SDU
University of Southern Denmark

OpenAlex à l'épreuve de la recherche scientifique

Maxime Holmberg Sainte-Marie
Syddansk Universitet, Danemark

Maxime Holmberg Sainte-Marie
Journée Francophone OpenAlex 2025

SDU
University of Southern Denmark

OpenAlex à l'épreuve de la recherche scientifique:

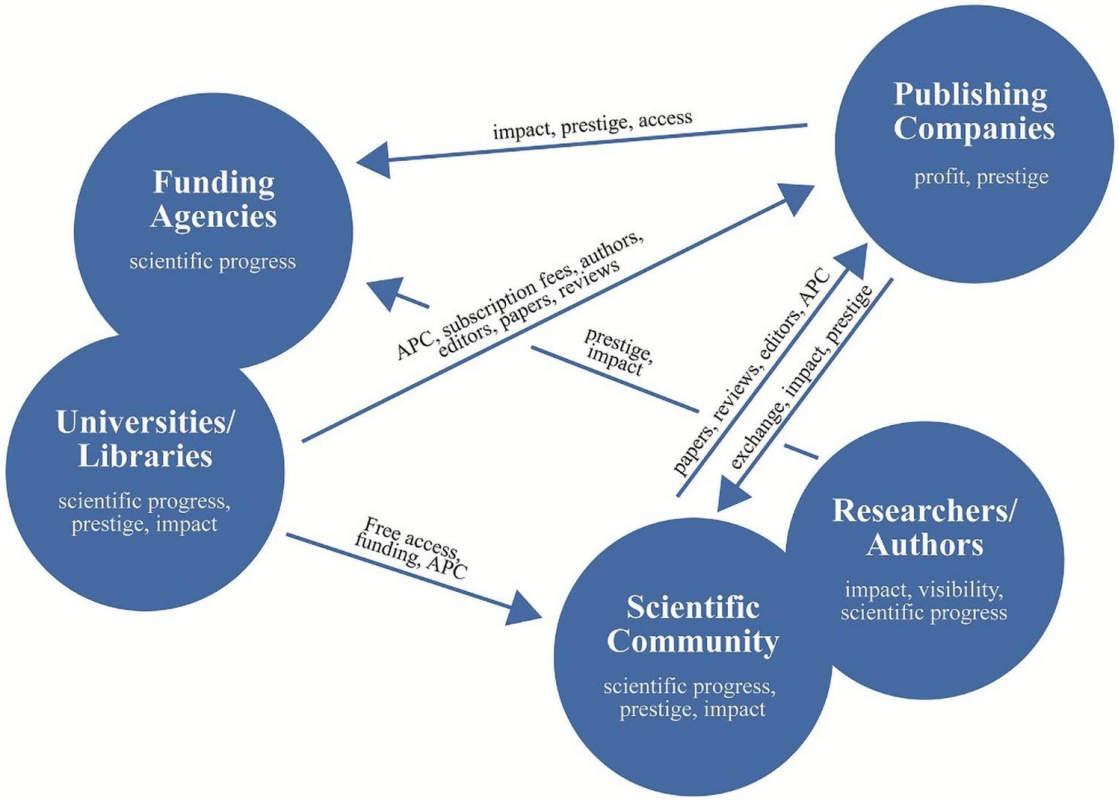
Le capitalisme savant

SDU
University of Southern Denmark

Globalisation, numérisation et publication savante

Attentes	Réalité	Source
Disparition des journaux savants	Prolifération des journaux (savants et prédateurs)	Hanson et al., 2023; Laakso&Matthias, 2020
Effondrement des hiérarchies basées sur le prestige	Oligarchies verticalement intégrées, propulsées par la mondialisation et la quantification	Hanson et al., 2023; Larivière et al., 2015
Libre circulation des connaissances	Les verrous d'accès payants (*paywall*) et les frais de traitement des articles (*APCs*) monétisent à la fois la diffusion et l'accès	Laakso&Matthias, 2020
Ouverture de l'évaluation de la recherche	Données bibliométriques privatisées et monétisées	Wilsdon et al., 2015

Un aperçu des dégâts



The Political Economy of Academic Publishing **(Puehringer et al., 2025)**

→ Estimation des fonds publics autrichiens versés aux maisons d'édition savantes

→ 4 canaux d'accès principaux aux fonds publics

1. Frais d'abonnement: **5-12 M€**
2. Frais de publications et de soumission: **1,25 – 1,5 M€**
3. Travail de révision et d'édition: **3,3 - 4,9 M€**
4. Production de connaissances: **57-85 M€**

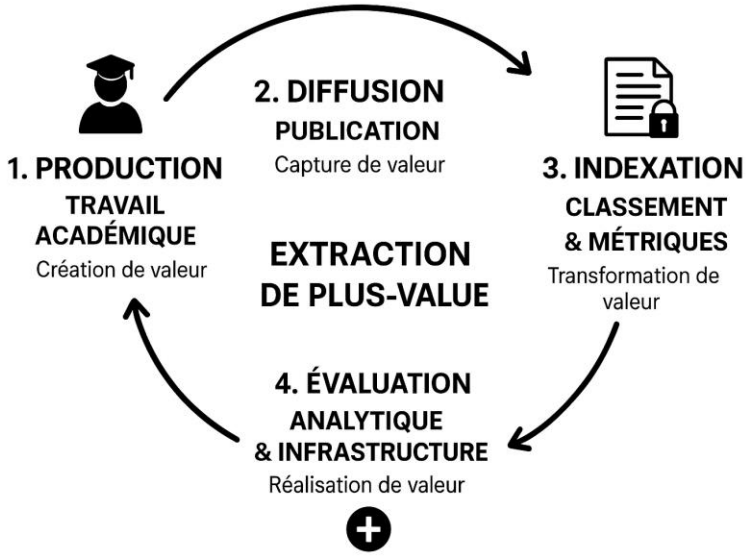
→ **66,55 - 103,2 M€** octroyé par le gouvernement autrichien aux maisons d'édition savantes

→ **~25%** du financement public alloué au champ

Maxime Holmberg Sainte-Marie
Journée Francophone OpenAlex 2025

SDU
University of Southern Denmark

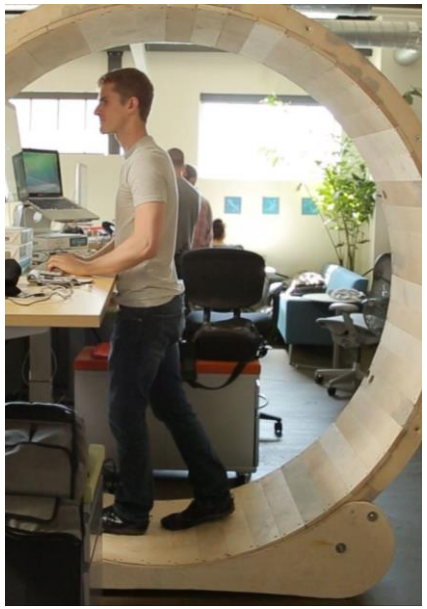
La capitalisation du cycle de publication



1. **Intrant: Travail savant (création de valeur d'usage)**
   - → Acteurs: chercheurs, réviseurs, éditeurs
   - → Acitivités: recherche, rédaction, évaluation
   - → Contexte institutionnel: financé par les université et agences publiques
   - → Extrant: Connaissances et données → matière première pour maisons d'édition savantes
2. **Transformation: Édition savante (Captation de valeur)**
   - → Acteurs: Maisons d'édition (Elsevier, Springer Nature, …)
   - → Activités: Contrôle d'édition, production, marquage, verrous d'accès payant
   - → Extrant: Marchandisation des articles (abonnements, frais de publication)
   - → Type de valeur: valeur d'échange (marchandisation des connaissances)
3. **Circulation: Indexation et métriques (marchandisation des données)**
   - → Acteur: Clarivate (Web of Science), Elsevier (Scopus), Digital Science (Dimensions)…
   - → Activités: Citation tracking, journal ranking, data analytics
   - → Extrant: Capital symbolique (impact, prestige) converti en produits et services de données commercialisables
   - → Type de valeur: Symbolique → capital données (plus-value bibliométrique)
4. **Realization: Evaluation & Infrastructure (capitalisme de plateforme (Srnicek, 2016)**
   - → Acteurs: Elsevier (Pure…), Clarivate (Converis…)...
   - → Activités: services d'analyse, tableaux de bord, outils de gestion de la recherche
   - → Extrant: les institutions de recherche paient pour accéder aux données générées à l'internet à des fins d'évaluation interne, conformisation aux classements universitaires
   - → Type de valeur: capital informationnel (extraction continue)
5. **Réinvestissement: Dépendance académique**
   - → Les mêmes institutions et chercheurs font l'objet de multiples pressions:
     - → Publier davantage dans les journaux savants sous la possession des oligarques
     - → Utilisation des produits et services d'évaluation de la recherche produits par les oligarques (validation du système)

SDU
University of Southern Denmark

La prolétarisation des chercheurs



- → **Ils génèrent de la valeur, mais ne la possèdent pas**
  - → Researchers generate knowledge outputs (articles, datasets, citations) that are the foundation of scholarly publishing.
  - → Publishers own the copyrights, platforms, and data infrastructures that turn those outputs into profitable commodities.
  - → Researchers' intellectual labor becomes alienated
  - → They produce knowledge but have no control over its use, circulation, or monetization.
- → **Ils travaillent en contexte de dépendance croissante**
  - → Most researchers rely on public funding, precarious contracts, and performance metrics that measure productivity, not autonomy.
  - → Their employability and prestige depend on publishing in high-impact journals — which are owned by a handful of corporate publishers.
  - → This creates a relation of subordination: the academic must produce within a system whose rules are dictated by capital (impact factors, APCs, rankings).
- → **Ils sont assujettis à une surveillance et un contrôle informatiques**
  - → Platforms like Elsevier's Pure track and evaluate researchers' productivity.
  - → These systems convert academic labor into data commodities — which are then sold back to institutions.
  - → The academic's very identity and reputation become forms of data capital, quantified and traded within the metrics economy.
- → **Ils effectuent une bonne part de travail non rémunéré**
  - → Peer review, editorial work, and conference organization are essential to the publishing system, yet almost entirely unpaid.
  - → This "free labor" sustains the infrastructure of scholarly capitalism

Maxime Holmberg Sainte-Marie
Journée Francophone OpenAlex 2025

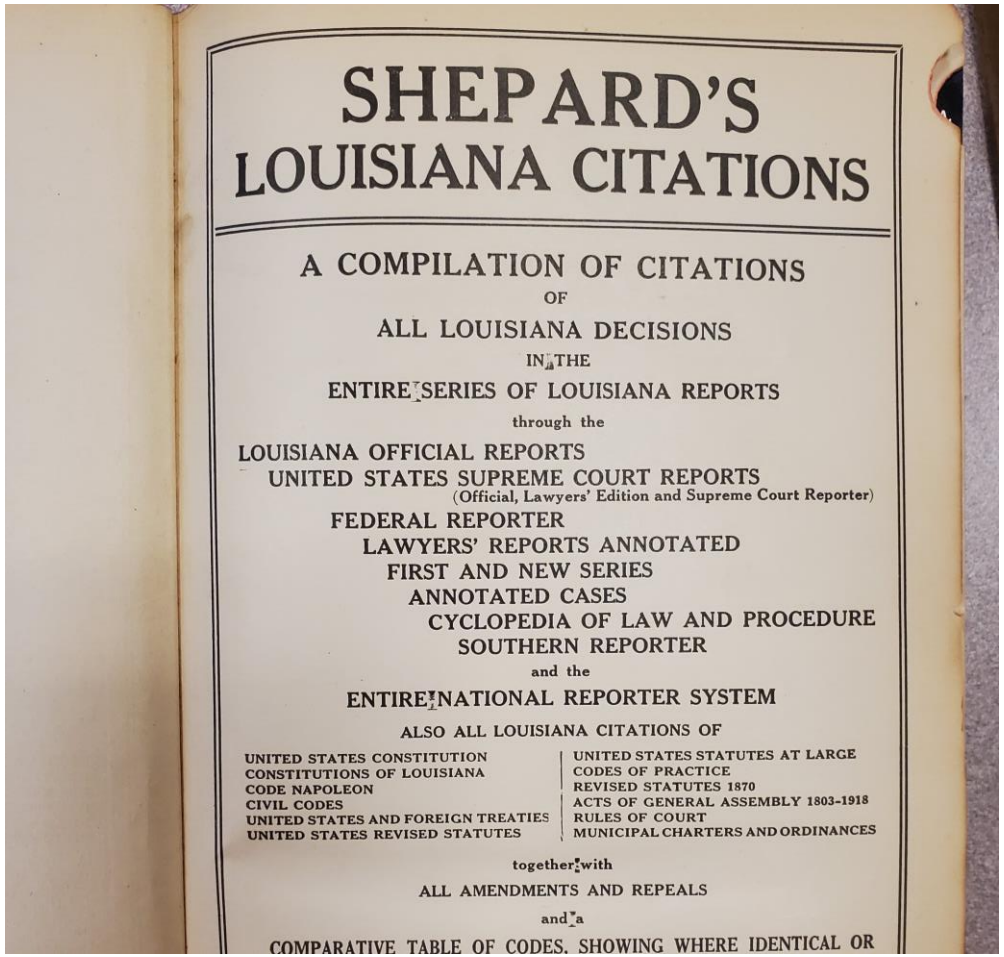
SDU
University of Southern Denmark

OpenAlex à l'épreuve de la recherche scientifique:

La part de la scientométrie

SDU
University of Southern Denmark

L'origine conceptuelle de la scientométrie



SHEPARD'S
LOUISIANA CITATIONS

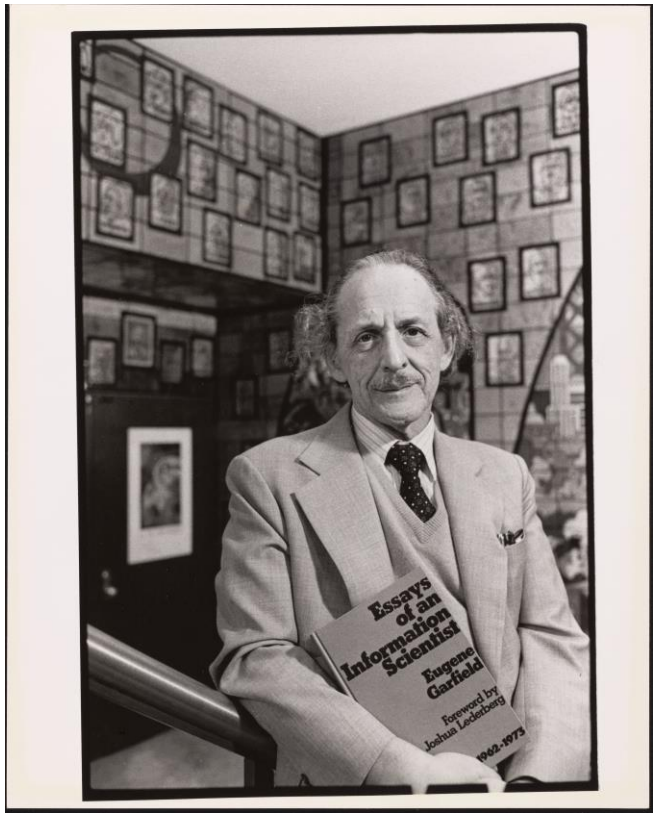
A COMPILATION OF CITATIONS
OF
ALL LOUISIANA DECISIONS
IN THE
ENTIRE SERIES OF LOUISIANA REPORTS
through the
LOUISIANA OFFICIAL REPORTS
UNITED STATES SUPREME COURT REPORTS
(Official, Lawyers' Edition and Supreme Court Reporter)
FEDERAL REPORTER
LAWYERS' REPORTS ANNOTATED
FIRST AND NEW SERIES
ANNOTATED CASES
CYCLOPEDIA OF LAW AND PROCEDURE
SOUTHERN REPORTER
and the
ENTIRE NATIONAL REPORTER SYSTEM
ALSO ALL LOUISIANA CITATIONS OF

UNITED STATES CONSTITUTION	UNITED STATES STATUTES AT LARGE
CONSTITUTIONS OF LOUISIANA	CODES OF PRACTICE
CODE NAPOLEON	REVISED STATUTES 1870
CIVIL CODES	ACTS OF GENERAL ASSEMBLY 1803-1918
UNITED STATES AND FOREIGN TREATIES	RULES OF COURT
UNITED STATES REVISED STATUTES	MUNICIPAL CHARTERS AND ORDINANCES

together with
ALL AMENDMENTS AND REPEALS
and a
COMPARATIVE TABLE OF CODES, SHOWING WHERE IDENTICAL OR

- → Shepard's Citation Service
  - → Originellement Shepard's Adhesive Annotations (1873)
  - → Fournit une liste de toutes les autorités citant une affaire, une loi ou une autre source juridique particulière.
  - → Permet de suivre l'évolution des décisions judiciaires, des lois et autres ressources juridiques invoquées à différentes époques et à diverses fins.
- → Shepardization : constuler Shepard's pour vérifier si une affaire a été infirmée, confirmée, remise en question ou citée par des décisions ultérieures.

Eugene Garfield (1925-2017)



- → "Citation Indexes for Science: A New Dimension in Documentation Through Association of Ideas" (Science, 1955)
  - → Proposition d'un nouvel outil de documentation basé sur les liens de citation
    - → En supposant que les références incluses dans une source pertinente soient elles-mêmes pertinentes, les chercheurs seraient mieux à même d'identifier les travaux liés à leurs propres recherches.
    - → Complément à la recherche bibliographique traditionnelle par titre ou auteur
    - → RIEN SUR L'ÉVALUATION DE LA RECHERCHE!
  - → L'idée de Garfield a reçu le soutien de plusieurs personnes et organisations.
    - → Une indexation-test est effectuée en génétique.
    - → Utilisation de fiches perforées et de bandes magnétiques.

Le *Science Citation Index* (1963)

- → Volumes de métadonnées d'articles provenant de 613 revues de sciences naturelles et médicales
  - → Métadonnées de publication usuelles (auteur, titre, année, etc.)
    - → Références
    - → Citations
  - → Critères de sélection des revues : Loi de dispersion de Bradford (1934)
    - → Distribution de Pareto : une minorité de revues reçoit la majorité des citations.
    - → L'objectif de Garfield était d'indexer le contenu des revues principales.
    - → Optimisation de la gestion des abonnements de la bibliothèque
    - → Conséquence: biais de couverture
      - → Sciences naturelles et médicalesCouverture accrue des sciences naturelles et médicales
      - → Variations disciplinaires dans la proportion d'Indexation des références d'articles
      - → Langue anglaise

De la recherche d'information à l'évaluation de la recherche

1972: Invention du facteur d'impact

→ Créé afin d'aider les bibliothèques au niveau de leur politique d'abonnement (Garfield, 1972)

→ Pas un outil d'évaluation de la recherche!

~1975: Premières utilisations évaluatives

→ Les décomptes de citation et le facteur d'impact peuvent servir d'outils quantitatifs d'évaluation de rendement en matière de recherche (Narin & Carpenter, 1975)

1975-1980: Institutionnalisation de l'évaluation de la recherche

→ Premières applications de l'analyse de citations à des fins de politique scientifique

→ Premiers Indicateurs nationaux de production scientifique (Narin, 1976)

→ Premières utilisation d'indicateurs basés sur le décompte des citations par les universités et organismes de financement (Garfield, 1979).

1980+: « Disciplinarisation » de la scientométrie

→ Nouvelle discipline portant sur l'analyse quantitative de la science

→ 1978: Lancement du journal Scientometrics

→ Le Science Citation Index devient la source de données principale pour l'analyse de la productivité en recherche, la conception et d'utilisation des indicateurs d'impact et la cartographie scientifique

SDU
University of Southern Denmark

Complexification des systèmes scientométriques

1. **Bases de données bibliographiques (PubMed)**
   - Titres, auteurs, lieux, résumés
   - Fonctionnalités de recherche et de filtrage
   - Relations simples (et souvent implicites)

2. **Index de citations (Science Citation Index)**
   → Graphe orienté de citations
   → Mesures de productive et d'impact
   → Permet la suive des filiations savantes
   → Sophistication: mise en relation des publications

3. **Systèmes de métadonnées savantes: intégration et uniformisation de données (Crossref)**
   - Agrégation et harmonisation de métadonnées provenant de multiples fournisseurs externes
     - Standardization, déduplication et désambiguïsation des enregistrements
     - Identifiants pérennes: auteurs, ouvrages, pays
   - Réutilisabilité et inter-opérabilité
   - « Infrastructuralisation » des données de

4. **Graphes de connaissances savants: sémantisation (Microsoft Academic Graph, OpenAlex)**
   → Modèle de données doté d'une structure ontologique riche
     → Différents types d'entités: Œuvres, auteurs, institutions, concepts, lieux, financeurs…
     → Différents types de relations sémantiques: auteur_de, affilié_à, financé_par, réfère à, cité_par…
   → Sémantisation des données de publication

Maxime Holmberg Sainte-Marie
Journée Francophone OpenAlex 2025

La scientométrisation du capitalisme savant

→ La scientométrie joue un rôle de premier plan au niveau du développement et du maintient du capitalisme savant

→ **Transforme le travail savant en valeur quantifiable**

→ Permet une évaluation basée sur des indicateurs pour le recrutement, la titularisation et le financement (Biagioli, 2016).

→ **Incite à une course effrénée à la publication *(« Publier ou périr »)***

→ Favorise la quantité, l'incrémentalisme et les stratégies d'optimisation (Franssen & Wouters, 2024).

→ **Encourage la publication stratégique et la prise de risques éthiques.**

→ Choix stratégique des revues, saucissonage *(salami-slicing)*, publication prédatrice (Hoffmann et al., 2024).

→ **Renforce les inégalités mondiales et institutionnelles.**

→ Privilégie la recherche en langue anglaise, dans les pays du Nord et dans les institutions d'élite (Silva et al., 2023).

→ **Fonctionne comme infrastructure de gouvernance et de classement.**

→ Intègre le travail académique dans des systèmes de contrôle managérial, d'indicateurs de performance et de financement compétitif (Hicks et al., 2015).

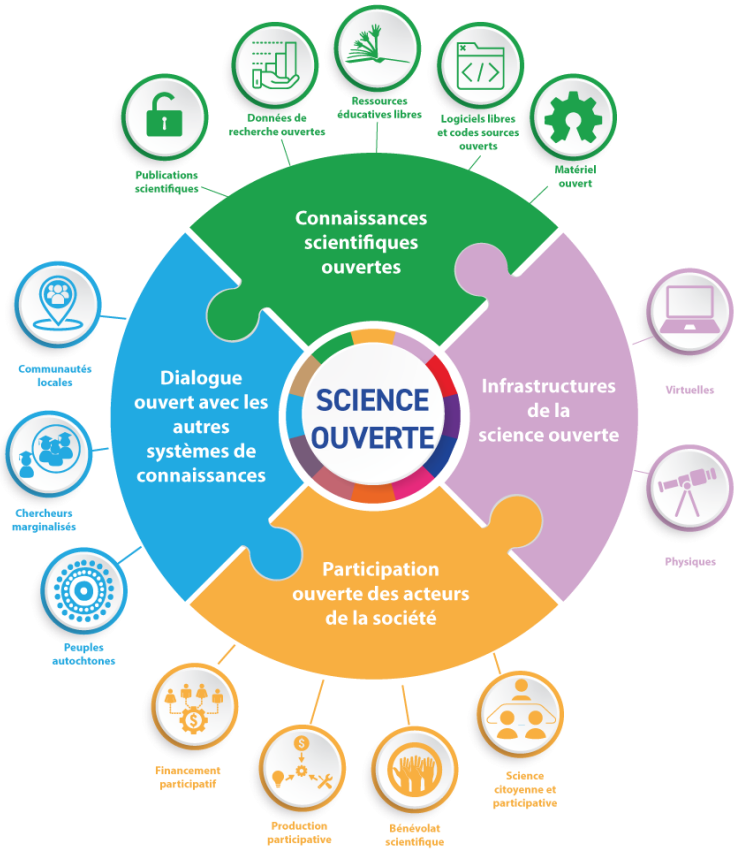
→ **Conclusion** : La bibliométrie agit comme mécanisme d’évaluation, de validation, d’incitation et de gouvernance du capitalisme académique.

Maxime Holmberg Sainte-Marie
Journée Francophone OpenAlex 2025

SDU
University of Southern Denmark

OpenAlex à l'épreuve de la recherche scientifique:

La science ouverte contre le capitalisme savant



→ Suppression des barrières payantes et démocratisation du savoir
  → Rend la recherche accessible gratuitement à tous.
  → Affranchissement face au modèle d'édition à but lucratif (Bibliothèque nationale de Suède, 2025 ; Tennant et al., 2022).
→ Le partage des données et des infrastructures favorisent la collaboration et la réutilisation.
  → Permet de considérer la science comme un bien public plutôt que comme un produit concurrentiel et privatisé (Tennant et al., 2022).
→ L'ouverture de l'évaluation par les pairs et l'auto-archivage décentralise le prestige et le contrôle d'accès (gatekeeping)
  → Affaiblit la domination de l'édition à but lucratif (PKP, 1998).
→ La science ouverte soutient l'équité, l'inclusion et la diversité.
  → Donne aux chercheurs issus d'institutions aux ressources limitées et de régions sous-représentées un accès à la littérature, aux données et aux opportunités de publication (Tennant et al., 2022).
→ Les infrastructures ouvertes et les citations ouvertes brisent le contrôle commercial des métadonnées et des indicateurs,
  → Réduisent la dépendance aux bases de données scientométriques à but lucratif
  → Favorisent l'émergence d'écosystèmes scientifiques gérés par la communauté (Initiative for Open Citations, 2017).
→ En somme, la science ouverte freine la marchandisation de la recherche, réduit les monopoles des éditeurs et encourage un système scientifique plus démocratique, coopératif et équitable.

SDU
University of Southern Denmark

La révolution OpenAlex

- → Ouverture
  - → Premier outil scientométrique en source libre
    - → Toutes les métadonnées et tous les identifiants sont librement accessibles, téléchargeables et réutilisables via son API et ses vidage dynamiques sélectifs (*shapshot dumps*)
      - → API de type REST, sans abonnement et avec des limites de requêtes généreuses.
    - → Gratuit pour tous les usages, y compris commerciaux
- → Couverture
  - → Premier outil scientométrique à visée compréhensive
    - → Prépublications, archives ouvertes, revues savantes, données de financement, logiciels et jeux données…
    - → Premier outil scientométrique véritablement multilingue
- → Sémantisation
  - → Graphe de connaissances offrant une représentation ontologique formelle des différents types d'entités et de relations peuplant l'écosystème de la recherche
- → Transparence
  - → Affiche la provenance des données, ce qui accroît la transparence en scientométrie
  - → Doté d'un système de versionnage, permettant la reproductibilités des analyses scientométriques
- → Dynamisme
  - → Mis à jour quotidiennement et encourage les retours de la communauté via un système de suivi des problèmes similaire à GitHub.

Vers une bibliométrie des communs

→ Renforce l'ouverture des flux de travail en recherche ouverte (Tennant et al., 2020 ; OCDE, 2021)
  → Données scientométriques ouvertes
  → Intégration d'identifiants ouverts (DOI, ORCID, ROR, etc.) et de métadonnées en libre accès.
→ Facilite la reproductibilité des recherches et évaluations
→ Brise les monopoles sur les métadonnées scientifiques
  → Affaiblit le contrôle commercial de la recherche (Fyfe et al., 2017 ; Larivière et al., 2015; Hicks et al., 2015 ; Wouters et al., 2019)
  → Rend possible le développement d'indicateurs, de produits et de services bibliométriques transparents et gérés par la communauté (Bilder et al., 2015; Toelch et Ostwald, 2018)
→ **Favorise l'équité mondiale**
  → **Permet aux chercheurs du monde entier, et notamment des pays du Sud, d'accéder à des métadonnées scientifiques de haute qualité (Chan et al., 2019 ; Babini, 2020).**
  → **Plus grande couverture régionale et linguistique**
→ Accroît la transparence de l'écosystème de la recherche.
  → Met en lumière les réseaux de citations, les modes d'accès et les relations de financement, révélant ainsi les inégalités structurelles renforcées par les éditeurs commerciaux (Mirowski, 2011 ; Posada et Chen, 2018).

Maxime Holmberg Sainte-Marie
Journée Francophone OpenAlex 2025

SDU
University of Southern Denmark

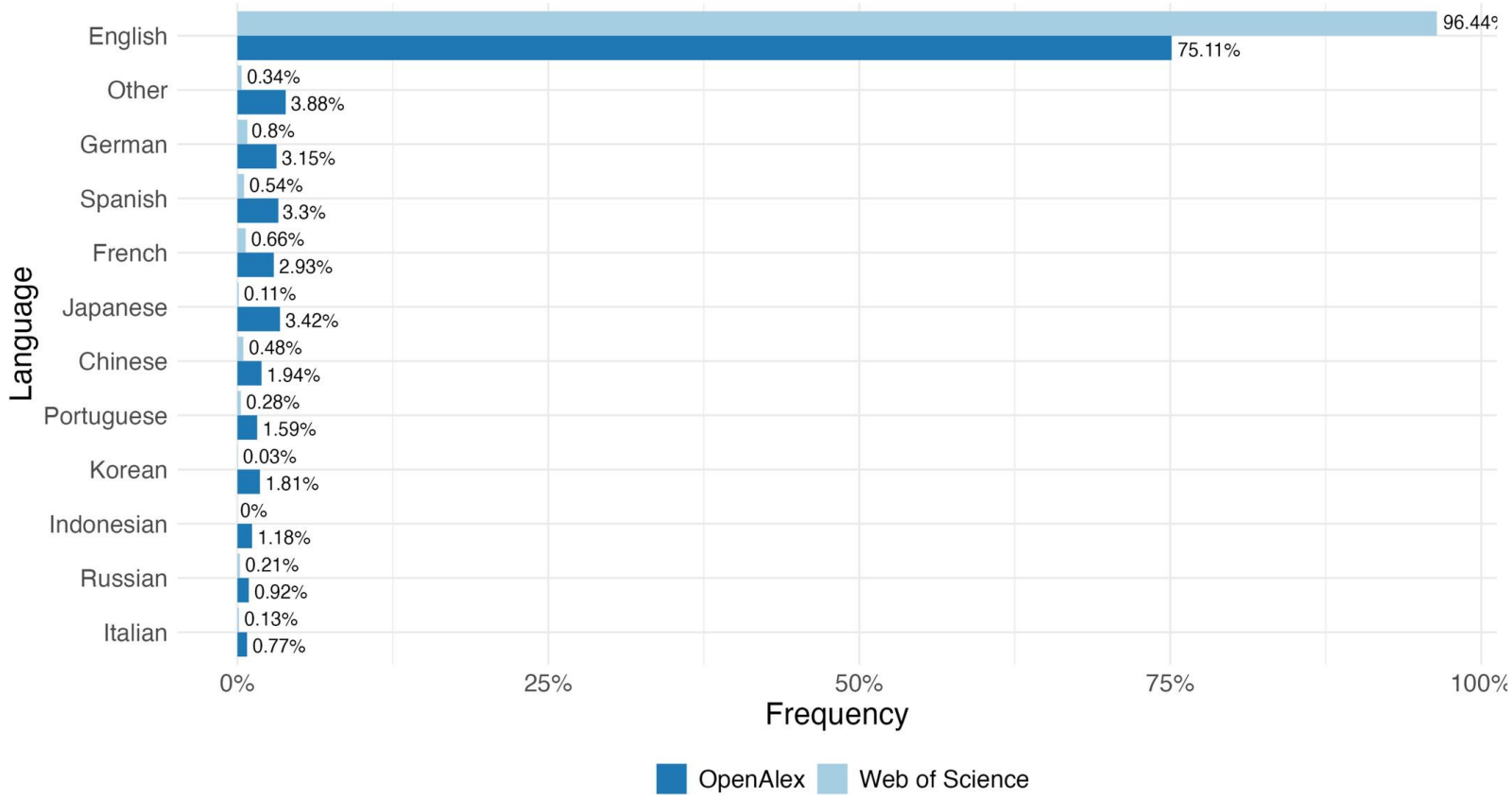
OpenAlex à l'épreuve de la recherche scientifique:

Évaluation de la qualité des métadonnées linguistiques d'OpenAlex (Céspedes et al., 2025)

SDU
University of Southern Denmark

OpenAlex et la babélisation de la scientométrie

- → Couverture linguistique plus équilibrée
  - → Anglais
    - → WoS: 96 % des articles
    - → OpenAlex: 75 % des articles
  - → Augmentation en proportion des articles allophones
    - → Allemand: 4x
    - → Coréen: 60x
    - → Français: 4x
    - → Japonais: 31x
  - → Indonésien
    - → WoS: Pratiquement inexistant
    - → OpenAlex: 1 %.



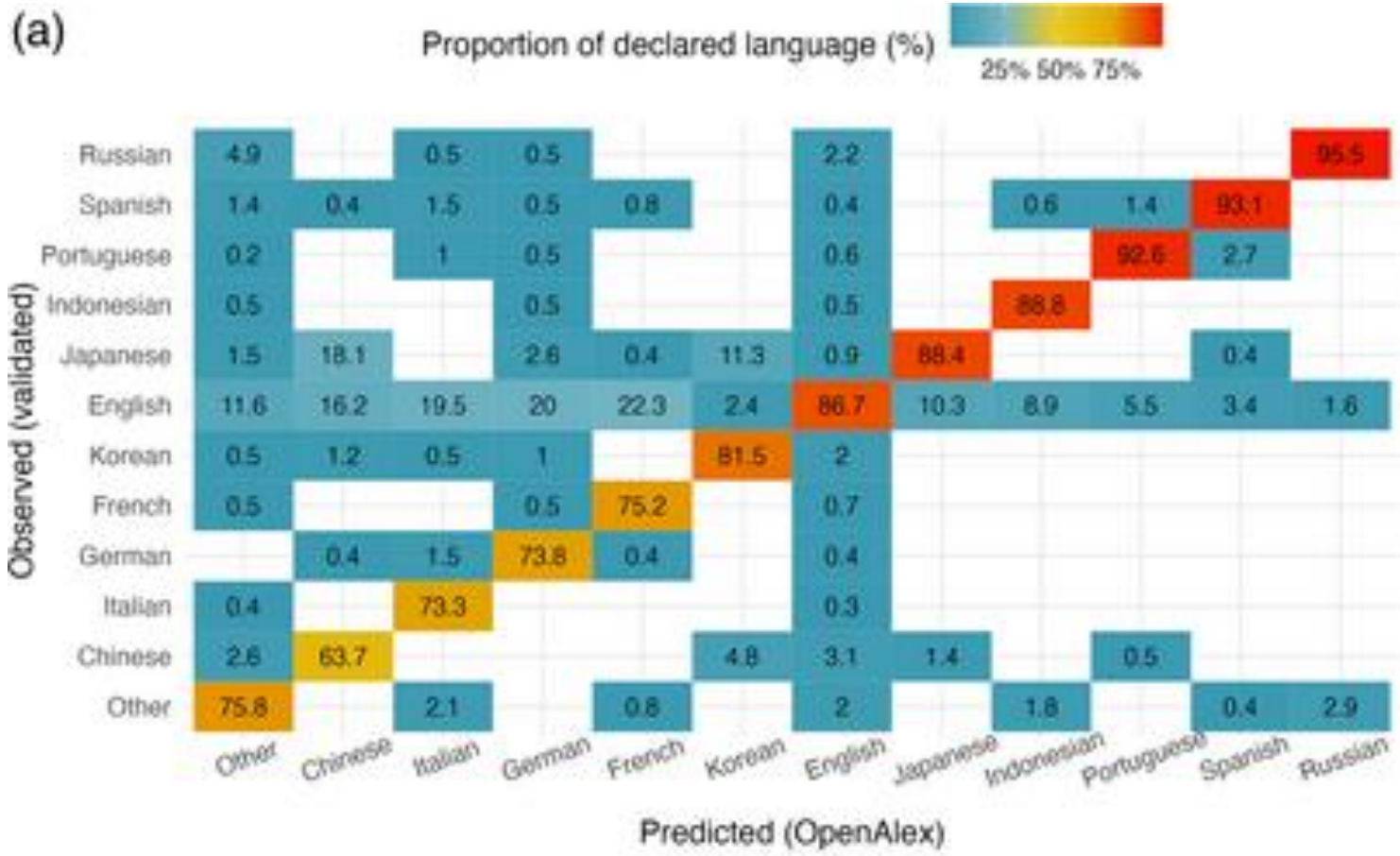
Évaluation par échantillonage stratifié

→ La langue du contenu de chaque article a été manuellement et directement vérifiée par 14 codeurs

→ 3 vagues d'échantillonnage stratifié

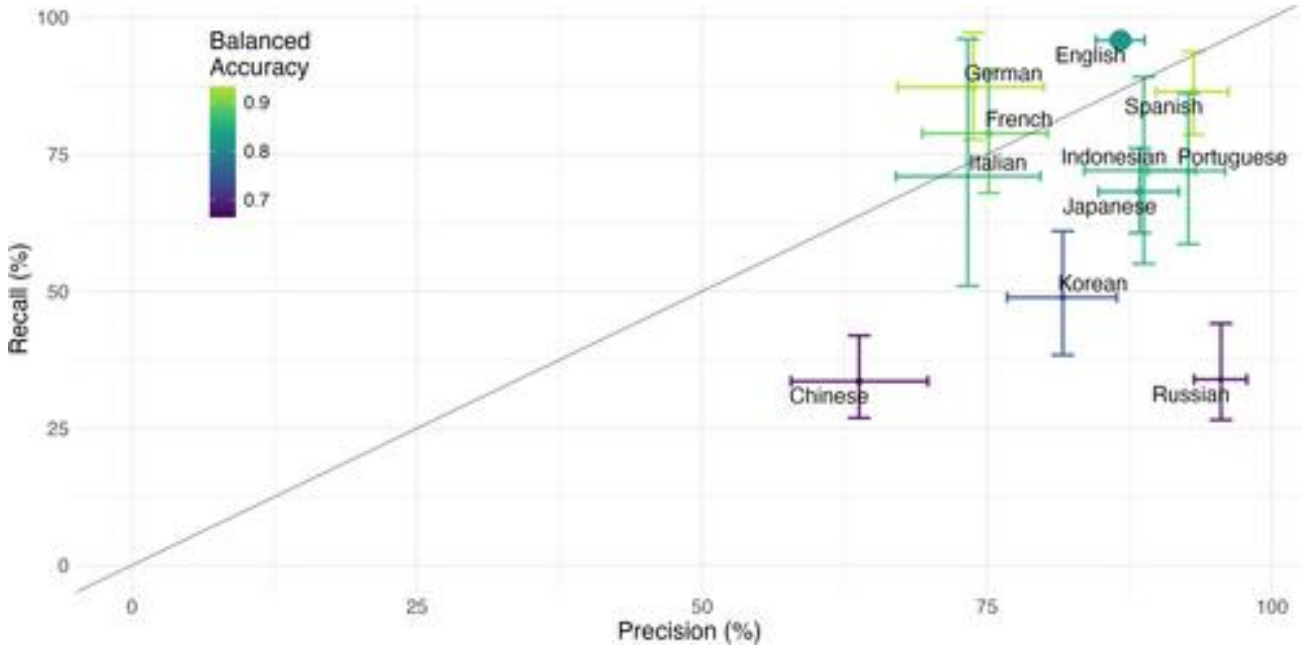
1. Jusqu'à 50 articles pour chacune des 55 langues indexées entre 2000 à 2020.
   → 2701 articles
2. 285 articles pour chacune des 11 langues les plus indexées
   → 95% des documents indexés
   → 11 x 285 = 3135 articles
3. 1000 articles indexés en anglais
   → Garantit représentation plus adéquate

→ Des 6836 documents échantillonnés, 5747 se sont avérés être de véritables articles et ont été inclus dans le jeu de données final



Estimation de la qualité des métadonnées linguistiques d'OpenAlex

- → Performance sur l'échantillon
  - → Haute précision et rappel: anglais, espagnol
  - → Haute précision, rappel modéré: indonésien, japonais, portugais
  - → Pire rappel: chinois, russe, coréen
- → Estimation par facteurs d'expansion
  - → Rapport entre fréquence documentaire totale pour chaque language et fréquence d'observations vérifiées
- → OpenAlex a tendance à surestimer la place de l'anglais tout en sous-estimant celle des autres langues
  - → Proportion de l'anglais estimée à 68 % (contre 75 % indexés par OpenAlex et 96,4 % indexés par WoS)
    - → Expliquent la plupart des faux négatifs ailleurs
  - → Russe: 186% du total indexé
  - → Chinois: +93 % du total indexé



Maxime Holmberg Sainte-Marie
Journée Francophone OpenAlex 2025

SDU
University of Southern Denmark

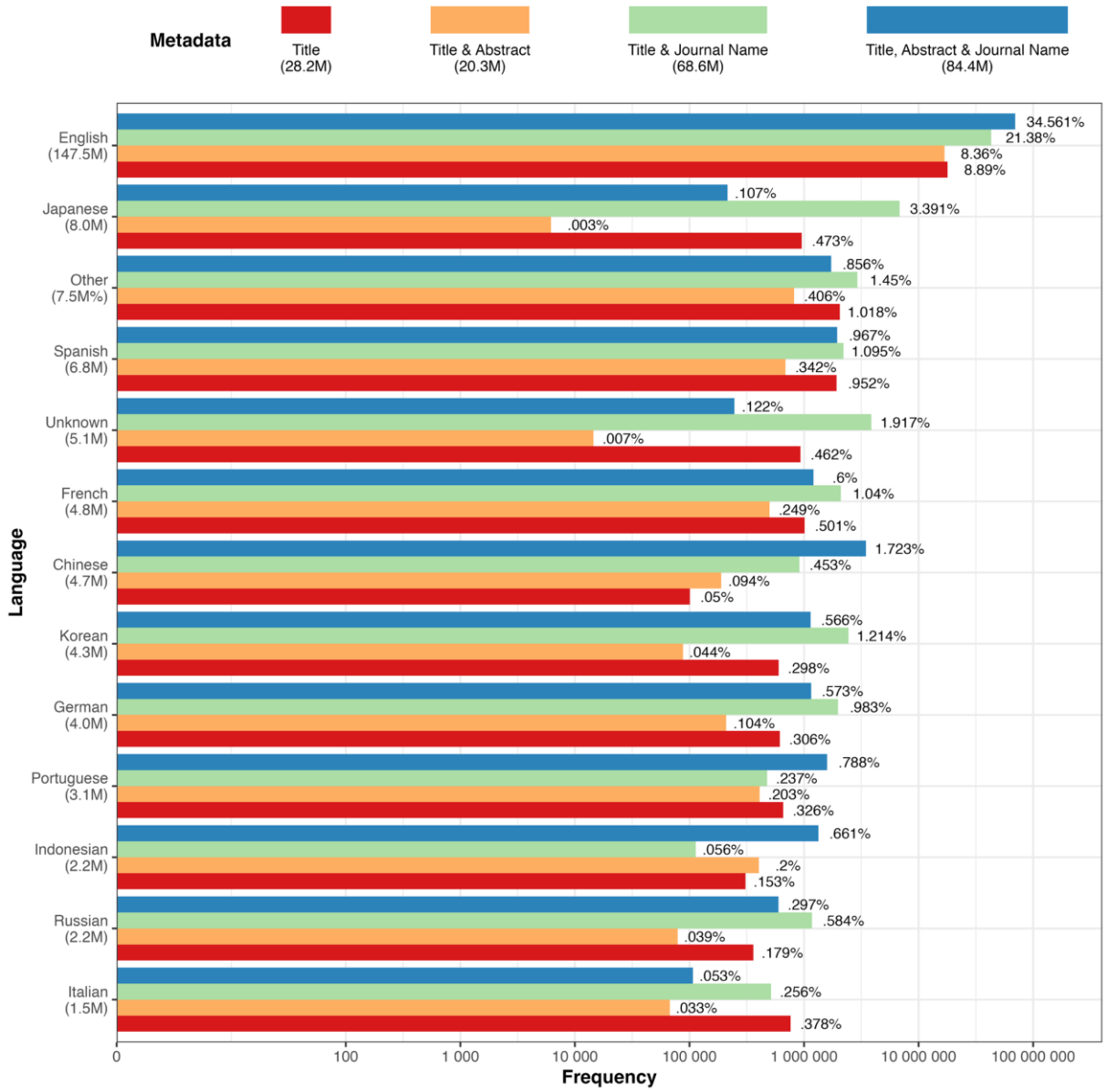
OpenAlex à l'épreuve de la recherche scientifique:

Optimization de la qualité des métadonnées linguistiques d’OpenAlex (Sainte-Marie et al., 2025)

SDU
University of Southern Denmark

Méthodologie

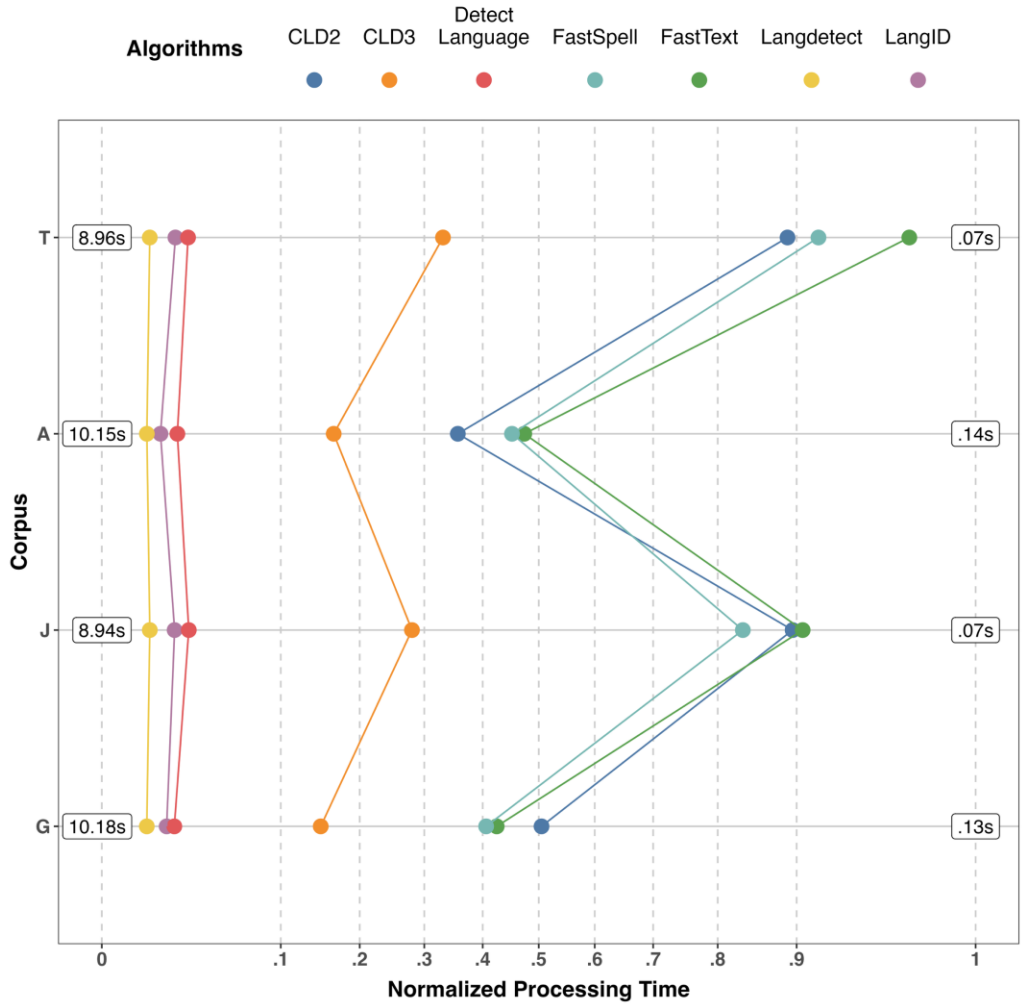
- → Comparaison systématique de la performance de différentes procédures de classification linguistique
  - → Combinaisons d'algorithmes d'identification de langue et de corpus de métadonnées
    - → Algorithmes (7)
      - → CLD2, CLD3, DetectLanguage, FastText, FastSpell, Langdetect, LangID
    - → Corpus de métadonnées (4)
      - → Titre
      - → Titre et résumé, si disponible
      - → Titre et Nom de journal
      - → Algorithme glouton

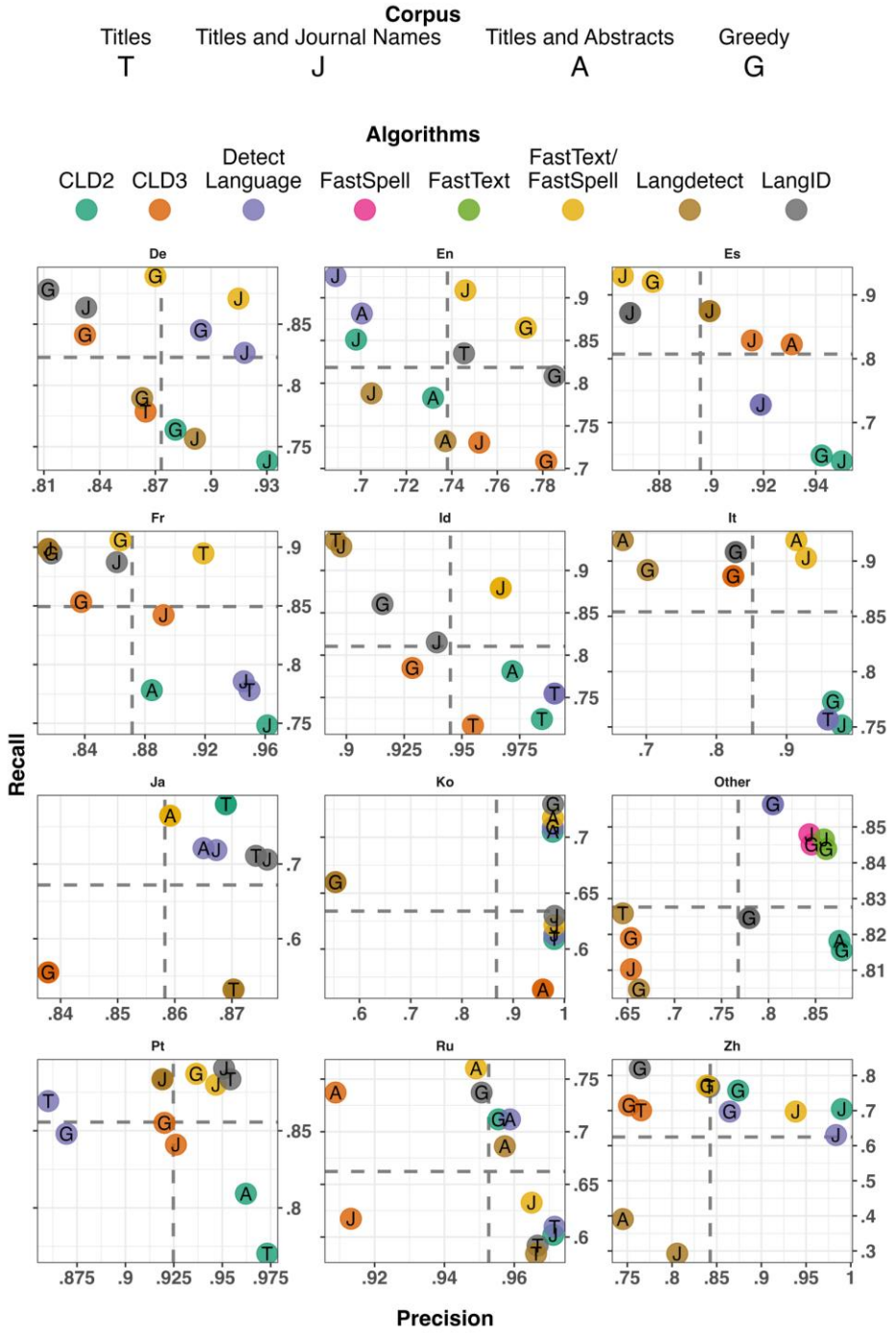


Maxime Holmberg Sainte-Marie
Journée Francophone OpenAlex 2025

SDU University of Southern Denmark

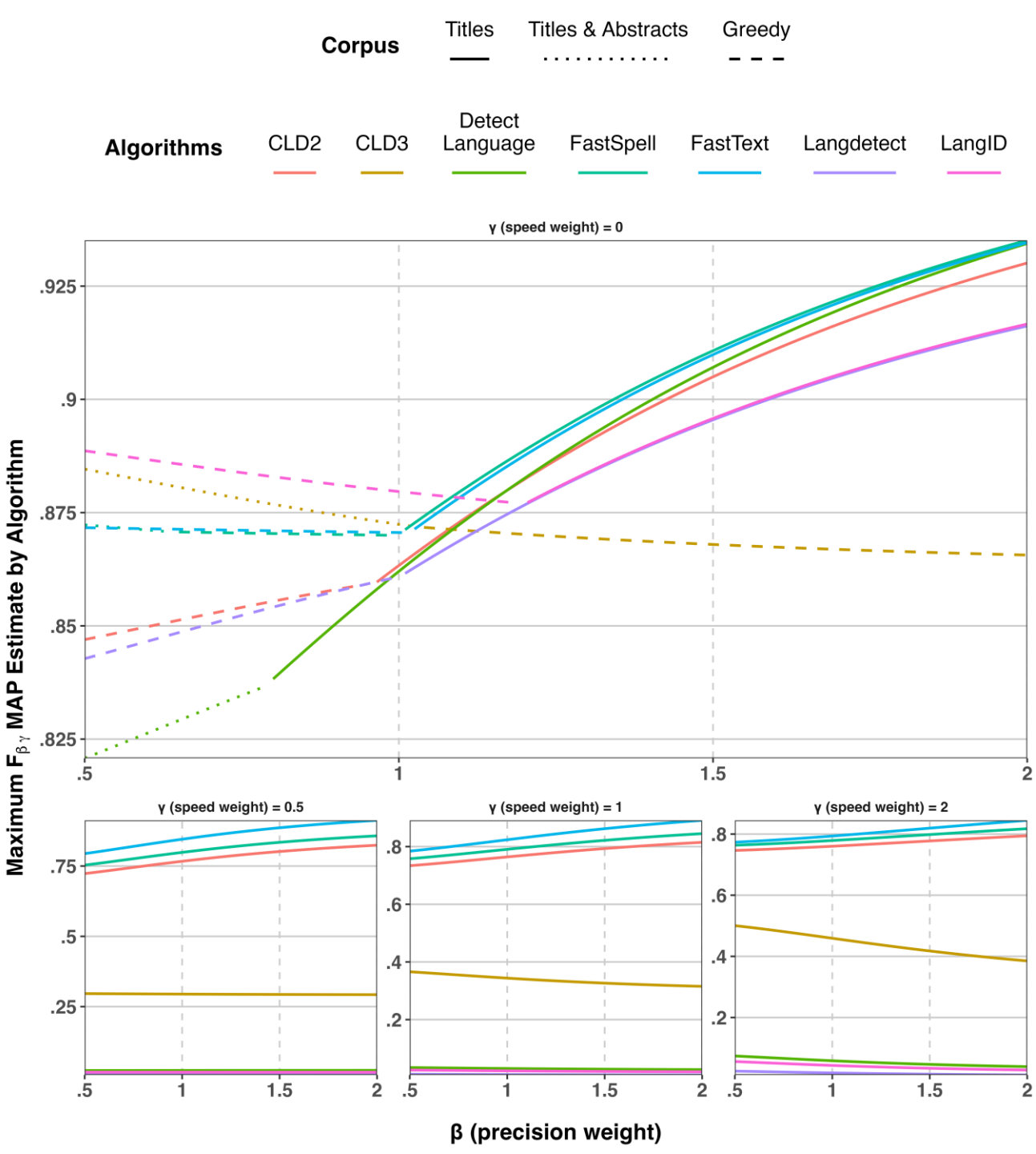
Résultats





Inférence bayésienne

- → Les résultats observés sur l'échantillon peuvent être trompeurs
  - → Robustesse à l'incertitude
    - → Fréquences linguistiques relatives distordues
    - → Métadonnées incomplètes et changeantes
  - → Données déséquilibrées
- → Simulation de matrices de confusion probabilistes
  - → Lois bêta conjuguées a priori pour la précision et le rappel
  - → Lois de Dirichlet conjuguées a priori pour les fréquences linguistiques relatives
- → La mesure agrégée **Fβ,γ** est utilisée pour mesurer l'importance relative de la précision (β), du rappel et de la vitesse de traitement (γ) sous différents régimes de pondération
- → Le choix de la meilleure procédure dépend fortement de la pondération accordée aux mesures d'évaluation
  - → Si la précision est privilégiée: **LangID + corpus glouton**



•En retour, OpenAlex pourrait intégrer certains de ces indicateurs en natif, renforçant ainsi son utilité pour les instituts de

Collaborer avec OpenAlex en tant que chercheur

→ Pour les chercheurs
- → **Amélioration de la qualité des données**
  - → Chercheurs et institutions peuvent signaler des erreurs, des doublons, ou des affiliations incorrectes.
  - → OpenAlex pourrait ainsi affiner ses méthodes de désambiguïsation (auteurs, institutions, domaines).
- → **Enrichissement et ouverture des connaissances**
  - → Les chercheurs pourraient contribuer à compléter des champs manquants dans les notices (mots-clés, catégorisation, citations manquantes).
- → **Meilleure interopérabilité avec les outils de recherche**
  - → Les équipes académiques pourraient aider à améliorer les API, les formats de données et les intégrations avec les logiciels de veille, de bibliométrie ou de gestion de références.
- → **Développement d'indicateurs responsables**
  - → En collaborant avec des experts en scientométrie, OpenAlex pourrait concevoir des indicateurs plus éthiques et moins biaisés.
  - → Cela contribuerait à des pratiques d'évaluation plus transparentes, en accord avec les principes DORA ou la science ouverte.
- → **Construction d'un écosystème plus équitable**
  - → La participation active des chercheurs du Sud global, des disciplines SHS ou de petites institutions permettrait de réduire la partialité souvent observée dans les données bibliographiques.

Collaborer avec OpenAlex en tant qu'institution

→ **Assurance-qualité**
  → Les institutions peuvent corriger ou enrichir leurs informations
    → Des données plus exactes (affiliations, structures internes, laboratoires) renforcent la présence de l'institution dans les moteurs de recherche académiques.
    → Une représentation fidèle dans la base améliore la visibilité des chercheurs, des unités et des productions scientifiques.

→ **Facilitation de la gestion stratégique de la recherche**
  → Un accès propre et standardisé aux données bibliographiques permet :
    → le suivi des performances scientifiques,
    → l'identification de collaborations existantes ou potentielles,
    → la veille disciplinaire et géographique.
  → Les institutions peuvent intégrer ces données dans leurs outils internes (tableaux de bord, SIGB, CRIS, CRM).
→ **4. Soutien à l'évaluation responsable**
→ En collaborant avec la plateforme, les institutions peuvent encourager le développement d'indicateurs plus transparents et moins dépendants des métriques propriétaires.
→ Cela aligne la stratégie institutionnelle avec les principes DORA et la science ouverte.
→ **5. Réduction des coûts**