LE PROJET FULLAB

LISIS (UPEM, INRA, ESIEE, CNRS)

& DIRECTION DE LA DOCUMENTATION (**ECOLE DES PONTS**)

& UNIVERSITÉ DE LORRAINE

SÉMINAIRE ISTEX – JUIN 2017

NANCY

FRÉDÉRIQUE BORDIGNON

LIANA ERMAKOVA

MARIANNE NOËL

NICOLAS TURENNE

OBJECTIFS

AMBITIEUX MAIS ATTEIGNABLES GRÂCE AUX DONNÉES D'ISTEX

OBJET D'ÉTUDE : L'ARTICLE **ET** SON ABSTRACT

- Contexte : évolution des modalités de publication
- Objet d'étude : l'article et son abstract
- Domaine : sciences environnementales

ABSTRACT : SIMPLE *TEASER* ?

- Comparer la quantité d'informations de l'abstract avec celle de l'article qu'il résume en utilisant la structure du full-text (FT)

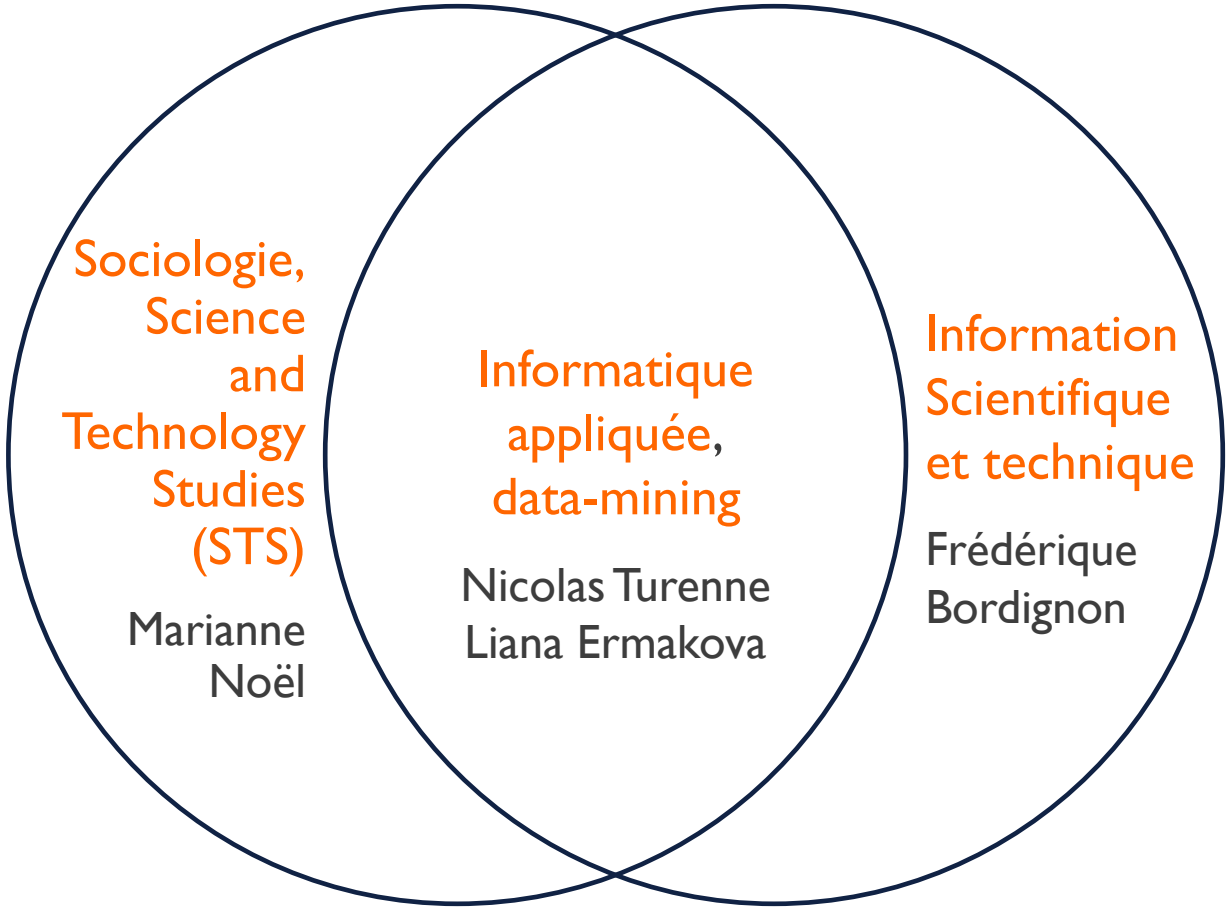
- Pour un calcul du **taux de générosité**

→ **Des perspectives d'étude**

EXPLOITER LE TAUX DE GÉNÉROSITÉ

- Selon la discipline : différences de comportement des chercheurs ? ✓
- Selon l'exigence des revues ?
- Selon le mode de diffusion : en Open Access ou pas ?
- Corrélation Taux de générosité/nombre de citations ? ✓
- Prédominance de certains types d'entités nommées ?
- Existence d'une éventuelle évolution au fil du temps ? ✓

L'OPPORTUNITÉ D'UN TRAVAIL INTERDISCIPLINAIRE



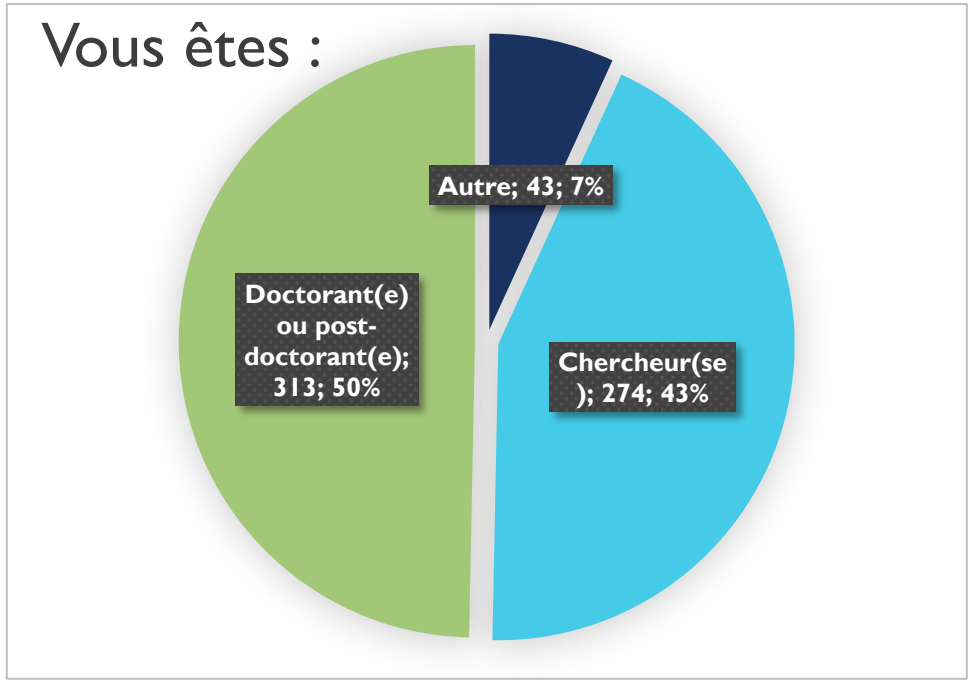
Un noyau informatique, des productions connexes

ANALYSE DE LA LITTÉRATURE : SEGMENTATION EN 7 PARTIES DU FULL-TEXT

- **Introduction - Contexte**
  Cette partie décrit ce qui est déjà connu du sujet d'une façon compréhensible par les chercheurs de tous les secteurs.
- **Objectif**
  Il s'agit de décrire ici ce qui n'est pas encore connu mais qui pourra être comblé par la recherche ou le raisonnement développé dans l'article.
- **Méthode – Design**
  Cette section informe le lecteur des techniques et stratégies mises en œuvre pour mener la recherche et prouver sa validité. (exemples : matériel utilisé, cadre méthodologiques retenu, population étudiée, process de collecte des données, taille de l'échantillon, etc.)
- **Résultats - Observations - Constats**
  Les principaux résultats sont présentés ici et sont accompagnés des données (éventuellement chiffrées) qui ont permis de les caractériser. Il peut aussi s'agir de résultats négatifs qui ne confortent pas l'hypothèse de départ.
- **Conclusions**
  Cette partie reprend le message principal de l'article. Elle montre comment sont interprétés les résultats et comment il est ainsi répondu à la question initiale.
- **Limites**
  Si des limites à l'étude ont été identifiées, elles sont présentées ici.
- **Perspectives**
  Il s'agit ici de positionner les résultats de l'étude dans un contexte plus général afin de montrer en quoi le papier est un progrès et comment d'autres études pourraient constituer de nouvelles avancées.

SONDAGE (1)

630 répondants



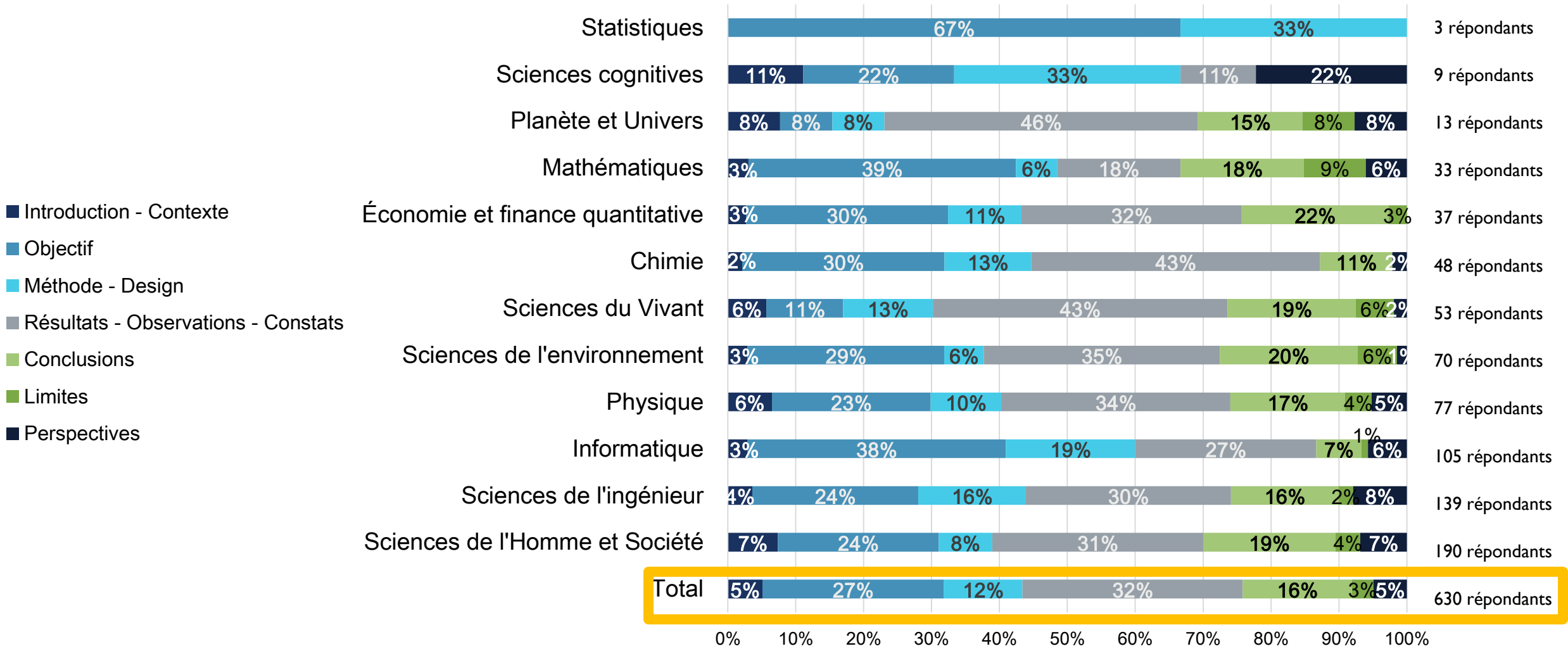
Dans quel(s) champ(s) disciplinaire(s) s'inscrivent vos recherches ? (2 max)

Sciences de l'Homme et Société	190
Sciences de l'ingénieur	139
Informatique	105
Physique	77
Sciences de l'environnement	70
Sciences du Vivant	53
Chimie	48
Économie et finance quantitative	37
Mathématiques	33
Planète et Univers	13
Sciences cognitives	9
Statistiques	3

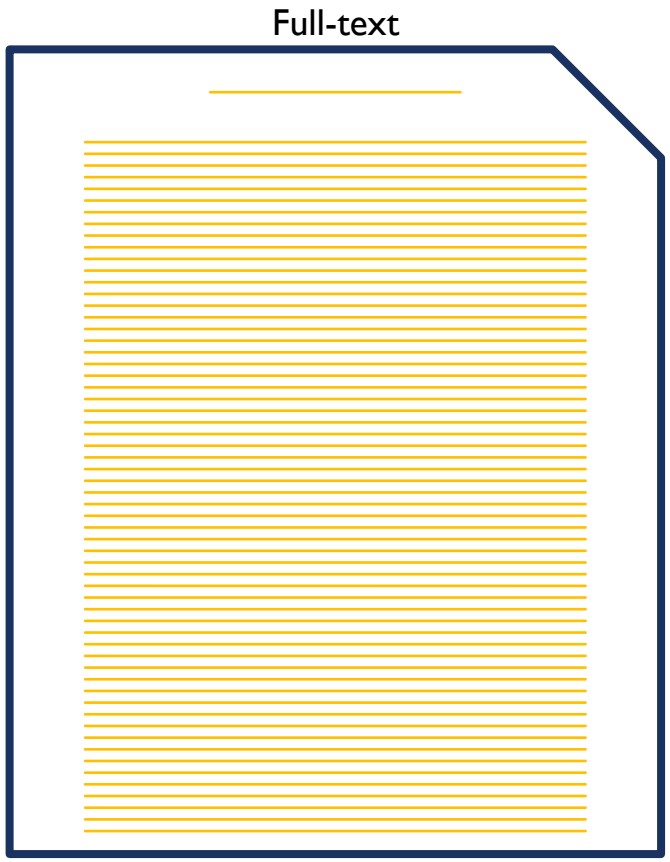
Nombre de répondants par discipline

SONDAGE (2)

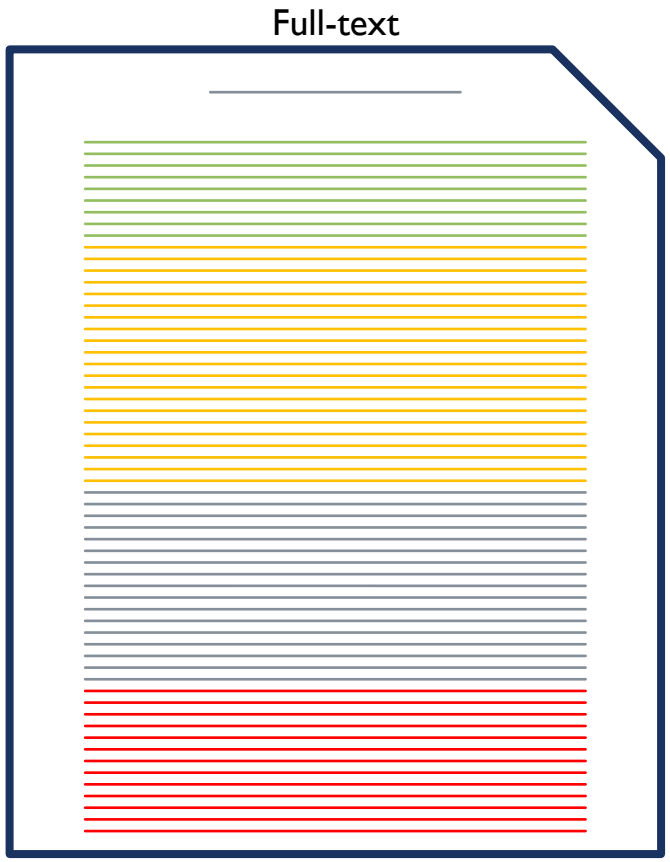
Selon vous, quelle section doit impérativement être présente dans l'abstract pour qu'il puisse être qualifié de "généreux" :



TAUX DE GÉNÉROSITÉ



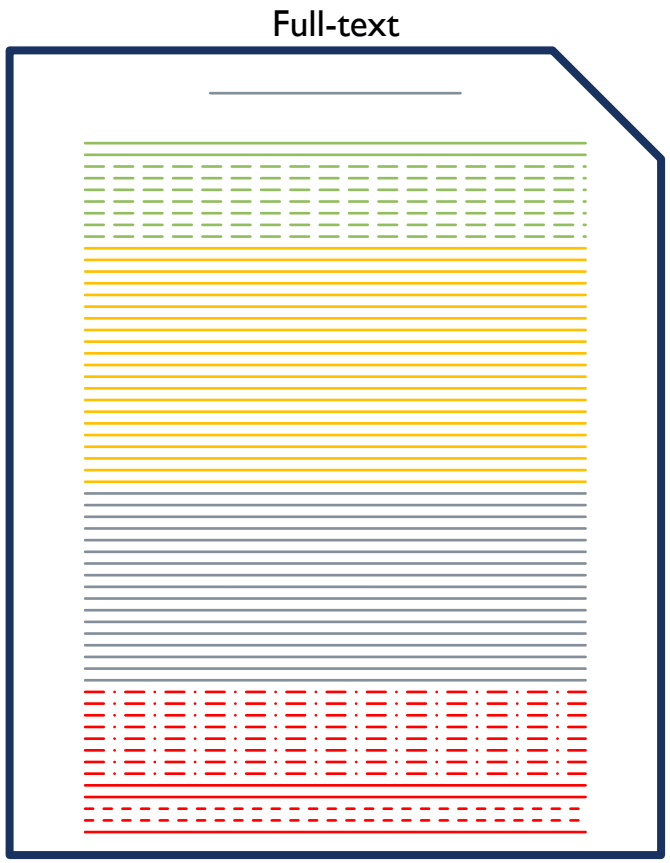
TAUX DE GÉNÉROSITÉ I



Recherche des sections :

- Introduction – Contexte
- Méthode – Design
- Résultats – Observations – Constats
- Conclusions

TAUX DE GÉNÉROSITÉ 2

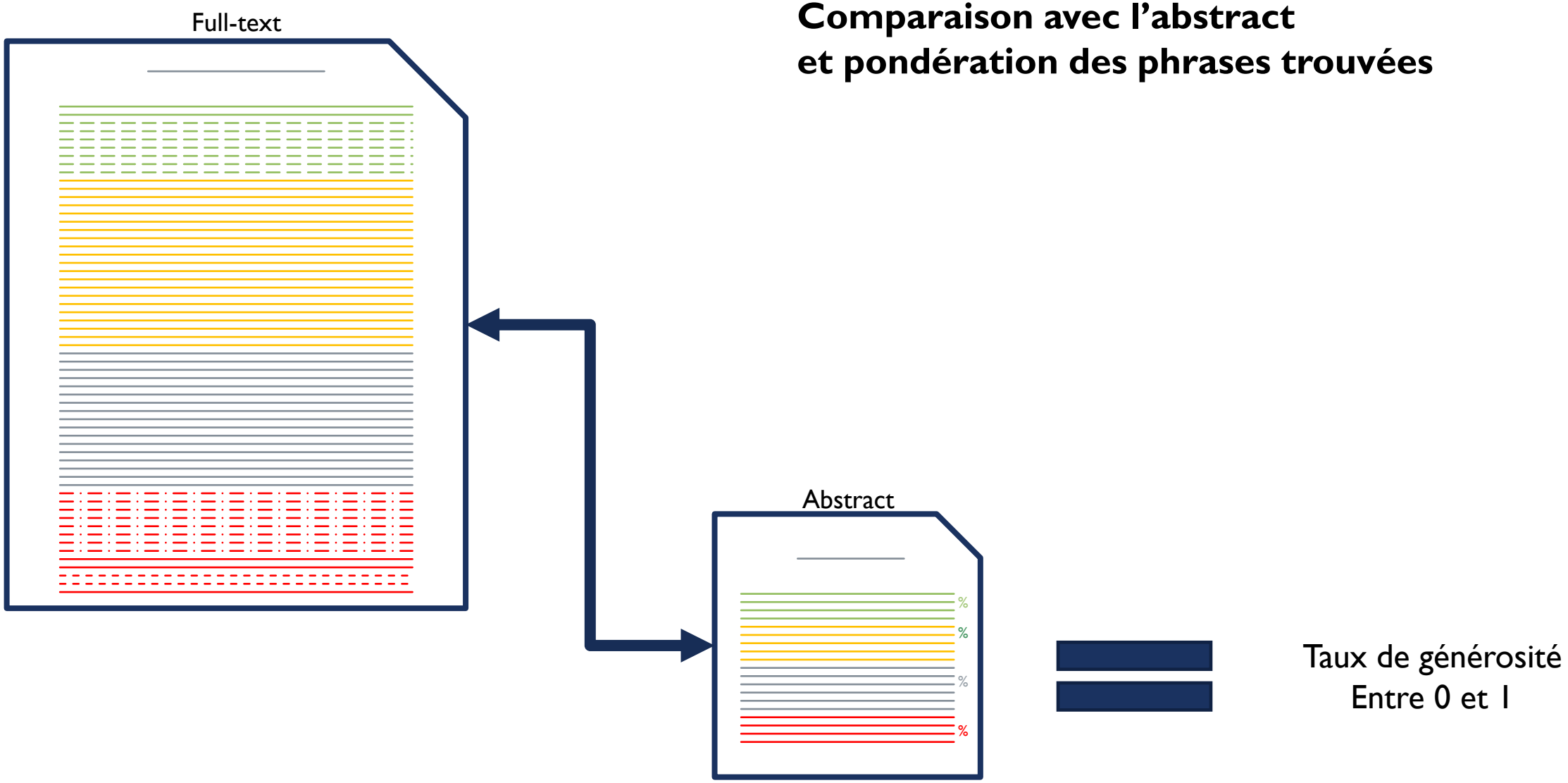


Recherche des sections :

- **Objectif dans Introduction – Contexte**
- **Limites dans Conclusions**
- **Perspectives dans Conclusions**

TAUX DE GÉNÉROSITÉ 3

Comparaison avec l'abstract
et pondération des phrases trouvées



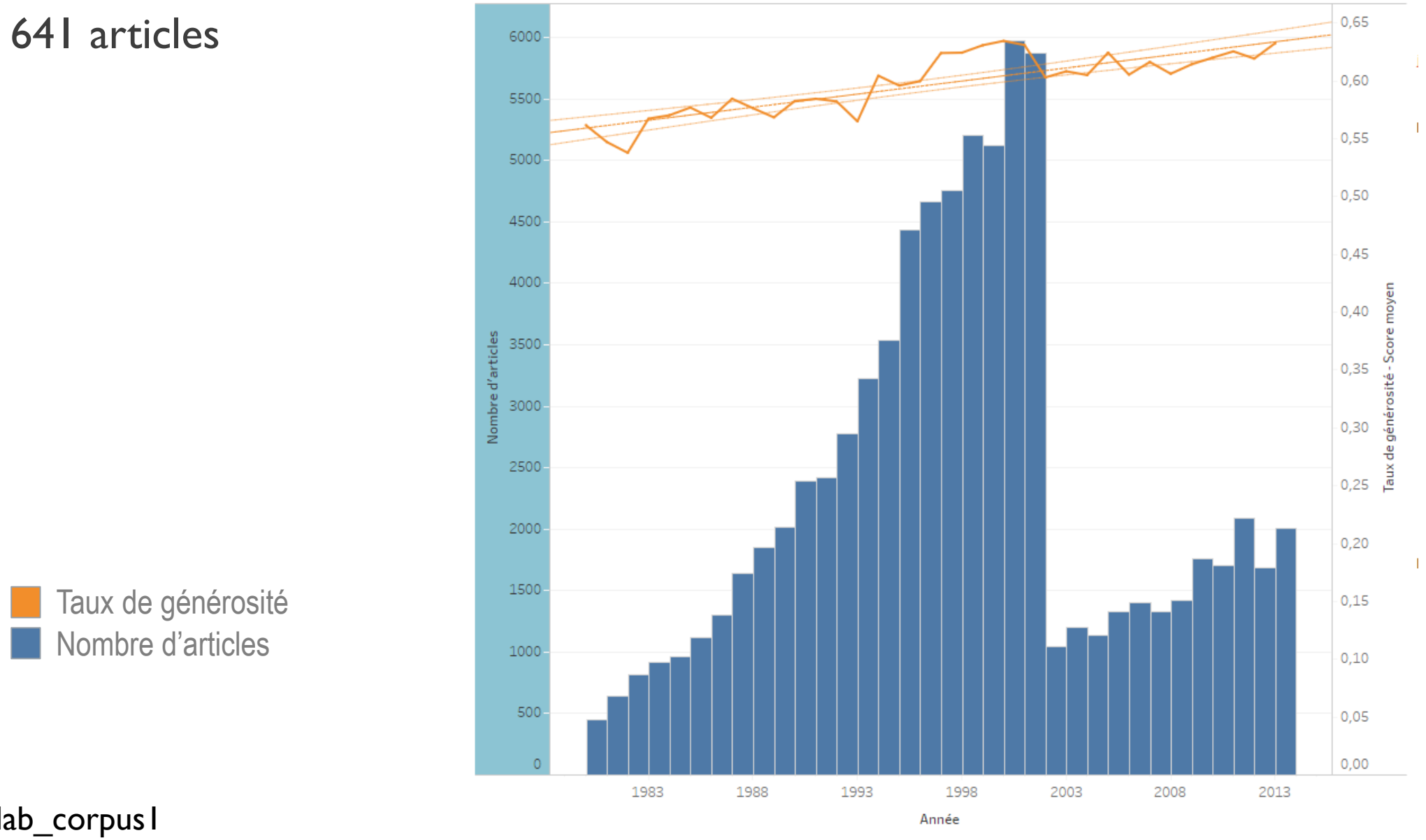
Taux de générosité
Entre 0 et 1

PREMIÈRE APPLICATION : ANALYSE DU CORPUS DE SCIENCES ENVIRONNEMENTALES

DESCRIPTION DU CORPUS

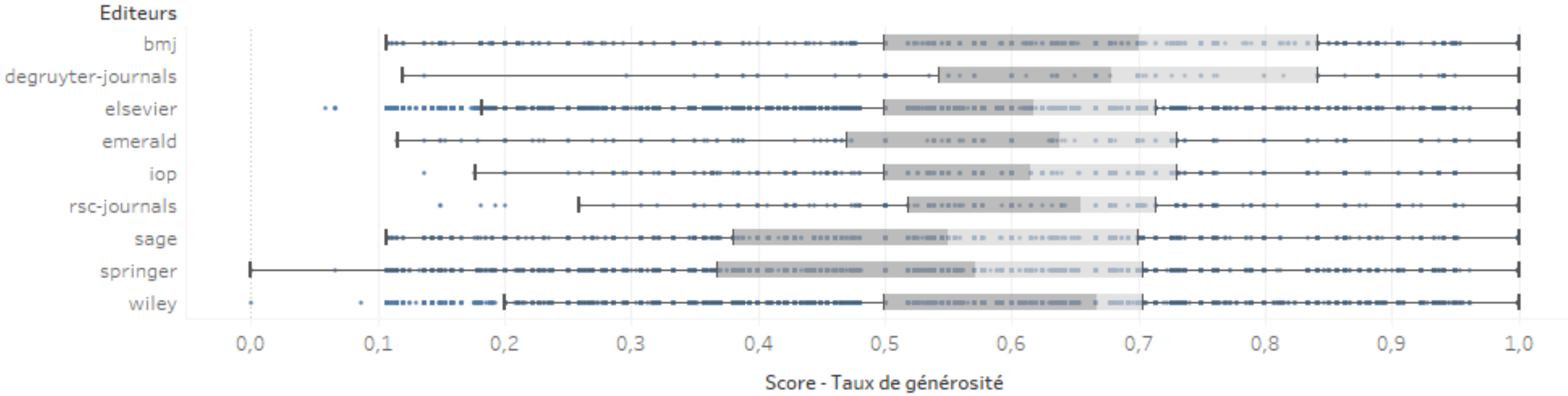
EVOLUTION DU TAUX DE GÉNÉROSITÉ AU FIL DU TEMPS

n = 35 641 articles



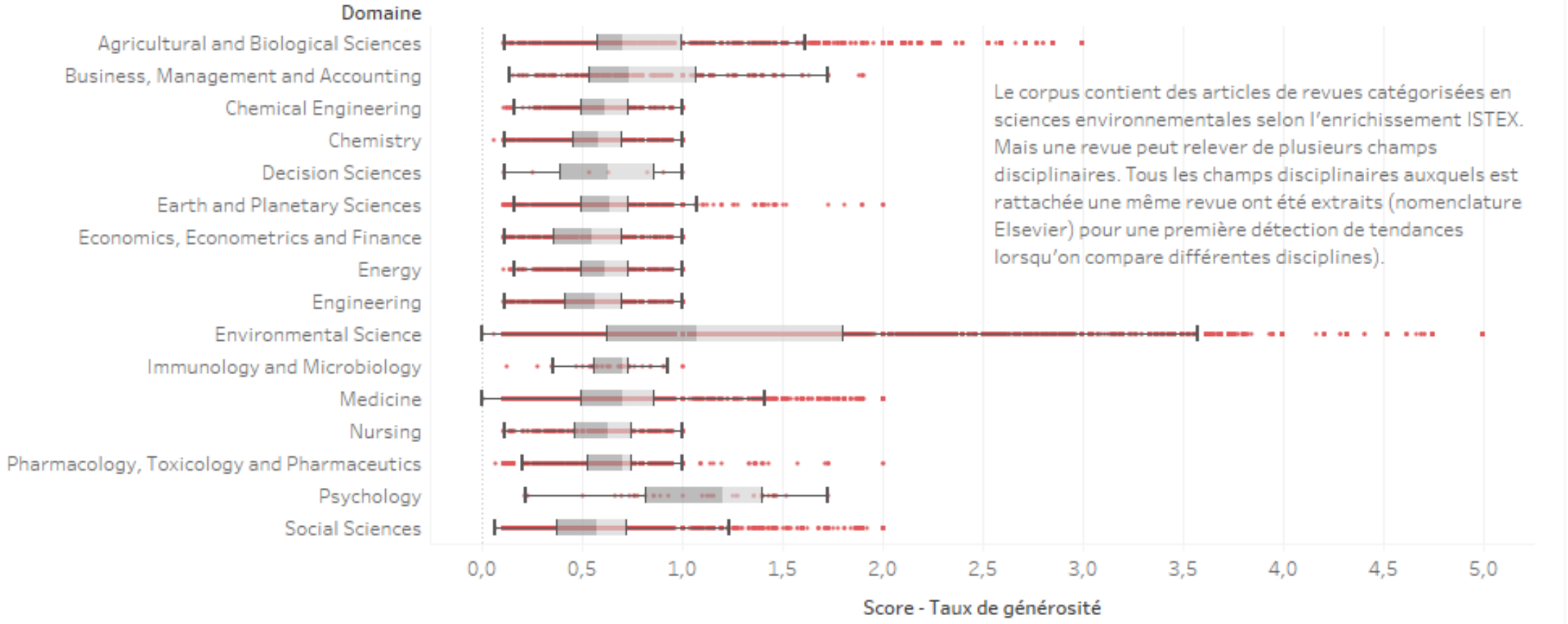
http://bit.ly/Fullab_corpus1

TAUX DE GÉNÉROSITÉ PAR ÉDITEUR



http://bit.ly/Fullab_corpus1

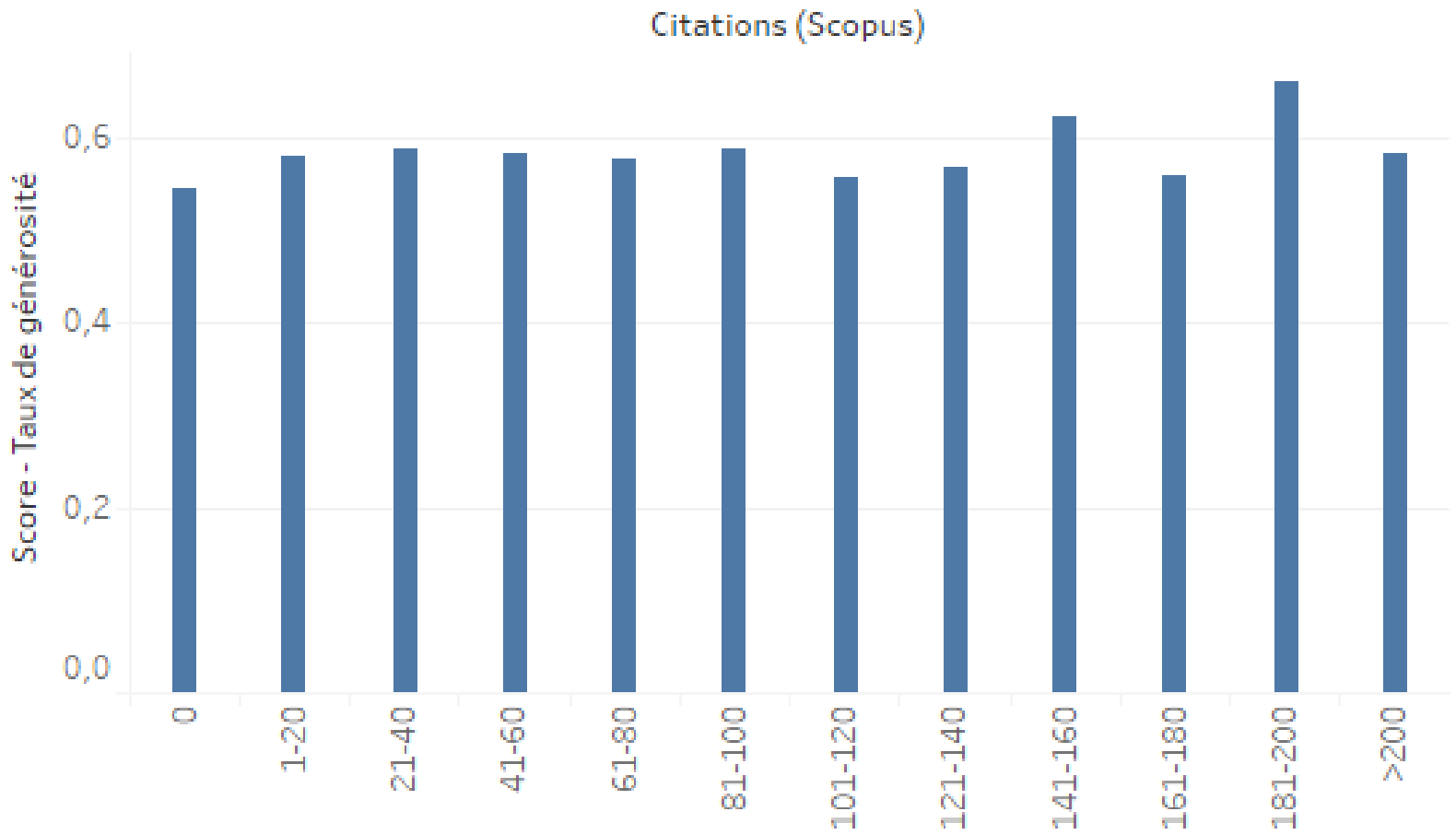
TAUX DE GÉNÉROSITÉ PAR DISCIPLINE



http://bit.ly/Fullab_corpus1

TAUX DE GÉNÉROSITÉ ET NOMBRE DE CITATIONS (SCOPUS)

n = 7745 articles



http://bit.ly/Fullab_corpus1

CONCLUSIONS ET PERSPECTIVES

CONCLUSION

- Du point de vue informatique
  - GEM est une métrique interprétable contrairement aux mesures d'informativité existantes (ROUGE, INEX, Pyramide ...)
  - GEM donne un score absolu contrairement aux mesures d'informativité existantes
- Du point de vue de l'exploitation d'ISTEX
  - ISTEX = limites (que sur des PDFs avec texte ; 95792 docs perdus)
  - Les statistiques sont parfois erronées (nombre de mots…)
  - Les fichiers TEI ne correspondent pas aux XML (découpage en sections)
  - Les problèmes fréquents avec l'API (extraction de fichiers, perte de résultats, stabilité)

PERSPECTIVES DE TRAVAIL

- Évaluation à la grande échelle à faire
- Comparer des corpus de domaines différents pour voir si on peut améliorer le calcul du taux de générosité (ex : noms de techniques)
- Intégrer GEM dans ISTEX
- Finir les interfaces web et desktop

PERSPECTIVES D'APPLICATIONS

- Pour le professionnel de l'IST :
  - Meilleure connaissance de la littérature scientifique
  - Meilleure connaissance de ses modalités de diffusion
  - Progrès dans la méthodologie de recherche
  - Meilleure stratégie économique (choix des abonnements)

- Pour le chercheur :
  - Clés pour la rédaction efficace d'un abstract
  - Génération automatique

- Pour les pays en voie de développement :

  Des clés pour décider s'il est raisonnable ou non de se passer du full-text

MERCI DE VOTRE ATTENTION

Frédérique Bordignon

Liana Ermakova

Marianne Noël

Nicolas Turenne