SAE S5.C.01 - Rapport de gestion du projet Proposer une solution optimisée à partir de données internes et externes - Equipe Pantone448C

Yasmine Ben Youssef
Amélia Ben Youssef
Marielle Vallée
Matthias Trupin
Baptiste Blanchet
Gaspard Pons
Lucas Liger
Adrien Pons
Ayoub Lamchichi
Kylian Lasik

Janvier 2025

Table des matières

Table des figures

Liste des tableaux

Introduction

La collaboration est au cœur de l'innovation scientifique, et l'évolution de ces dynamiques peut offrir des perspectives fascinantes sur la manière dont la recherche se transforme au fil du temps.La pandémie de COVID-19 a bouleversé les modes de travail traditionnels, apportant son lot de défis et de transformations. Cette crise mondiale a soulevé de nouvelles questions sur la manière dont les chercheurs interagissent, collaborent et innovent dans un contexte en constante mutation. Ce projet s'inscrit dans un processus d'exploration visant à répondre à la question suivante : **comment la pandémie de COVID-19 a-t-elle impacté les collaborations de recherche ?** Loin de chercher à démontrer une hypothèse préétablie, l'objectif est d'examiner les données et de découvrir, à travers une analyse ouverte, les changements survenus dans les réseaux de collaboration avant et après la crise sanitaire.

Pour ce faire, nous utiliserons DBLP (Digital Bibliography and Library Project), une base de de données bibliographiques dans le domaine de l'informatique, qui recense plus de 7 millions de publications scientifiques et plus de trois millions d'auteurs. Créée dans les années 1980 par l'Université de Trèves en Allemagne, DBLP existe depuis plus de 30 ans et s'est imposée comme une ressource clé pour l'analyse des collaborations scientifiques. Grâce à la richesse de ses métadonnées, cette base de données nous permettra de suivre l'évolution des collaborations, d'explorer les transformations dans la fréquence des interactions, la géographie des partenariats, ainsi que l'émergence de nouvelles pratiques collaboratives.

L'objectif principal de ce projet est de modéliser et de visualiser les évolutions des collaborations scientifiques pendant cette période de transition. En exploitant les données de DBLP, nous examinerons les changements dans la fréquence des collaborations, la géographie des partenariats et l'émergence de nouvelles pratiques collaboratives. La modélisation des réseaux de recherche, accompagnée de visualisations interactives, permettra de dresser un panorama des effets de la crise sanitaire sur la coopération scientifique, tout en offrant des perspectives sur les tendances futures de la recherche académique.

La modélisation des réseaux, enrichie de visualisations interactives, nous offrira une vue d'ensemble des impacts de la crise sur les pratiques collaboratives. Toutefois, il ne s'agit pas de prouver un résultat précis, mais plutôt d'ouvrir un champ d'investigation, d'identifier les tendances émergentes et de poser des questions dont les réponses se dessineront au fur et à mesure de l'analyse des données. Cette approche permettra non seulement de mieux comprendre les dynamiques actuelles de la recherche, mais aussi d'anticiper les évolutions possibles de la collaboration scientifique dans les années à venir.

Chapitre 1

Préparation des données

La préparation des données est une étape essentielle pour garantir leur qualité et leur pertinence. Ce chapitre décrit le processus d'extraction, de nettoyage et de structuration des données issues du fichier `dblp.xml`. L'objectif est de transformer ce vaste corpus bibliographique en un ensemble exploitable, prêt pour l'analyse et effectuer des requêtes. En identifiant les entités clés, les relations critiques et en éliminant les anomalies, nous jetons les bases d'une modélisation efficace des collaborations scientifiques.

1.1 Caractéristiques du fichier source

Le fichier *dblp.xml* possède les caractéristiques suivantes :

— **Structure hiérarchique :** Chaque élément XML représente un type d'entrée bibliographique, par exemple `article`, `procedings` ou `data`.
— **Volume élevé :** 4Go - La taille considérable du fichier impose une gestion optimisée des ressources.
— **Qualité des données :** Les champs contiennent parfois des entités HTML (`&`, `<`), ainsi que des caractères invisibles (`200b`) nécessitant une normalisation.
— **Hétérogénéité des entrées :** Les données présentent une grande variété de formats pour les différents champs, comme les noms d'auteurs (initiales, noms complets, pseudonymes) et les titres de publications (usage mixte de majuscules et minuscules).
— **Multiplicité des langues :** Bien que majoritairement en anglais, certaines entrées contiennent des titres ou des résumés dans d'autres langues, ce qui peut compliquer l'analyse automatique.
— **Temporalité des données :** Le fichier couvre plusieurs décennies de publications, impliquant des variations dans les formats et standards au fil du temps.

1.2 Description du DataSet

1.2.1 Les différentes catégories de documents

A l'aide de la documentation disponible sur le site et de nos recherches personnelles, nous avons pu identifier neuf catégories de documents principaux, chacun ayant des attributs spécifiques.

— **article** : Un article publié dans un journal scientifique
— **inproceedings** :Un papier spécifique présenté lors d'une conférence.
— **proceedings** :Le volume des actes d'une conférence
— **book** : Une monographie rédigée par un auteur ou un recueil d'articles édité.
— **incollection** : une partie ou un chapitre d'un monographe
— **phdthesis** : Thèse de doctorat (réalisée par un doctorant dans une école).
— **mastersthesis** : Thèse de master (réalisée par un étudiant dans une école).
— **www** : Page web.
— **data** : Données (jeu de données, logiciels...).

Remarque : De nombreux articles de type 'inproceedings' appartiennent à un 'proceeding', tout comme de nombreuses 'incollections' appartiennent à un 'book'. Les publications 'proceedings' et 'book' sont respectivement les parents des 'inproceedings' et 'incollection' .

1.2.2 Caractéristiques d'une publication

Chaque publication contient un ensemble d'attributs :

Balise	Information	Format	Récurrent
`key`	Identifiant unique d'un enregistrement	Chaîne de caractères	Oui
`mdate`	Date de dernière modification d'une publication et/ou de ses métadonnées	YYYY-MM-DD	Oui
`title`	Libellé pour chaque enregistrement	Chaîne de caractères	Oui
`crossref`	Contient la `key` de la publication parent	Chaîne de caractères	Non
`ee`	Lien électronique, souvent un DOI*	URL	Non
`url`	Lien vers le répertoire local	Chaîne de caractères	Non
`pubtype`	Information lié au statut/type de la publication	Chaîne de caractères	Non
`author`	Nom(s) des auteurs, dans l'ordre de contribution (inclut ORCID* si disponible)	Chaîne de caractères	Non
`year`	Année de publication	YYYY (norme ISO)	Non
`pages`	Généralement une plage de pages	Chaîne de caractères	Non
`note`	Informations complémentaires (ex : affiliation, récompenses, pays...)	Chaîne de caractères	Non

TABLE 1.1 – Balises communes aux publications dans DBLP

Notes :

— ***DOI** : Digital Object Identifier - Il s'agit d'un identifiant numérique unique attribué aux objets numériques, comme les articles scientifiques, les chapitres de livre, les thèses ou même des ensembles de données. Il permet de localiser et d'accéder facilement à ces ressources via une url permanente.*

— ***ORCID** : Open Researcher and Contributor ID - Il s'agit d'un identifiant unique utilisé qui garantit l'identité des chercheurs, indépendamment des variations de leur nom ou affiliation.*

1.2.3 Caracéristiques spécifiques à chaque catégorie de publication

Articles

Les articles peuvent aussi contenir

— **journal** : Titre du journal dans lequel l'article est publié.

— **volume** : Numéro du volume.

— **number** : Numéro de l'édition spécifique au volume.

```
<article key="journals/cacm/Szalay08" mdate="2008-11-03">
    <author>Alexander S. Szalay</author>
    <title>Jim Gray, astronomer.</title>
    <pages>58-65</pages>
    <year>2008</year>
    <volume>51</volume>
    <journal>Commun. ACM</journal>
    <number>11</number>
    <ee>http://doi.acm.org/10.1145/1400214.1400231</ee>
```

```
    <url>db/journals/cacm/cacm51.html#Szalay08</url>
</article>
```

Listing 1.1 – Exemple d'article

Proceedings

Les attributs peuvent inclure :
— **Editor** : Le(s) éditeur(s) responsable(s) de la compilation des actes.
— **Publisher** : Le nom de l'entité responsable de la publication.
— **ISBN** : Le numéro ISBN, identifiant unique de l'ouvrage.
— **Series** : La collection ou série à laquelle appartient l'ouvrage.
— **Volume** : Le numéro de volume dans la série, si applicable.
— **Booktitle** : Le titre abrégé de la conférence ou de l'atelier.

```
<proceedings key="conf/cbc/2020" mdate="2020-07-28">
    <editor>Marco Baldi</editor>
    <editor>Edoardo Persichetti</editor>
    <editor>Paolo Santini</editor>
    <title>Code-Based Cryptography - 8th International Workshop, CBCrypto 2020,
    Zagreb, Croatia, May 9-10, 2020, Revised Selected Papers</title>
    <year>2020</year>
    <booktitle>CBCrypto</booktitle>
    <publisher>Springer</publisher>
    <series>Lecture Notes in Computer Science</series>
    <volume>12087</volume>
    <isbn>978-3-030-54073-9</isbn>
    <isbn>978-3-030-54074-6</isbn>
    <ee>https://doi.org/10.1007/978-3-030-54074-6</ee>
    <url>db/conf/cbc/cbcrypto2020.html</url>
</proceedings>
```

Listing 1.2 – Exemple de proceedings

Inproceedings

Les attributs peuvent inclurent :
— **Booktitle**

```
<inproceedings key="conf/cbc/BartzYBL20" mdate="2020-07-28">
    <author>Hannes Bartz</author>
    <author>Emna Ben Yacoub</author>
    <author>Lorenza Bertarelli</author>
    <author>Gianluigi Liva</author>
    <title>Protograph-Based Decoding of Low-Density Parity-Check Codes with Hamming
     Weight Amplifiers.</title>
    <pages>80-93</pages>
    <year>2020</year>
    <booktitle>CBCrypto</booktitle>
    <crossref>conf/cbc/2020</crossref>
    <ee>https://doi.org/10.1007/978-3-030-54074-6_5</ee>
    <url>db/conf/cbc/cbcrypto2020.html#BartzYBL20</url>
</inproceedings>
```

Listing 1.3 – Exemple d'inproceedings en XML

Books

Les books peuvent aussi contenir :
— **Publisher**
— **Series**
— **ISBN**
— **School** : le nom de l'école

```xml
<book key="books/daglib/0001587" mdate="2021-07-17">
    <author>Wolfgang Heidrich</author>
    <title>High-quality shading und lighting for hardware-accelerated rendering - revision 1.1.</title>
    <pages>I-XVIII, 1-148<pages>
    <year>1999</year>
    <publisher>Utz</publisher>
    <school>University of Erlangen-Nuremberg, Germany<school>
    <series>Informatik</series>
    <isbn>978-3-89675-624-4</isbn>
    <ee>https://d-nb.info/957538103</ee>
</book>
```

Listing 1.4 – Exemple de livre en XML

Incollections

Les incollections peuvent aussi avoir
— **Booktitle**

```xml
<incollection key="books/acm/17/CohenO17" mdate="2020-12-16">
    <author>Philip R. Cohen</author>
    <author>Sharon L. Oviatt</author>
    <title>Multimodal speech and pen interfaces.</title>
    <pages>403-447</pages>
    <year>2017</year>
    <booktitle>The Handbook of Multimodal-Multisensor Interfaces, Volume 1 (1)</booktitle>
    <crossref>books/acm/17/OSCSPK2017</crossref>
    <ee>https://doi.org/10.1145/3015783.3015795</ee>
    <url>db/books/collections/OSCSPK2017.html#CohenO17</url>
</incollection>
```

Listing 1.5 – Exemple d'incollection en XML

Phdthesis

Les phdthesis contiennent
— **school**

```xml
<phdthesis key="books/daglib/0000492" mdate="2021-07-17">
    <author>Michael Jaedicke</author>
    <title>New Concepts for Parallel Object-Relational Query Processing.</title>
    <pages>1-184</pages>
    <year>1999</year>
    <school>University of Stuttgart, Germany</school>
    <ee>https://d-nb.info/958406758</ee>
</phdthesis>
```

Listing 1.6 – Exemple de thèse de doctorat en XML

Mastersthesis

Comme les phdthesis, les mastersthesis ont aussi :
— **school**

```xml
<mastersthesis key="ms/Yurek97" mdate="2018-06-13">
    <author>Tolga Yurek</author>
    <title>Efficient View Maintenance at Data Warehouses.</title>
    <year>1997</year>
    <school>University of California at Santa Barbara, Department of Computer Science, CA, USA</school>
</mastersthesis>
```

Listing 1.7 – Exemple de mémoire de master en XML

data

```xml
<data key="data/10/AbayomiAlliAAMA22" mdate="2024-06-18">
    <author orcid="0000-0002-3875-1606" >Adebayo Abayomi-Alli</author>
    <author>Odeyinka Abiola</author>
    <author orcid="0000-0001-9338-491X" >Oluwasefunmi 'Tale Arogundade</author>
    <author orcid="000-0002-3556-9331" >Sanjay Misra</author>
    <author orcid="0000-0003-2513-5318" >Olusola Oluwakemi Abayomi-Alli</author>
    <title>Large Dataset of Nigeria Covid-19 Tweets for Sentiment Analysis and
    Opinion Mining Tasks.</title>
    <year>2022</year>
    <month>August</month>
    <ee>https://doi.org/10.5281/zenodo.4748715</ee>
    <url>streams/repo/zenodo</url>
</data>
```

Listing 1.8 – Exemple de jeu de données en XML

www

```xml
<www key="homepages/00/10012-3" mdate="2019-09-24">
    <author>Shalini Jain 0003</author>
    <title>Home Page</title>
    <note>Rutgers University, USA</note>
</www>
```

Listing 1.9 – Exemple de page web en XML

Enfin, il existe d'autres attributs complémentaires, qui seront étudiées dans la partie suivante (nettoyage).

1.3 Analyse préliminaire des données

1.3.1 Critères d'analyse

Pour structurer ce vaste ensemble de données et répondre efficacement à la problématique il y a trois critères à prendre en compte :

- **la perspective métier** : Quelles données sont déterminantes dans l'identification d'un document ?
- **la perspective analytique** : Quelles données sont exploitables pour des analyses futures ?
- **occurences des balises** : Est-ce qu'un attribut est suffisamment représenté pour être significatif/valorisable ?

Ainsi, en combinant ces trois dimensions : la valeur métier, l'exploitabilité analytique et la représentativité des données il devient possible d'évaluer la pertinence de chaque attribut. Ce croisement des perspectives constitue le fondement d'un modèle de données structuré, pour répondre à la problématique. Le tableau ci-dessous répertorie les différentes catégories de documents ainsi que leur fréquence dans la DBLP :

Type	**Occurrence**
`inproceedings`	3 592 280
`proceedings`	60 268
`incollection`	70 705
`data`	9 686
`book`	20 857
`article`	3 705 372
`mastersthesis`	27
`phdthesis`	138 700
`www`	3 664 534
TOTAL	**11 262 429**

Table 1.2 – Types de documents et leurs occurrences.

- **Documents prédominants** : Les types `inproceedings`, `articles` , et `www` représentent la majorité des données, avec chacun plusieurs millions d'occurrences.
- **Occurrences limitées** : Certains types comme `mastersthesis` `data` sont très peu représentés, limitant leur pertinence dans une analyse globale. Cela peut s'expliquer par leur nature spécialisée pour les `mastersthesis` ou leur faible intégration dans cette base.
- **Taille totale de la base** : Avec 11 262 429 enregistrements, la structure hiérarchique est cohérente : les types spécifiques (`inproceedings`, `articles`) dominent, tandis que les types consolidés (`books`, `proceedings`) regroupent plusieurs contributions.

1.3.2 Fouille et tri des balises

Une fois les proportions des types de publications établies, nous analyserons les balises et sous-balises présentes dans les données. Leur tri repose sur deux critères principaux : leur fréquence d'apparition et leur utilité analytique. Ce tri permet de distinguer les balises **"à conserver"**, qui apportent une valeur significative à l'analyse, de celles **"à retirer"**, car elle sont rares ou non pertinentes dans ce contexte.

1.3.3 Eléments à conserver

Ces balises ont été sélectionnées en fonction de leur capacité à enrichir les réseaux de collaboration, structurer les informations ou à permettre des analyses avancées.

Tableau : Balises à conserver et justifications

Balise	Exemple	Occurrence	Pertinence	Justification
`key`	*"conf/er/Norrie08"*	11262429 (100%)	Très pertinent	Identifiant unique d'un enregistrement.
`title`	*"JimGray, astro nomer."*	11262429 (100%)	Très pertinent	Peut donner des informations sur le domaine de recherche.
`year`	*"2008"*	7597594 (67%)	Très pertinent	Année de publication, primordial pour analyser les dynamiques de collaboration.
`mdate`	*"2008-10-20"*	11262429 (100%)	Pertinent	Date de dernière modification en base, peut être utile pour analyser les dynamiques de collaboration.
`booktitle`	*"Computer Vision, A Reference Guide"*	3724314 (33%)	Très pertinent	Peut donner des informations sur le domaine et identifier les publications faisant partie de la même œuvre.
`author`	*"Moira C. Norrie"*	28830620 (255%)	Très pertinent	Fondamental pour identifier le/les auteur(s) et les collaborations.
`crossref`	*"conf/er/2008"*	3661481 (32%)	Très pertinent	Relie un document à un autre via la key. Nécessaire pour contextualiser les liens entre les publications.
`ee`	*"http ://-doi.acm.org/ 10.1145/ 1400214.1400231"*	7419161 (65%)	Pertinent	Permet d'accéder directement à la publication (peut être utile pour la visualisation).
`pages`	*"58-65"*	6401361 (56%)	Pertinent	Donne la plage de pages pour une publication, représente quantitaviement une collaboration/recheche.
`orcid`	*"0000-0002-1163-8988"*	5856433 (52%)	Très Pertinent	Permet de différencier les homonymes et de s'assurer de l'intégrité des auteurs
`editor`	*"Qing Li"*	153777 (1,3%)	Très pertinent	Information utile pour réprésentater la logique éditoriale dans les collaborations de recherche.
`publtype`	*"encyclopedia"*	784497 (6%)	Pertinent	Information sur le type/statut de publication. (peut être utile pour l'analyse/visualisation).
`publisher`	*"Springer"*	94958 (0,8%)	Pertinent	Permet d'identifier les responsables de la publication et de la diffusion des articles et peut être intéressant pour l'analyse.
`school`	*"University of Southampton, UK"*	142035 (1,2%)	Très pertinent	Permet de retrouver les affiliations des auteurs et primordial pour l'analyse.
`notes`	*"to be published by Cambridge University Press"*	265534 (2,3%)	Pertinent	Informations complémentaires sur le document, contient des affilitation et données géographiques à extraire.
`journal`	*"EAI Endorsed Trans. Ubiquitous Environ."*	3705439 (32%)	Pertinent	Porteur d'informations sur le domaine, nécessaire pour décrire les articles.

TABLE 1.3 – Synthèse complète des balises à conserver

1.3.4 Elements à retirer

D'autres balises n'apportent pas de valeur significative pour répondre à la problématique. Certaines sont redondantes, d'autres sont trop peu fréquentes pour être exploitées efficacement.

Tableau : Balises à retirer et synthèses des justifications

Balise	Exemple	Occurrence	Pertinence	Justification
`url`	*"db/journals/cacm/cacm51.html#Szalay08"*	7631556 (67%)	Inutile	Permet de connaître le chemin d'accès dans le répertoire local du DBLP. Inutile pour répondre à la problématique.
`isbn`	*"978-3-540"*	91869 (0,8%)	Inutile	Code unique d'un livre. Permet de l'identifier, mais inutile pour l'analyse.
`series`	*"Lecture Notes in Business Information Processing"*	36855 (0,3%)	Pertinent	Fournit des informations liées quand on n'a pas de `crossref`. Trop peu de balises pour être exploité.
`volume`	*"51"*	3737064 (33%)	Inutile	Numéro de volume du journal où l'article est publié, inutile pour l'analyse.
`number`	*"11"*	2393031 (21%)	Inutile	Numéro du journal où l'article est publié, inutile pour l'analyse.
`month`	*"June"*	285074 (2,5%)	Pertinent	Mois de publication. Intéressant mais trop peu fréquent pour être utilisé.
`rel`	*"type="versionOf" uri="https://doi.org/10.18419/darus-3044" label="1.0" sort="0""*	4225 (0.03%)	Inutile	Uniquement présent pour Data. Difficile pour l'interprétation avec beaucoup de sous-balises.
`cite`	*"51"*	172835 (1.5%)	Inutile	sens divers : dans un cas, il s'agissait des lectures recommandées sur le site en dessous du document. Dans un autre des références (sources) utilisées → Inutile pour l'analyse et non exploitable.
`publnr`	*"TR06-1271"*	4410 (0.04%)	Inutile	sens divers : Numéro de publication, souvent utilisé dans des contextes académiques, scientifiques ou techniques → Inutile pour l'analyse et non exploitable.
`address`	*"New York"*	3 (0%)	Pertinent	Seulement 3 occurrences, avec une seule information "New York" → Inutile pour l'analyse et pas exploitable.
`chapter`	*"21"*	2 (0%)	Inutile	sens divers :Seulement 2 occurrences → Inutile pour l'analyse et pas exploitable.
`cdrom`	*"SIGIR1988/P117.pdf"*	12933 (0.1%)	Inutile	Version PDF souvent accessible via Ee, avec plus de détails.
`stream`	*"treams/repo/zenodo"*	178120 (1.5%)	Inutile	Indique le répertoire local où a été classé le jeu de données → Inutile pour l'analyse.

TABLE 1.4 – Synthèse complète des balises à supprimer

1.3.5 Synthèse

Dans le cadre de l'analyse des données, nous avons classifié les publications en différents types (articles, conférences, livres, etc.) et identifié les balises spécifiques associées à chaque type, ainsi que les informations essentielles permettant de caractériser une publication. Cependant, certaines anomalies ont été relevées, notamment des incohérences dans les jeux de données.

1.4 Anomalies identifiées dans les données

1.4.1 Anomalies structurelles

1. **Lignes quasiment vides :** Certaines entrées contiennent uniquement une clé (`key`) et une date de modification (`mdate`), sans aucune autre information. Ces lignes, bien que valides dans la structure XML, ne fournissent aucune donnée exploitable pour l'analyse.
2. **Doublons :** Certaines publications sont comptabilisées deux fois en raison de variations dans les informations. Par exemple :
   — Une nouvelle édition d'un livre, avec un éditeur différent, peut apparaître comme une publication distincte.
   — Une mise à jour ou un ajout à un article peut être interprété comme une nouvelle publication.

1.4.2 Anomalies liées aux auteurs

1. **Noms incomplets ou abrégés :** Certains noms d'auteurs apparaissent de manière incomplète (par exemple, uniquement les initiales du prénom ou du nom).
2. **Identifiant ORCID manquant :** L'absence d'un identifiant `ORCID` complique l'identification unique des auteurs, surtout dans les cas d'homonymie. Cela peut générer des doublons et compromettre l'intégrité des données. D'autant plus qu'un même auteur peut apparaître sous différents noms en raison de variations comme des accents, l'inversion prénom/nom, ou l'ajout d'un second prénom (exemple : "Juan" et "Júan", ou "Alexandre Gramfort" et "Gramfort Alexandre" qui représente le même individu apparaissent comme des personnes différentes).
3. **Homonymes :** De nombreux noms et prénoms sont très courants, tels que *Chen Li*, et reviennent fréquemment dans les données. La distinction entre les différents individus portant ces noms repose souvent sur un numéro ajouté (ex. : *Chen Li 0001*, *Chen Li 0002*). Cependant, ce système de numérotation est problématique :
   — Il ne garantit pas que toutes les occurrences d'un même auteur soient correctement regroupées.
   — Il complique l'analyse, car il devient difficile d'assurer que chaque numéro correspond bien à un auteur unique.

1.4.3 Anomalies liées aux contenus des balises

1. **Titres atypiques :** Certains titres de publications ne respectent pas les conventions habituelles et incluent des valeurs telles que *error*, *was never published* ou *...*, qui peuvent indiquer des erreurs de saisie ou des publications incomplètes.
2. **Informations géographiques dans la balise `school` :** La balise `school` contient des informations géographiques précieuses, telles que le nom des villes et des universités, permettant d'analyser les collaborations selon leur localisation géographique. Cependant, ces données sont parfois incomplètes ou manquantes.
3. **Informations diverses dans la balise `note` :** La balise `note` regroupe des informations diverses et annexes sur les publications. Dans certains cas, elle inclut des données similaires à celles présentes dans la balise `school` (pays, ville, université). Cette balise ajoute une complexité dans l'extraction des affiliations.

1.4.4 Synthèse

Les anomalies observées dans les données, qu'elles soient structurelles, liées aux auteurs ou aux contenus des balises, mettent en évidence des limites qui compromettent la qualité des analyses. Ces problèmes soulignent l'importance d'un nettoyage et d'une structuration préalable pour assurer la fiabilité des résultats.

1.5 Représentation du dataset : modèle entité/association

Après l'extraction et l'analyse des données bibliographiques du fichier dblp.xml, les données ont été représentées à l'aide d'un modèle entité-association (E/A). Ce modèle permet d'identifier les entités clés et les relations associées, tout en offrant une vision claire de la structure des données sources. Il constitue une étape essentielle pour concevoir, par la suite, un modèle adapté pour répondre à la problématique :

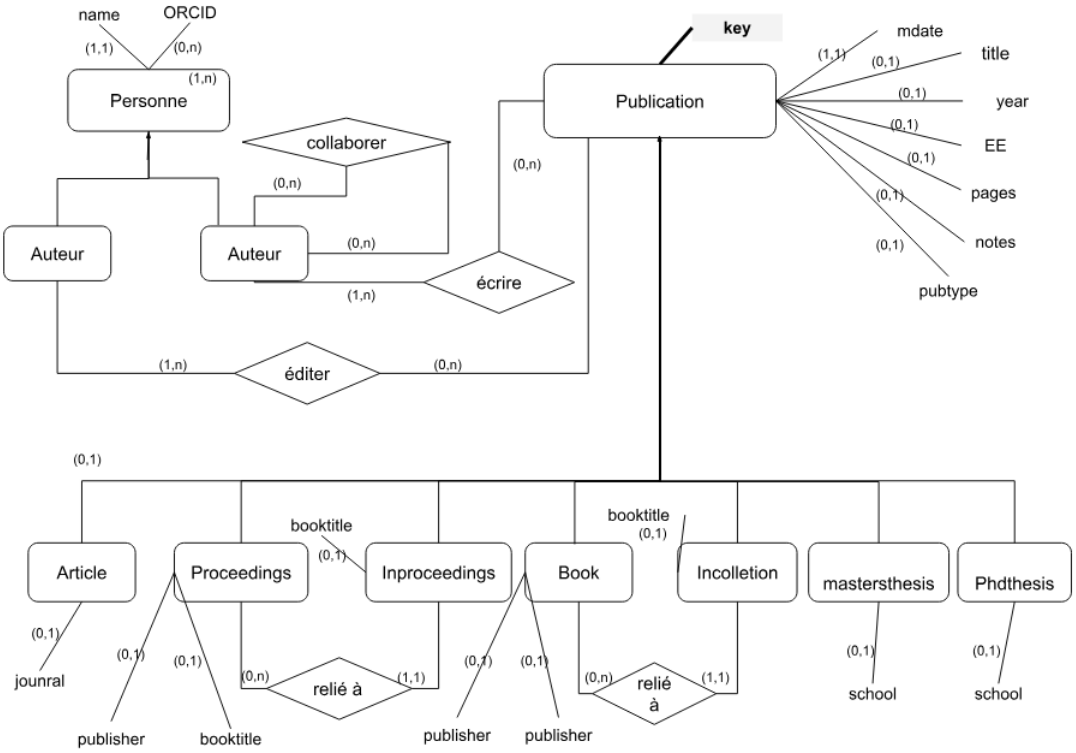


FIGURE 1.1 – Représentation de la structure du XML du DBLP

1.5.1 Identification des Entités

Nous avons identifié les entités suivantes :

- **Personne** : Représente les individus participant aux publications, tels que les auteurs et les éditeurs. Ses attributs sont : le nom et parfois un identifiant ORCID. Cette entité est centrale pour modéliser les contributions humaines aux publications.
- **Publication** : Regroupe toutes les productions scientifiques, identifiées par une clé unique et décrites par des métadonnées communes (titre, année, pages, etc.). Elle se décline en sous-catégories adaptées à chaque type de document (article, livre, conférence, thèse), qui partagent ces métadonnées tout en intégrant des détails spécifiques propres à leur nature, comme un journal pour un article ou une école pour une thèse.

1.5.2 Identification des relations

Nous avons identifié les relations suivantes :

— **Écriture (écrire)** : Relie une ou plusieurs *Personne* (auteurs) à une *Publication* à laquelle ils ont contribué.
  — Chaque *Personne* peut écrire plusieurs publications.
  — Une *Publication* peut avoir plusieurs auteurs, voire aucun.

— **Édition (éditer)** : Relie une ou plusieurs Personne (*éditeurs*) à une *Publication* qu'ils ont supervisée.
  — Chaque *Personne* peut éditer plusieurs publications.
  — Une *Publication* peut être édité par plusieurs auteurs, voire aucun.

— **Lien entre publication parent/enfant (relié à)** : Associe les publications enfant à leur publication parent.
  — Chaque entité enfant (par exemple, *Inproceedings*, *Incollection*) doit être liée à un seul parent.
  — Une entité parent (par exemple, *Proceedings*, *Book*) peut avoir plusieurs enfants.

1.5.3 Limites du modèle E/A

La problématique posée vise à analyser comment la pandémie de COVID-19 a influencé les collaborations de recherche scientifique. Cela nécessite une représentation structurée des données qui inclut :

— Les collaborations entre chercheurs (ou groupes de chercheurs).
— Les institutions ou organisations affiliées, leur localisation géographique.
— Le contexte temporel des publication.

Bien que le modèle E/A identifie les entités principales ainsi que leurs relations, il présente plusieurs limites :

— **Manque de structuration pour des analyses spécifiques** : Le modèle ne détaille pas suffisamment les dimensions temporelles nécessaires pour examiner l'impact avant, pendant et après la pandémie.
— **Absence explicite des dimensions géographiques et institutionnelles** : Bien que les informations sur les institutions ou localisations géographiques soient présentes dans les données, elles ne sont ni normalisées ni suffisamment mises en évidence dans ce modèle.
— **Focalisation sur les entités au détriment des relations enrichies** : Si les relations essentielles entre les entités sont définies, le modèle reste centré autour des publications en tant qu'éléments principaux. Il ne permet pas de représenter explicitement les collaborations entre chercheurs, leurs affiliations institutionnelles ou géographiques. Cette limitation est inhérente à la nature de la DBLP, qui est avant tout un répertoire bibliographique et non un modèle conçu pour l'analyse des réseaux de collaboration.

Ces limitations rendent nécessaire une transition vers un modèle conceptuel plus détaillé (MCD), intégrant des ajustements pour répondre directement à la problématique posée.

1.6 Extraction des données

1.6.1 Approche adoptée

Nous avons défini une stratégie structurée pour transformer le fichier extitdblp.xml en une base de données relationnelle, organisée et prête pour des requêtes analytiques. Voici les principales étapes de ce processus :

1. **Extraction des données pertinentes :** Convertir les informations XML en un format tabulaire (étape de transformation en CSV) pour faciliter leur exploration initiale.
2. **Nettoyage et normalisation :** Éliminer les anomalies, traiter les doublons et standardiser les données dans un entrepôt intermédiaire.
3. **Modélisation relationnelle :** Construire un schéma relationnel adapté à nos objectifs analytiques, en identifiant les relations clés entre les entités (auteurs, publications, conférences, etc.).
4. **Implémentation dans SQL :** Charger les données dans une base relationnelle pour permettre des requêtes complexes et des analyses approfondies.

Voici un schéma récapitulatif :

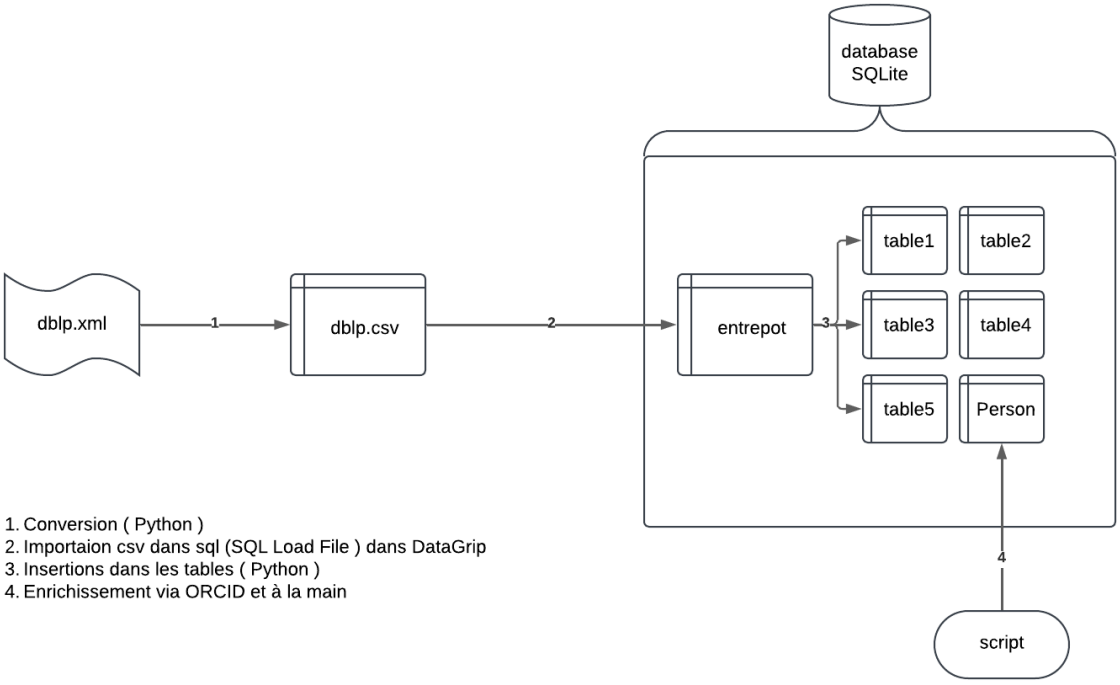


FIGURE 1.2 – Processus d'extraction : .xml to .sql

1.6.2 Extraction préliminaire du xml au csv

Pour amorcer le processus d'analyse, nous avons décidé de parser le fichier XML afin de le convertir en un format tabulaire (CSV). Cette extraction préliminaire nous permet d'obtenir une vue d'ensemble sur les données et de réaliser un premier tri. Voici les grandes lignes de cette étape :

— **Identification des informations essentielles :** Extraire les champs critiques comme les noms d'auteurs, les titres de publications, les années et les conférences associées.
— **Transformation en format tabulaire :** Représenter les données XML sous forme de tableau avec des colonnes clés pour simplifier leur manipulation et leur ingestion dans les étapes suivantes.

— **Gestion des erreurs :** Traiter les anomalies observées, telles que les champs manquants ou mal formatés, et corriger les incohérences lors de l'étape d'exportation.
— **Pré-sélection des données :** Effectuer un tri initial pour isoler les éléments les plus pertinents en vue des analyses ultérieures.

Cette méthode nous aide à structurer efficacement le processus tout en garantissant que chaque transformation appliquée aux données soit claire et traçable. Ainsi, les données préparées pourront être utilisées de manière optimale pour modéliser les relations et effectuer des analyses approfondies.

1.6.3 Techniques de nettoyage, normalisation et correction des données

Afin de garantir la qualité des données, plusieurs techniques ont été mises en place pour éliminer les erreurs et les incohérences. Ce processus a permis de normaliser, nettoyer et corriger les informations brutes extraites du fichier XML, en vue de leur intégration dans le modèle relationnel.

1. Normalisation et suppressions des données incorrectes/invalides : La suppression des publications invalides a été réalisée avec `pandas` pour filtrer efficacement les données, grâce à sa capacité à manipuler de grands ensembles de données. `lxml` a été utilisé pour parser le fichier XML de manière efficace sans charger l'intégralité du fichier en mémoire, ce qui est essentiel pour les fichiers volumineux. La normalisation des données a également été effectuée avec `pandas`, permettant d'harmoniser rapidement les valeurs. Les expressions régulières (`re`) ont été employées pour corriger les erreurs de format dans les noms et dates.

2. Correction des anomalies : Les anomalies ont été corrigées grâce à `pandas` pour remplacer par exemple les valeurs manquantes par des valeurs par défaut. Tandis que `logging` a été utilisé pour suivre et enregistrer les erreurs, assurant ainsi la traçabilité des problèmes rencontrés et facilitant la correction.

3. Enrichissement des données : Les requêtes HTTP qui permettent de récupérer les informations complémentaires via l'API ORCID ont été réalisées avec `requests`, une bibliothèque permettant d'interroger intelligemment une API Web en construisant des requêtes HTTP avec des paramètres définis. Ce processus a ainsi permis d'enrichir les données sur les auteurs de manière automatisée.

1.6.4 Implémentation des données dans le modèle SQL

Après le nettoyage et la normalisation des données, celles-ci ont été intégrées dans le modèle relationnel à l'aide de technologies spécifiques pour garantir leur cohérence et optimiser les performances des requêtes :

1. Création des tables relationnelles : Les tables ont été créées dans une base de données SQLite, un système léger et local adapté pour gérer les données issues d'un projet collaboratif.

2. Insertion des données : Les données nettoyées ont été insérées dans la base de données en utilisant des scripts Python avec la bibliothèque `sqlite3`. Cette bibliothèque permet d'interagir facilement avec la base SQLite pour exécuter des requêtes `INSERT INTO` et insérer les informations dans les tables respectives. Des scripts de traitement par lots ont été utilisés pour insérer efficacement les données en grande quantité, optimisant ainsi le temps de traitement qui peut prendre beaucoup de temps dans cette situation.

3. Gestion des relations : Les relations entre les entités (par exemple, publications, auteurs, catégories) ont été gérées en utilisant des clés étrangères gérées par le programme Python. Ces relations ont ensuite été intégrées au modèle en utilisant des requêtes SQL.

4. Optimisation des requêtes : Pour garantir des performances élevées lors des requêtes analytiques, des index ont été créés sur les colonnes fréquemment utilisées dans les jointures, comme les identifiants des publications et des contributeurs. La création d'index dans SQLite a été réalisée en utilisant des instructions `CREATE INDEX`, ce qui permet de réduire significativement le temps d'exécution des requêtes complexes.

Conclusion

La préparation des données a permis de nettoyer et normaliser les informations extraites du fichier XML, en corrigeant les anomalies structurelles et en enrichissant les données, notamment grâce à l'API ORCID. Les différentes catégories de publications ont été analysées et les données invalides ont été supprimées. L'utilisation d'outils comme pandas, lxml et les expressions régulières a facilité ces transformations. Ce travail a ainsi assuré la qualité et la cohérence des données, garantissant la fiabilité des résultats des requêtes futures.

Chapitre 2

Modélisation de la base de données relationnelle

Ce chapitre présente la structuration des données pour répondre à la problématique : comment la pandémie de COVID-19 a-t-elle impacté les collaborations scientifiques ? La problématique étant complexe et multidimensionnelle, elle impose une approche analytique structurée en plusieurs axes clés, fondée sur les objectifs suivants :

1. **Analyser et comparer les réseaux de collaboration avant et après la pandémie.**
2. **Identifier les tendances scientifiques et leurs évolutions.**
3. **Proposer des visualisations dynamiques des collaborations.**

2.1 Structurer les données : une approche analytique

Dans cette section, nous présenterons le raisonnement ayant conduit à la conception du modèle relationnel (MCD) pour répondre à la problématique posée. Le point de départ repose donc sur les résultats attendus précedemment identifié. À partir de ces objectifs, nous décomposerons la problématique en 4 axes d'analyse stratégiques, chacun correspondant à une thématique spécifique nécessitant une étude approfondie. Ces axes vont permettre de structurer la réflexion et de poser des questions ciblées, divisant ainsi la problématique globale en sous-problématiques exploitables.

En nous appuyant sur le modèle entité/association (E/A) établi en amont, nous identifierons le potentiel des données existantes pour répondre à ces sous-problématques et nous identifierons les limites associées. Afin de pallier ces insuffisances, nous proposerons des enrichissements, qu'ils soient internes (réorganisation ou extraction des données existantes) ou externes (intégration de données complémentaires).

Le modèle relationnel (MCD) a été conçu à partir de cette base méthodologique. Il structure les données pour répondre aux besoins identifiés, tout en garantissant la performance des requêtes analytiques . Ces requêtes permettront de produire des visualisations exploitables. En finalité, ce modèle relationnel constitue un outil central pour analyser les résultats, répondre aux sous-problématiques soulevées par les axes, et apporter des éléments concrets à la problématique globale. Nous détaillerons ainsi une démarche rigoureuse et méthodique, guidée par les objectifs finaux, pour justifier chaque étape de la construction du modèle et son utilité dans l'analyse des données.

2.1.1 Axe 1 : Analyse géographique

L'axe géographique explore la localisation des auteurs et des institutions impliquées dans les publications scientifiques, pour comprendre l'évolution des collaborations locales et internationales. Cette analyse repose sur les affiliations des auteurs fournies par le dataset initial.

Questions principales :

— Quelle est la répartition des collaborations par pays avant, pendant et après la pandémie ?
— La pandémie a-t-elle renforcé ou affaibli les collaborations internationales ?

Données disponibles :

— **Affiliations des auteurs** : Chaque affiliation correspond à une institution (université, institut ou autre organisation) liée à un chercheur. Ces informations reflètent un contexte géographique (ville et pays) et organisationnel pouvant influencer leurs travaux et leurs collaborations. Elles sont issues des balises *school* et *note* du dataset.

Limites des données disponibles

Malgré leur richesse, les données présentent plusieurs limites :
— **Représentation partielle :** Les balises *school* (1,2 % des entrées) et *note* (2,3 % des entrées) couvrent une faible partie du dataset.
— **Incohérences :** Les noms des institutions peuvent être incomplets ou mal renseignés (absence de ville et/ou de pays).
— **Diversité des formats :** Les affiliations sont présentées sous différents formats (« Spain » vs « España », « USA » vs « United States ») et peuvent contenir des erreurs typographiques (« Tunisia » vs « Tunesia »).
— **Structure centrée sur les publications :** Les affiliations sont associées aux publications et non directement aux auteurs, compliquant la représentation des collaborations.
Cette absence de normalisation rend l'analyse des réseaux de collaboration complexe et peuvent biaiser les résultats.

Stratégies d'enrichissement

Pour enrichir cet axe, il serait intéressant d'obtenir les affiliations manquantes en exploitant les identifiants ORCID des chercheurs disposant de cet identifiant. L'ORCID permet de valider et/ou de compléter les informations relatives aux affiliations, améliorant ainsi la qualité et la fiabilité des données.

Justification de la modélisation :

Modélisation des affiliations

Pour modéliser les affiliations, nous avions 2 options :

1. **Associer les affiliations aux publications** : Cette approche aurait permis une centralisation des données par document publié, mais elle aurait également dilué la valeur des informations géographiques en ne liant pas directement les chercheurs à leur contexte local et national.
2. **Associer les affiliations aux auteurs/éditeurs** : Cette option, retenue, offre une vision centrée sur les chercheurs et leurs dynamiques collaboratives.

Ce choix repose sur plusieurs arguments :

— **Focus sur les dynamiques de collaboration :** L'objectif principal est de comprendre comment les chercheurs interagissent, évoluent dans le temps et se regroupent en clusters. Ce type d'analyse exige une représentation directe des chercheurs et de leurs affiliations.

— **Dimension géographique et culturelle :** L'affiliation géographique porte une forte valeur contextuelle pour le chercheur, en intégrant des dimensions locales, culturelles et nationales qui influencent leurs collaborations. Associer ces informations directement aux chercheurs permet de capturer cette valeur ajoutée.
— **Limitation des biais :** Associer les affiliations aux publications pourrait entraîner des imprécisions, notamment dans les cas de publications multi-auteurs où plusieurs institutions et localisations sont impliquées.

Intégration des affiliations enrichies à notre modèle

Identifier une affiliation unique pour chaque chercheur pose des problèmes :
— **Multiplicité des affiliations :** Les chercheurs peuvent avoir plusieurs affiliations au cours de leur carrière. L'affiliation présente dans le dataset DBLP peut ne pas être actuelle.
— **Hétérogénéité :** Les chercheurs avec ORCID peuvent avoir plusieurs affiliations (3 ou 4), tandis que ceux sans ORCID n'en auraient qu'une seule, voire aucune.
— **Manque de chronologie :** Les affiliations ORCID ne contiennent pas toujours les dates associées, rendant la temporalité difficile à modéliser.
— **Complexité relationnelle :** Intégrer plusieurs affiliations pour un même chercheur augmenterait considérablement la complexité du modèle.

En réponse à ces contraintes, nous avons choisi de représenter deux affiliations par chercheur :
— **Affiliation DBLP :** Issue des balises *school* ou *note* du dataset.
— **Affiliation ORCID :** Correspondant à la dernière affiliation connue pour les chercheurs disposant de cet identifiant.

Cette stratégie permet un enrichissement des données tout en limitant la complexité relationnelle.

Décomposition des informations géographiques issues des affiliations

Chaque affiliation est décomposée en trois colonnes directement associées à chaque chercheur : *nom de l'institution*, *ville*, et *pays*, pour les données issues de DBLP et ORCID. Ce choix offre plusieurs avantages :
— Une clarté structurelle en liant directement les informations géographiques aux chercheurs, rendant les données plus compréhensibles.
— Une réduction de la complexité en évitant une normalisation stricte (3NF). Vu le volume important des données, cette décision évite des dépendances transitives et simplifie la gestion.
— Une flexibilité analytique accrue, facilitant les requêtes et explorations des dynamiques locales, nationales et internationales.

Ce modèle simplifié équilibre précision et exploitabilité, permettant de suivre les dynamiques de collaboration tout en maintenant une architecture cohérente pour l'analyse.

2.1.2 Axe 2 : Analyse des collaborations

L'axe des collaborations vise à explorer la manière dont les chercheurs interagissent et collaborent à travers les publications scientifiques. Après avoir examiné l'axe géographique, qui identifie où travaillent les chercheurs, cet axe s'intéresse à comment ils travaillent ensemble, en étudiant la diversité et la densité de leurs interactions au fil du temps et selon les périodes.

Questions principales :

- Quels réseaux de collaborations scientifiques ont émergé avant, pendant et après la pandémie ?
- La pandémie a-t-elle modifié la densité ou la diversité des interactions entre chercheurs ?

Définition d'une collaboration :

Dans le contexte de la recherche scientifique, la notion de collaboration soulève plusieurs questions. Une collaboration, par définition, implique l'association de deux personnes ou plus travaillant ensemble vers un objectif commun. Cependant, lorsqu'il s'agit de caractériser une collaboration scientifique, il est nécessaire de préciser les rôles et contributions des différents acteurs impliqués.
Dans ce cadre, nous pouvons envisager deux options principales pour définir la collaboration scientifique. La première consiste à inclure toute association entre deux personnes ou plus, sans distinction de leur rôle spécifique. Cela engloberait à la fois les auteurs et les éditeurs. La seconde option, plus restreinte, se concentre exclusivement sur les auteurs. Nous avons choisi de définir la collaboration scientifique comme une **association entre deux auteurs ou plus**, ces derniers étant directement impliqués dans l'élaboration et la production du contenu scientifique.
Il est important de noter que, bien que les éditeurs participent également au processus, leur rôle relève davantage de la logique éditoriale que de la collaboration scientifique au sens strict. Si leur contribution est indéniablement précieuse, elle ne concerne généralement pas la construction du contenu scientifique lui-même. Par conséquent, il semble pertinent de restreindre la définition de la collaboration scientifique aux auteurs .Ainsi, pour cette analyse, nous utiliserons le terme "collaboration scientifique" pour désigner toute association entre deux auteurs ou plus, laissant de côté les éditeurs et d'autres acteurs éditoriaux.

Données disponibles :

- **Chercheurs et leurs rôles** : Le dataset contient des informations sur les auteurs et éditeurs ayant contribué aux publications. Ces données permettent de relier chaque chercheur à un rôle précis dans une publication.
- **Publications** : Les publications multi-auteurs permettent d'identifier les relations entre chercheurs, formant des réseaux implicites de collaboration.

Limites :

Malgré les données disponibles, les collaborations sont déduites des publications et non directement représentées dans les données, rendant la représentation des collaborations plus complexe.

Justification de la modélisation :

La conception des corrélations dans le modèle relationnel repose sur deux entités principales et deux association :

1. **Person** : Chaque chercheur est identifié par une nom, un identifiant ORCID (lorsqu'il est disponible) et des affiliatons.
2. **Publication** : Les publications représentent les travaux des chercheurs (*Person*).

3. **Contribution** : Identifie le rôle d'une personne dans une publication. Cette association lie donc l'entité *Person* à l'entité *Publication*.
4. **Collaboration** : Représente l'association entre deux auteurs (*Person*) pour l'élaboration de contenu scientifique dans une *Publication*. Les collaborations sont implicites et doivent être déduites des relations n :m (many-to-many) entre chercheurs et publications via l'assiciation *Contribution*. Une publication avec **n auteurs** génère $\sum(n-1)$ **collaborations**

2.1.3 Axe 3 : Axe temporel

L'axe temporel explore l'évolution de la production scientifique et des collaborations à travers le temps. Pour analyser efficacement cette dimension, il est nécessaire de définir des périodes pertinentes en tenant compte des contextes historiques, technologiques et scientifiques.

Suppositions et problématique

Pour comprendre la répartition temporelle des publications dans la base de données DBLP, plusieurs suppositions s'appuient sur l'évolution de la diffusion des travaux scientifiques au fil des décennies.

Avant 1980 : Des travaux sur supports physiques

Avant 1980, les modalités de recherche étaient caractérisées par :
- **Production manuelle** : Les travaux scientifiques étaient documentés à l'aide de machines à écrire (sur papier). Les échanges se faisaient lors de réunions physiques internationales ou via courrier postal.
- **Diffusion limitée** : Les rapports et résultats de recherche étaient imprimés en quantités restreintes, souvent distribués dans les bibliothèques spécialisées ou lors de conférences académiques, rendant l'accès difficile pour les chercheurs éloignés.
- **Exemple des limites de coordination et diffusion dans la recherche :**
    - ***[ALGOL 60]** : Le développement d'ALGOL 60, un langage de programmation conçu dans les années 1950-60, illustre les limites de la recherche collaborative de l'époque. Les spécifications étaient discutées lors de réunions physiques comme celle de Zurich (1958) ou Paris (1960), suivies d'échanges lents par courrier postal. La diffusion se limitait à des copies imprimées et à des publications dans des revues académiques spécialisées (Communications of the ACM), accessibles uniquement aux institutions majeures. Ces contraintes ralentissaient l'adoption universelle et standardisée d'ALGOL 60, pourtant fondamental pour la programmation moderne.*

A partir des années 1980 : Transformation numérique et début de la mondialisation scientifique

Les années 1980 ont marqué une transition majeure avec l'émergence des ordinateurs personnels et des bases de données numériques, révolutionnant les pratiques de recherche. Les scientifiques ont progressivement adopté des outils numériques pour rédiger, archiver et diffuser leurs résultats.

- **Apparition des bases de données numériques** : Des systèmes comme PubMed (1971, pour les sciences biomédicales) et IEEE Xplore (1988, pour l'ingénierie et l'informatique) ont permis de centraliser et de structurer la production scientifique. Ces plateformes, comme DBLP, ont facilité l'accès aux publications à une échelle internationale.

- **Invention du World Wide Web (WWW)** : En 1989, Tim Berners-Lee a introduit le concept du WWW, qui a révolutionné la recherche scientifique en connectant des bases

de données à travers le monde, permettant un accès instantané à une multitude de ressources.

— **Accélération de la collaboration internationale** : Les échanges par e-mail et les conférences téléphoniques ont permis aux chercheurs de collaborer en temps réel, réduisant les délais de communication.

— **Essor des outils collaboratifs, notamment les outils de visioconférence** : À partir des années 2000, les outils de visioconférence, comme Skype (lancé en 2003), ont progressivement permis des réunions virtuelles en temps réel. Leur usage s'est démocratisé avec l'arrivée de plateformes avancées comme Zoom (2011) ou Microsoft Teams (2017). Ces outils ont permis de multiplier les échanges scientifiques, d'organiser des webinaires et de renforcer les collaborations internationales.

— **Exemple de l'impact des avancées technologiques sur la recherche scientifique**

  — ***[Le Projet Génome Humain (1990-2003)]*** *: Le Projet Génome Humain (1990-2003) est l'un des exemples les plus emblématiques de collaboration scientifique internationale des années 1990, reflétant l'impact des outils numériques sur la recherche. Ce projet ambitieux, impliquant des chercheurs de plusieurs pays, visait à séquencer l'intégralité du génome humain.Grâce à des plateformes comme GenBank, les données génétiques générées par des laboratoires mondiaux étaient partagées en temps réel, accélérant leur exploitation par la communauté scientifique. La version 155 de GenBank, datée d'août 2006, contenait plus de 65 milliards de bases de nucléotides (consituant de l'ADN), reflétant l'ampleur des données partagées.*

Depuis 2020 : Suppositions sur l'impact de la pandémie de COVID-19

La pandémie de COVID-19 a transformé les pratiques scientifiques, bien que ces observations reposent sur des **suppositions** nécessitant validation/réfutation.

1. **Collaboration numérique accrue** : Face aux contraintes de distanciation physique et à l'urgence d'une crise sanitaire inédite, les chercheurs ont été contraints d'adopter massivement des outils numériques pour collaborer. D'une part, la gravité de la situation a nécessité la mobilisation rapide de tous les experts compétents, quels que soient leur domaine ou leur localisation géographique. Cette mobilisation mondiale a exigé un échange immédiat d'idées, d'hypothèses, de résultats de recherche et d'analyses, dans le but de fournir aux gouvernements des stratégies éclairées pour lutter contre la pandémie. Ainsi, des plateformes comme Zoom, Microsoft Teams ou Slack sont devenues des outils indispensables, marquant une transition vers une nouvelle manière de collaborer imposée par les circonstances.
2. **Augmentation des publications** : Une hausse significative des publications sur des sujets comme la santé publique ou l'épidémiologie est observée. Par exemple, une augmentation de 50 % des articles sur les vaccins a été rapportée en 2020-2021 (Nature, une revue scientifique hebdomadaire de renommée mondiale).

Justification des plages temporelles définies

Pour structurer l'analyse des dynamiques de recherche scientifique, les données ont été segmentées en périodes équilibrées, cohérentes avec le contexte technologique et géopolitique.

— **Before (avant 2017)** : Cette période regroupe des publications et collaborations issues de contextes très différents, allant des travaux du siècle dernier, marqués par des pratiques scientifiques traditionnelles, aux collaborations plus récentes caractérisées par une adoption croissante des technologies numériques. Bien que ces contextes technologiques et historiques soient distincts, nous avons choisi de les regrouper dans une même plage temporelle pour des raisons méthodologiques. Ce choix repose sur la volonté de structurer notre analyse autour de plages de trois ans, en cohérence avec la durée de la pandémie

de COVID-19 (2020-2022), déclarée par l'OMS. Cette approche permet de comparer les dynamiques de collaboration scientifique sur des périodes équilibrées et pertinentes.

— **3 Years Before (2017-2019)** : Ces trois années avant la pandémie constituent une base de référence pour analyser les dynamiques de collaboration scientifique dans des conditions "normales". Cette période est marquée par un contexte où les outils numériques étaient déjà largement adoptés, mais sans les perturbations majeures provoquées par la pandémie. Elle est essentielle pour évaluer les changements induits par les périodes suivantes.

— **During (2020-2022)** : Cette période couvre les années de la pandémie de COVID-19. Cette période est unique en raison de l'urgence sanitaire mondiale et de l'impact direct de la pandémie sur les dynamiques de recherche et de collaboration.

— **After (2023-2024)** : Cette période post-pandémique reflète la stabilisation progressive des pratiques scientifiques, tout en intégrant les transformations accélérées par la crise. Bien qu'elle partage de nombreuses similitudes avec la période During en termes d'interconnectivité et de technologies numériques, elle se distingue par l'absence de l'urgence sanitaire mondiale.

— **Exclusion de 2025** : L'année 2025 a été volontairement exclue de l'analyse, car elle est encore en cours. Inclure des données partielles pourrait fausser les observations et compromettre la cohérence des graphiques.

Ces choix permettent de comparer les dynamiques avant, pendant, et après la pandémie, en mettant en lumière les impacts de cet événement mondial sur la recherche scientifique, tout en garantissant une répartition cohérente et équilibrée des données.

Questions principales :

— Quelles sont les variations dans les collaborations et les thématiques en fonction des périodes clés, comme avant, pendant et après la pandémie ?
— Y a-t-il des pics de publications ou de collaborations liés à des périodes spécifiques ?

Données disponibles

Dans le dataset, deux types principaux de données temporelles sont disponibles :
— **l'année de publication**
— **la date de dernière modification d'une publication**

Limites

Dans 67 % des cas, l'année de publication est explicitement indiquée dans les métadonnées des publications. Cette information est directement représentative de la date à laquelle une recherche a été finalisée et rendue publique, ce qui en fait un indicateur fiable pour analyser les dynamiques de publication et de collaboration. Néanmoins, cette donnée ne peux pas refléter le temps réel nécessaire à la révision, à l'étude, à la recherche et à la construction des travaux scientifiques. Ainsi, il est possible qu'une publication post-COVID (après 2022) ait mobilisé la collaboration de plusieurs auteurs durant la période de pandémie, sans que cela transparaisse dans les données. Cette limite peut introduire une certaine imprécision dans l'étude des dynamiques de collaborations, en raison du décalage naturel entre la phase de recherche et la date de publication finale.

La `mdate`, bien que systématiquement présente, n'est pas retenue comme base pour l'analyse temporelle, car elle reflète davantage des modifications techniques ou administratives que des changements liés au contenu de la publication. Par exemple, une modification des métadonnées, telle que l'ajout d'un ORCID pour un auteur ou une mise à jour de l'éditeur, peut entraîner une

mise à jour de la mdate sans que cela ne reflète une réelle modification du contenu scientifique. De ce fait, utiliser la mdate pour classer les publications selon les périodes définies pourrait introduire des biais et ne pas représenter fidèlement la durée ou la temporalité des collaborations scientifiques.

Ainsi nous nous interesserons uniquement aux publications qui ont une date de publication renseignée.

Justification de la modélisation

L'attribut *year* est central dans l'entité publication, car il reflète le moment où les résultats scientifiques ont été finalisés et partagés avec la communauté. Ce choix s'appuie sur sa pertinence en tant que base temporelle claire et directement exploitable pour examiner les dynamiques de recherche et de collaboration.
Pour approfondir cette analyse, chaque publication est associée à une des périodes définies : Before, 3 Years Before, During, et After, via son année de publication. Par exemple, une publication datée de 2018 sera classée dans "3 Years Before". Cette catégorisation apporte plusieurs avantages significatifs. Elle permet de structurer l'analyse des dynamiques de recherche dans un cadre temporel cohérent, facilite l'identification des tendances majeures dans la collaboration scientifique.
En excluant la mdate de ce processus, ce choix méthodologique privilégie une approche rigoureuse et alignée avec l'objectif d'identifier les véritables dynamiques temporelles des collaborations et des publications.

2.2 Modèle conceptuel de données : base relationnelle

Modèle Conceptuel de données

Voici le MLD issu du MCD pour notre base de données :

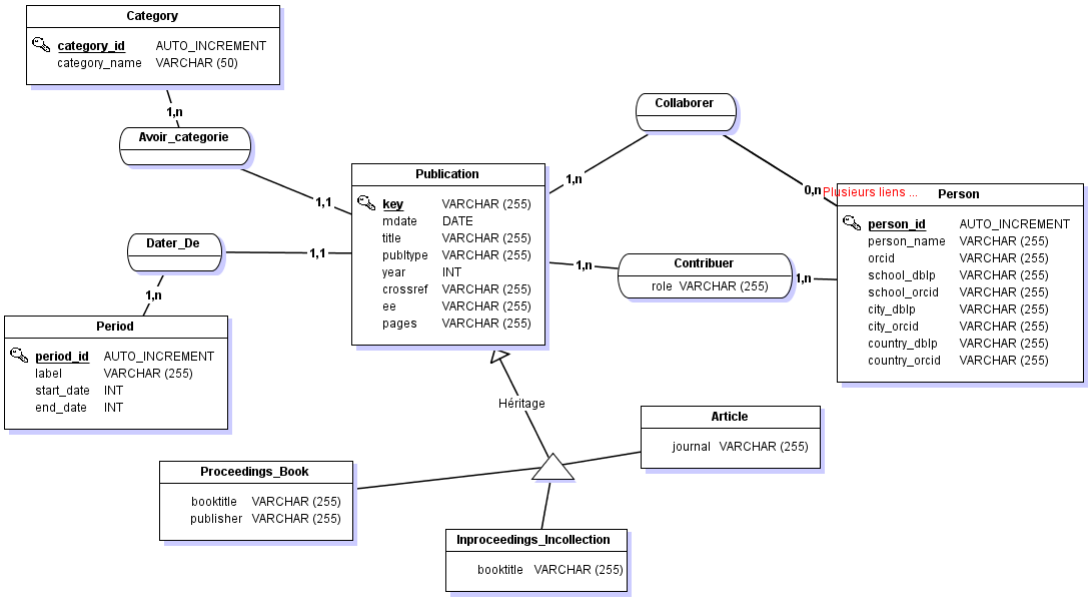


FIGURE 2.1 – MCD

2.2.1 Table `Publication`

La table `Publication` regroupe les attributs essentiels permettant de définir une publication scientifique :

- `key` : l'identifiant unique
- `title` : le libellé de la publication
- `mdate` : date de dernière modification de la publication
- `year` : date de publication
- `pubtype` : information complémentaire relative à la publication
- `crossref` : pour relier les publications enfant à leur parent.
- `ee` : lien externe de la publication
- `pages` : nombre de pages de la publication

Nous avons décidé de ne conserver que ces attributs, car ils offrent un niveau de précision nécessaire et suffisant pour décrire une publication de manière pertinente. Ce choix garantit une structure optimisée et exploitable.

2.2.2 Justification de l'héritage

Lors de la conception de notre modèle, plusieurs options ont été envisagées pour structurer les données des publications :

1. **Créer une table par catégorie de publication** :
   Cette solution aurait impliqué de créer des tables distinctes pour chaque type de publication (`Article`, `Proceedings`, `Inproceedings`, `Book` etc.). Cependant, cette approche aurait engendré une importante redondance d'information, car un grand nombre d'attributs (comme `title`, `mdate`, `year`, etc.) sont communs à toutes les publications. Cela aurait non seulement alourdi le modèle, mais aussi réduit son efficacité.
2. **Créer une table unique pour toutes les publications** :
   Une autre possibilité aurait été de regrouper tous les attributs dans une seule table, avec une colonne pour chaque information pertinente. Cependant, cette option aurait conduit à de nombreuses valeurs nulles, car certains attributs sont spécifiques à un type de publication. Par exemple :
   - L'attribut `journal` est exclusif aux articles et donne du sens à l'article en indiquant le journal où il est publié. Cet attribut n'aurait pas de sens pour un livre ou un chapitre.
   - De même, un livre (`Proceeding-Book`) est publié en tant qu'entité complète, contrairement à un chapitre (`Inproceedings-Incollection`) qui est inclus dans un livre.

   En regroupant tout dans une table unique, on aurait perdu le sens et la logique de chaque type de publication.

Choix de la généralisation distincte et incomplète

Pour surmonter ces inconvénients, nous avons opté pour une généralisation distincte et incomplète. La relation de généralisation permet de faire ressortir les propriétés communes (attributs, associations, opérations) et les différences entre les classes.

- Pourquoi Disjointe ? Une généralisation distincte signifie qu'un objet de la superclasse ne peut appartenir à plus d'une sous-classe, et c'est notre cas : on ne veut pas pouvoir avoir une publication qui est a la fois article et livre et chapitre.
- Pourquoi incomplète ? Une généralisation incomplète signifie qu'un objet peut appartenir à la superclasse sans appartenir à une des sous-classes. Et c'est notre cas : les `mastersthesis` ou `phdthesis` bien qu'ils aient quasiment tous une information sur les affiliations des contributeurs, sachant qu'on veut associer l'affiliation à l'auteur, la table publication a elle seule suffit pour représenter ces catégories de publication, idem pour les publications de type `data` ou encore `www`.

Ce choix présente plusieurs avantages : tout d'abord, la centralisation des attributs communs dans la table `Publication`. Ensuite, la spécialisation des sous-classes permet de répondre aux

besoins propres à chaque type de publication. Enfin, la structure hiérarchique optimise les requêtes. Les informations communes à toutes les publications peuvent être directement interrogées dans la table `Publication`, simplifiant ainsi les requêtes globales. De même, pour des analyses plus spécifiques, chaque sous-classe peut être consultée séparément, assurant ainsi des requêtes plus ciblées.

Regroupement des types `InProceeding` et `InCollection`

Dans le cadre de la généralisation distincte et incomplète choisie pour notre modèle, les entités `InProceeding` et `InCollection` ont été regroupées en une seule sous-classe. Ce choix s'appuie sur leurs similitudes structurelles et fonctionnelles :Ces deux types de publications se caractérisent par leur appartenance à un volume parent, généralement identifié par un `crossref` et un `bookTitle`, et elles partagent également l'attribut `publisher`, indiquant l'éditeur de leur volume parent. Ce regroupement permet de réduire la complexité du modèle tout en préservant la logique des relations entre les entités. En regroupant ces deux types sous une même sous-classe, nous simplifions la gestion des publications enfants.

Regroupement des types `Proceedings` et `Book`

De manière analogue, les entités `Proceedings-Book` et `Book` (livres standards) ont également été regroupées, car elles possèdent des attributs similaires, notamment : l'attribut `bookTitle`, qui identifie le titre de l'ouvrage. En regroupant ces types dans une même sous-classe, nous simplifions la gestion des publications parents.

Isolement de la catégorie `Article`

À l'inverse, la catégorie `Article` a été isolée dans une sous-classe distincte en raison de son attribut exclusif `journal`, qui spécifie la revue dans laquelle l'article a été publié. Cet attribut est essentiel pour capturer la singularité des articles et ne peut être partagé avec d'autres types de publications.

Représentation des types `Mastersthesis` et `Phdthesis`

La table `Publication` suffit à représenter ces types de données de manière cohérente et complète, sans nécessiter de sous-classe spécifique.

Représentation des types `Data` et `www`

Ces catégories sont représentées directement par la table `Publication`.

2.2.3 La table `Category`

La table `Category` a été créée pour structurer et isoler les types de publications. En extrayant les catégories de publication dans une table distincte, nous avons cherché à simplifier les requêtes. En effet, associer chaque publication à une catégorie définie facilite les recherches et les analyses par type de publication.

Cette table contient :

— `category-id` : un identifiant unique pour chaque catégorie.

— `category-name` : le nom de la catégorie (par exemple, `Article`, `Book`,`Data`, etc.).

En séparant les catégories de publication dans une table dédiée, nous avons également évité les duplications d'informations, rendant le modèle plus propre et efficace.

2.2.4 La table `Period`

La table `Period` a été créée à partir des plages temporelles précédemment définies et conçue pour faciliter les requêtes liées aux analyses temporelles. Plutôt que de manipuler directement des dates pour chaque publication, cette table permet de regrouper les publications par périodes cohérentes, définies en amont.

Elle contient :
- `period-id` : un identifiant unique pour chaque période.
- `label` : le nom de la période (Par exemple : 'During').
- `start-date` et `end-date` : les dates de début et de fin de la période.

Cette table permet de labéliser les publications par période, va simplifier les requêtes et permettre d'analyser et de rechercher par période sans manipuler les dates, ce qui peut être intéressant pour analyser des tendances sur des intervalles précis.

2.2.5 La table `Person` et la relation `Contribuer`

La table `Person` a été créée pour regrouper toutes les informations utiles à l'identification des auteurs (leur nom complet) et à leurs situations géographiques. Elle contient :
- `person-id` : un identifiant unique pour chaque auteur.
- `person-name` : le nom complet de l'auteur (par exemple : 'Laurel Haak').
- `orcid` : l'identifiant unique orcid de l'auteur. À titre de rappel, 20% des auteurs ont un identifiant orcid référencé.
- `school-dblp` : l'université dans laquelle les auteurs sont affiliés selon les informations disponibles dans la base de données DBLP. Ce champ contient le nom de l'institution ou de l'université.
- `city-dblp` et `country-dblp` : la ville et le pays de l'institution dans laquelle travaille l'auteur.
- `city-orcid` et `country-orcid` : la ville et le pays où vit l'auteur, obtenus grâce à l'enrichissement de données orcid.

Notre modèle vise à pallier le manque d'informations géographiques sur les auteurs de la base DBLP en ajoutant un enrichissement externe via orcid, afin de maximiser l'échantillon géographique sur lequel nous pouvons effectuer des requêtes.

La relation `Contribuer` lie un auteur à une publication en spécifiant son rôle, ce qui permet d'identifier sa contribution.
Grâce à cette relation et cette entité, les requêtes peuvent ainsi analyser efficacement la collaboration des scientifiques en intégrant la dimension géographique.

2.2.6 La table `Collaborations`

Structure de la table

La table `Collaboration` se compose des attributs suivants :
- `person1_id` : une clé étrangère référencée dans la table `Person`, représentant le premier collaborateur.
- `person2_id` : une clé étrangère référencée dans la table `Person`, représentant le second collaborateur.
- `publication_id` : une clé étrangère référencée dans la table `Publication`, identifiant la publication associée à la collaboration.

La clé primaire (`person1_id`, `person2_id`, `publication_id`) garantit l'unicité de chaque collaboration dans le contexte d'une publication donnée.

Justification du choix

La création de la table `Collaboration` répond à plusieurs objectifs précis :
- **Représentation explicite des collaborations** : chaque relation entre deux individus dans le cadre d'une publication est explicitement modélisée, ce qui facilite les analyses et les visualisations des réseaux de collaboration.
- **Pré-calcul pour des performances optimales** : en effectuant les calculs nécessaires sur une machine performante avant le déploiement, la table `Collaboration` est précalculée et incluse directement dans la base de données. Cela réduit considérablement le temps de traitement des requêtes ultérieures, rendant possible l'exploitation de la base même sur des infrastructures modestes.

— **Simplification du suivi des affiliations** : La structure de la table `Collaboration` facilite également le suivi des affiliations des contributeurs. Chaque collaboration étant explicitement définie, il devient plus simple d'identifier les collaborations inter-institutionnelles, suivre les affiliations des contributeurs dans le temps, effectuer des analyses croisées entre les affiliations, les publications et les collaborations.

Limites

Plusieurs chercheurs peuvent être associés à plusieurs publications, et réciproquement, une publication peut avoir plusieurs auteurs. Alors, la table Collaboration pourrait atteindre des dizaines de millions de lignes, car une publication avec **n auteurs** génère $\sum(n-1)$ **relations**.Cette explosion combinatoire représenterait une part importante du stockage totale et alourdit la base.

Exemple d'utilisation

Considérons l'exemple suivant :

person1_id	person2_id	publication_id
1	2	P101
1	3	P101
2	3	P101

TABLE 2.1 – Exemple d'enregistrements dans la table `Collaboration`

Dans cet exemple, les individus `1`, `2`, et `3` ont tous collaboré sur la publication `P101`. Grâce à la pré-calculation des collaborations, ces informations sont immédiatement disponibles sans nécessiter de traitement complexe à la volée.
En conclusion, la table `Collaboration` est un élément fondamental de notre modèle relationnel. En étant pré-calculée et bien structurée, elle offre une représentation claire, extensible et optimisée des relations entre les contributeurs d'une publication scientifique, tout en simplifiant la gestion et le suivi des affiliations.

2.3 Modèle en graphes : modélisation en Neo4j

Dans le cadre de l'analyse de l'impact de la pandémie de COVID-19 sur les collaborations scientifiques, nous avons adopté une double approche : une modélisation relationnelle classique et une modélisation en graphes avec **Neo4j**. Ces deux modèles sont complémentaires, chacun apportant une perspective différente et palliant les limites de l'autre. Cependant, l'approche en graphes offre des avantages uniques, notamment dans la compréhension des relations complexes et dynamiques entre les entités.

2.3.1 Complémentarité entre modèle relationnel et modèle en graphes

Le modèle relationnel a permis une structuration rigoureuse des données issues de *dblp.xml*, garantissant une gestion efficace des informations et des requêtes analytiques basées sur des dimensions bien définies, comme les statistiques géographiques ou temporelles. Cependant, certaines limites inhérentes à ce modèle ont justifié le recours à Neo4j :

— **Analyse des relations complexes** : Les collaborations entre chercheurs, qui impliquent souvent plusieurs dimensions (géographiques, institutionnelles, temporelles), sont difficiles à représenter et à interroger dans un modèle relationnel.
— **Navigation relationnelle** : Les relations indirectes, telles que les collaborations impliquant des auteurs affiliés à différentes institutions, sont plus intuitives à explorer dans un graphe.
— **Flexibilité des données** : La structure semi-structurée et hiérarchique du fichier *dblp.xml* s'intègre plus naturellement dans un modèle graphique.

En combinant les deux approches, nous avons pu profiter des forces respectives de chaque modèle :

— Le **modèle relationnel** excelle dans les requêtes agrégées et les analyses statistiques à grande échelle.
— Le **modèle en graphes** est particulièrement performant pour analyser les réseaux complexes et identifier des comportements/dynamiques.

2.3.2 Modélisation en graphes avec Neo4j

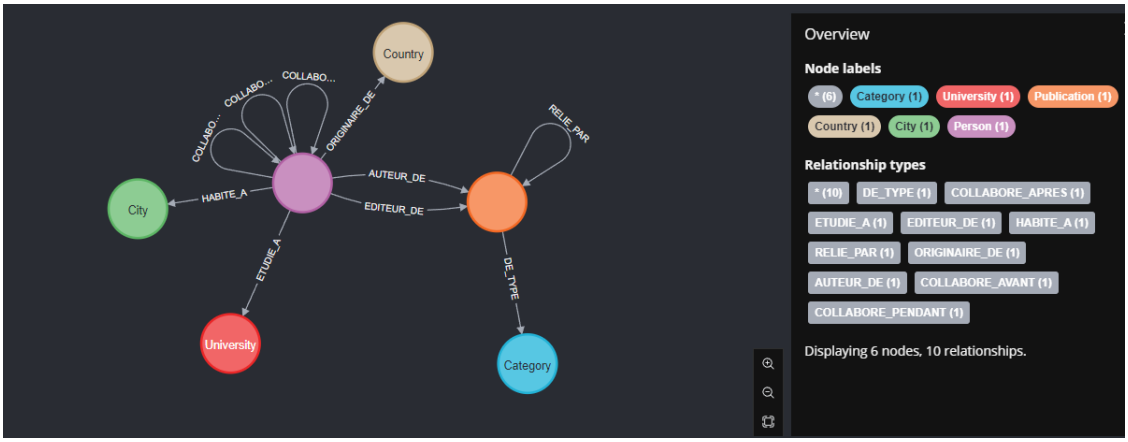


FIGURE 2.2 – schema bd neo4j

La modélisation en graphes dans Neo4j repose sur une représentation intuitive des entités et de leurs relations, mettant en avant les interactions complexes entre les auteurs, les publications et leurs affiliations. Voici les composants principaux :

Noeuds (Nodes)

— **Person** : Représente un auteur ou un éditeur.
  — *Propriétés* : `name`, `ORCID`.
— **Publication** : Une publication scientifique (article, conférence, livre, etc.).
  — *Propriétés* : `key`, `title`, `year`.
— **School** : Institution affiliée (université ou laboratoire).
  — *Propriétés* : `school_name`.
— **City** : Ville associée à l'institution.
  — *Propriétés* : `city_name`.
— **Country** : Pays correspondant à l'affiliation.
  — *Propriétés* : `country_name`.
— **Category** : Type ou catégorie de publication.
  — *Propriétés* : `category_name`.

Relations (Edges)

— **(Person)-[:AUTEUR_DE]->(Publication)** : Indique qu'un auteur a contribué à une publication.
— **(Person)-[:EDITEUR_DE]->(Publication)** : Relie un éditeur à une publication.
— **(Person)-[:DE_TYPE]->(Publication)** : Associe une publication à une Category.
— **(Person)-[:COLLABORER_AVANT->(Person)** : Modélise une collaboration explicite entre deux auteurs avant la pandémie (2017-2019).
— **(Person)-[:COLLABORER_PENDANT->(Person)** : Modélise une collaboration explicite entre deux auteurs pendant la pandémie (2020-2022).

— **(Person)-[:COLLABORER_APRES->(Person)** : Modélise une collaboration explicite entre deux auteurs après la pandémie (2023-2024).
— **(Person)-[:ETUDIE_A]->(School)** : Associe un auteur à une institution.
— **(Person)-[:HABITE_A]->(City)** : Associe un auteur à une ville.
— **(Person)-[:ORIGINNAIRE_DE]->(City)** : Associe un auteur à un pays.
— **(Publication)-[:RELIE_PAR]->(Publication)** : Relie une publication "enfant" (chapitre, inproceedings) à sa publication "parente" (livre, proceedings).

2.4 Synthèse et perspectives

La modélisation en graphes complète l'approche relationnelle en offrant une nouvelle perspective sur les collaborations scientifiques. Alors que le modèle relationnel reste pertinent pour les analyses agrégées, Neo4j brille par sa capacité à modéliser et analyser les interactions détaillées entre chercheurs, publications et dimensions géographiques. Cette complémentarité garantit une analyse approfondie et multidimensionnelle des données, en exploitant au mieux les points forts des deux approches.

2.4.1 Conclusion

Avant la pandémie

Les réseaux de collaboration scientifique avant le COVID-19 se développaient dans un cadre structuré et stable. Ces collaborations étaient principalement locales ou nationales, reposant sur des relations institutionnelles établies de longue date. Les thématiques scientifiques abordées dans ces réseaux étaient bien définies et stables, avec une progression graduelle des interactions internationales. Malgré leur croissance régulière, ces collaborations étaient souvent limitées par les contraintes géographiques et dépendaient fortement des rencontres en présentiel, comme les conférences et les ateliers. Pendant et après la pandémie

La pandémie a imposé des changements significatifs dans les réseaux de collaboration. Les restrictions physiques, comme les interdictions de voyage et la suspension des événements en présentiel, ont fragmenté certains réseaux établis. Cependant, les outils numériques ont permis aux chercheurs de maintenir leurs interactions, adaptant ainsi leurs pratiques à cette nouvelle réalité. Cette période a également vu la création de nouveaux réseaux centrés sur des thématiques directement liées à la crise, comme la santé publique, l'épidémiologie, et les technologies numériques.

Après la pandémie, ces nouvelles dynamiques se sont consolidées. Les collaborations initiées pendant cette période ont évolué pour former des communautés scientifiques plus larges et plus diversifiées. Ces réseaux post-pandémie se distinguent par leur résilience et leur adaptabilité, témoignant de la capacité de la communauté scientifique à répondre efficacement à des défis mondiaux. Comparaison globale

En comparant les réseaux avant et après la pandémie, plusieurs évolutions majeures se dessinent. Sur le plan de l'échelle, les collaborations internationales ont surpassé les interactions locales, devenant le moteur principal des dynamiques scientifiques. En ce qui concerne les thématiques, les réseaux post-pandémie sont marqués par une focalisation accrue sur des problématiques interdisciplinaires, impliquant des échanges entre plusieurs domaines scientifiques. Enfin, sur le plan structurel, la période post-pandémie a favorisé l'émergence de réseaux plus flexibles et agiles, adaptés aux nouvelles pratiques numériques, contrastant avec les réseaux rigides qui prévalaient avant la pandémie.

Chapitre 3

Analyse des évolutions des collaborations

Les données nettoyées, les modèles établis, les bases de données créées et les insertions réalisées, nous procéderons dans ce chapitre à l'identification et l'analyse des différents réseaux de collaboration en fonction des axes préalablement définis.

3.1 Vision globale : fonction d'agrégation sur la base de données relationnelle

Cette section est dédiée à des observations élémentaires et nécessaires pour se faire une vision d'ensemble de la base. Elle repose sur des requêtes utilisant majoritairement des fonctions d'agréation, `count(*)`, `avg()`, `max()`

COUNT(*)

Nombre total de publications

Nous disposons de **7 578 943** publications, la plus ancienne datant de 1936 et les plus récentes de 2025. Nous sommes face à un volume massif qui reflète la richesse de la production scientifique depuis le début de XXe siècle.

Proportions des catégories de publications

Voici les proportions par catégorie dans la BD :

- **article** : 3 706 535 (48,91%)
- **inproceedings** : 3 600 854 (47,51%)
- **phdthesis** : 139 895 (1,85%)
- **incollection** : 59 875 (0,79%)
- **proceedings** : 43 606 (0,58%)
- **book** : 18 777 (0,25%)
- **data** : 9 363 (0,12%)
- **mastersthesis** : 27 (<0,01%)
- **www** : 11 (<0,01%)

- **Observations** : Les catégories de publications montrent une forte disparité. Les articles et inproceedings dominent largement, représentant 96,4 % des 7 578 943 publications. Les articles (3 706 535, soit 48,9 %) sont les plus nombreux, tandis que des catégories comme `www` et `phdthesis` restent marginales. Le volume des publications « enfants » dépasse nettement celui de leurs « parents », reflétant une activité intense dans des formats fractionnés, notamment pour les conférences.

— **Analyse** : Les articles et actes de conférences sont surreprésentés par rapport aux thèses de doctorat. Les articles, essentiels pour la reconnaissance académique, permettent de diffuser des recherches détaillées et validées. Les inproceedings, quant à eux, favorisent une diffusion rapide et les échanges en conférences, surtout dans des domaines comme l'informatique. À l'inverse, les thèses sont peu visibles car rarement publiées sous leur format original, leurs résultats étant souvent transformés en articles ou présentés en conférences. Cela montre que la communauté scientifique mondiale a une préférence pour des formats courts, collaboratifs et adaptés à un rythme de recherche rapide et une diffusion plus large des résultats.

Nombre de contributeurs (auteurs et/ou editeurs)

Nous avons **4 591 275** chercheurs et parmi ces chercheurs nous observons que
— **4 534 751** sont uniquement auteurs
— **2 230** sont uniquement éditeurs
— **54 294** sont auteurs et éditeurs

AVG(*)

Moyenne de collaborations par publication $\simeq 7{,}5$

Une moyenne de 7,5 collaborations par publication indique une forte tendance aux travaux collectifs, reflétant la complexité croissante des projets scientifiques nécessitant des expertises variées.

Moyenne de publications par auteur $\simeq 5{,}5$

Un auteur contribue en moyenne à 5,5 publications. Cela pourrait indiquer un équilibre entre les chercheurs très prolifiques et ceux ayant une production plus modeste.

MAX(*)

Année avec le plus de publications : 2023

nombre : 516 634

Année avec le plus de collaborations : 2024

nombre : 5 415 572

Pays avec le plus de collaborations : Etats-Unis

nombre : X

3.2 Présentation des résultats : méthodologie

L'identification des axes et des sous-problématiques associées nous a permis de concevoir des vues et des requêtes adaptées. Ces dernières ont permis d'extraire des résultats sous forme de fichiers CSV. À partir de ces fichiers, nous avons développé nos propres scripts Python pour produire des graphiques. Ces graphiques serviront de support à nos analyses.

3.3 Études des tendances générales

3.3.1 Évolution du nombre de publications dans le temps

Question de départ : De manière générale, le nombre de publications a-t-il augmenté à travers le temps ?

Hypothèse : En ce qui concerne l'effet spécifique de la pandémie de Covid-19, nous anticipons un pic significatif du nombre de publications durant cette période, principalement dû à une mobilisation sans précédent dans des domaines tels que la santé publique et l'épidémiologie.

Graphique : Évolution du nombre de publications (2017-2024) Ce graphique présente ci-dessous l'évolution du nombre de publications par année, classées selon les périodes définies en amont :

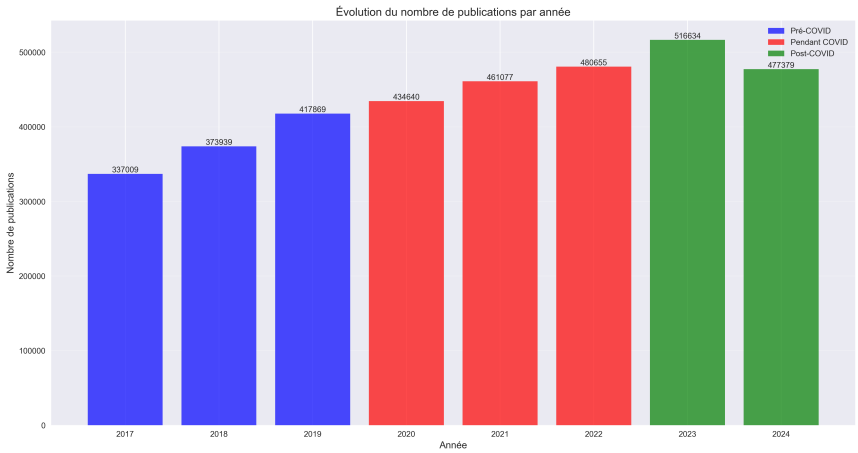


FIGURE 3.1 – Évolutions du nombre de publications par périodes, de 2017 à 2024

Description des résultats observés

Période	Observations
3 ans avant (2017-2019)	Ces années pré-pandémiques montrent une accélération des publications. En moyenne 376 272 publications par an.
Covid (2020-2022)	Un pic marqué dans les publications, en ligne avec l'hypothèse d'une intensification des efforts scientifiques pendant la pandémie. En moyenne 458 791 publications par an.
Post-Covid	Une légère augmentation suivi d'une stabilisation du nombre de publications. En moyenne 497 006 publications par an.

Bilan : La Covid-19 a intensifié les efforts scientifiques, lié à une mobilisation exceptionnelle pour répondre à la crise. L'augmentation du nombre de publications post-covid suivie d'une diminution légère, mais supérieure à la période covid suggère 2 choses : d'une part, un retour à la normale et d'autre part, que le covid a entrainé une augmentation de la diffusion de contenu

scientifique des sur des plateformes collaboratives comme la DBLP .

3.3.2 Évolution des catégories des publications dans le temps

Question de départ : Comment ont évolué les formats des contenus scientifiques publiés à travers le temps ?

Hypothèses :

1. Certaines catégories ont pu être impactées par la Covid-19.
2. Les data et formats numériques devraient augmenter dans le temps.

Graphique : Ce graphique présente ci-dessous l'évolution du nombre de publications par année, classées selon les périodes définies en amont :

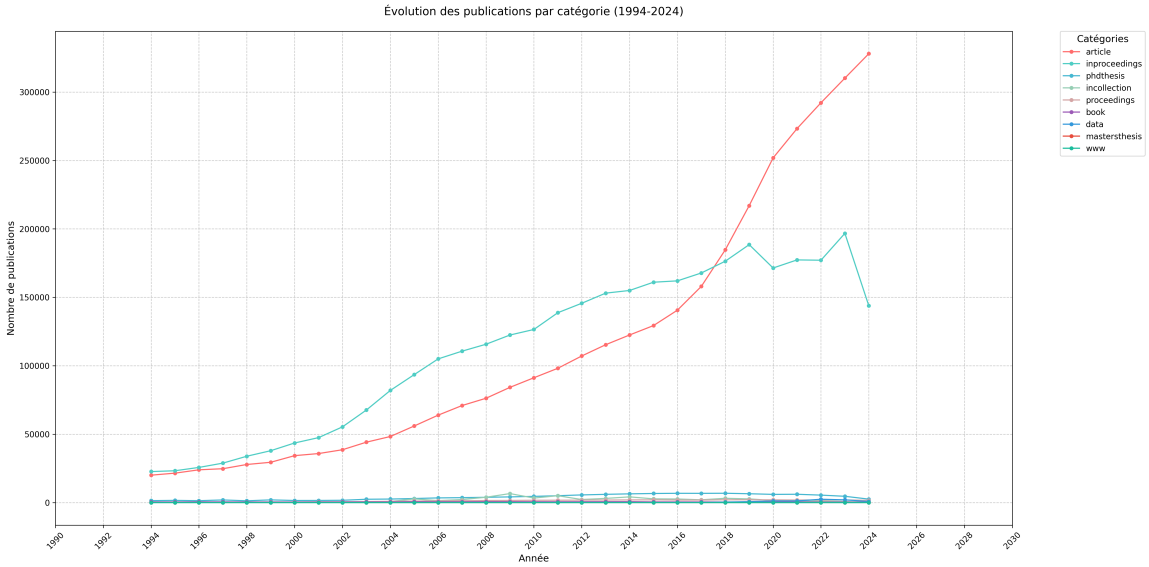


FIGURE 3.2 – Évolution des catégories de publications dans le temps

Description des résultats observés

On observe que le COVID a eu un impact sur les formats de publications scientifiques :

Évolutions majeures

1. **Formats en augmentation**
   - **Articles** :
     → Le nombre d'articles publiés a connu un pic pendant la période du COVID (**825 000** articles contre **565 000** les trois années précédentes).
     → Stabilisation après la pandémie (**650 000** articles).
   - **Données (Data)** :
     → période pré-pandémie : **700 publications** en 2019.
     → pandémie : pic à **4 900 publications** en 2022.
     → post-pandémie : augmentation à **2 800** en 2023.

2. **Formats en diminution**
   - **Incollection** : Baisse importante pendant le COVID (**7 900 sur les 3 années précédentes, 3 500 pendant le COVID, 1 000 depuis la fin du COVID**).
   - **PhD Thesis (Thèses de doctorat)** : Diminution constante depuis 2021 (**-1 000 par an environ**).

— **Inproceedings** : On observe une stagnation du nombre de publications pendant le COVID puis une augmentation de 2022 à 2023 puis une baisse importante.

3. **Diminution déjà amorcée avant le COVID**
   — **Livres** : Déclin progressif depuis 2015.
   — **Proceedings)** : Stabilité observée avant, pendant et après (**1 800-2 100 par an**).
4. **Stabilisation** :
5. **Inproceedings** : Malgré les restrictions liées au COVID, les publications dans cette catégorie sont restées globalement stables.

Bilan : Le COVID-19 a profondément transformé les dynamiques de publication scientifique, et ces changements ont perduré bien au-delà de la pandémie, témoignant d'une adoption durable de nouvelles pratiques par la communauté scientifique. Les augmentations observées, notamment pour les articles et les publications de données (data), reflètent l'urgence de partager rapidement des résultats de recherche pendant la pandémie. Ces formats, favorisés par les outils collaboratifs et le télétravail, ont continué à croître après le COVID, montrant que ces pratiques ont été intégrées et intériorisées par les chercheurs. À l'inverse, les diminutions observées pour les incollections et les thèses de doctorat se sont poursuivies après la pandémie. Cela suggère que la communauté scientifique a conservé les nouvelles habitudes de travail initiées pendant le COVID, privilégiant des formats plus adaptés à la collaboration à distance et à la diffusion rapide des résultats. Les inproceedings, en revanche, illustrent une résilience particulière. Malgré les perturbations, la communauté scientifique a su s'adapter rapidement en maintenant les conférences grâce à des solutions numériques. Cette capacité d'adaptation a permis de limiter l'impact du COVID sur ce type de publication. Ces évolutions, tant dans les augmentations que dans les diminutions, montrent clairement que les pratiques adoptées pendant le COVID ne sont pas restées temporaires. Elle ont durablement transformé les modes de publication scientifique, en favorisant des formats plus collaboratifs et agiles, adaptés aux contraintes et opportunités du travail numérique.

3.3.3 Évolution du nombre de publications et évolution du nombre de publications collectives

Question de départ : Visualiser la proportion des publications collaboratives et son évolution en fonction des périodes stratégiques.

Hypothèse : La pandémie de COVID-19 a considérablement influencé le nombre de publications collaboratives, provoquant soit un ralentissement, soit une augmentation.

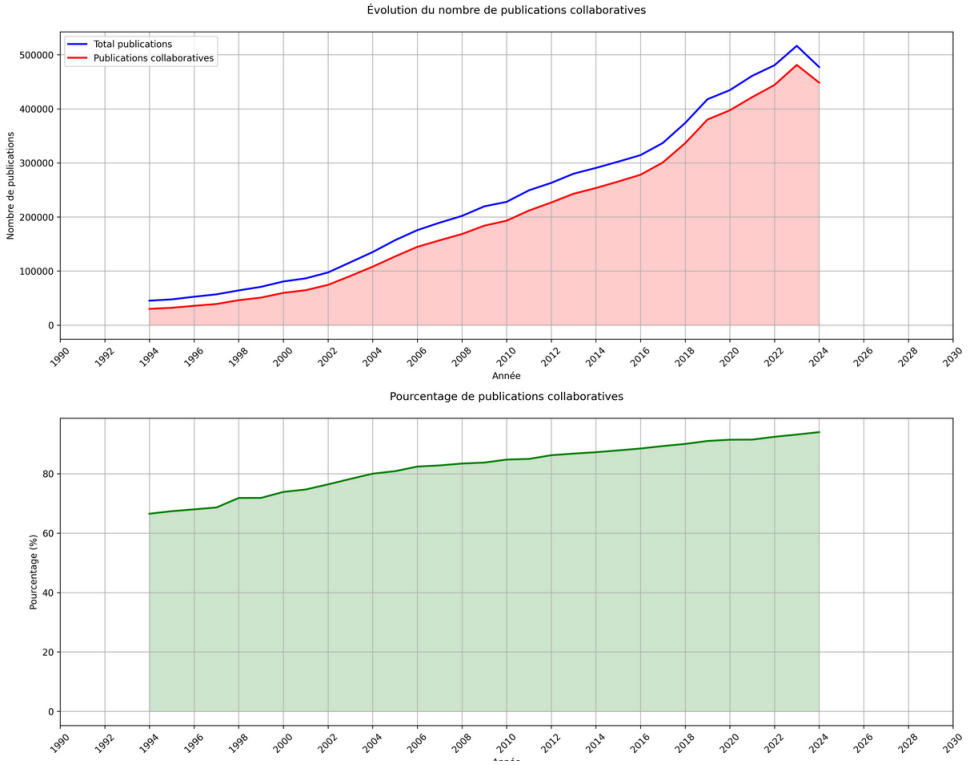


FIGURE 3.3 – Évolution du nombre de publications collaboratives dans le temps (1994-2024)

Graphiques : Évolution du nombre de publications collaboratives dans le temps (1994-2024) : Ce graphique illustre l'évolution du nombre de publications collaboratives (en rouge) par rapport au total des publications (en bleu). Le pourcentage de publications collaboratives est également présenté sur un second graphique pour montrer la part croissante des collaborations dans la production scientifique globale.

Observations : De manière générale, on a un nombre de publications d'un seul auteur bien moins important que le nombre de publications collaboratives car on observe que les courbes sont très proches. Le nombre de publications en collaboration suit les tendances des évolutions du nombre de publications totales

Description des résultats observés

Période	Observations pour les deux courbes (bleue et rouge)
3 ans avant (2017-2019)	Les courbes montrent une **croissance régulière et linéaire**. Le nombre total de publications (courbe bleue) et les premières collaborations (courbe rouge) progressent parallèlement, avec une proportion stable autour de **90 %** de collaborations dans le total des publications.
Covid (2020-2022)	Une **augmentation modérée** est visible sur les deux courbes. La courbe bleue (total des publications) et la courbe rouge (collaborations) continuent de croître, mais sans bond significatif. La proportion de collaborations augmente légèrement.
Post-Covid (2023-2024)	Les deux courbes atteignent un **pic en 2023**, marquant le maximum historique du nombre total de publications en collaboration. En 2024, une **stabilisation à un niveau élevé** est observée, avec les collaborations représentant toujours plus de **95 %** des publications totales.

Bilan : Le COVID-19 n'a pas entraîné de ralentissements majeurs dans les dynamiques de collaborations scientifiques, qui ont continué de croître de manière stable pendant la pandémie. Le pic observé en 2023 peut être interprété comme un regain de rassemblement post-pandémie, favorisé par une reprise complète des activités scientifiques et une meilleure disponibilité des données. Ce pic pourrait également s'expliquer par un ralentissement des processus de publication pendant la pandémie, certaines collaborations ayant pu se concrétiser plus tôt mais n'être publiées qu'après la crise. Cela illustre un effort scientifique accru pour étudier les impacts de la pandémie et explorer de nouveaux axes collaboratifs.

3.3.4 Proportion d'auteurs par pays

L'objectif de cette visualition est d'identifier les zones géographiques actives : nombre d'auteurs par pays.
Question : Quelle est la répartition géographique des contributeurs ?
Hypothèse : Le COVID-19 a pu entrainer une augmentation de nouvelles collaborations scientifiques.

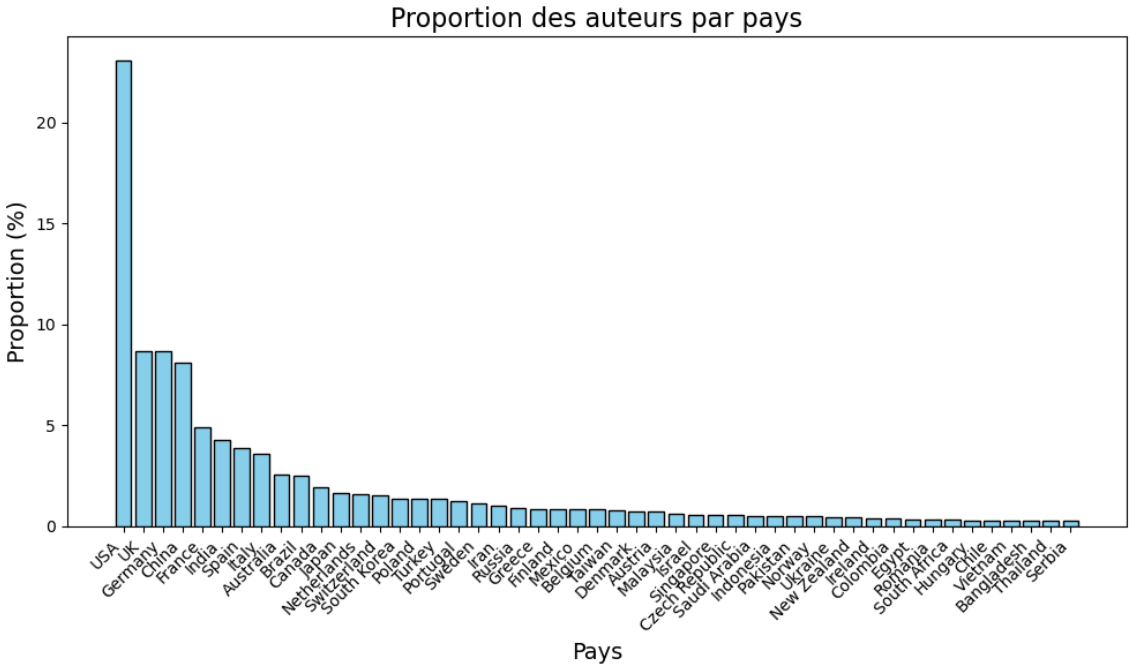


FIGURE 3.4 – Proportion des affiliations des auteurs par pays

Graphique : Proportion des affiliations des auteurs par pays. Deux points sont importants à noter avant l'analyse :

1. **Taille de l'échantillon** : Nous disposons de 15 % des affiliations des auteurs soit 700 000 auteurs, et cet échantillon est considéré comme représentatif des affiliations totales.
2. **Limitation** : Le graphique se limite aux 50 pays les plus représentés, et n'inclut donc pas tous les pays contributeurs.

Description des résultats observés

La répartition des affiliations des auteurs met en lumière des tendances géographiques et économiques significatives :

— **Dominance des États-Unis :** Les États-Unis représentent environ 23 % des affiliations (160 000 auteurs), reflétant leur position de leader mondial en recherche scientifique. Ce

résultat s'explique par la présence d'universités prestigieuses (Harvard, MIT, Stanford) et d'institutions majeures (NIH, NSF, NASA), ainsi que par des financements massifs. Le pays attire des chercheurs internationaux grâce à ses infrastructures avancées et son environnement propice à l'innovation.

— **Le Royaume-Uni et les pays d'Europe occidentale :** Le Royaume-Uni et l'Allemagne occupent chacun environ 10% (70 000 auteurs) des affiliations, juste derrière les États-Unis.
  → Le Royaume-Uni est porté par ses universités prestigieuses, comme Oxford et Cambridge.
  → L'Allemagne s'appuie sur des instituts renommés tels que la Max-Planck-Gesellschaft et la Fraunhofer-Gesellschaft.

  D'autres nations européennes comme la France, l'Italie et l'Espagne figurent également dans le graphique, bien que dans des proportions légèrement inférieures (moins de 5 % chacune, 56 000 auteurs) , occupent des positions importantes, reflétant une tradition scientifique riche et des efforts constants pour soutenir la recherche. Ensemble, ces pays renforcent le rôle de l'Europe occidentale comme un pôle clé de la production scientifique mondiale.

— **Les BRICS dans la recherche scientifique :**
  → La Chine, 8% (56 000 auteurs), se distingue par son développement rapide en innovation technologique et ses investissements massifs.
  → L'Inde, grâce à ses centres d'excellence (IIT) et sa population importante, montre une croissance significative dans la production scientifique.
  → Le Brésil, principal acteur en Amérique du Sud, se classe dans le top 10 grâce à ses efforts pour développer ses capacités scientifiques.
  → La Russie, moins visible, se concentre sur des secteurs stratégiques (défense, énergie, espace) peu présents dans DBLP, et souffre des répercussions des conflits géopolitiques.

— **L'Afrique :** Les contributions africaines sont limitées. L'Égypte et l'Afrique du Sud figurent néanmoins dans le classement, reflétant une production scientifique concentrée dans certaines régions spécifiques du continent.

Bilan : En résumé, les affiliations des auteurs dans la base DBLP mettent en évidence une forte concentration autour des grandes puissances scientifiques : les États-Unis dominent avec 23 % des affiliations, suivis par le Royaume-Uni, l'Allemagne (10 % chacun), et la Chine (8 %). Les autres pays, majoritairement en dessous de 2 %, soulignent les disparités importantes entre les leaders de la recherche et le reste du monde.

3.4 Etudes liées aux collaborations

Nous étudierons dans cette partie les dynamiques de collaborations entre les chercheurs.

3.4.1 Évolution du nombre de collaborations

Rappellons que nous comptabilisons comme collaboration toute association de deux auteurs pour la création de contenu scientifique. Ainsi une publication avec **n auteurs** génère $\sum(n-1)$ collaborations.

Question de départ : Comment le nombre de collaborations scientifiques a-t-il évolué dans le temps ?
Hypothèse : Le COVID-19 a eu un impact significatif sur le nombre de collaborations scientifiques, provoquant une augmentation notable pendant cette période.
Graphique : Le graphique repésente l'évolution du nombre de collaborations (locales et internationnales confondues) sur les vingt dernières années.

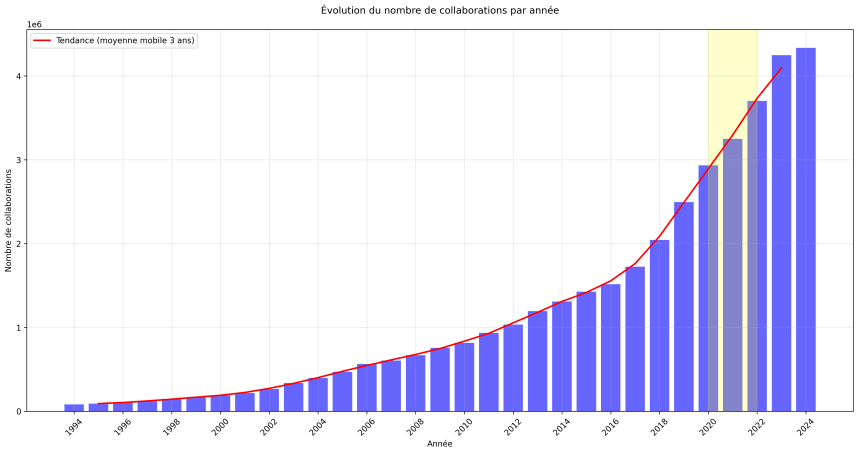


FIGURE 3.5 – Evolution du nombre de collaborations

Description des résultats observés

Période	Observations sur les collaborations scientifiques
3 ans avant (2017-2019)	Le nombre de collaborations montre une **croissance régulière**, atteignant une moyenne de **2,8 millions de collaborations par an**. Cette période est marquée par une dynamique scientifique déjà bien établie.
Pendant la pandémie (2020-2022)	Une **augmentation modérée** est observée, avec une moyenne annuelle d'environ **3,7 millions de collaborations**, soit une hausse de **30 %** par rapport à la période précédente. Cela reflète une mobilisation accrue pour répondre à l'urgence sanitaire.
Post-Covid (2023-2024)	Le nombre de collaborations atteint un **pic en 2023**, avec environ **4,4 millions de collaborations par an**. En 2024, une **stabilisation à ce niveau élevé** est visible. Par rapport à la période avant COVID, on observe une augmentation globale de **57 %** en moyenne.

Bilan : L'analyse des collaborations scientifiques au cours de la période étudiée révèle des tendances intéressantes, notamment en lien avec la pandémie de Covid-19. L'augmentation continue du nombre de collaborations observée pendant la pandémie montre que ces dernières

se sont non seulement maintenues malgré les contraintes liées à cette période, mais qu'elles ont suivi une tendance déjà établie d'augmentation progressive. Après le Covid-19, cette dynamique s'est encore renforcée, avec une augmentation du nombre de collaborations par rapport à la période précédente, ce qui illustre la résilience et l'adaptabilité des chercheurs face aux défis de la période pandémique.

Le pic observé en 2023 peut s'expliquer de deux manières. Si l'on considère que les collaborations répertoriées pour des publications parues en 2023 dans la base DBLP reflètent des travaux menés en amont (compte tenu du temps nécessaire à la production scientifique), il est possible qu'une partie significative de ces collaborations ait été initiée pendant la pandémie. Cela pourrait suggérer que 2022, en réalité, aurait été une année de forte activité collaborative, où les scientifiques, malgré les contraintes, ont intensifié leurs efforts pour produire du contenu scientifique. Cette hypothèse implique que le Covid-19 aurait favorisé les collaborations pendant cette période.

Pour vérifier cette hypothèse, nous avons effectué un test du chi^2 pour évaluer la dépendance entre le fait qu'une publication soit co-éditée et qu'elle ait été publiée après le Covid-19.

Test statistique d'indépendance :

Explication du calcul : Le **test du chi-deux** permet d'analyser la relation entre deux variables catégorielles en comparant les fréquences observées à celles attendues si les deux variables étaient indépendantes. Dans ce contexte, nous examinons la relation entre la co-écriture des publications (co-écrite ou non) et la période de publication (avant ou après la pandémie de COVID-19). Le test calcule une statistique pour déterminer si les différences entre les fréquences observées et attendues sont suffisamment grandes pour être considérées comme significatives. Ce test est particulièrement utile pour déterminer s'il existe une association entre ces deux variables, mais il peut être sensible à la taille des échantillons et aux petites catégories.

Objectif : L'objectif est d'analyser si la co-écriture des publications est liée à la période de publication (avant ou après la pandémie de COVID-19).

Question : La co-écriture des publications est-elle significativement liée à la période avant ou après la pandémie de COVID-19 ?

Hypothèse : L'hypothèse est que la pandémie de COVID-19 a influencé la co-écriture des publications, créant une relation significative entre la co-écriture et la période de publication.

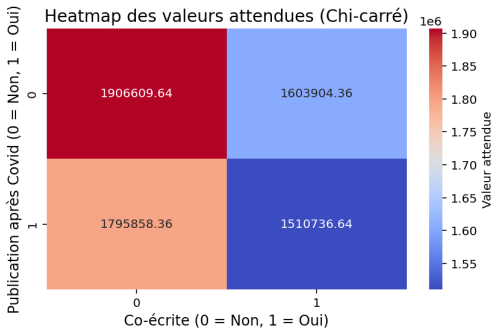


FIGURE 3.6 – Tableau de contingence du test χ^2

Statistique	Valeur
Chi-deux (χ^2)	23656.06
p-value	0.00
Degré de liberté (df)	1

TABLE 3.1 – Résultats du test du chi-deux

Observations : Le test du chi-deux a révélé une statistique χ^2 de 23656.06 avec une p-value quasi-nulle de 0.00, ce qui est bien inférieur au seuil de significativité classique de 0.05.Cette valeur de p-value s'explique car l'échantillon est très grand. Ces résultats indiquent que la co-écriture des publications est significativement liée à la période de publication (avant ou après la pandémie de COVID-19). En d'autres termes, il existe une association statistiquement significative entre ces deux variables, suggérant que la pandémie a influencé la manière dont les publications ont été co-écrites.

Bilan : Le test du chi-deux révèle que la co-écriture des publications est significativement influencée par la période de publication. En effet, les publications co-écrites sont nettement plus fréquentes après la pandémie, ce qui suggère que la crise sanitaire a facilité ou renforcé la collaboration internationale. Ce phénomène pourrait être dû à une plus grande coopération mondiale en réponse aux défis imposés par la pandémie. Les résultats montrent une tendance marquée vers l'augmentation des publications collaboratives à l'échelle internationale.

3.4.2 Evolution du nombre de premières collaborations avant, pendant et après le COVID

Question : A quelle période les collaborateurs ont-ils travaillé avec d'autres pour la première fois ?
Hypothèse : Le COVID-19 a pu entrainer une augmentation de nouvelles collaborations scientifiques.
Graphique : Évolution du nombre de nouveaux collaborateurs dans le temps

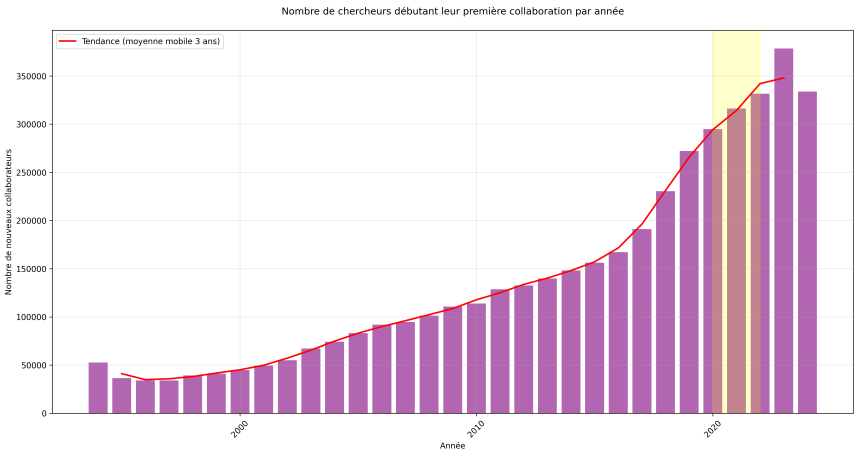


FIGURE 3.7 – Évolution du nombre de nouveaux collaborateurs dans le temps

Description des résultats observés

Ce graphique présente nombre de nouveaux collaborateurs par année, en mettant en évidence les périodes avant, pendant et après le COVID.

Période	Observations sur les premières collaborations
3 ans avant (2017-2019)	Durant cette période, le nombre de chercheurs débutant leur première collaboration est stable, avec une moyenne d'environ **250 000 à 275 000 nouveaux collaborateurs par an**. Cette stabilité reflète une dynamique bien installée avant la pandémie.
Pendant le Covid (2020-2022)	Une **augmentation notable** est observée, atteignant une moyenne de **300 000 à 325 000 nouveaux collaborateurs par an**. Cela représente une hausse d'environ **50 000 collaborateurs supplémentaires par an** par rapport à la période précédente.
Post-Covid (2023-2024)	Un **pic est atteint en 2023**, avec environ **350 000 nouveaux collaborateurs**, marquant un maximum historique. En 2024, une légère baisse est observée, mais le nombre reste élevé, autour près de **380 000 nouveaux collaborateurs**.

Bilan Le pic de 2023 s'inscrit dans une tendance générale à l'augmentation du nombre de nouvelles collaborations, une dynamique qui s'est maintenue malgré la crise sanitaire. Pendant le COVID, les chercheurs ont continué à initier des collaborations grâce aux outils numériques, favorisant ainsi une augmentation stable des premières collaborations même en période de restrictions.Ce pic peut également s'expliquer par la concrétisation, après la pandémie, de nombreuses collaborations initiées mais retardées par les contraintes sanitaires. Le retour des événements en présentiel en 2023 a aussi permis de renforcer ces dynamiques et d'encourager de nouvelles rencontres scientifiques. Enfin, les données accumulées pendant la pandémie ont permis d'impliquer davantage de chercheurs dans des projets collaboratifs. En résumé, le COVID n'a pas freiné la tendance à l'augmentation des premières collaborations et a même contribué à la préparer.

3.4.3 Évolution du nombre de collaborations internationales

Question : Comment les collaborations internationales ont-elles évoluées ?
Hypothèse : Avec le développement des moyens d'échanges numériques, les collaborations internationales ont certainement augmenté fortement. De plus, la crise sanitaire, ayant créé un défi commun aux pays du monde entier, a certainement boosté le nombre de ces collaborations.
Graphique : Évolution du nombre de publications issues de collaborations internationales dans le temps (1994-2024). Il s'agit de l'ensemble du nombre de publications co-écrites provenant de différents pays par an de 1994 à 2024.

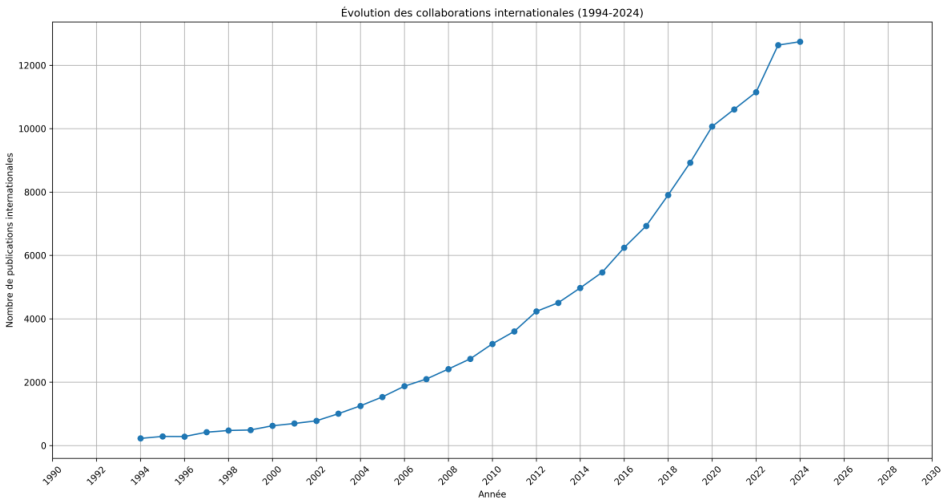


FIGURE 3.8 – Évolution du nombre de collaborations internationales

Observations :

- Entre 1994 et 2002, la croissance est relativement lente. Le nombre de collaborations internationales augmente de manière modérée
- A partir de 2002, on observe une croissance accélérée du nombre de collaborations internationales, avec une pente plus raide, ce qui coïncide avec l'adoption massive d'Internet et le développement des outils de collaboration en ligne (mails, conférences virtuelles, bases de données numériques).
- Ensuite, on observe une pente encore plus raide qu'avant à partir de 2016, soit peu avant la période du Covid jusqu'en 2020. Cela peut s'expliquer car les outils de visioconférence et de collaboration numérique (comme Zoom, Microsoft Teams) deviennent plus performants et largement adoptés. De même les investissements dans la recherche mondiale, en particulier dans des domaines stratégiques comme l'intelligence artificielle, augmentent, comme mentionné lors de notre séminaire sur l'IA générative et ChatGPT.
- De 2020 à 2022, période du Covid, on s'aperçoit que la croissance des collaborations se poursuit mais avec un rythme moins soutenu qu'avant 2020 et se stabilise de 2023 à 2024.

Bilan : On peut donc en conclure que c'est principalement le développement des moyens de télécommunications et des plateformes d'échanges numériques qui ont contribué à l'augmentation du nombre de collaborations internationales et que ces moyens, encore en développement continuent à les favoriser. Cependant, on s'aperçoit que, contrairement à l'hypothèse qu'on avait au début, le Covid a été un frein temporaire aux publications des travaux de collaboration internationale.
A noter que sur cette analyse, on regarde le nombre de publication de collaborations internationales entre pays de l'ONU. La remontée post-Covid du nombre de publication en collaborations internationales peut également s'expliquer par la politique post-Covid de l'ONU dont notamment le financement de la communauté scientifique mondiale.

3.4.4 Evolution du nombre de publications internationales et locales

Objectif : Cette visualisation est d'observer l'évolution des collaborations locales par rapport aux collaborations internationales dans le temps.

Question : Y-a-t-il eu une augmentation plus importante des collaborations internationales par rapport aux contributions locales pendant le Covid ?
Hypothèse : On suppose qu'il y a eu une augmentation des collaborations internationales pendant le Covid car la crise a représenté un défi à l'échelle mondiale.
Graphique : Évolution du nombre de publication issues de collaborations internationales dans le temps (1994-2024). Il s'agit de l'ensemble du nombre de publications co-écrites provenant de différents pays par an de 1994 à 2024.

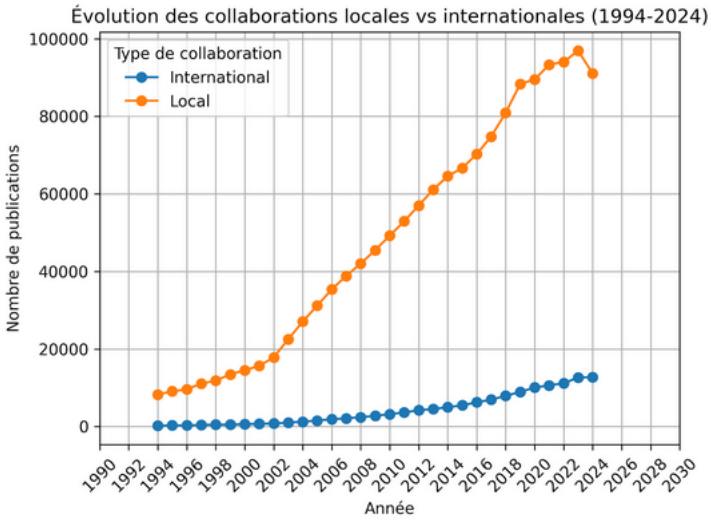


FIGURE 3.9 – Evolution du nombre de publications internationales et locales (1994-2024)

Observations :

1. **Échelle locale : collaborations**
   - Depuis 2002, les collaborations locales ont connu une croissance significative, passant de 20 000 publications issues de collaborations entre auteurs d'un même pays à 80 000 en 2018.
   - Cependant, cette progression ralentit à partir de 2020, pour finalement observer une diminution du nombre de collaborations locales en 2023.
2. **Échelle internationale**
   - Bien que moins nombreuses, les collaborations internationales affichent une progression constante depuis 2008, passant de 0 en 1994 à près de 15 000 en 2024.
   - En 2018, on comptabilisait 8 000 publications issues de collaborations entre auteurs de pays différents, un chiffre qui a doublé après la période de la pandémie de COVID-19.

Bilan : Une corrélation semble se dessiner entre ces dynamiques : à partir de 2020, la stagnation, voire la légère diminution des publications locales, coïncide avec l'augmentation des collaborations internationales. Cette tendance indique l'émergence d'une nouvelle dynamique dans la communauté scientifique, qui privilégie davantage les collaborations internationales pour la production de contenus scientifiques, particulièrement à partir de la période post-COVID.

3.4.5 Diversité géographique des collaborations

Indice de diversité de Simpson :

Explication du calcul : L'indice de diversité de Simpson mesure la probabilité que deux collaborations choisies au hasard appartiennent à des catégories géographiques différentes. Il met davantage l'accent sur les catégories dominantes, ce qui peut exagérer l'importance des régions ayant une forte proportion de collaborations. En conséquence, il peut sous-estimer la contribution des petites catégories géographiques. Cela en fait un outil utile mais pouvant être biaisé dans des contextes de forte disparité.

Objectif : L'objectif est d'analyser l'évolution de la diversité géographique des collaborations scientifiques, en mesurant l'indice de diversité de Simpson pour chacune des périodes identifiées.
Question : Y a-t-il une diversité internationale distincte dans les collaborations ?

Hypothèse : L'hypothèse est que la pandémie de COVID-19 a limité les collaborations internationales, au profit des collaborations nationales, et faisant ainsi reculer la diversité internationale des collaborations.

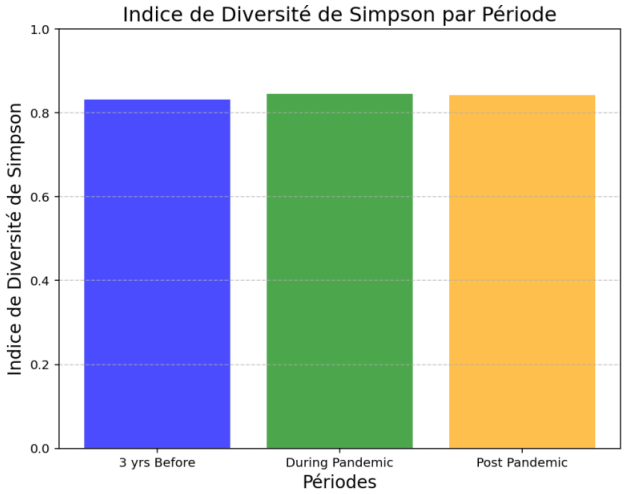


FIGURE 3.10 – Histogramme des indices de diversité de Simpson par période.

Bilan : Nos observations viennent contredire notre hypothèse de départ qui stipulait que la covid affectait la diversité internationale des collaborations scientifiques. Cependant, étant donné les faibles différences entre les indices de diversité de Simpson, nous ne pouvons pas conclure que la covid ait beaucoup impacté la diversité internationale des collaborations.

Indice de diversité de Shannon :

Explication du calcul : L'indice de diversité de Shannon évalue la diversité d'un ensemble en tenant compte à la fois du nombre de catégories et de la répartition des éléments entre elles. Contrairement à l'indice de Simpson, il n'accorde pas une importance particulière aux catégories dominantes, ce qui permet de mieux capturer la diversité dans des ensembles plus équilibrés. Cet indice est particulièrement utile pour observer la diversité globale, mais il peut être sensible aux petites catégories, qui peuvent influencer plus ou moins fortement le score selon leur proportion dans l'ensemble.

Objectif : L'objectif est de mesurer différemment la diversité internationale des collaborations par période.
Question : La diversité internationale des collaborations varie-t-elle ?
Hypothèse : L'hypothèse est que la pandémie de COVID-19 a limité les collaborations internationales, la diversité internationale diminue donc lorsque la pandémie advient.

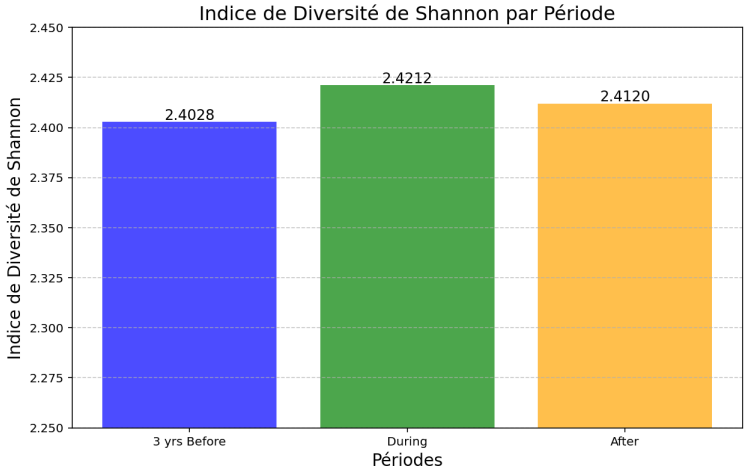


FIGURE 3.11 – Histogramme des indices de diversité de Shannon par période.

Observations : Nous observons une différence plus prononcée dans ce test de diversité de Shannon que dans celui de Simpson. On remarque bien sur cet histogramme que la diversité internationale a augmenté significativement lors de la pandémie, puis a diminué une fois l'épisode terminé.

Bilan : Pendant la pandémie, la diversité internationale a significativement augmenté, comme le montre l'indice de Shannon, probablement en raison de l'adoption des outils numériques et de la mobilisation exceptionnelle des compétences internationales sur des enjeux de santé. Après la pandémie, une diminution est observée, suggérant un retour progressif à des thématiques traditionnelles. Toutefois, les niveaux restent supérieurs à ceux d'avant la pandémie.

Test de proportion Z :

Explication du calcul : Le test de proportion Z permet de comparer les proportions d'un phénomène (ici, les collaborations internationales) entre différentes périodes. Il évalue si la différence observée entre les proportions avant, pendant et après la pandémie est significative. Ce test prend en compte l'échantillon de données et les écarts-types des proportions pour déterminer si la variation observée dépasse ce qui pourrait être attribué au hasard. Il est particulièrement utile pour tester des hypothèses sur les changements dans les proportions.

Objectif : L'objectif est d'analyser l'évolution des propotions internationales, avant, pendant et après la pandémie de COVID-19.
Question : La pandémie a-t-elle entraîné une réduction significative des collaborations internationales ?
Hypothèse : L'hypothèse est que la pandémie de COVID-19 a réduit les collaborations internationales, du fait des restrictions et confinements.

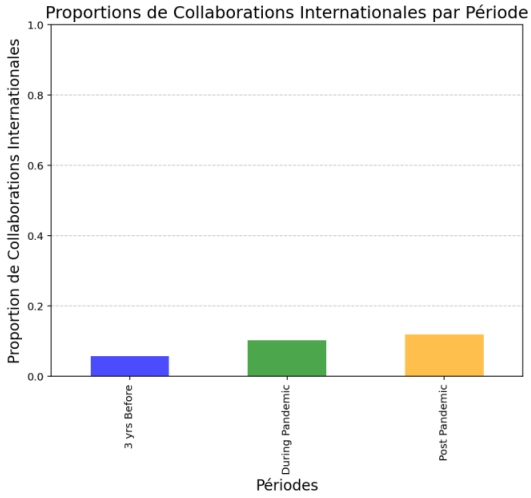


FIGURE 3.12 – Histogramme des proportions des collaborations internationales par période.

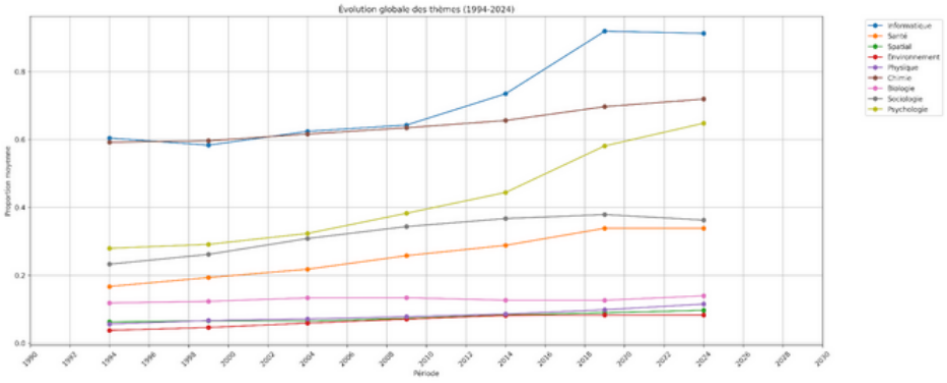
Observations :
- Trois ans avant la pandémie, la proportion de collaborations internationales est plutôt faible, aux alentours de 6%.
- Durant la pandémie, cette proportion a presque doublé pour atteindre 11%.
- Après cet épisode pandémique, la proportion de collaborations internationales continue d'augmenter pour atteindre 13%.

Bilan : Ce test de proportion Z montre que la COVID-19 a accéléré dans un premier temps, puis renforcé l'internationalisation des collaborations. Les scientifiques ont donc plus tendance à collaborer à l'international depuis la pandémie.

3.4.6 Etude de l'évolution des themes des publications

Question : Comment les priorités scientifiques et les thématiques de recherche ont-elles évolué au fil du temps, et est-ce que le COVID-19 a eu un impact sur ces priorités ?
Hypothèses :
- La pandémie de COVID-19 a entraîné une augmentation significative des publications liées à la santé, en particulier sur les maladies infectieuses, les vaccins, et la gestion des crises sanitaires.
- Le domaine de l'informatique doit etre dominant au vue de la nature de la DBLP.
- La psychologie a vu un intérêt accru pour les problématiques de santé mentale, notamment liées à l'isolement social, au stress, et à l'anxiété, exacerbés par les confinements.

Graphiques :Évolution globale des thèmes (1994-2024) : Le graphique linéaire en haut illustre les proportions de publications pour différents thèmes au fil des années.



Préparation des graphiques

Analyse des titres des publications

Pour cette étude, nous avons commencé par examiner **les titres des publications scientifiques**. Cette approche a permis d'identifier les mots et expressions qui revenaient fréquemment dans les titres. Ces observations ont servi de base pour définir une structure claire des thèmes et sous-thèmes.

Création des thèmes, sous-thèmes et mots-clés

À partir des mots récurrents identifiés dans les titres, nous avons élaboré une catégorisation en **thèmes généraux**, eux-mêmes divisés en **sous-thèmes**. Pour chaque sous-thème, une liste de mots-clés a été créée afin de capturer au mieux les sujets abordés dans les publications.
Les listes de mots-clés ont été **composées manuellement**, afin de garantir leur pertinence et leur adaptabilité aux particularités des données. Par exemple, pour les termes comme *COVID*, le script a été ajusté pour inclure toutes les variations pertinentes (e.g., *covid-19*, *COVID-19*, *sars-cov*). Pour les autres thèmes, la détection repose sur la présence d'une chaîne minimale dans les titres, incluant tout ce qui contient ou dérive du mot-clé.

Exemple de thème, sous-thème et mots-clés

- **Thème : Informatique**
  - **Sous-thème : Termes généraux**
  - **Exemples de mots-clés :** *algorithm, computer, software, programming, database, network, security, artificial, intelligence, machine, learning, data, web, cloud, system.*
- **Exemple de titre correspondant :** *"Machine learning and cloud systems for secure data processing"*
  - Dans ce titre, plusieurs mots-clés associés au sous-thème **Termes généraux** apparaissent : *machine, learning, cloud, system, et data.* Cela permet de classer ce titre sous le thème **Informatique** et le sous-thème **Termes généraux**.

Analyse de la fréquence des mots-clés

L'analyse s'est concentrée sur la **fréquence d'apparition des mots-clés dans les titres**, Chaque occurrence des mots-clés a été comptabilisée et associée à un thème correspondant et que ce thème est lié au thème de la publication

Observations : En observant les proportions moyennes des mots clés dans les titres des publications au cours des 20 dernières années on peut dégager des tendances :

1. des thèmes déjà présent qui se popularisent :

→ **Informatique :** Avant la pandémie, les publications liée à informatique représentaient 70% des publications.Pendant la pandémie, les publications ont atteint environ 92% (+36%).Cette augmentation peut s'expliquer par plusieurs facteurs : La base DBLP, utilisée pour cette analyse, répertorie principalement des publications liées à l'informatique et aux sciences connexes, ce qui reflète naturellement une dominance de ce domaine. De plus, les innovations technologiques constantes, telles que l'intelligence artificielle et les systèmes de sécurité avancés, continuent de stimuler les recherches dans ce domaine.

→ **Chimie :** Avant la pandémie, les publications liée au thème de la chimie ont augmentée passant de 60% avant le Covid-19 a 65% pendant le covid et plus de 70% après le covid. Cette tendance pourrait être liée à des efforts pour analyser le virus SARS-CoV-2 et ses variants, ainsi qu'à l'élaboration et la production de vaccins (par exemple, Pfizer et Moderna). Ces recherches, essentielles pour répondre à la crise sanitaire mondiale, pourraient avoir stimulé la croissance des publications dans ce domaine.

2. des thèmes émergeant :

→ **Psycologie :** Progression de 50% en 2017 à 62.% après le covid. Cette progression reflète probablement l'intérêt croissant pour les effets psychologiques du COVID-19. Les bouleversements induits par les confinements, l'isolement social et les pertes d'emploi ont conduit à une augmentation des troubles tels que le stress, l'anxiété et la dépression.

Déduction sur les sujets étudiés : Ces observations, bien qu'appuyées sur les proportions et les mots-clés, restent des suppositions. Elles montrent que les publications scientifiques tendent à refléter à la fois des préoccupations fondamentales et des priorités contextuelles. Les thèmes dominants comme l'informatique et la chimie mettent en évidence des recherches nécessaires pour répondre à des défis globaux, tels l'étude d'innovations techniques et la gestion de la pandémie. En parallèle, les recherches en psychologie témoignent d'un intérêt accru pour les conséquences humaines et sociales des crises, en particulier sur le plan psychologique. Ces tendances suggèrent que la pandémie a non seulement accéléré les recherches dans des domaines déjà bien établis, mais a aussi ouvert la voie à de nouveaux axes d'études, façonnés par les besoins sociétaux et les défis spécifiques de cette période.

3.5 Analyse à partir de notre base orientée graphes

3.5.1 Contexte : La connexité d'un graphe de collaboration

Rappel - Définition Connexité

- Soit un graphe G est défini comme $G = (V, E)$, où :
  - V est l'ensemble des sommets.
  - E est l'ensemble des arêtes reliant ces sommets.
- Un graphe est dit **connexe** si, pour tous les sommets $u, v \in V$, il existe un chemin entre u et v. Si ce n'est pas le cas, le graphe peut être décomposé en plusieurs **composantes connexes**, où chaque composante est un sous-graphe connexe maximal.

Réseaux de collaborations : composantes fortement connexes

Dans le cadre des collaborations scientifiques, notre objectif est d'étudier les collaborations entre auteurs, représentées par des arêtes dans un graphe, et d'identifier les **réseaux de collaboration** qui en découlent. Cette approche conduit naturellement à s'intéresser aux **composantes fortement connexes**.
Dans ce contexte, une collaboration est **non orientée** par définition. Cela signifie qu'une relation entre deux auteurs est réciproque : si AA collabore avec BB, alors BB collabore également avec AA. Les réseaux de collaborations peuvent donc être modélisés par des **graphes non orientés**, où les arêtes traduisent des relations de co-publication.
Dans un graphe non orienté, une composante connexe regroupe tous les auteurs directement ou indirectement connectés par des collaborations. Par exemple :

- Si A collabore avec B, et B collabore avec C, alors A, B, et C appartiennent à la même composante connexe, même si A et C n'ont pas collaboré directement. Ce concept peut donc élargir artificiellement le réseau, car les connexions indirectes sont inclus et représentées dans le même réseau.

3.5.2 Analyse des composantes fortement connexes

Dans les réseaux de collaborations scientifiques, les composantes fortement connexes (CFC) jouent un rôle clé pour structurer le graphe en regroupant les auteurs directement ou indirectement connectés. Toutefois, ce concept peut conduire à des échelles très différentes de composantes, allant de très grandes à très petites, en fonction des liens de collaboration présents dans le réseau. Par exemple avant le COVID, La première composante fortement connexe regroupe une majorité écrasante d'auteurs, avec plus de 1 000 000 de nœuds.

pourquoi une taille si grande ?

- Cette énorme composante est le résultat de collaborations transversales, où certains auteurs ou groupes jouent un rôle de pont entre différentes équipes ou domaines. Par exemple, un petit groupe d'auteurs qui collabore simultanément avec plusieurs autres groupes connecte indirectement des milliers d'auteurs, élargissant ainsi le réseau.
- Résultat : L'échelle de connexité devient très large, incluant des auteurs qui n'ont parfois jamais collaboré directement entre eux, mais qui se retrouvent connectés via des liens indirects.

Ainsi pour palier à ces difficultés d'interprétion de l'interconnexion globale du réseau scientifique nous avons choisi de nous intéresser à **la deuxieme composante la plus fortement connexe regroupant 98 chercheurs.** Ce choix repose sur plusieurs considérations méthodologiques et interprétatives, qui visent à garantir une meilleure lisibilité et pertinence de nos résultats :

- **Hypothèse sur la nature de ce réseau** : la deuxième composante, avec ses 98 nœuds, offre une taille plus réduite, facilitant une analyse ciblée et détaillée des interactions.

— **Des connexions probablement plus directes** : La deuxième composante regroupe des auteurs qui collaborent probablement de manière plus directe, avec moins de connexions indirectes. Ce type de réseau est plus propice à observer et comprendre des dynamiques internes, comme la fréquence et l'intensité des collaborations entre les contributeurs.

— **Exploration des comportements isolés** : Bien que cette composante soit isolée de la première (et donc du réseau global), cet isolement peut être un avantage. Il permet d'explorer un réseau de collaboration autonome, où les relations sont concentrées et spécifiques, sans l'influence des liens transversaux ou interdisciplinaires caractéristiques des grandes composantes.

— **Hypothèse sur la nature de ce réseau** : Nous supposons que cette deuxième composante reflète des groupes homogènes de collaborateurs, travaillant dans des domaines de recherche spécifiques ou dans des équipes restreintes. Cela en fait un terrain d'analyse pertinent pour identifier les comportements de collaboration dans des environnements ciblés.

En limitant la deuxième composante, on peut mieux étudier les dynamiques de collaboration, notamment avant, pendant et après la pandémie.

3.5.3 Etude des évolutions d'un sous-graphe, avant, pendant, après la pandémie

Méthodologie

Nous procéderons de la manière suivante :

1. **Identification des réseaux de base :** Pour chaque période (*avant, pendant et après le COVID*), nous partons de la **deuxième composante fortement connexe la plus importante** dans le graphe global.
2. **Étude des dynamiques au sein des réseaux :** À partir de ces trois réseaux de base (un pour chaque période), nous examinerons leurs dynamiques internes en analysant :
   — **Le nombre de collaborations :** Cela permet d'évaluer l'intensité des interactions scientifiques dans chaque réseau.
   — **Le nombre de personnes :** Cela nous informe sur la taille du réseau et la participation des auteurs au fil du temps.
3. **Analyse des évolutions selon les périodes :** Nous comparerons les réseaux de chaque période pour explorer les changements dans les dynamiques de collaboration :
   — Nous observerons si le COVID-19 a eu un impact sur le volume des collaborations (augmentation, diminution ou stagnation).
   — Nous chercherons à comprendre comment cet impact s'est traduit dans les réseaux : réduction de leur taille, diminution des interactions, ou réorganisation des groupes de collaboration.

Objectif final : Avec cette étude nous espérons mettre en évidence les éventuels impacts du COVID-19 sur les comportements collaboratifs scientifiques. Cela permettra de comprendre si, et comment, la pandémie a transformé les dynamiques des réseaux de recherche.

Structure du tableau

Colonnes :

— **Réseau de départ :** Indique la période d'origine pour laquelle le réseau de collaboration est extrait (deuxième composante fortement connexe "Avant", "Pendant", ou "Après").
— **Période étudiée :** La période temporelle analysée par rapport au réseau de départ (*Avant, Pendant, ou Après*).
— **Nb collaborations :** Nombre total de collaborations observées dans le réseau étudié pour cette période.

Réseau de départ	Période étudié	Nb collaborations	Nb personnes
2e CFC Avant (2017-2019)	Avant	280	98
	Pendant	96	46
	Après	24	30
2e CFC Pendant (2020-2022)	Avant	35	27
	Pendant	335	92
	Après	2	4
2e CFC Après (2017-2019)	Avant	0	0
	Pendant	4	10
	Après	459	149

— **Nb personnes :** Nombre total de personnes impliquées dans les collaborations pour cette période.

Lignes :

— Chaque ligne regroupe les résultats d'un réseau de départ. Pour chaque réseau, les collaborations et le nombre de personnes sont étudiés sur trois périodes distinctes : *Avant*, *Pendant*, et *Après* le COVID-19.

3.5.4 Visulisations par les graphes

Remarque : Nous utiliserons l'acronyme CFC pour parler de composante fortement connexe.

3.5.5 Réseau de départ : 2e CFC, 3 ans avant (2017-2019)

Observations : étudiant le réseau issu de la **deuxième CFC** avant le COVID (2017-2019), nous constatons un impact significatif sur la dynamique des collaborations pendant et après la pandémie :

— **Pendant le COVID :** Le nombre de personnes connectées au réseau a été réduit de moitié, passant de **98 à 46 auteurs**, et le nombre de collaborations a chuté d'un tiers, passant de **280 à 96 collaborations**.Cette baisse indique une dispersion des auteurs vers d'autres réseaux ou une réduction des collaborations dans ce sous-ensemble. Bien que certains sous-graphes aient résisté et continué à collaborer, le réseau global a clairement été fragilisé par la pandémie, se fragmentant probablement en de plus petits clusters.

— **Après le COVID :** Cette dynamique de dispersion semble se poursuivre, avec un passage de **46 à 30 auteurs** et une diminution du nombre de collaborations de **96 à 24**. Cela reflète une perte continue de cohésion, les collaborations restantes semblant être maintenues par des sous-groupes résilients.Ce phénomène illustre l'impact prolongé du COVID sur ce réseau, avec un affaiblissement global et un éclatement progressif.

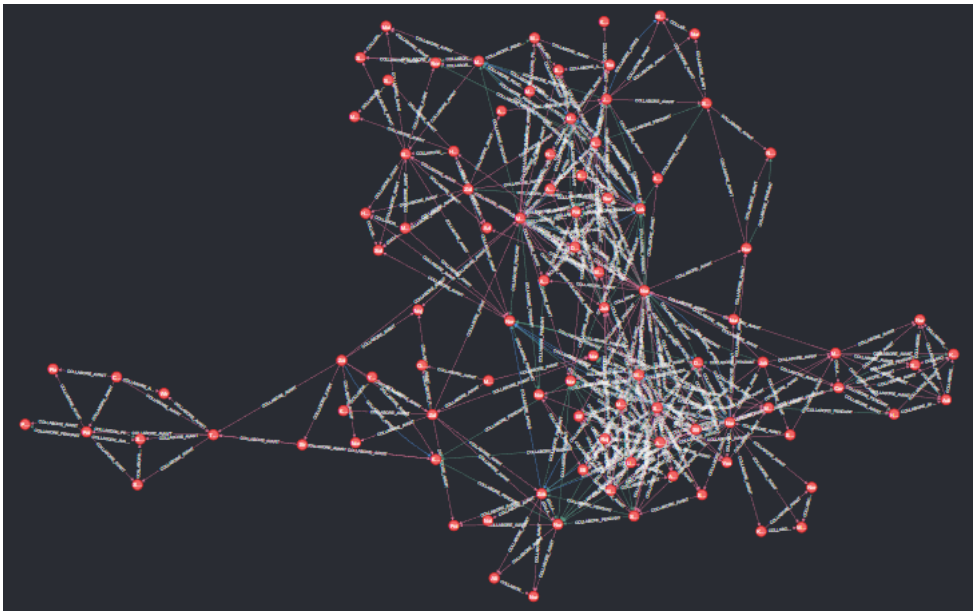


FIGURE 3.13 – Réseau de départ : 2e CFC, 3 ans avant (2017-2019)

Réseau pendant le Covid

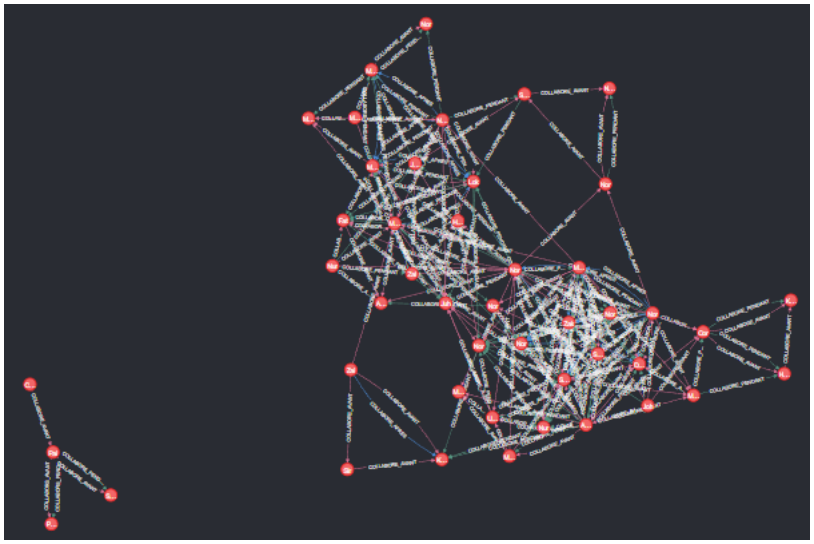


FIGURE 3.14 – Réseau pendant le Covid (2022-2023)

Réseau après le Covid

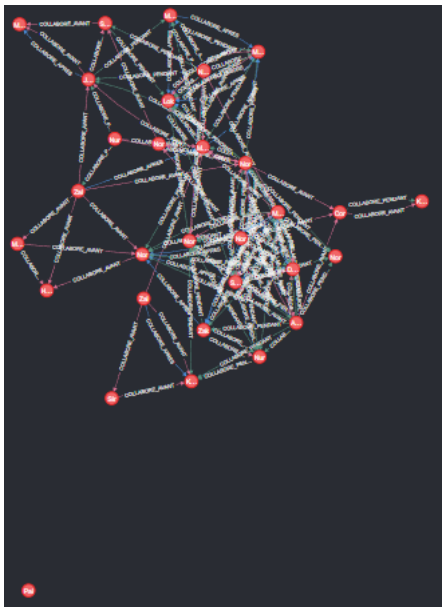


FIGURE 3.15 – Réseau après le Covid (2023-2024)

3.5.6 Réseau de départ : 2e CFC, pendant le covid (2020-2022)

Observations : Pour le réseau issu de la **deuxième CFC** pendant le COVID (2020-2022), les dynamiques observées sont radicalement différentes :

- **Avant le COVID :** Ce réseau était initialement petit, composé de **27 auteurs pour 35 collaborations**, indiquant une activité limitée dans ce sous-ensemble avant la pandémie.

- **Pendant le COVID :** Le réseau a connu une expansion spectaculaire, avec un triplement du nombre de personnes (passant à **92 auteurs**) et une multiplication par presque 10 du nombre de collaborations (**335 collaborations**). Cela reflète un **rassemblement des chercheurs** autour de thématiques spécifiques probablement liées au COVID, menant à une forte augmentation des interactions scientifiques.Ces résultats montrent que la pandémie a servi de catalyseur, poussant les auteurs de ce réseau à intensifier leurs collaborations pour répondre à des besoins scientifiques immédiats.

- **Après le COVID :** Une **quasi-disparition** de ce réseau est observée, avec seulement **4 collaborations pour 10 auteurs** restants.Cela suggère que les collaborations observées pendant le COVID étaient **de circonstance**, motivées par des besoins spécifiques liés à la pandémie. Une fois ces objectifs atteints, le réseau semble s'être désagrégé.

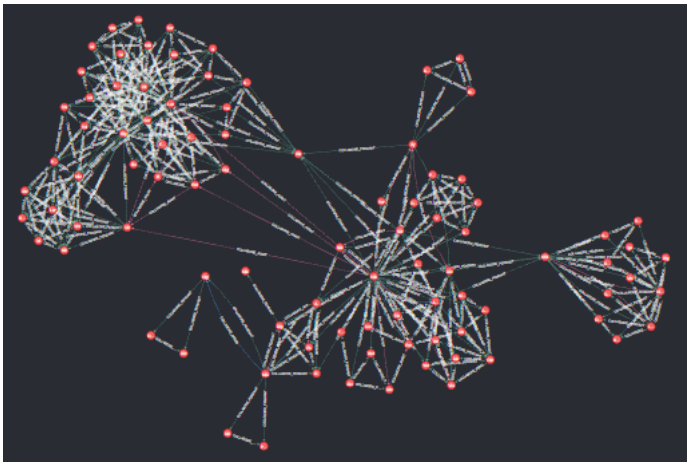


FIGURE 3.16 – Réseau de départ : 2e CFC, pendant (2017-2019)

Réseau avant le Covid

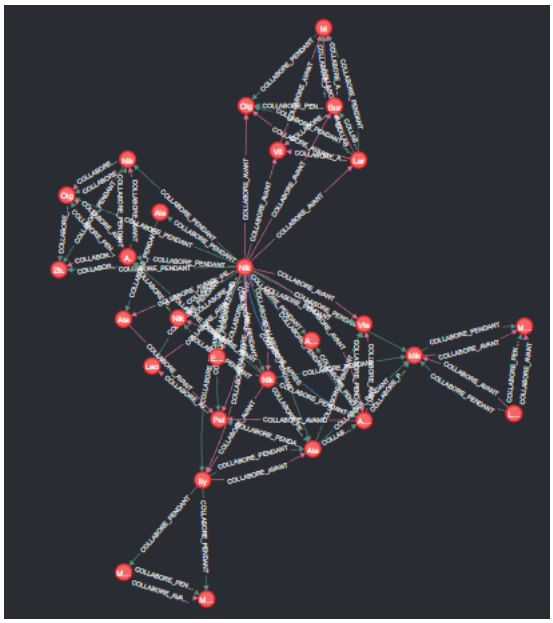


FIGURE 3.17 – Réseau avant le Covid (2017-2019)

Réseau après le Covid

FIGURE 3.18 – Réseau après le Covid (2023-2024)

3.5.7 Réseau de départ : 2e CFC, après le covid (2023-2024)

Observations : Le réseau issu de la **deuxième CFC** après le COVID présente une dynamique particulière, marquée par une forte densification post-pandémie :

— **Avant le COVID :** Ce réseau n'existait pas. Aucun auteur ni collaboration n'est identifié dans ce sous-ensemble avant la pandémie.

— **Pendant le COVID :** Ce réseau a commencé à émerger, bien que modestement, avec **10 auteurs et 4 collaborations**. Cette faible activité initiale montre qu'il s'agit d'un regroupement récent, probablement amorcé dans le contexte de la pandémie.

— **Après le COVID :** Une forte explosion est observée, avec le réseau atteignant **149 auteurs et 459 collaborations**, soit une moyenne d'environ **3 collaborations par chercheur**. Cette densification rapide reflète la formation d'une **nouvelle communauté scientifique** autour de thématiques ou projets spécifiques, qui a émergé pendant le COVID et s'est consolidée après.Ce réseau illustre la capacité de la pandémie à non seulement restructurer les collaborations existantes, mais aussi à créer de **nouvelles dynamiques scientifiques pérennes**.

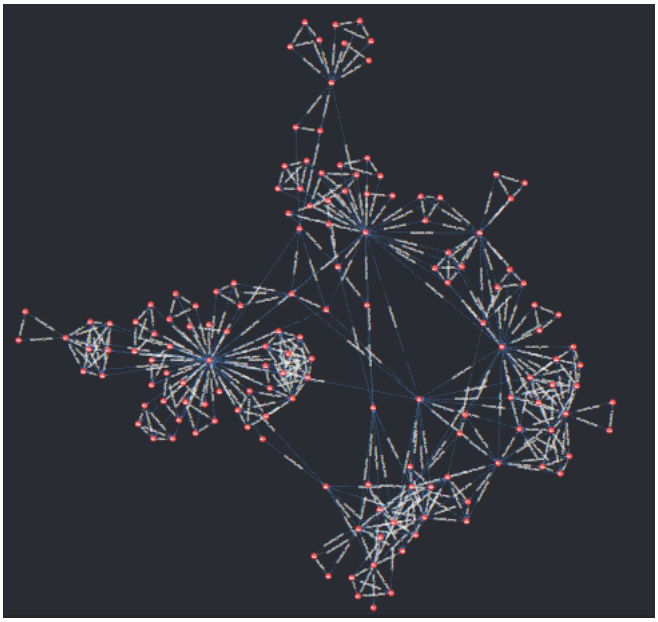


FIGURE 3.19 – Réseau de départ : 2e CFC, après le covid (2023-2024)

Réseau pendant le Covid

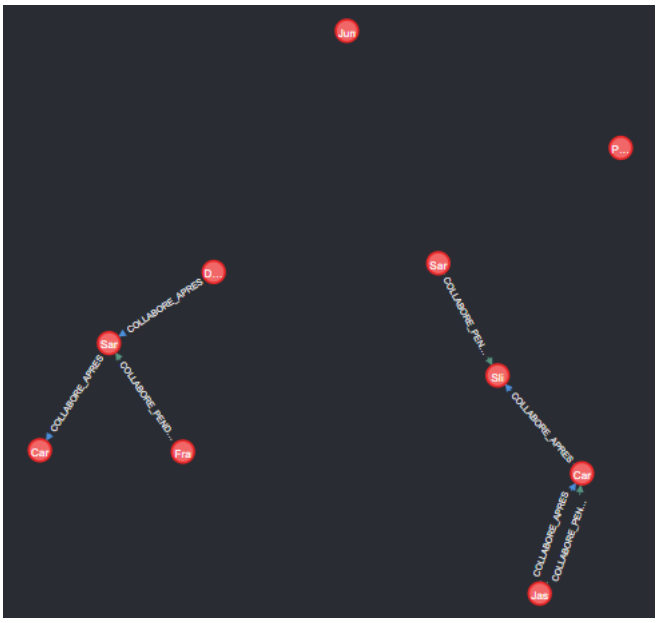


FIGURE 3.20 – Réseau pendant le Covid (2023-2024)

3.5.8 Synthèse des dynamiques observées

Bilan : Ces observations mettent en évidence trois types de comportements dans les réseaux de collaboration :

1. **Affaiblissement et dispersion (réseau avant le COVID) :** Un réseau initialement bien établi, mais fragmenté par la pandémie, montrant une perte de cohésion progressive.
2. **Collaboration circonstancielle (réseau pendant le COVID) :** Une forte intensification des collaborations pour répondre aux besoins immédiats liés à la pandémie, suivie d'une désagrégation rapide une fois ces objectifs atteints.

3. **Création et consolidation (réseau après le COVID) :** Une nouvelle communauté scientifique qui s'est formée en réaction à la pandémie et qui a connu une expansion rapide et durable.

Ces dynamiques permettent de comprendre l'impact de la pandémie sur les réseaux scientifiques, en montrant comment elle a non seulement fragilisé certains groupes, mais aussi favorisé la formation de nouvelles structures collaboratives.

Limites et nuances : Il est important de noter que ces observations reposent uniquement sur l'analyse des **deuxièmes composantes fortement connexes** , qui représentent les réseaux les plus cohérents mais pas nécessairement la totalité des collaborations. De plus, notre capacité à explorer les affiliations institutionnelles est limitée : nous disposons seulement de 15% des affiliations des auteurs. Cette faible couverture des affiliations a conduit à ne pas les incorporer dans l'analyse, car la probabilité que les auteurs d'une même composante fortement connexe (CFC) possèdent tous une affiliation est faible. Par conséquent, certaines dynamiques plus globales ou transversales, notamment en lien avec les affiliations, n'ont pas pu être capturées dans cette étude.

3.6 Corrélations entres les études

3.6.1 Corrélation entre l'augmentation du nombre de publications et l'évolution de formats privilégiés

Au cours de la période [2017-2024], marquée par une hausse progressive du nombre total de publications, un changement notable s'est également produit à partir de 2017 dans les formats privilégiés de publication. Une possible corrélation peut être établie entre cette augmentation et la popularité croissante des formats courts (articles), qui permettent une production plus rapide et plus accessible. Ces formats, particulièrement prisés durant la pandémie de Covid-19, semblent avoir répondu aux besoins des chercheurs dans un contexte de contraintes accrues, renforçant ainsi leur adoption généralisée et contribuant potentiellement à la hausse globale des publications.

3.6.2 Corrélation entre l'augmentation du nombre de publications , l'évolution de formats privilégiés et les collaborations globales

Malgré les évolutions dans les façons de travailler, avec l'adoption de nouveaux formats et une augmentation du nombre de publications annuelles entre 2017 et 2024, le travail collaboratif continue de croître, poursuivant son développement sans variations notables. Cela suggère que les changements dans les formats de publication ne semblent pas avoir eu d'effet sur les collaborations.

3.6.3 Corrélation entre le nombre de publications et la proportion des auteurs par pays

On peut donc s'attendre à avoir davantage de publications provenant des pays les plus représentés (USA, Royaume-Uni, Europe occidentale, Chine, Inde, Brésil) et des collaborations impliquants des auteurs de ces pays.

3.6.4 Corrélation entre le nombre de collaborations et de nouveaux collaborateurs

On observe une corrélation très forte entre l'évolution du nombre de collaborations par année et le nombre de collaborateurs qui collaborent pour la première fois. En effet, les deux courbes suivent la même forme de croissance (Les graphes sont quasiement identiques). De plus, on peut en conclure que l'augmentation du nombre de collaborations est directement impacté par l'arrivée de nouveaux collaborateurs en plus de ceux déjà présents.

Par exemple, en 2020, on a environ 2 900 000 de collaborations et 300 000 nouveaux collaborateurs et en 2019 nous avons eu 2 500 000 et si l'on fait la somme (2 500 000 + 300 000 = 2 800 000 valeur proche du nombre de collaborations totales comptabilisées l'année suivante) on remarque que les premieres collaborations ont contruibuées à l'augmentation du nombre total de collaborateurs :

Année	Nb total collaborations	Nouveaux collaborateurs
2020	Environ 2 900 000	300 000
2019	Autour de 2 500 000	270 000
2018	Environ 2 100 000	250 000

On observe également pour ces 2 courbes, un pic en 2023 post-covid avec plus de nouveaux collaborateurs et de collaborations. On peut en déduire que c'est la période post-covid qui a impacté le nombre de collaborations et incité des chercheurs à collaborer pour la première fois et pas pendant le Covid contrairement à ce à quoi on pouvait s'attendre. Ceci peut s'expliquer par les financements des gouvernements dans la recherche suite à la crise ou par la fin de la pandémie qui a potentiellement ralenti la croissance qui aurait pu être plus importante (bien qu'on ait pas de tendance permettant de l'affirmer).

3.6.5 Test du chi-deux

Calcul de la corrélation entre une publication co-écrite et le fait qu'elle soit publiée après le Covid. Le test du chi-deux révèle que la co-écriture des publications est significativement influencée par la période de publication. En effet, les publications co-écrites sont nettement plus fréquentes après la pandémie, ce qui suggère que la crise sanitaire a facilité ou renforcé la collaboration internationale. Ce phénomène pourrait être dû à une plus grande coopération mondiale en réponse aux défis imposés par la pandémie. Les résultats montrent une tendance marquée vers l'augmentation des publications collaboratives à l'échelle internationale. Ces résultats sont cohérents avec la dynamique observée lors de l'étude des composantes fortement connexes des réseaux de collaboration. Nous avions déduit des phénomènes de consolidation de réseaux après le covid.

3.6.6 Indices de diversité de Shannon

Indique l'équilibre de la diversité des pays dans des ensembles Pendant la pandémie, la diversité internationale a significativement augmenté, comme le montre l'indice de Shannon, probablement en raison de l'adoption des outils numériques et de la mobilisation exceptionnelle des compétences internationales sur des enjeux de santé. Après la pandémie, une diminution est observée, suggérant un retour progressif à des thématiques traditionnelles. Toutefois, les niveaux restent supérieurs à ceux d'avant la pandémie. Cette diversité internationale est probablement permise par l'importante augmentation des publications internationales, représentée ci-avant graphiquement.

3.6.7 Test de proportion Z

Tois ans avant la pandémie, la proportion de collaborations internationales est plutôt faible, aux alentours de 6— Durant la pandémie, cette proportion a presque doublé pour atteindre 11— Après cet épisode pandémique, la proportion de collaborations internationales continue d'augmenter pour atteindre 13

Chapitre 4

Développement de l'interface web de visualisation

4.1 Introduction et fonctionnalités proposées

Étant donné l'envergure du projet, le mode de restitution des analyses a fait l'objet d'une réflexion approfondie et d'une planification rigoureuse. Une présentation inappropriée pourrait en effet entraîner des interprétations erronées ou une incompréhension des résultats. C'est pourquoi nous avons opté pour une interface web moderne, à la fois ergonomique, dynamique et interactive. Cette interface met en valeur les résultats obtenus, les processus d'analyse suivis, ainsi que l'organisation globale du projet, tout en garantissant un accès clair et efficace aux informations essentielles.

4.2 Organisation du site web et pages principales

Pour présenter de manière optimale les résultats de nos recherches, nous avons choisi de développer un site web structuré en plusieurs pages, chacune dédiée à un aspect spécifique du projet. Ces pages sont accessibles à tout moment grâce à une navigation intuitive assurée par un header et un footer, permettant à l'utilisateur de passer aisément d'une section à l'autre.

4.2.1 Maquettage et validation

Le design du site a été réalisé à l'aide de Figma, une plateforme qui permet de créer des prototypes dynamiques et interactifs. Ces maquettes ont permis une simulation navigable du site, facilitant la validation par l'ensemble de l'équipe avant le développement final. Cette approche a assuré une cohérence visuelle et fonctionnelle, tout en permettant des ajustements rapides et itératifs. Figma a été choisi, car plusieurs membres de l'équipe l'ont déjà utilisé en entreprise, il est assez simple à prendre en main et est collaboratif de plus c'est l'un des outils actuels les plus utilisé sur le marché pour le maquettage des sites web.

4.2.2 Navigation et interactivité

— **Header et footer :** Capture illustrant leur organisation.
— **Interactivité :** Présente les pages du projet, éléments cliquables, navigation facile et intuitive.

4.2.3 Pages principales du site web

Le site web est structuré de manière logique et intuitive autour de plusieurs pages principales, chacune dédiée à un aspect spécifique du projet. Voici une description des pages et leurs illustrations :

1. Page d'accueil

- **Illustration associée :** Capture du KPI et du cadre de présentation.
- **Description :** Présente les chiffres clés du projet et met en contexte les données utilisées.

2. Pages des axes d'analyse

- **Illustration associée :** Capture du carrousel des axes.
- **Description :** Présentes les différents axes d'analyses : analyses géographiques, thématiques et temporelles. Nous affichons les différents axes par lesquels nous avons choisi d'interpréter le jeu de données, avec un bref éclaircissement sur leur contenu. Ces axes sont donc des éléments de réponses à la problématique : « Comment la pandémie de la COVID-19 a-t-elle impactée les collaborations scientifiques ? », Séparé tel quel par la nécessité de mettre en parallèle certaines observations pour pouvoir les interpréter. Nous avons ainsi une visualisation par la fenêtre de la géographie ou encore par les thèmes traités dans les sujets de recherche.

3. Page Axe : dédiée aux graphiques et visualisations

- **Illustration associée 1 :** Capture d'écran d'une page axe
- **Description 1 :** Les pages "axes" sont accompagnés par différents graphiques jugés comme étant les plus pertinents, mais également par des démonstrations mathématiques prouvant l'intérêt de tels axes. Il y est donc présent : une carte à bulles, des tree-map ou encore des histogrammes de plusieurs dimensions. Bien entendu, les justifications, arguments et une conclusion sont également présents dans les axes pour ne pas perdre le lecteurs dans des interprétations différentes de celles sur lesquelles nous avons aboutis.
- **Illustration associée 2 :** Capture d'un graphique interactif (carte à bulles, treemap, histogrammes, etc.).
- **Description 2 :** Permet une exploration approfondie grâce à des outils modulables.

4. Page de recherche

- **Illustration associée :** Capture de la page de recherche.
- **Description :** Permet aux utilisateurs de consulter et explorer les données stockées. Nous avons également choisi de mettre en place un « moteur de recherche » offrant l'opportunité d'observer les données stockées dans notre base de données. Celui-ci s'établit dans la continuité d'une transparence sur les données traitées, tout en laissant l'utilisateur constater par lui-même les données utilisées et le modèle de notre base de données. En effet, ce dernier pourra acquérir les données en les triant selon des critères de nationalité, de nom d'écoles ou encore par période. Ainsi, le lecteur n'a qu'à sélectionner la table désirée, la colonne et la valeur recherchée pour lancer le programme. Les résultats correspondant seront alors affichés dans la rubrique « Résultats » de la page.

5. Page dédiée à l'équipe du projet

- **Illustration associée :** Capture de la page équipe.
- **Description :** Présente les membres de l'équipe et leur organisation. La page est accessible depuis le footer. Nous avons incorporé au site une visualisation avec des graphiques de la répartition des contributions individuelles et collectives pour les équipes respectives.

4.2.4 Présentation des visualisations principales

Les axes d'analyse principaux sont présentés avec des graphiques soigneusement choisis pour illustrer les résultats de manière claire et compréhensible :

images de graphiques intéressants

- Des cartes à bulles interactives permettent de visualiser les collaborations géographiques de manière intuitive.
- Des treemaps détaillent les thématiques dominantes et révèlent les priorités et tendances de recherche dans les publications.
- Des histogrammes multidimensionnels explorent les tendances temporelles et mettent en évidence les évolutions au fil des années.

Certaines visualisations intègrent une dimension interactive, permettant aux utilisateurs de choisir une période ou un critère d'analyse pour observer les changements et obtenir des perspectives supplémentaires. Ces outils sont systématiquement accompagnés de démonstrations mathématiques et de justifications des choix de modélisation, offrant un contexte et une validation des données pour guider les lecteurs dans leur interprétation des résultats.

4.3 Choix de conception et justification des outils

La création et l'hébergement par nos soins d'un site web dynamique répondent à deux objectifs principaux : répondre aux besoins des utilisateurs et tirer parti des technologies disponibles. Nous avons d'abord abordé une problématique clé : **comment offrir une expérience utilisateur immersive et intuitive ?** Cela impliquait de concevoir une ergonomie adaptée et un parcours utilisateur optimisé. Comment rendre le site captivant ? Comment transmettre efficacement les enjeux du projet ? Ces réflexions nous ont conduits à choisir une navigation basée sur des onglets dynamiques, enrichie de graphiques interactifs et d'une transparence totale sur nos données.

4.3.1 Analyse des options et choix technologiques

Une fois ces axes définis, la question s'est posée : **quels outils et technologies adopter pour concrétiser ces idées ?** Nous avons exploré différentes solutions et pesé leurs avantages et inconvénients. Notre premier concept reposait sur un style visuel de type Dashboard, soutenu par Kibana, avec une liaison des données via ElasticSearch. Cependant, l'utilisation d'un outil tiers impliquait des contraintes d'intégration et de personnalisation. Après des recherches approfondies, nous avons choisi une autre approche : l'utilisation d'**Angular** combinée à la librairie de visualisations **D3.js**. Ce choix offrait une modularité totale pour concevoir une interface interactive et sur-mesure, en plus de s'aligner avec nos compétences acquises lors de nos formations académiques.

4.3.2 Backend et gestion des données

Pour la gestion des données, nous avons opté pour le développement d'une API en Python, un langage reconnu pour sa robustesse et sa simplicité dans la manipulation de bases de données SQL. Cette API assure une communication fluide entre le serveur (base de données) et le client (frontend) en permettant des requêtes spécifiques pour générer les visualisations demandées. Python s'est révélé être le choix le plus pertinent pour la connexion à une base de données, grâce à sa souplesse et à sa compatibilité avec divers systèmes, notamment SQL et Neo4j.

4.3.3 Architecture technique et hébergement

Le site repose sur une architecture composée de trois éléments interconnectés :

1. **Frontend** : développé avec Angular (HTML, CSS, TypeScript/JavaScript) pour offrir une interface utilisateur dynamique.
2. **Backend** : une API en Python construite avec Flask, servant de passerelle entre les données et le frontend.
3. **Base de données** : un système hybride combinant SQL et Neo4j pour répondre à des besoins variés de stockage et de requêtage.

Chaque composant est hébergé sur une machine virtuelle (VM) dédiée, exécutant **Ubuntu Live Server**. Les VM communiquent entre elles via des requêtes HTTP structurées selon le schéma suivant : `Frontend` ↔ `Backend` ↔ `Base de données`. Cette séparation garantit une meilleure gestion des ressources et une isolation des environnements, tout en rendant le site accessible partout via l'adresse suivante : https://sae.lliger.fr.

Conclusion

Analyse des réseaux de collaboration avant et après la pandémie

Avant la pandémie : Les réseaux de collaboration scientifique avant le COVID-19 étaient caractérisés par une croissance régulière et des structures bien établies. Les groupes de chercheurs collaboraient principalement dans des cadres locaux ou nationaux, même si les collaborations internationales progressaient doucement. Les réseaux travaillaient sur des thématiques scientifiques stables, souvent soutenus par des relations de longue date entre institutions notamment au Royaume-Uni, en Allemagne et aux Etats-Unis.

Pendant et après la pandémie : La pandémie a redéfini ces réseaux. Plusieurs dynamiques clés se dégagent, on observe leur fragmentation et adaptation durant cette période. Des réseaux établis ont été affaiblis par les contraintes de la pandémie, en observant une baisse de leur production. Nous persons cependant que les outils numériques ont permis de maintenir une grande partie des interactions, la baisse de production étant faible. Par ailleurs, on remarque que de nouveaux réseaux se sont créés. La pandémie a favorisé l'émergence de nouvelles collaborations, souvent internationales, autour de thématiques liées à la gestion de la crise (santé publique, épidémiologie, technologies numériques). Après la pandémie, les collaborations initiées ont évolué pour former des communautés scientifiques plus larges et plus diversifiées. Cela témoigne d'une résilience remarquable de la communauté scientifique, qui a su adapter ses pratiques pour répondre aux défis mondiaux.

Comparaison globale : En comparant les réseaux avant et après la pandémie, plusieurs évolutions majeures se dessinent. Sur le plan de l'échelle, les collaborations internationales ont surpassé les interactions locales, devenant le moteur principal des dynamiques scientifiques. En ce qui concerne les thématiques, les réseaux post-pandémie sont marqués par une focalisation accrue sur des problématiques interdisciplinaires, impliquant des échanges entre plusieurs domaines scientifiques. Enfin, sur le plan structurel, la période post-pandémie a favorisé l'émergence de réseaux plus flexibles et agiles, adaptés aux nouvelles pratiques numériques, contrastant avec les réseaux rigides qui prévalaient avant la pandémie.

Sources

Bases de données bibliographiques

— **DBLP (Digital Bibliography and Library Project)** : https://dblp.org
Plateforme principale pour les publications en informatique, regroupant plus de 6 millions de publications.
— **Google Scholar** : https://scholar.google.com
Moteur de recherche académique couvrant divers domaines.

Outils et frameworks techniques

— **Python (Documentation officielle)** : https://docs.python.org
— **Flask (Framework Python)** : https://flask.palletsprojects.com
— **Angular** : https://angular.io
— **D3.js** : https://d3js.org
— **Neo4j** : https://neo4j.com
— **SQLite** : https://www.sqlite.org
— **ElasticSearch et Kibana** : https://www.elastic.co
— **PostgreSQL** : https://www.postgresql.org

Bibliothèques et outils spécifiques

— **ORCID** : https://orcid.org : Pour retrouver les affiliations des auteurs et leurs publications.
— **Requests (Bibliothèque Python pour les requêtes HTTP)** : https://docs.python-requests.org/en/latest/
— **Pandas (Analyse et nettoyage de données)** : https://pandas.pydata.org
— **LXML (Parsing XML/HTML)** : https://lxml.de

Sources pour données globales et recherche géographique

— **Wikipedia** : https://www.wikipedia.org : Utile pour la recherche d'informations générales

Données et standards bibliographiques

— **Digital Object Identifier (DOI)** : https://doi.org
— **CrossRef** : https://www.crossref.org

Ressources éducatives et généralistes

— **Stack Overflow** : https://stackoverflow.com
— **W3Schools** : https://www.w3schools.com

Correction orthographique, reformulation

— **ChatGPT** : https://openai.com