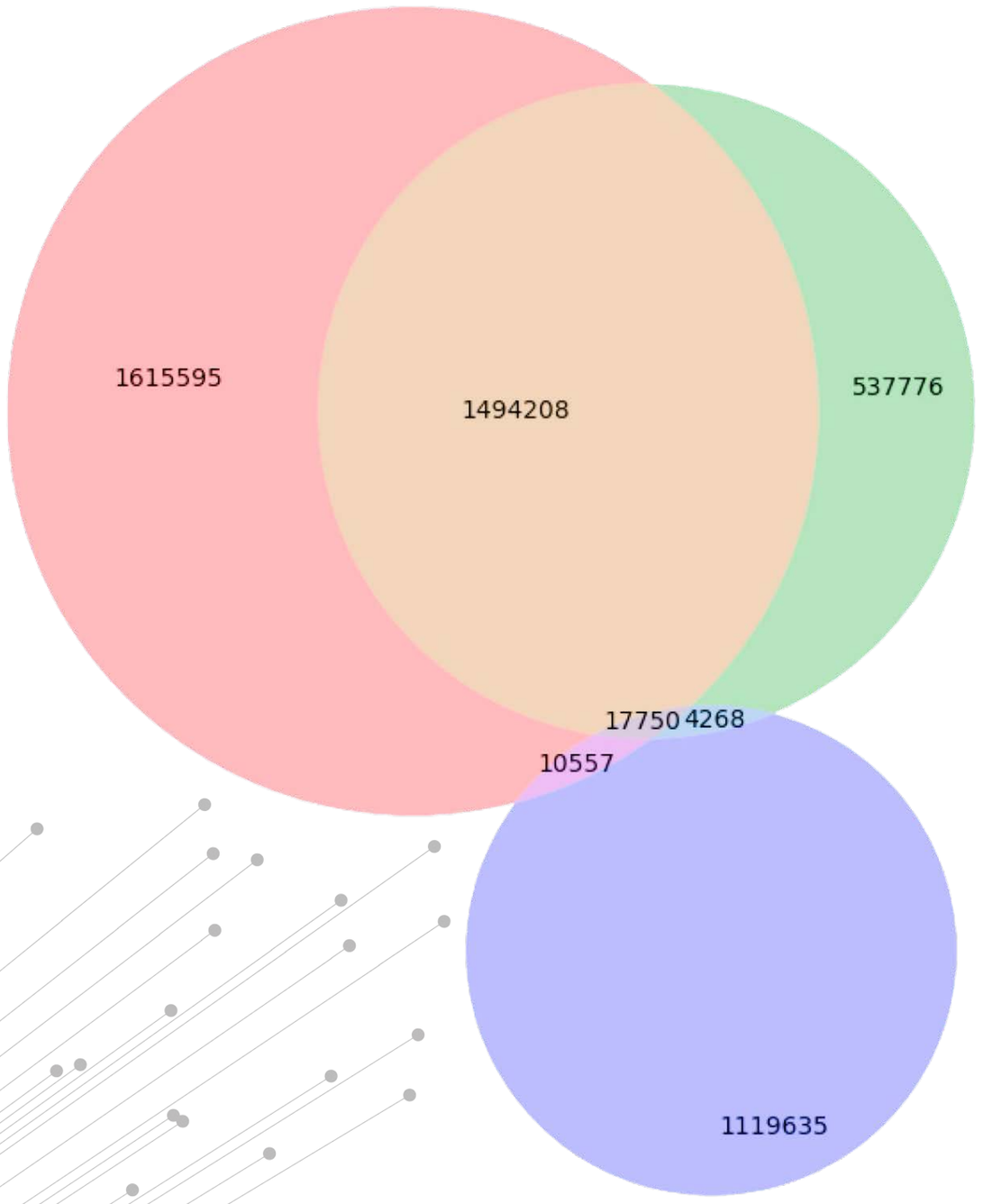


https://pps.irit.fr

https://www.irit.fr/~Guillaume.Cabanac/problematic-paper-screener

PPS - France

Recouvrement avec les données de recherche française



- PPS x HAL : **9 834** publications de HAL sont dans le PPS
- PPS x Open Alex : **28,320** publications présentes dans le PPS (>x 2 HAL)
- PPS x BSO : **22,017** publications présentes dans le PPS

PPS : une base de données de « *fingerprints* »

➡ **Fingerprint** = trace/indice de mauvaises pratiques = présence d'**éléments artificiels** :

- **syntaxiques** : paraphrasage et expressions torturées, expressions automatiques de prompts d'IAs génératives, expressions liées à des générateurs de text, …
- **référentiels** : fausses références bibliographiques, références rétractées, références à des articles de journaux « highjackés », …
- **techniques** : problème de notation et d'usage de séquences nucélotidiques, fausses formules mathématiques, …

➡ Les « fingerprints » sont regroupées en **détecteurs (n = 10):**

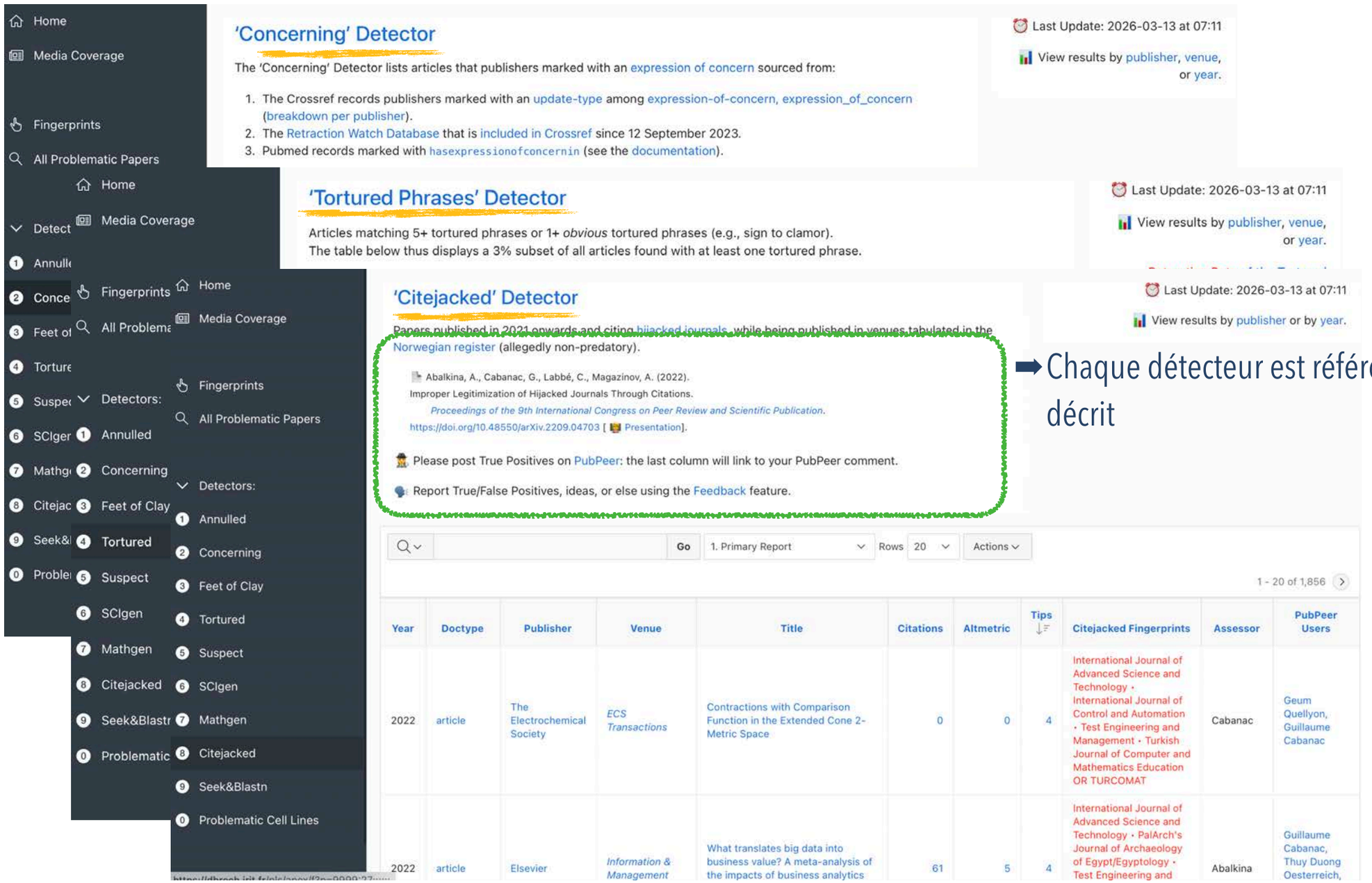
1. Annulled
2. Concerning
3. Feet of Clay
4. Tortured
5. Suspect
6. SCIgen
7. Mathgen
8. Citejacked
9. Seek&Blastn
0. Problematic Cell Lines

➡ Papiers rétractés ou annulés ou suspendus	**137 283** papiers
➡ Papiers avec « *expressions of concerns* » de l'éditeur	**6 812** papiers
➡ Papiers citant des références rétractées et/ou présentes sur le PPS	**1 172 911** papiers
➡ Papiers possédant des expressions torturées (1 discriminante ou 5)	**25 405** papiers
➡ Papiers possédant 5 fingerprints (diverses)	**6 909** papiers
➡ Papiers générés avec le logiciel SciGen	**301** papiers
➡ Papiers générés avec le logiciel Mathgen	**53** papiers
➡ Papiers citant des papiers publiés dans des revues piratées	**1 856** papiers
➡ Papiers citant des séquences de nucléotides problématiques	**791** papiers
➡ Papiers citant de mauvaises lignées cellulaires	**1 596** papiers

PPS : une base de données de « *fingerprints* »

Présentation de l'interface

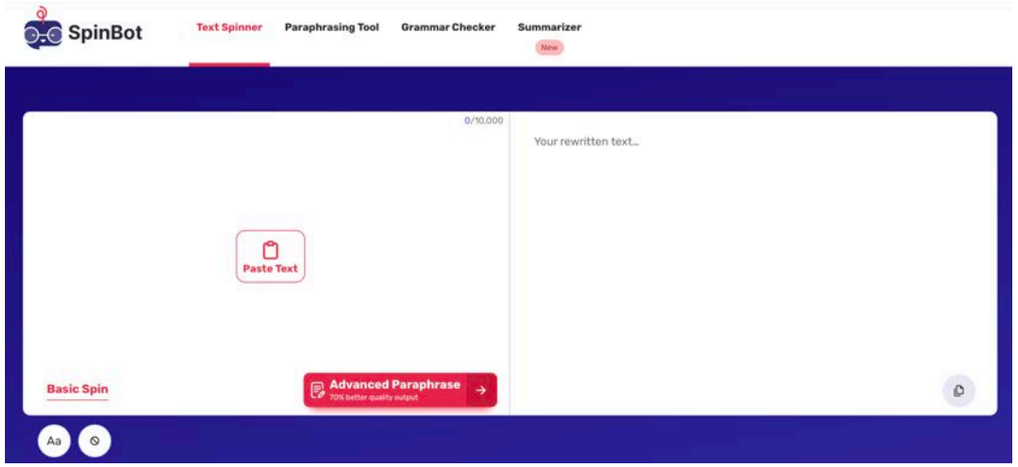
➡Chaque détecteur possède son propre tableau:



➡Chaque détecteur est référencé et décrit

PPS : Zoom sur le détecteur de phrases torturées

Expression torturée	Expression académique
Neural organization	Neural network
Bosom disease|malignancy	Breast cancer
Lactose Bigotry	Lactose Intolerance
Overwhelming metal	Heavy metal
…	…



- **Expression torturée** = expressions non-conventionnelles mimant une expression scientifique conventionnelle par substitution de synonymes aux termes attendus (spinbot).
- En termes linguistiques: une désintégration des unités lexicalisées du langage scientifique par synonymisation.

« [Our study introduces the] concept of tortured phrases: ***unexpected weird phrases in lieu of established ones****, such as 'counterfeit consciousness' instead of 'artificial intelligence. » Cabanac et al., Tortured phrases: A dubious writing style emerging in science, 2021*

« We define a tortured phrase as an ***expression resulting from the use of a spinner on a well-established scientific expression*** *with a specific and fixed meaning. » Lentshat et al., Detection of tortured phrases in scientific literature, 2024.*

- **Pour quoi faire?** Ne pas être détecter par des logiciels d'anti-plagiat

PPS : une base de données de « *fingerprints* »

Zoom sur le détecteur de « phrases torturées »

« Big data offers unique opportunities to study human behavior on a large scale. It has been increasingly used in social and psychological research to explore individual differences and group dynamics. However, there are several theoretical and methodological challenges in big data research that deserve attention. Here, we highlight four key challenges: the balance between data-driven and hypothesis-driven approaches, the validity of measurements, multi-level longitudinal analysis, and data integration. These represent fundamental issues that social scientists often face when working with big data. We provide examples of these challenges and propose potential solutions. »

Texte généré par paraphrasage inversé par prompt – Modèle Anthropic, Poe,

«Big data presents unprecedented opportunities to understand human behavior on a large scale. It has been increasingly used in social and psychological research to reveal individual differences and group dynamics. There are a few theoretical and methodological challenges in big data research that require attention. In this paper, we highlight four issues, namely data-driven versus theory-driven approaches, measurement validity, multi-level longitudinal analysis, and data integration. They represent common problems that social scientists often face in using big data. We present examples of these problems and propose possible solutions. »

Résumé original de l'article intitulé « Big data in social and psychological science: theoretical and methodological issues », paru en janvier 2018 dans le Journal of Computational Social Science, par Lin Qiu et al.

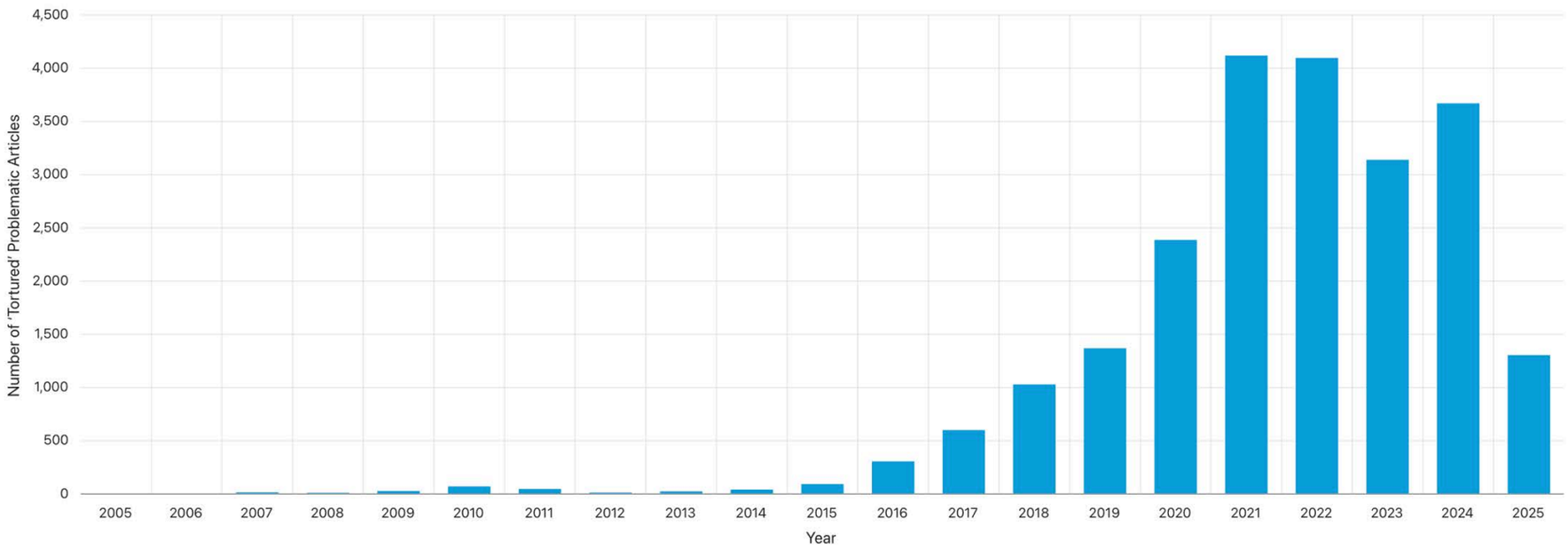
PPS : une base de données de « *fingerprints* »

Zoom sur le détecteur de « phrases torturées »

➡ Catalogue des fingerprints de phrases torturées (n = 8 076):

Fingerprint - Tortured Phrase	Expected Text	Nb Retrieved Papers ↓
"simulated intelligence"	artificial intelligence	1,177
"computerized reasoning"	artificial intelligence	1,160
"computer-based intelligence"	artificial intelligence (or possibly \"machine learning\")	1,016
"man-made brainpower"	artificial intelligence	664
"man-made intelligence"	artificial intelligence	661
"man-made consciousness"	artificial intelligence	659
"man-made reasoning"	artificial intelligence	633
"synthetic intelligence (AI)"	artificial intelligence (AI)	481
"fake knowledge"	artificial intelligence	393
"administered AI"	supervised artificial intelligence	231
"manufactured insights"	artificial intelligence	207
"fake insight"	Artificial Intelligence	199

L'expression « *man-made brainpower*» est recherchée dans l'ensemble de la littérature scientifique mondiale (Dimensions.ai), 664 papiers comportent cette expression



Évolution du nombre de publications identifiées dans le PPS par le détecteur de phrases torturées

PPS : une base de données de « *fingerprints* »

Zoom sur le détecteur de « phrases torturées »

Publisher	Nb. Publications
Institute of Electrical and Electronics Engineers (IEEE)	7,968
Springer Nature	5,067
Elsevier	2,901
Taylor & Francis	1,333
Wiley	1,182
AIP Publishing	863
Hindawi	848
SAGE Publications	469
IOP Publishing	416
MDPI	325
EDP Sciences	235

Classement des maisons d'édition par nombre de publications identifiées dans le PPS par le détecteur de phrases torturées - top 10

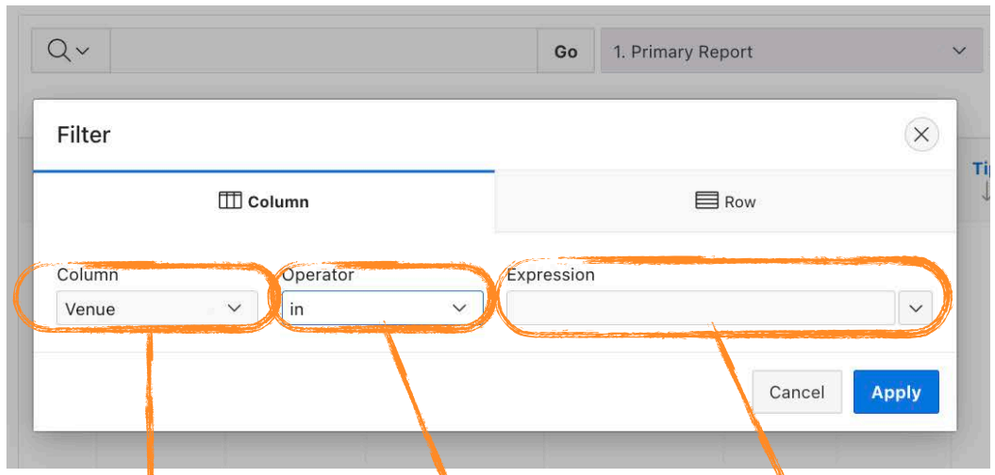
Venue	Nb. Publications
AIP Conference Proceedings	858
Lecture Notes in Networks and Systems	688
Materials Today Proceedings	623
Lecture Notes in Electrical Engineering	469
2024 International Conference on Optimization Computing and Wireless Communication (ICOCWC)	417
2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)	416
Microprocessors and Microsystems	407
Advances in Intelligent Systems and Computing	349
SSRN Electronic Journal	336
Research Square	230

Classement des revues par nombre de publications identifiées dans le PPS par le détecteur de phrases torturées - top 10

PPS : Atelier

➡PPS >> Détecteur « Tortured » >> Actions >> Filter

- Rechercher des revues de vos domaines



Go!!!

Sélection de la colonne à filtrer

Opérateur booléen de filtrage

Pattern(s) à rechercher

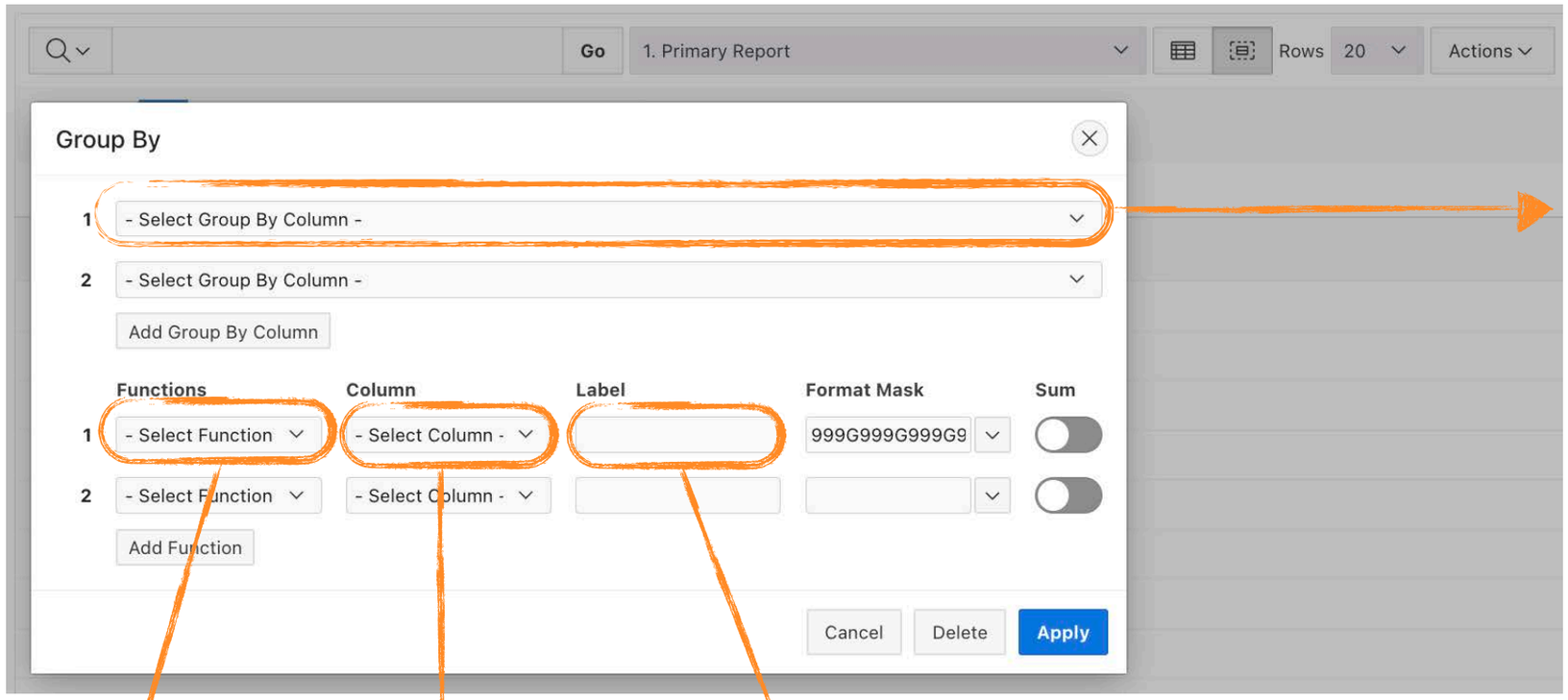
⚠

- « **=** » : égalité exacte, case-sensitive (nombres et chaînes de caractères)
- « **!=**»: différent de(nombres et chaînes de caractères)
- « **in** »: recherche l'ensemble des expressions séparées par une virgule (nombres et chaînes de caractères)
- « **like** »: Pattern SQL,% comme wildcard (chaînes de caractères)
- « **contains** »: sous-chaîne présente (chaînes de caractères)
- « **not contains** »: sous-chaîne absente (chaînes de caractères)
- « **Is null** » : valeurs nulles

PPS : Atelier

➡**PPS >> Détecteur « Concerned » >> Actions >> Group By**

- Quelles sont les revues qui ont reçues le plus d'expressions of concerns?
- Parmi les revues ayant reçues des « expressions of concerns » quelles sont celles qui ont reçues le plus de commentaires PubPeer?



Colonne que l'on groupe

Type d'opération de groupement

Colonne sur laquelle l'opération va porter

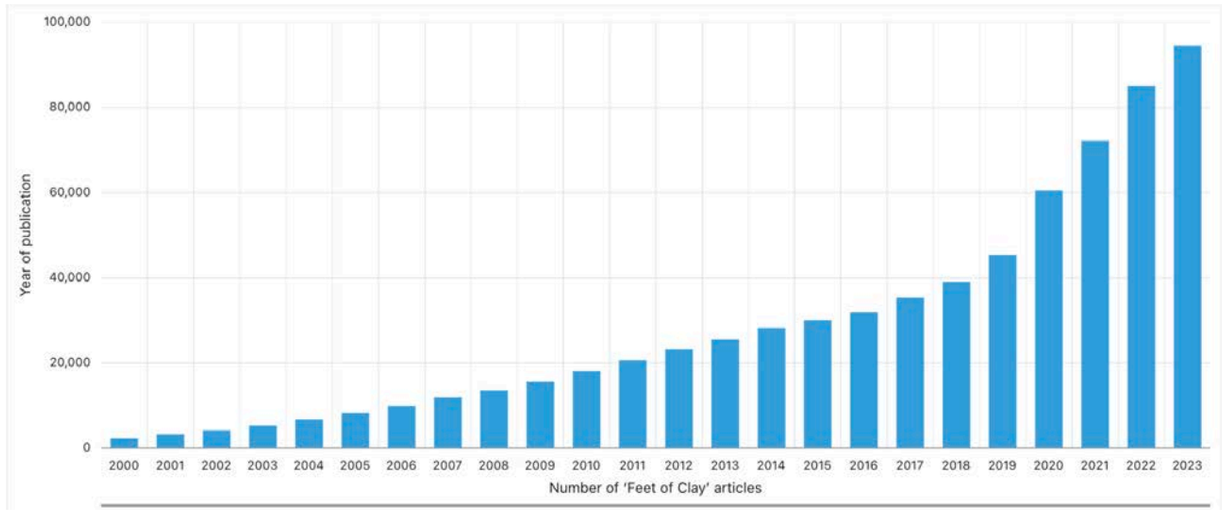
Nom de la colonne Dde résultats

PPS : une base de données de « *fingerprints* »

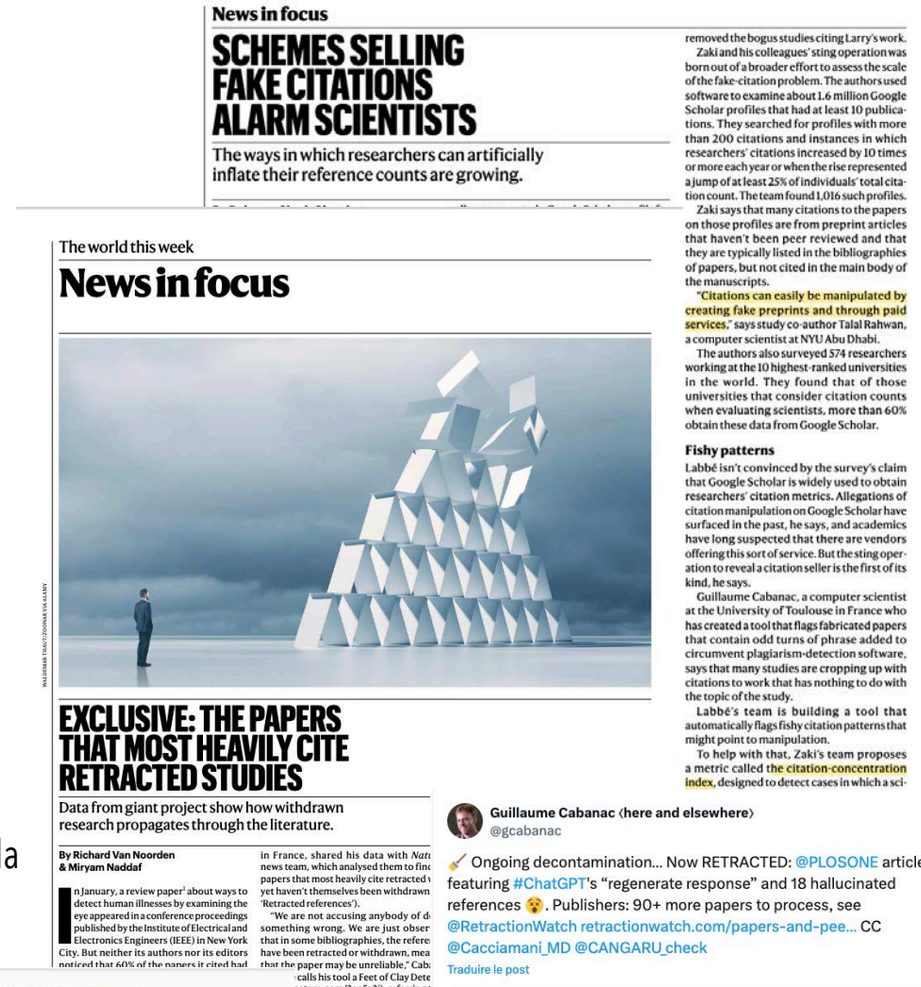
Zoom sur le détecteur « Feet of Clay »

- « Feet of Clay » = détecteur de **citations problématiques**
- **Citation problématique** =

Citation rétractée et/ou citation d'un article identifié par un ou plusieurs autres détecteurs du PPS



Evolution du volume de papiers comportant des citations problématiques par an sur la période 2000-2023 – IR PPS « Yearly distribution [2000-2023] »



News in focus

SCHEMES SELLING FAKE CITATIONS ALARM SCIENTISTS

The ways in which researchers can artificially inflate their reference counts are growing.

removed the bogus studies citing Larry's work.

Zaki and his colleagues' sting operation was born out of a broader effort to assess the scale of the fake-citation problem. The authors used software to examine about 1.6 million Google Scholar profiles that had at least 10 publications. They searched for profiles with more than 200 citations and instances in which researchers' citations increased by 10 times or more each year or when the rise represented a jump of at least 25% of individuals' total citation count. The team found 1,016 such profiles.

Zaki says that many citations to the papers on those profiles are from preprint articles that haven't been peer reviewed and that they are typically listed in the bibliographies of papers, but not cited in the main body of the manuscripts.

"Citations can easily be manipulated by creating fake preprints and through paid services," says study co-author Talal Rahwan, a computer scientist at NYU Abu Dhabi.

The authors also surveyed 574 researchers working at the 10 highest-ranked universities in the world. They found that of those universities that consider citation counts when evaluating scientists, more than 60% obtain these data from Google Scholar.

Fishy patterns

Labbé isn't convinced by the survey's claim that Google Scholar is widely used to obtain researchers' citation metrics. Allegations of citation manipulation on Google Scholar have surfaced in the past, he says, and academics have long suspected that there are vendors offering this sort of service. But the sting operation to reveal a citation seller is the first of its kind, he says.

Guillaume Cabanac, a computer scientist at the University of Toulouse in France who has created a tool that flags fabricated papers that contain odd turns of phrase added to circumvent plagiarism-detection software, says that many studies are cropping up with citations to work that has nothing to do with the topic of the study.

Labbé's team is building a tool that automatically flags fishy citation patterns that might point to manipulation.

To help with that, Zaki's team proposes a metric called the citation-concentration index, designed to detect cases in which a sci-

The world this week

News in focus

EXCLUSIVE: THE PAPERS THAT MOST HEAVILY CITE RETRACTED STUDIES

Data from giant project show how withdrawn research propagates through the literature.

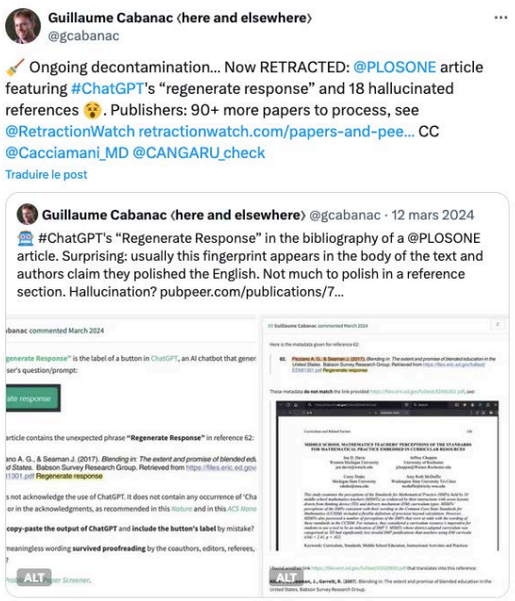
By Richard Van Noorden & Miryam Naddaf

In January, a review paper[1] about ways to detect human illnesses by examining the eye appeared in a conference proceedings published by the Institute of Electrical and Electronics Engineers (IEEE) in New York City. But neither its authors nor its editors noticed that 60% of the papers it cited had

in France, shared his data with *Nat* news team, which analysed them to find papers that most heavily cite retracted yet haven't themselves been withdrawn 'Retracted references').

"We are not accusing anybody of d something wrong. We are just obser that in some bibliographies, the refere have been retracted or withdrawn, mea that the paper may be unreliable," Cab calls his tool a Feet of Clay Dete nature.com/3ap5a3i), referring t originally from the Bible, about difices that collapse because of y foundations.

EEE paper is the second-highes ssembled by *Nature*, with 18 of th



Guillaume Cabanac ‹here and elsewhere› @gcabanac

🧹 Ongoing decontamination… Now RETRACTED: @PLOSONE article featuring #ChatGPT's "regenerate response" and 18 hallucinated references 😵. Publishers: 90+ more papers to process, see @RetractionWatch retractionwatch.com/papers-and-pee… CC @Cacciamani_MD @CANGARU_check

Traduire le post

Guillaume Cabanac ‹here and elsewhere› @gcabanac · 12 mars 2024

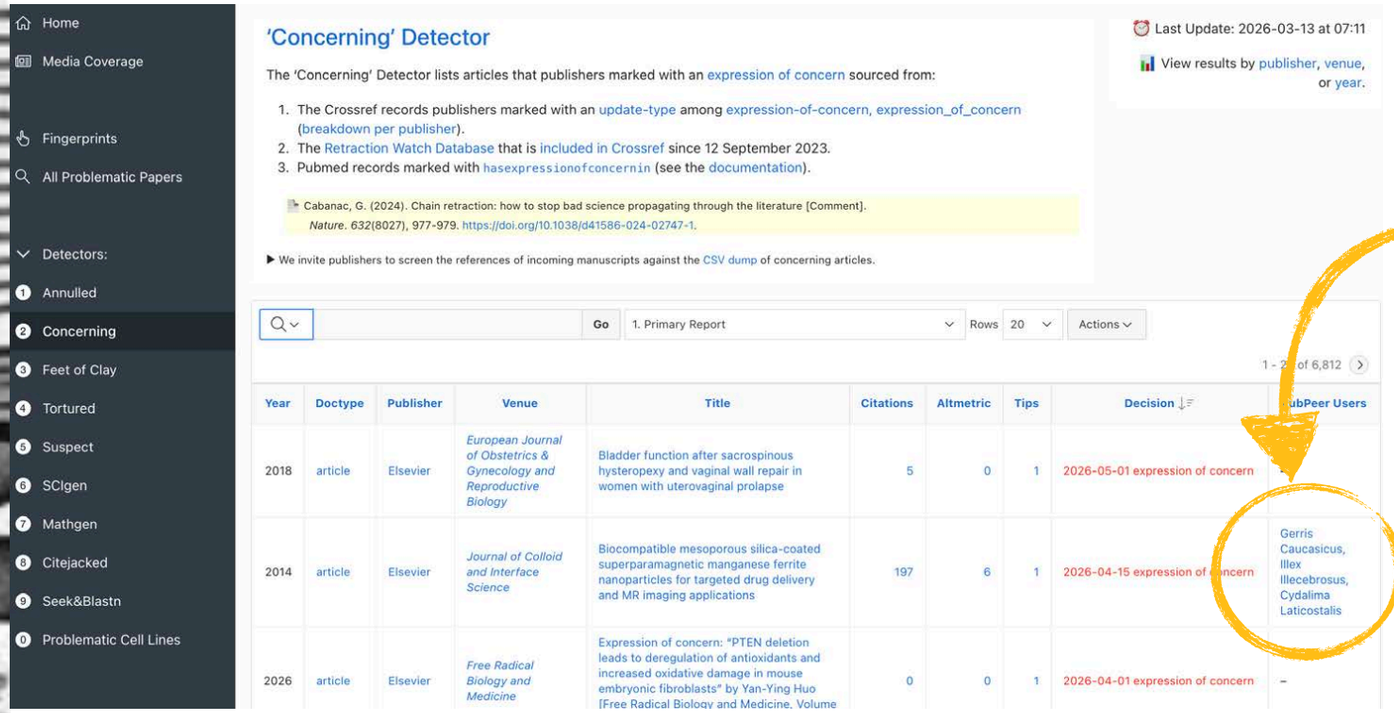
#ChatGPT's "Regenerate Response" in the bibliography of a @PLOSONE article. Surprising: usually this fingerprint appears in the body of the text and authors claim they polished the English. Not much to polish in a reference section. Hallucination? pubpeer.com/publications/7…

Publisher	Nb. Publication
Elsevier	277,419
Springer Nature	205,593
Wiley	96,702
MDPI	65,863
Taylor & Francis	58,069
Institute of Electrical and Electronics Engineers (IEEE)	40,614
Frontiers	33,303
SAGE Publications	25,556
Wolters Kluwer	24,106
Oxford University Press (OUP)	23,849
Hindawi	20,644

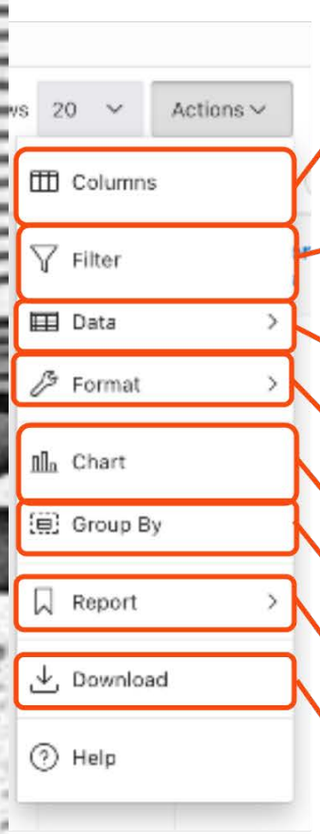
PPS : une base de données de « *fingerprints* »

Présentation de l'interface

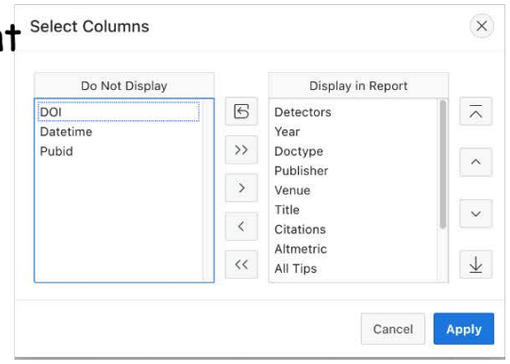
➡Chaque tableau est connecté à la plateforme de révision post-publication PubPeer:



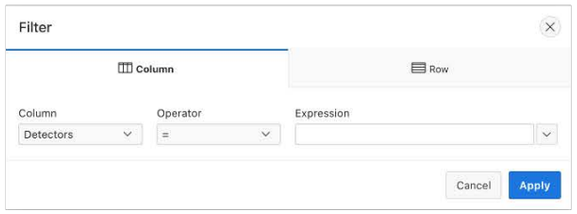
PPS : Toutes les opérations possibles



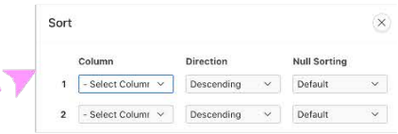
Modifier l'affichage courant en sélectionnant/dé-sélectionnant les colonnes apparentes.



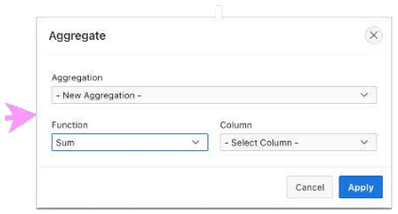
Filtrer les résultats à l'aide d'opérateurs logique.



Trier par ordre asc./desc.



Aggréger à l'aide d'une fonction

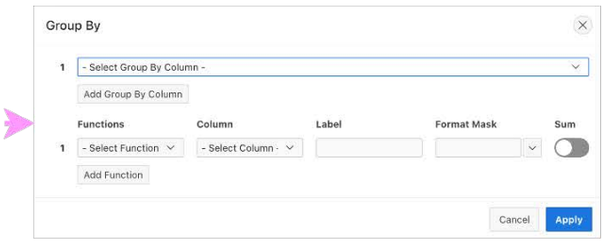


Calculer des résultats intermédiaires

Outil de formatage des données

Création de graphiques

Aggrégation complexe de données



Réinitialisation de la vue

Télécharger les données courantes(csv)

PPS : Atelier

➡ **Retrouver tous les papiers du Publisher Hindawi sur le PPS:**

- Quelles sont les revues les plus problématiques (top 10)?
- Quelles sont les revues pour lesquelles on décompte le plus grand nombre d'articles rétractés (top 10)?
- Quelles sont les revues pour lesquelles les articles contenant des phrases torturées sont les plus nombreux (top 10)?

Hindawi reveals process for retracting more than 8,000 paper mill articles

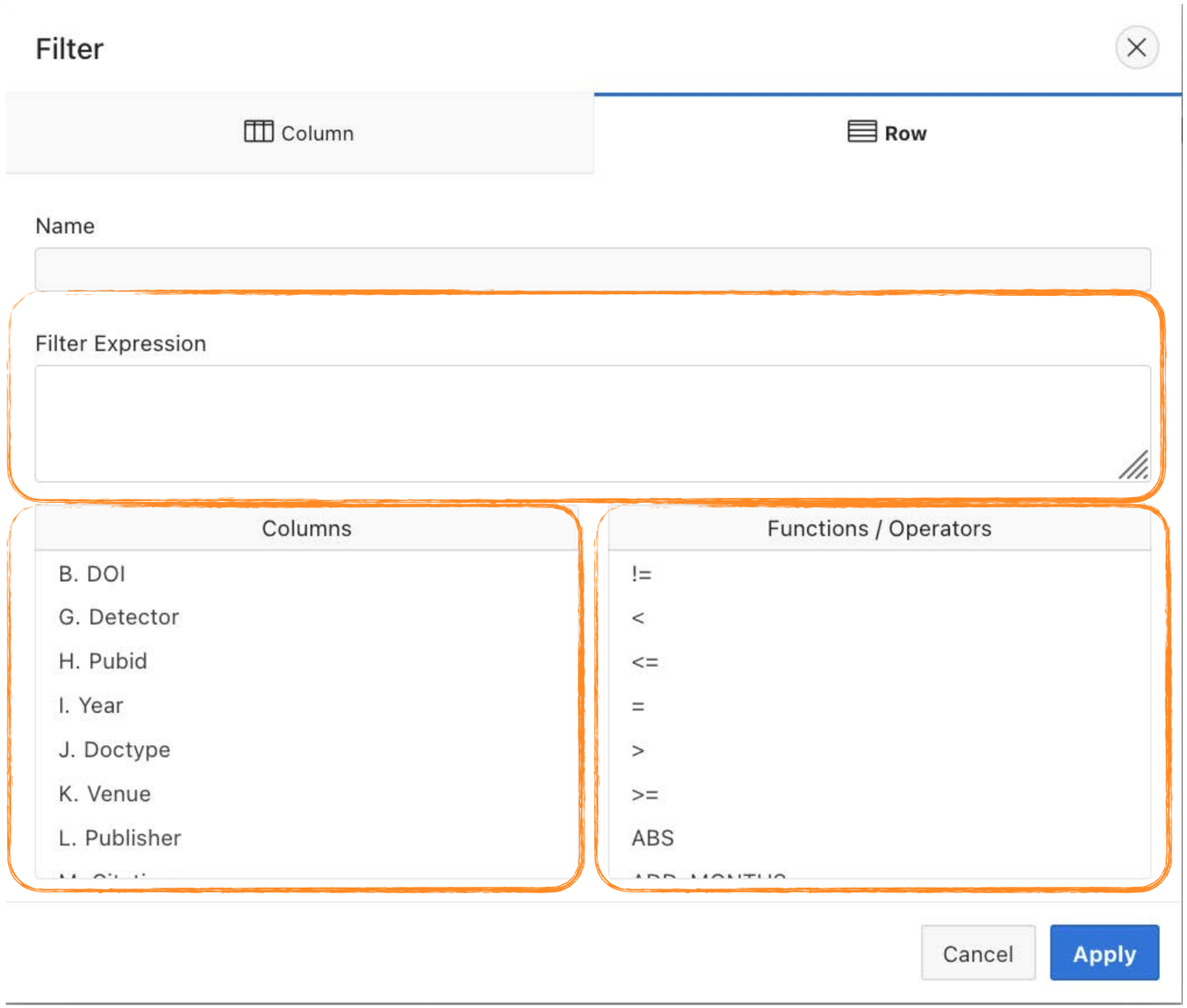
Over the past year, amid announcements of thousands of retractions, journal closures and a major index delisting several titles, executives at the troubled publisher Hindawi have at various times mentioned a "new retraction process" for investigating and pulling papers "at scale." The publisher has declined to provide details – until now.

So far in 2023, Hindawi has retracted over 8,000 articles – more than we've ever seen in a single year from all publishers combined. And Hindawi is not done cleaning up from paper mills' infiltration of its special issues, according to a new report from its parent company, Wiley.

Reckoning with Hindawi's paper mill problem has cost Wiley, which bought the open-access publisher in 2021, an estimated $35-40 million in lost revenue in the current fiscal year, Matthew Kissner, Wiley's interim president and CEO, said on the company's most recent earnings call. Wiley will stop using the "Hindawi" name next year, Kissner told investors.

PPS : Filtrage complexe

➡ Filtrage par ligne



Variables de filtres (colonnes)

Espace de composition des requêtes avancées (expressions régulières, multi-critères de filtres, …)

Opérateurs relationnels et logiques

PPS : Filtrage complexe

Règle de filtrage	Fonction
B IN ('10.1016/j.ijantimicag.2020.105932','10.1016/j.micpath.2020.104228','10.1016/j.tmaid.2020.101663')	Afficher les résultats pour la liste de doi compris entre ()
((K='Sustainability') AND (I>=2020) AND (M>=10) AND (O >=5))	Afficher toutes les publications de la revue « Sustainability » publiés après 2020 qui ont reçu au moins 10 citations et qui ont été flaggés par le PPS pour au moins 5 raisons différentes.
(REGEXP_LIKE(K, 'socio|anthropo|econom|histo', 'i') AND NOT REGEXP_LIKE(K, 'bioeconomy','i'))	Afficher toutes les revues (colonne K) dont le titre contient soit « socio », soit « histo », soit « econom », soit « anthropo », mais pas « bioeconomy ». La règle est « case insensitive (‘i’).

PPS : Filtrage complexe - Cas pratique

➡ **Je veux vérifier les productions de mon unité de recherche**

1️⃣ Je récupères les identifiants HAL de mon laboratoire/structure

2️⃣ Je récupère la liste des publications HAL

http://api.archives-ouvertes.fr/search/?
q=structId_i:**1001883**&fl=docid,doiId_s,publicationDate_s,publicationDateY_i,title_s&rows=2000&wt=csv

- **Docid de votre structure**
- Infos (colonnes) à récupérer
- Nb. Docs à récupérer (adaptez)
- Exporter en csv directement

3️⃣ Je **concatène** la liste des dois de mes publications au **format suivant**:

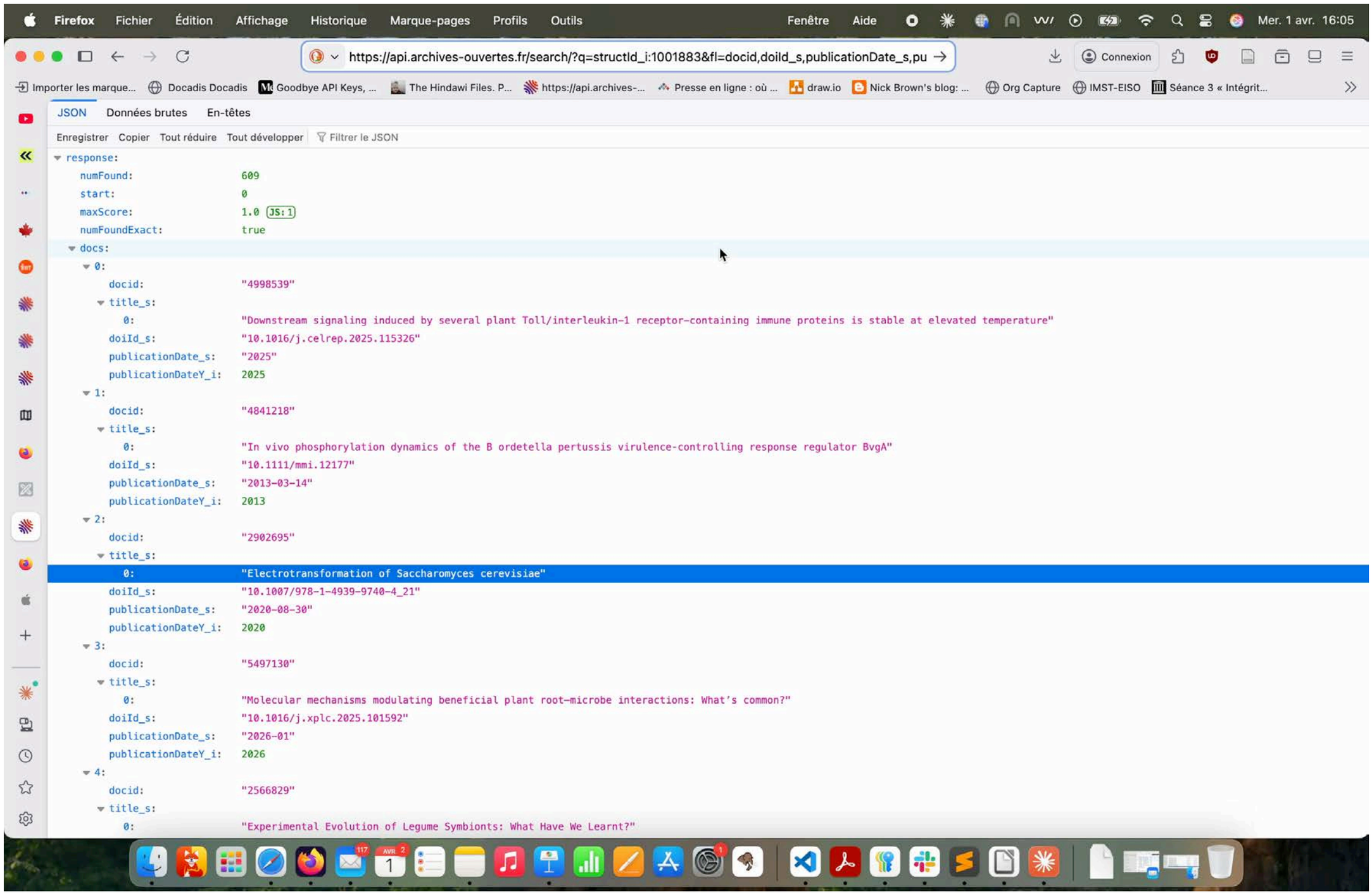
('doi1','doi2', 'doi3',.....'doi*n*')

⚠️ **Pas plus de 150/155 dois à la fois**

4️⃣ Dans la fenêtre de filtre, je filtre par « **row** »:

B. IN ('10.7554/eLife.90170.3','10.7554/eLife.59384','10.7554/eLife.25114','10.64898/2025.12.15.694456',...)

PPS : Filtrage complexe - Cas pratique



PPS : Participez!

Une plateforme collaborative

➡ Le PPS est une plateforme nourrie par:

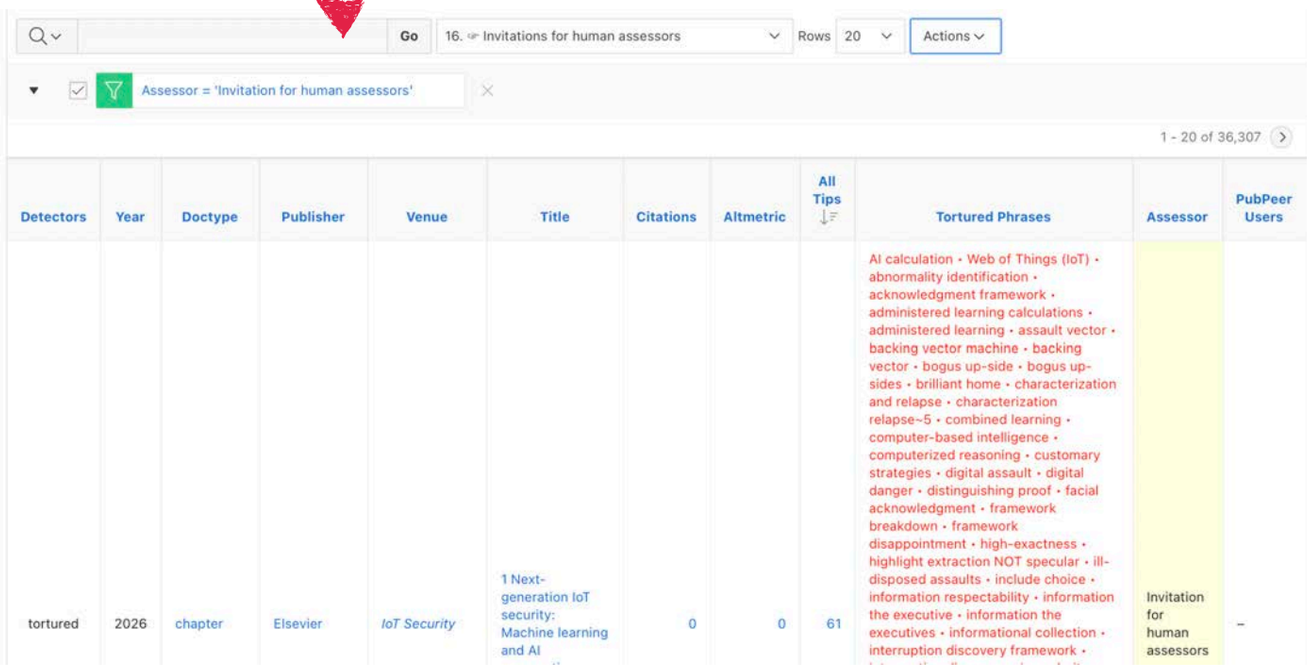
- ✓ des retours d'experts,
- ✓ de la base de données PubPeer,
- ✓ des retours **utilisateurs**!

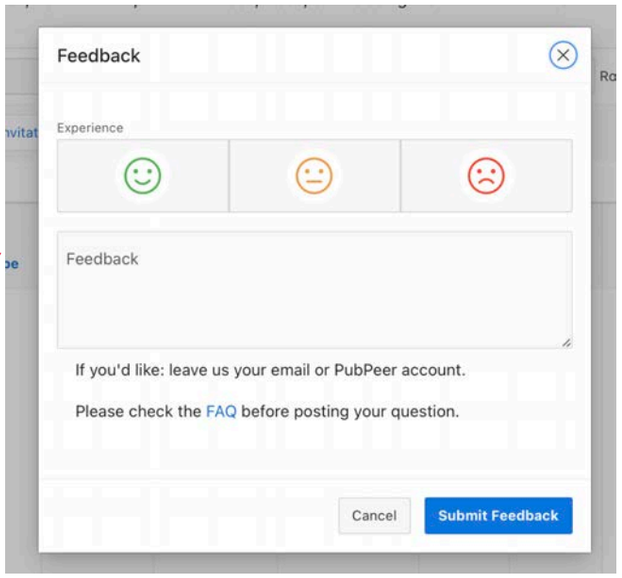
All Problematic Papers Found with the Detectors

👮 Check potentially problematic papers and report on PubPeer: the last column will link to your PubPeer comment.

🗣️ Report True/False Positives, new *tortured phrases* to track, ideas, or else using the Feedback feature.

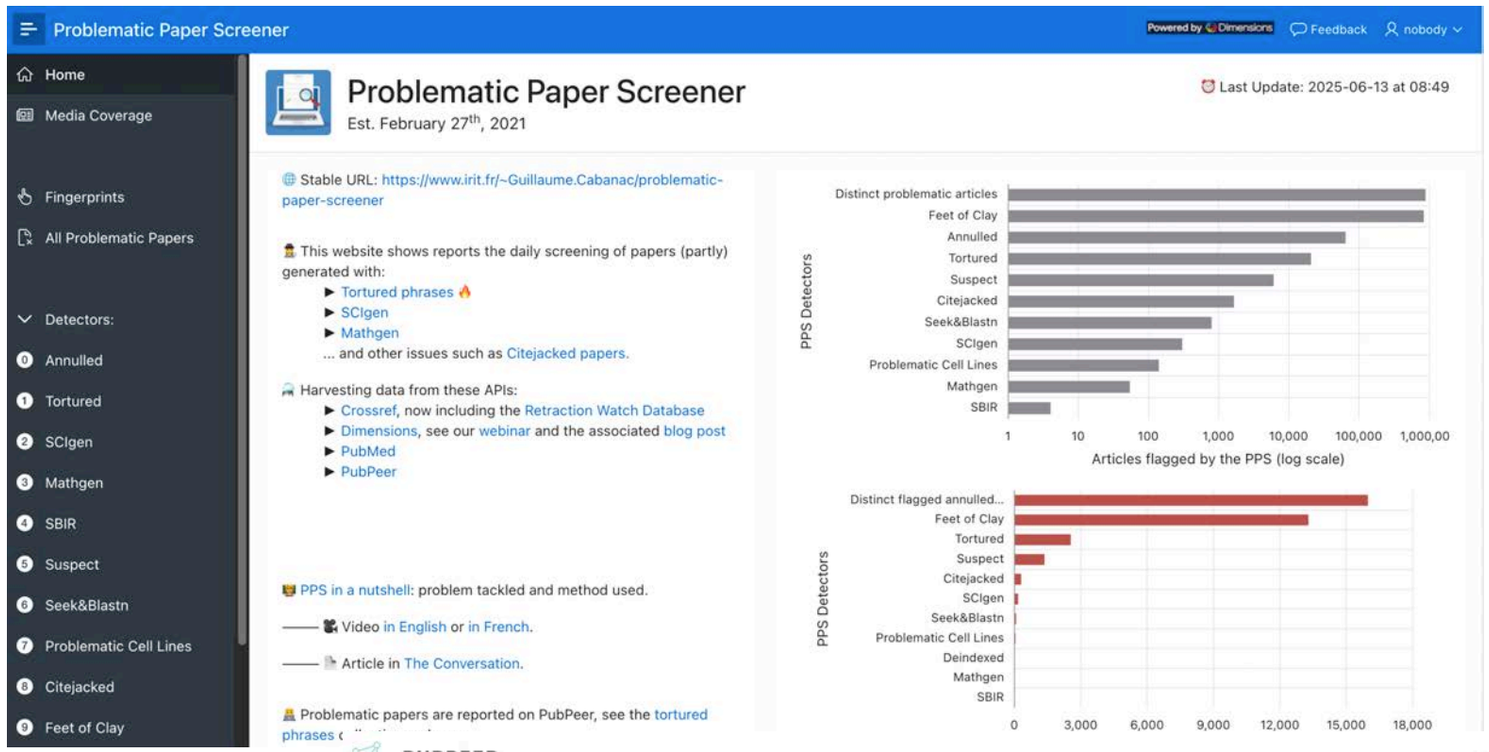
Papiers en cours d'évaluation (n = 36 307)

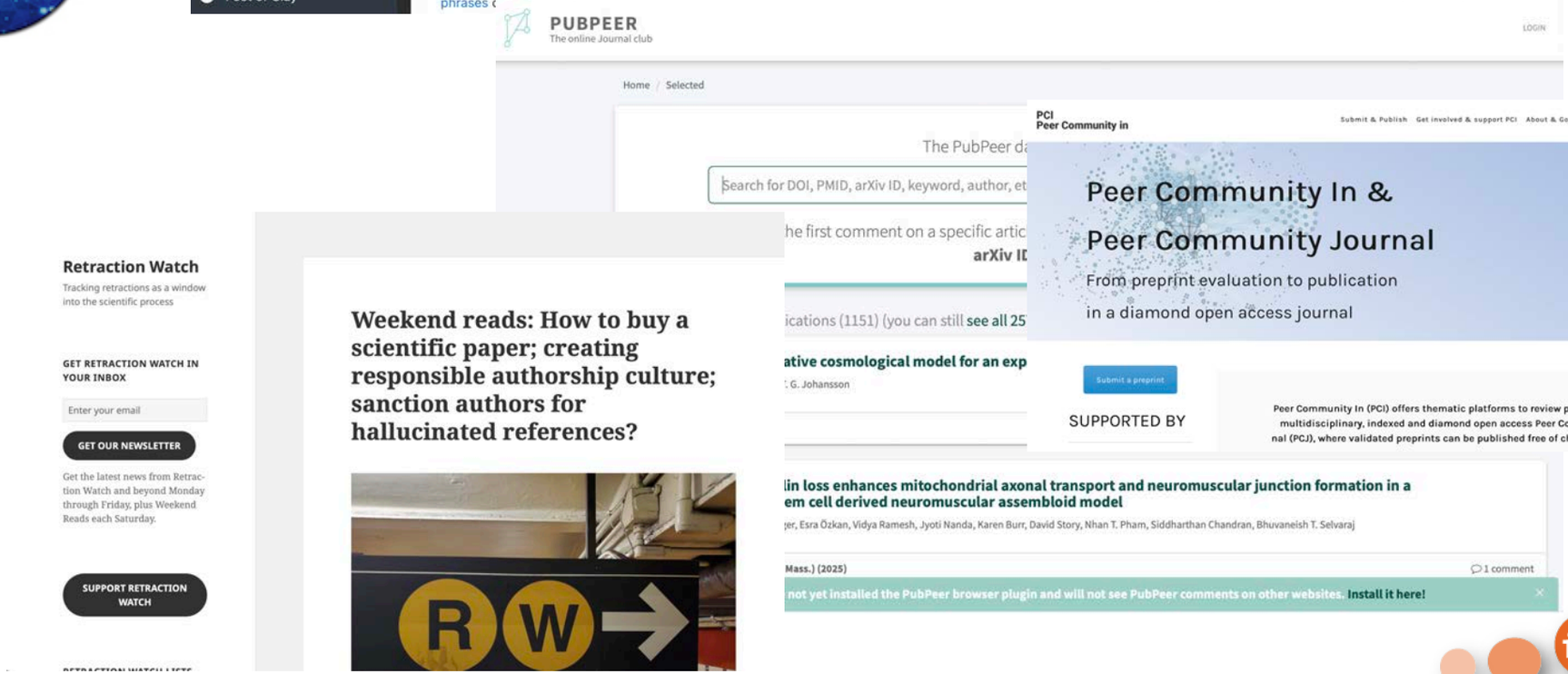




Outil de reporting pour la plateforme

Conclusion


Problematic Paper Screener
Powered by Dimensions
Feedback
nobody
Home
Media Coverage
Fingerprints
All Problematic Papers
Detectors:
0 Annulled
1 Tortured
2 SCIgen
3 Mathgen
4 SBIR
5 Suspect
6 Seek&Blastn
7 Problematic Cell Lines
8 Citejacked
9 Feet of Clay
Problematic Paper Screener
Est. February 27th, 2021
Last Update: 2025-06-13 at 08:49
Stable URL: https://www.irit.fr/~Guillaume.Cabanac/problematic-paper-screener
This website shows reports the daily screening of papers (partly) generated with:
Tortured phrases
SCIgen
Mathgen
... and other issues such as Citejacked papers.
Harvesting data from these APIs:
Crossref, now including the Retraction Watch Database
Dimensions, see our webinar and the associated blog post
PubMed
PubPeer
PPS in a nutshell: problem tackled and method used.
Video in English or in French.
Article in The Conversation.
Problematic papers are reported on PubPeer, see the tortured phrases
PPS Detectors
Distinct problematic articles
Feet of Clay
Annulled
Tortured
Suspect
Citejacked
Seek&Blastn
SCIgen
Problematic Cell Lines
Mathgen
SBIR
1
10
100
1,000
10,000
100,000
1,000,00
Articles flagged by the PPS (log scale)
Distinct flagged annulled...
Feet of Clay
Tortured
Suspect
Citejacked
SCIgen
Seek&Blastn
Problematic Cell Lines
Deindexed
Mathgen
SBIR
0
3,000
6,000
9,000
12,000
15,000
18,000


PUBPEER
The online Journal club
Home / Selected
Search for DOI, PMID, arXiv ID, keyword, author, et
Retraction Watch
Tracking retractions as a window into the scientific process
GET RETRACTION WATCH IN YOUR INBOX
Enter your email
GET OUR NEWSLETTER
Get the latest news from Retraction Watch and beyond Monday through Friday, plus Weekend Reads each Saturday.
SUPPORT RETRACTION WATCH
Weekend reads: How to buy a scientific paper; creating responsible authorship culture; sanction authors for hallucinated references?
R W
PCI
Peer Community in
Submit & Publish
Get involved & support PCI
About & Governance
Peer Community In &
Peer Community Journal
From preprint evaluation to publication
in a diamond open access journal
Submit a preprint
SUPPORTED BY
Peer Community In (PCI) offers thematic platforms to review preprints and the multidisciplinary, indexed and diamond open access Peer Community Journal (PCJ), where validated preprints can be published free of charge for authors
1 comment
not yet installed the PubPeer browser plugin and will not see PubPeer comments on other websites. Install it here!

Une « externalisation » du contrôle social de la science?

Ce que l'on peut faire:

> « *We propose that scholarly publishing needs to be **re-communalized**. Universities, libraries, funders and other members of the academic community need to build a **system that is community-led and managed**, and which works to further research and education. This should involve active support for federated open infrastructures (such as LA Referencia), and investment in community-based publication platforms and non-commercial journals.* » Crosetto & al., *The Drain of Scientific Publishing*, 2025

Équiper et animer une re-communalisation des produits de la recherche :

- Une réflexion à mener sur les produits de recherche comme « communs » (Ostrom, Hesse),
- (Re)définir les communautés d'ayant-droits = accords couperin
- Un commun = des espaces encastrés de réciprocités = des structures fédératives, construites par le bas = les SCD et structures IST, PubPeer, PCI, Unjournal,….
- Des modalités de gestion et de gouvernance polycentriques et coopératives,

Bibliographie

Ben-Yehuda, N., 1986a. Deviance in Science - Towards the Criminology of Science. The British Journal of Criminology 26, 1–27. https://doi.org/10.1093/oxfordjournals.bjc.a047577

Cabanac, Guillaume, 'Chain Retraction: How to Stop Bad Science Propagating through the Literature', *Nature*, 632.8027 (2024), pp. 977–79, doi:10.1038/d41586-024-02747-1

Cabanac, Guillaume, Cyril Labbé, and Alexander Magazinov, 'Tortured Phrases: A Dubious Writing Style Emerging in Science. Evidence of Critical Issues Affecting Established Journals' (arXiv, 12 July 2021), doi:10.48550/arXiv.2107.06751

Clausse, Alexandre, and others, 'Mining Tortured Abbreviations from the Scientific Literature', October 2023 <https://hal.science/hal-04311600>

Clausse, Alexandre, and Fidan Badalova, 'Post Publication Peer Review to (Re)Assess Scientific Articles Using the Problematic Paper Screener and PubPeer', February 2025 <https://hal.science/hal-04892301>

Conroy, Gemma, 'Scientific Sleuths Spot Dishonest ChatGPT Use in Papers', *Nature*, 2023, doi:10.1038/d41586-023-02477-w

December 2021, 15, 'Nature's 10' <https://www.nature.com/immersive/d41586-021-03621-0/index.html> [accessed 25 June 2025]

DUBOIS, Michel, 'L'intégrité scientifique et l'éthique de la recherche 2024: une enquête par questionnaire auprès du personnel de l'Inserm', *GEMASS*, 3 March 2024 <https://www.gemass.fr/enquete-inserm/>

Fanelli, Daniele, 'How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data', ed. by Tom Tregenza, PLoS ONE, 4.5 (2009), p. e5738, doi:10.1371/journal.pone.0005738

Fraisier-Vannier, Ophélie, and Guillaume Cabanac, '(Més)Usages Des LLM Pour La Production Scientifique' (unpublished Doctoral, Institut de Physique du Globe de Paris, France, April 2025) <https://hal.science/hal-04840973>

Haven, Tamarinde, and René van Woudenberg, 'Explanations of Research Misconduct, and How They Hang Together', Journal for General Philosophy of Science, 52.4 (2021), pp. 543–61, doi:10.1007/s10838-021-09555-5

'Https://Arxiv.Org/Pdf/2107.06751v1' <https://arxiv.org/pdf/2107.06751v1> [accessed 25 June 2025]

Ibrahim, Hazem, and others, 'Citation Manipulation through Citation Mills and Pre-Print Servers', Scientific Reports, 15.1 (2025), p. 5480, doi:10.1038/s41598-025-88709-7

Labbé, Cyril, 'ONE OF THE GREAT STARS IN THE SCIENTIFIC FIRMAMENT', . . NR.

Lay, Puthineath, Martin Lentschat, and Cyril Labbé, 'Investigating the Detection of Tortured Phrases in Scientific Literature' (arXiv, 24 October 2022), doi:10.48550/arXiv.2210.13024

Martel, Eléna, Martin Lentschat, and Cyril Labbé, 'Detection of Tortured Phrases in Scientific Literature' (arXiv, 2 February 2024), doi:10.48550/arXiv.2402.03370

Bibliographie

Orduña-Malea, Enrique, and Álvaro Cabezas-Clavijo, 'ChatGPT and the Potential Growing of Ghost Bibliographic References', *Scientometrics*, 128.9 (2023), pp. 5351–55, doi:10.1007/s11192-023-04804-4

Orrall, Avery, 'Sleuths Spur Cleanup at Journal with Nearly 140 Retractions and Counting', *Retraction Watch*, 22 August 2024 <https://retractionwatch.com/2024/08/22/sleuths-spur-cleanup-at-journal-with-nearly-140-retractions-and-counting/>

Pourret, Olivier, 'On the Emergence of Tortured Phrases: A Threat to Scientific Integrity—the Example of “Heavy Metal&rdquo';, *European Science Editing*, 50 (2024), p. e131771, doi:10.3897/ese.2024.e131771

Singh Chawla, Dalmeet, 'The Citation Black Market: Schemes Selling Fake References Alarm Scientists', *Nature*, 632.8027 (2024), pp. 966–966, doi:10.1038/d41586-024-01672-7

Sovacool, Benjamin K., 'Exploring Scientific Misconduct: Isolated Individuals, Impure Institutions, or an Inevitable Idiom of Modern Science?', *Journal of Bioethical Inquiry*, 5.4 (2008), pp. 271–82, doi:10.1007/s11673-008-9113-6

Travis, Kate, 'A New Journal Record: Sage Title Retracts 678 More Papers, Tally over 1,500', *Retraction Watch*, 17 April 2025 <https://retractionwatch.com/2025/04/17/sage-journal-intelligent-fuzzy-systems-retracts-678-more-papers/>

Van Noorden, Richard, and Miryam Naddaf, 'Exclusive: The Papers That Most Heavily Cite Retracted Studies', *Nature*, 633.8028 (2024), pp. 13–15, doi:10.1038/d41586-024-02719-5

Wilhite, Allen W., and Eric A. Fong, 'Coercive Citation in Academic Publishing', *Science*, 335.6068 (2012), pp. 542–43, doi:10.1126/science.1212540

Zitt, Michel, and Elise Bassecoulard, 'Delineating Complex Scientific Fields by an Hybrid Lexical-Citation Method: An Application to Nanosciences', *Information Processing & Management*, Special Issue on Informetrics, 42.6 (2006), pp. 1513–31, doi:10.1016/j.ipm.2006.03.016

Zuckerman, H., 1977. Deviant Behavior and Social Control in Science », in: Sagarin, E. (Ed.), Deviance and Social Change, Sage, Londres. pp. 87–138.