

Projet de Master

**Construction d'une base de données historique sur la  
concentration de la propriété intellectuelle**

Extraction et analyse comparative

*Allemagne, France, Royaume-Uni, États-Unis  
1900-1945*

Naël Dillenbourg  
27.06.2025

LHST Lab  
Professeur Jérôme Baudry  
Superviseur Dr. Nicolas Chachereau



## Résumé

Au cours de ce projet, nous avons construit une banque de données fondée sur des documents d'archives de brevets anglais, français, allemands et américains entre 1900 et 1945, dans le but de mesurer la concentration de la propriété intellectuelle durant cette période. Nous avons mené une étude comparative des méthodes d'extraction nécessaires au calcul de cette concentration et avons traité près de 4 millions de brevets, comptant plus de brevets que les alternatives existantes.

L'exploitation des données révèle une tendance nette à la concentration croissante de la propriété intellectuelle au cours de la première moitié du XX<sup>e</sup> siècle. Nous analysons l'effet des crises économiques et politiques sur cette dynamique, et montrons qu'elles correspondent majoritairement à des phases d'augmentation de la concentration.

Ce projet permet également une analyse sectorielle à l'aide de la classification CPC. Les études de cas sur la chimie et la radio illustrent des configurations diverses de l'économie de la propriété intellectuelle selon les pays.

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Objectifs du projet . . . . .	3
1.2	Bornes temporelles et géographiques . . . . .	3
1.2.1	La première moitié du XX <sup>e</sup> siècle : une période charnière pour la propriété intellectuelle . . . . .	3
1.2.2	Comparaison transnationale : les brevets en Allemagne, France, au Royaume-Uni et aux États-Unis . . . . .	3
1.3	Analyser la concentration industrielle par les brevets . . . . .	5
1.3.1	La concentration industrielle . . . . .	5
1.3.2	L'usage des brevets par les grandes entreprises . . . . .	5
1.3.3	Compiler des données sur la propriété intellectuelle . . . . .	5
<b>2</b>	<b>Dataset, consolidation et post-processing</b>	<b>6</b>
2.1	Source de données . . . . .	6
2.2	Manque de brevets . . . . .	6
2.2.1	Patentcity . . . . .	6
2.3	Manque de clés . . . . .	7
2.3.1	Date de publication . . . . .	7
2.3.2	Classification . . . . .	8
2.3.3	Labellisation . . . . .	11
2.3.4	OCR . . . . .	12
2.3.5	Extraction . . . . .	12
2.4	Validation de la consolidation des données . . . . .	15
2.5	Post-processing . . . . .	17
2.6	Limites . . . . .	18
2.6.1	Réflexion sur le clustering . . . . .	19
<b>3</b>	<b>Expérimentation</b>	<b>19</b>
3.1	Classifications . . . . .	19
3.2	Cooccurrence . . . . .	20
3.3	Concentration . . . . .	22
3.3.1	Mesure de la concentration . . . . .	23
3.3.2	Vue globale . . . . .	24
3.3.3	Étude temporelle . . . . .	25
3.3.4	Par Section . . . . .	31
3.3.5	Par Classe . . . . .	32
3.4	Entités nommées . . . . .	40
3.4.1	Changements législatifs . . . . .	40
3.4.2	Centre de recherche et développement . . . . .	41
<b>4</b>	<b>Conclusion</b>	<b>42</b>
4.1	Apports et objectifs atteints . . . . .	42
4.2	Innovation méthodologique . . . . .	43
4.3	Implications théoriques . . . . .	43
4.4	Perspectives critiques . . . . .	43
4.5	Perspectives de recherche . . . . .	44

# 1 Introduction

## 1.1 Objectifs du projet

Ce projet de master vise à développer une banque de données permettant d'analyser la concentration industrielle par le biais des brevets d'invention entre 1900 et 1945 en France, au Royaume-Uni, en Allemagne et aux États-Unis. Pour ce faire, nous explorons plusieurs méthodes d'extraction de données sur des documents d'archives afin de composer une banque de données dûment affinée des brevets déposés. Cette période, marquée par des bouleversements économiques et géopolitiques majeurs, est particulièrement importante pour mieux comprendre le rôle de la propriété intellectuelle dans la structuration des marchés et son évolution. En se concentrant sur les acteurs qui ont façonné et profité du système des brevets, ce projet vise à fournir des outils pour analyser la manière dont la propriété intellectuelle a influencé l'économie et sa relation aux crises traversées durant la période étudiée. Nous souhaitons ainsi pouvoir voir l'évolution de la concentration industrielle à travers la première moitié du siècle et comparer celle-ci sur plusieurs pays.

Dans le but de montrer les potentiels usages de cette banque de données pour les travaux historiques, nous expérimenterons avec ces données en développant quelques études de cas, dans le domaine de la radio et de la chimie.

## 1.2 Bornes temporelles et géographiques

### 1.2.1 La première moitié du XX<sup>e</sup> siècle : une période charnière pour la propriété intellectuelle

La période de 1900 à 1945 s'inscrit dans la continuité de la "première mondialisation", amorcée dès 1870, associée à une expansion rapide du commerce et de l'investissement à l'international. [Faubert \[2012\]](#) Au niveau de la propriété intellectuelle, cela se manifeste par le processus d'internationalisation des brevets qui a été formalisé par des accords multilatéraux, tels que la Convention de Paris de 1883 qui visait à faciliter l'obtention et l'exploitation des brevets à travers les frontières. [\[Lapointe, 2000, p. 16\]](#)

Cette période est profondément impactée par les deux guerres mondiales, des événements qui ont non seulement interrompu le développement de la mondialisation économique, notamment la Première Guerre mondiale, et qui entraînent également une interruption de l'internationalisation croissante de la propriété intellectuelle. [\[Chachereau, 2022, p. 18\]](#). Les guerres sont présentées par certains historiens comme des catalyseurs, "[encourageant] la propension à innover et finalement à favoriser le processus de croissance économique lui-même". [\[Diebolt and Pellier, 2012, p. 1\]](#). D'autres voient par exemple dans la fin de la Seconde Guerre mondiale une période de régularisation accrue d'économies enlisées dans des dynamiques monopolistiques telles qu'aux États-Unis et en Angleterre où les cartels étaient nombreux. Des nouvelles lois antitrust créent les conditions nécessaires à un nouvel investissement dans l'innovation, un domaine sous-investi par l'excès de monopole. [\[Christophers, 2016, p. 169-172\]](#)

En addition du contexte économique volatile, le début du XX<sup>e</sup> siècle est également une période charnière pour la propriété intellectuelle, du fait de changements dans l'approche des brevets, tels que la création de centres de recherche et développement dans certaines industries. [Lamoreaux \[1999\]](#) Ces changements de pratiques des entreprises actives dans le dépôt de brevets peuvent impacter la mesure de la concentration<sup>1</sup> que nous cherchons à développer dans ce projet, rendant cette période potentiellement très intéressante à analyser. C'est également le cas des changements législatifs, tels que les réformes effectuées en Allemagne sous le régime nazi. [\[Gispen, 2007, p. 70\]](#)

### 1.2.2 Comparaison transnationale : les brevets en Allemagne, France, au Royaume-Uni et aux États-Unis

---

1. Ces changements dans la forme et l'assignation des inventeurs et des entreprises sur les brevets peuvent impacter les données que nous pouvons extraire des documents d'archive.

Ces 4 pays représentent également plus de 50% des brevets délivrés entre 1900 et 1945, comme nous pouvons le constater dans la figure 1. [Diebolt and Pellier \[2012\]](#)

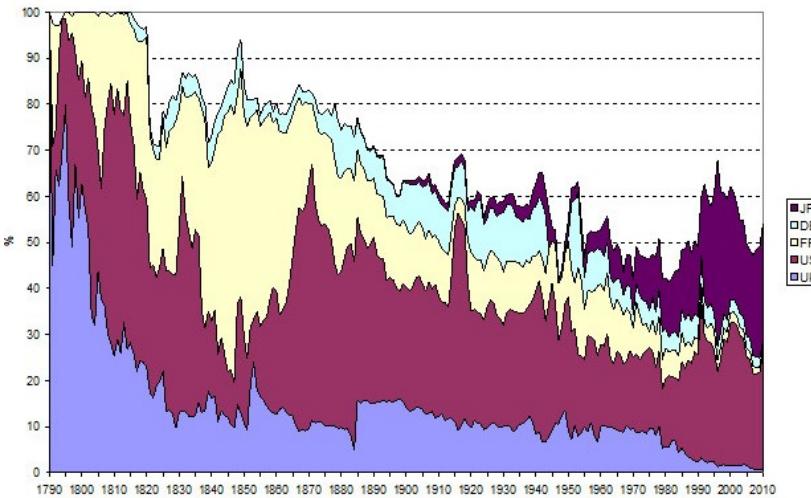


FIGURE 1 – Part de brevets délivrés par les US, DE, UK, FR, JP dans le total des brevets délivrés (146 pays) [Diebolt and Pellier \[2012\]](#)

Le nombre total de brevets publiés varie selon le pays, avec 1'700'000 aux États-Unis, 670'000 au Royaume-Uni, 610'000 en France et en Allemagne durant la période sélectionnée.

- En plus de cette différence numérique, la gestion de la propriété intellectuelle varie de pays en pays.
- Allemagne : Le système allemand des brevets (mis en place en 1877) se caractérise par une durée limitée des brevets due à des annuités croissantes, et par des recherches extensives. [[Gispenn, 2007](#), p. 60] La loi allemande de 1877 était influencée par de grands industriels tel que Werner Siemens et prévoyait par exemple l'obligation de fournir une licence. [[Gispenn, 2007](#), p. 61] [[Götting et al., 2012](#), p. 93]. L'Allemagne nazie promulgue ensuite des changements dans sa loi en faveur des inventeurs. [[Gispenn, 2007](#), p. 75]
  - France : La France a une longue histoire de protection par les brevets, avec une loi dès 1791. [[Baudry, 2014](#), p. 462] La France possède également d'un système d'annuités progressives depuis 1844. [France \[1844\]](#) La moitié des brevets n'atteint pas deux ans. [[Chachereau, 2022](#), p. 204] La France a un modèle sans examen préalable des brevets (*prior examination*). [[Chachereau, 2022](#), p. 347]
  - Royaume-Uni : C'est en 1624, avec l'introduction du "Statute of Monopolies" qu'est introduite l'une des premières lois sur les brevets. [[Chachereau, 2022](#), p. 54] Comme l'Allemagne, le Royaume-Uni pratique un examen de fond des critères de brevetabilité (*prior examination*). [[Diebolt and Pellier, 2012](#), p. 614] Le pays possède également un paiement d'annuités dès 1852. [Khan \[2008\]](#)
  - États-Unis : Le système de brevets américain est perçu, par certains, comme un idéal "démocratique" et efficace en raison de son coût modique et de son accès peu discriminant, contribuant au développement économique du pays. [[Khan, 2005](#), p. 5] L'introduction d'un examen préalable en 1836 a significativement réduit l'incertitude sur la validité des brevets et a favorisé le développement d'un marché des brevets. [[Chachereau, 2022](#), p. 245] Comme nous aurons l'occasion de le voir, les firmes américaines comme United Shoe Machinery Co. et Westinghouse sont des acteurs majeurs de l'acquisition de brevets tout en étant également connues pour leur domination du marché.

Il est intéressant de noter tant les différences que les similitudes entre les systèmes de brevets tels que l'existence ou non de "*prior-examination*", de protection des inventeurs, de systèmes d'annuités croissantes qui limitent la durée des brevets, celles-ci pouvant avoir des conséquences sur l'usage des brevets et donc sur les résultats mesurés à travers nos données.

## 1.3 Analyser la concentration industrielle par les brevets

### 1.3.1 La concentration industrielle

La concentration industrielle désigne habituellement la concentration de la production industrielle par quelques grandes entreprises. [Shughart \[2002\]](#) Dans notre cas, nous cherchons à calculer la concentration vis-à-vis du nombre de brevets publié par entreprise.

C'est un sujet déjà abordé par plusieurs économistes critiques de la forte concentration de brevets aux mains d'un faible nombre d'entreprises. [Watson and Holman \[1970\]](#) Wilcox, un économiste étatsunien, parle de "monopolisation des monopoles de brevets". [\[Wilcox, 1966, p. 172\]](#) L'économiste Corwin D. Edwards, argumente que l'accumulation de brevet constitue une monopolisation considérable de l'industrie. Dans *The Political Economy of Monopoly*, Fritz Machlup décrit que les conséquences de la possession de milliers de brevets par un nombre restreint d'entreprises mènent à un degré de monopole encore jamais envisagé. [\[Machlup, 1952, p. 284\]](#) Enfin Kaysen et Turner analysent que la forte concentration des brevets empêche le développement technologique due aux risques juridiques encourus et la charge additionnelle d'innover autour de brevets déposés. [\[Kaysen and Turner, 1959, p. 165\]](#)

La majorité de ces auteurs mesurent la concentration industrielle aux États-Unis pour une ou quelques années. Notre projet permet de quantifier et de développer une mesure à travers le temps pour étudier l'évolution de cette concentration durant la première partie du XX<sup>e</sup> siècle.

### 1.3.2 L'usage des brevets par les grandes entreprises

La littérature économique a privilégié l'étude des usages du brevet par les grandes entreprises. Il semble établi que la tendance à breveter augmente significativement avec la taille de l'entreprise, et le nombre de brevets déposés semble également corrélé positivement à celle-ci. [\[Baudry, 2014, p. 463\]](#) Les grandes entreprises, dotées de ressources financières, mènent même des "stratégies offensives", cherchant à dominer le marché, grâce à leurs "portefeuilles de brevets étendu[s]". [\[Baudry, 2014, p. 468,469\]](#)

Les brevets occupent une place centrale dans les stratégies de grandes firmes de certains secteurs. [\[Chachereau, 2022, p. 313\]](#) Elles développent des "stratégies propriétaires" visant à différencier leurs produits et à conserver l'exclusivité des variantes techniques brevetées, parfois même en adoptant des "pratiques plus monopolistiques". [\[Chachereau, 2022, p. 466\]](#) L'exemple de United Shoe Machinery Corporation illustre l'intégration des brevets dans une stratégie monopolistique et le contrôle des procédés de fabrication. [Time \[1953\]](#)

Certaines entreprises font également usage de brevets "sans inventions" et avec peu de chance de subsister à une bataille légale mais permettant tout de même à développer un monopole. D'autres pratiques consistent à développer des brevets sur une technologie sans la mettre sur le marché mais dans le but d'empêcher tout concurrent de le faire. [\[Wilcox, 1966, p. 170-172\]](#)

Enfin d'autres stratégies existent telles que la *mise en valeur*, c'est-à-dire la recherche de revenus par licences de brevets ou encore la *logique défensive*, cherchant à éviter les procès et voir leurs activités bloquées par des concurrents ayant déposé des brevets avant eux. Nous pouvons également mentionner les *patent pools*, regroupement de plusieurs entreprises mettant leurs brevets en commun. [\[Chachereau, 2022, p. 272\]](#)

Différentes autres recherches se concentrent sur des thèmes liés à la place des grandes entreprises et des inventeurs indépendants dans le système de propriété intellectuelle mais sans nécessairement regarder l'état de la concentration des brevets et son évolution durant le XX<sup>e</sup> siècle au travers des crises de cette période. [Lamoreaux and Sokoloff \[1999\]](#) [Lamoreaux et al. \[2009\]](#) [\[Chachereau, 2022, ch. 4,6\]](#) [Nicholas \[2011\]](#)

### 1.3.3 Compiler des données sur la propriété intellectuelle

Une recherche de Watson et Holman se saisit du calcul de la concentration dans la propriété intellectuelle. Cependant, elle n'aborde que les données aux États-unis sans aborder l'évolution de celle-ci durant le XX<sup>e</sup> siècle. Watson et Holman mesurent ici la concentration des brevets pour les plus grandes entreprises américaines et cherchent à comparer cette concentration avec la concentration des performances des centres de recherche et développement ainsi que la concentration des actifs de grandes entreprises. [Watson and Holman \[1970\]](#)

De nombreuses recherches ont déjà été effectuées dans la compilation des données de la propriété intellectuelle, cependant sans avoir conçu de méthode pour le calcul de concentration des brevets.

Les travaux de Michael J. Andrews, qui comparent différentes sources de données, ont été très utile durant ce projet. Cependant, ces différentes sources se concentrent sur les brevets aux États-Unis, ne nous permettant pas de réutiliser simplement ces données. [Andrews \[2021\]](#) Durant ce rapport, nous faisons usage de ces différents travaux dans la partie [2.4](#), incluant KPSS [Kogan et al. \[2017\]](#), USPC [Berkes \[2018\]](#), SAZ [Andrews et al. \[2020\]](#).

En addition, les travaux de Bergeaud et Verluise seront utilisés comme référence à plusieurs reprises dans ce projet. PatentCity a l'avantage de contenir les données françaises, allemandes, américaines et anglaises durant la période que nous étudions. Cependant, cette banque de données a été construite pour différents objectifs, rendant la réutilisation de ces données plus compliquée. [Bergeaud and Verluise \[2024\]](#)

## 2 Dataset, consolidation et post-processing

### 2.1 Source de données

Nous faisons principalement usage de données provenant de l'office européen des brevets. Nous avons pu y obtenir des métadonnées contenant les champs nécessaires à notre analyse. Ces données étaient au format json ou txt, selon la méthode de téléchargement employée.

Une partie des métadonnées a été obtenue via le logiciel Patmonitor, et une seconde, à travers l'API de l'office européen des brevets.

Ces deux sources n'utilisant pas le même *schemes* de métadonnées, nous avons dû adapter deux pipelines pour l'importation de ces métadonnées en trouvant, là où c'était possible, les équivalences entre les clés utilisées. Pour un certain nombre de brevets, où les métadonnées étaient insuffisantes, nous utilisons de scans obtenus également au travers de l'office européen des brevets.

Enfin un travail constant de complétion des données a été nécessaire, dû à des lacunes initiales dans les fichiers fournis. Nous avons supplémenté aux données déjà existantes, plus de 327'000 brevets, principalement allemands, manquant à l'appel.

Par gain de temps, nous avons aussi obtenu, là où l'alternative était trop coûteuse en ressources de calcul, les textes de certains brevets américains à travers Google Patent. Il s'agissait souvent d'une version de mauvaise qualité, mais que nous avons limité à quelques centaines de brevets seulement.

### 2.2 Manque de brevets

Malgré plusieurs millions de métadonnées déjà collectées lors du lancement du projet, nous avons remarqué des manques importants. Nous nous sommes donc chargés de collecter ces données supplémentaires afin de nous assurer que cette lacune ne crée pas un angle mort dans notre analyse.

Nous avons tout d'abord comparé l'ensemble de notre corpus de métadonnées à l'ensemble du corps d'images de brevets téléchargées, pour découvrir les éléments manquants. Deux types de données qui avaient été téléchargées de manière indépendante.

Nous avons également comparé nos données à d'autres travaux similaires, afin de trouver des brevets dont nous n'avions pas les métadonnées ou les images d'archives. Nous comparons ces différentes sources et dans la partie [2.4](#).

#### 2.2.1 Patentcity

La base de données PatentCity, élaborée par Antonin Bergeaud et Cyril Verluise, couvrant notre période historique et le domaine géographique de notre problématique, nous a permis d'avoir un point de comparaison essentiel à la validation de nos données. [Bergeaud and Verluise \[2024\]](#)

Cependant cette banque de données n'a pas les mêmes objectifs que notre projet, se concentrant sur l'extraction des données géographique des brevets. Celle-ci ne contient qu'un seul inventeur/déposant par brevet, alors que de multiples noms peuvent être mentionnés dans un même brevet.

Nous avons tout de même été en mesure de comparer notre base de données à PatentCity. A l'aide de celle-ci, nous avons identifié de nombreux brevets figurant chez PatentCity mais introuvable dans les données initialement téléchargées. Une fois identifiés, nous avons ensuite récupéré ces métadonnées à travers l'API de l'office européen de brevets.

Après ce travail de complémentation de notre banque de données, nos données partagent 3'517'219 brevets avec PatentCity. En outre, nous possédons les métadonnées de 201'729 brevets supplémentaires ne figurant pas sur PatentCity. De son côté, PatentCity semble toujours posséder 2'229 brevets introuvables dans nos données.

## 2.3 Manque de clés

Nous avons constaté un manque de clés important sur de nombreux brevets. Afin de compléter ces manques, nous avons cherché à construire une pipeline permettant l'extraction de ces données à partir des brevets scannés et d'autres informations à notre disposition.

### 2.3.1 Date de publication

Nous faisons usage des dates de publication de brevet, celles-ci étant plus représentées dans nos métadonnées que les dates de dépôt de brevet. Ce choix est une concession nécessaire, elle nous nous permet d'exploiter plus de données, mais implique un retard entre la décision de breveter une invention et le moment où nous la comptabilisons dans nos données. Une différence de temps qui n'est pas uniforme dans le temps, entre le pays et même entre les brevets.

Certaines dates de publications manquent cependant à l'appel. C'est particulièrement le cas en Allemagne entre 1900 et 1920. Les numéros de brevets allemands étant incrémentaux, comme nous le montre la figure 2, nous devrions être en mesure d'obtenir l'année de publication.

Pour vérifier la précision de cette méthode, nous essayons de prédire chaque année de publication de brevet à partir de son numéro de publication. Nous vérifions pour chaque date à prédire si 5 brevets publiés chronologiquement avec et après sa publication partagent la même année de publication.

Comme nous sommes en mesure de le voir sur la figure 3, centré sur les années 1900 à 1920, où nous manquons le plus de dates, nous sommes en mesure d'extrapoler la date correcte une majorité de cas, avec au total une précision de 95.47%.

Cependant, cette méthode n'est naturellement pas appropriée pour les dates de demande de brevet, avec seulement une précision de 28.4%. Ceci est probablement dû aux délais variables entre le dépôt du brevet et la création du numéro de brevet.

PatentCity ont adopté une autre méthode : ils itèrent vers l'arrière pour remplir l'année de publication jusqu'à atteindre le dernier numéro de publication de l'année précédente. [Bergeaud and Verluise \[2024\]](#)

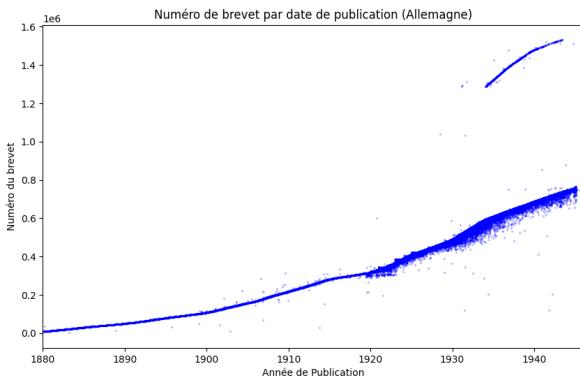


FIGURE 2 – Date de publication de brevets allemands

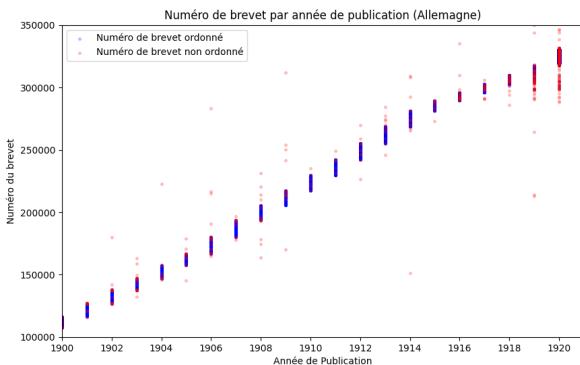


FIGURE 3 – Précision de l'extrapolation de l'année de publication à partir du numéro de publication

### 2.3.2 Classification

Certaines des métadonnées sont accompagnées d'une classification USPC, en usage entre 1899 et 2013. Cependant nous faisons usage de la classification CPC, utilisé par l'office européen des données. Cette classification possède une hiérarchie emboîtée, avec un niveau de détail très poussé. Chaque brevet peut correspondre à plusieurs classifications. Nous analysons l'usage des classifications plus en détail dans la partie [3.1](#) de ce rapport.

Nous souhaitons comprendre au travers des catégories de cette classification quels sont les acteurs principaux faisant usage du droit intellectuel. A cette fin, nous devons donc nous assurer que nous possédons la classification d'un maximum de brevets.

Comme il est visible sur la figure [5](#), il manque pour certains pays et pour certaines périodes, une classification CPC. C'est le cas pour la quasi-entière des brevets anglais dans la période 1900 à 1909.

Afin de classifier ces brevets, nous avons cherché à créer un multioutput classifier. Déterminer les sections d'un brevet sur la base de leur titre est possible au moyen de nombreuses techniques de classification.

Nous faisons usage du score F1 Micro et Macro pour évaluer les résultats de ces modèles, une mesure plus adéquate pour les données avec un déséquilibre. Bien que le score macro nous permet de calculer un score en considérant chaque classe comme égale, quel que soit leur nombre réel d'utilisation, le weighted F1 score nous permet de tenir compte du fait que de nombreuses classes sont très peu représentées. Nous calculons également le micro F1 score pour comparer les résultats avec d'autres ressources faisant usage de cette mesure.

Symbol	Niveau		description
	Section	Classe	
		Sous-classe	sucré
		Groupe	production de sucre
		Sous-groupe	purification de jus sucré
			par électricité
C 13	B 20	18	

FIGURE 4 – Hiérarchie d'une classification CPC

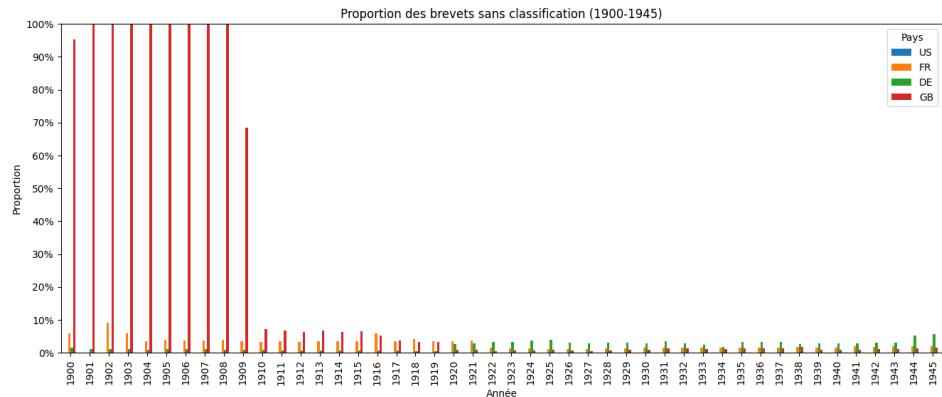


FIGURE 5 – Proportion de classifications manquantes par année et par pays

Nous avons étudié plusieurs possibilités et validé leur performance avant d'appliquer un modèle capable de classifier ces brevets :

**CV + LR**

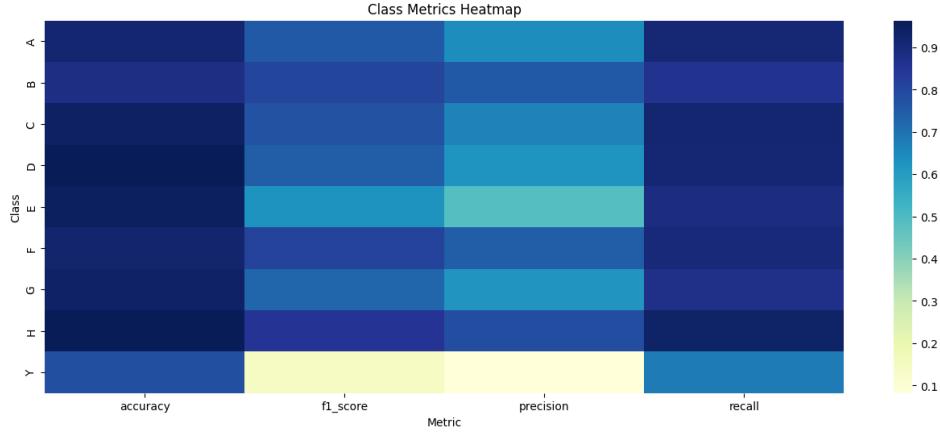


FIGURE 6 – Accuracy, F1-Score, Precision et Recall

Nous obtenons de bons résultats, comme illustré dans la figure 6, en implémentant une régression logistique à partir de la représentation matricielle des tokens de titres de brevets (Count Vectorizer). Nous mesurons un score macro F1 score de 69.38% et micro de 71.58%. La section Y, de par sa nature de section généraliste, possède une faible précision, un résultat attendu.

#### TF-IDF + MultinomialNB

Les résultats semblent légèrement meilleurs en utilisant une matrice TF-IDF et d'une classification naïve bayésienne. Les résultats de cette méthode indiquent un score macro F1 de 71.53%, et micro de 77.30%. La figure 7 illustre cependant un taux de recall inférieur au modèle précédent, démontrant que ce modèle possède de nombreux faux négatifs.

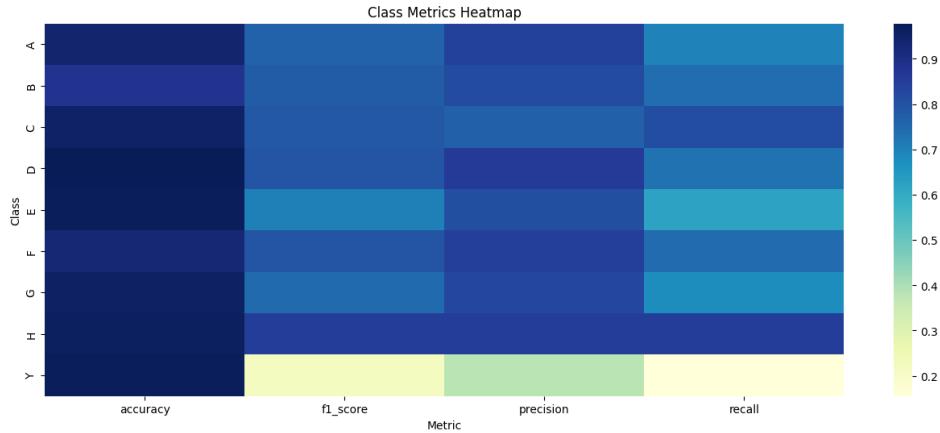


FIGURE 7 – Accuracy, F1-Score, Precision et Recall

#### Sentence Embedding + LR

Nous avons également exploré la possibilité de classifier les sections à l'aide de plongement lexical (sentence embedding) et d'une régression logistique. Cette méthode obtient un macro score F1 de 82% et une score micro de 87%. Ce modèle, surpassant les scores des deux versions précédentes nous semble être la meilleure option, même si le taux de valeur prédictive positive (precision sur la figure 8), indique trop de prédictions positives.

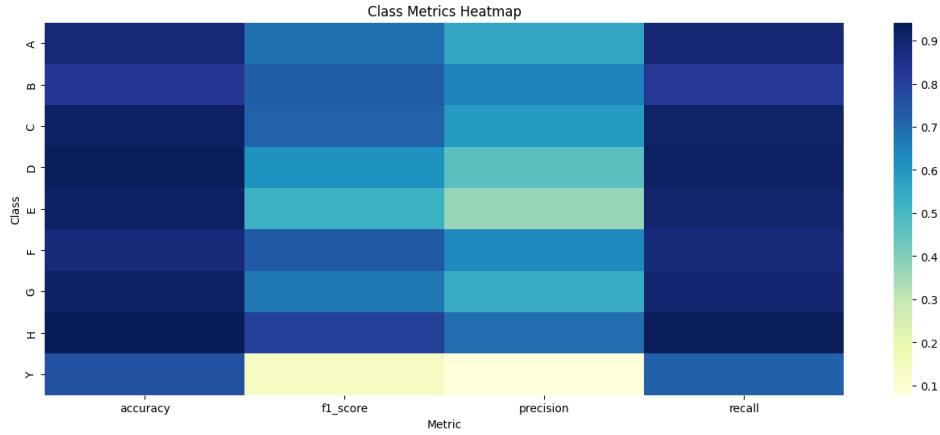


FIGURE 8 – Accuracy, F1-Score, Precision et Recall

Cependant, la difficulté réside dans la prédiction non seulement des sections de brevets mais aussi des classes et sous-classes. Ceci augmente significativement le nombre de labels possible, de 9 à 128 classes et plus de 600 sous-classes. Nous avons alors tenté d'appliquer différentes méthodes basées sur notre meilleur modèle (en utilisant du Sentence Embedding) pour non seulement classifier les sections de brevets mais également les classes ainsi que les sous-classes.

### Classification Sous-Classe

Plusieurs chercheurs se sont déjà plongés sur le problème de la classification des brevets au niveau des sous-classes. Certains chercheurs font usage d'une régression logistique sur la base d'informations supplémentaires obtenues sur les brevets et d'autres méthodes pour la sélection des features (grid ranking et linear cut-off). [Ghosh et al. \[2017\]](#) Ils atteignent ainsi un score F1-micro de 69.89% [Lee and Hsiang \[2020\]](#) Certaines recherches font usage à la fois d'information complémentaire sur les brevets ainsi que d'une variation de TF-IDF et un modèle de classification naïve bayésienne. [Li et al. \[2016\]](#) Une implémentation qui a atteint une précision de 87.2% (sur la base d'une mesure de précision personnalisée) Un projet exploite un réseau de neurones avec un extracteur de feature n-gram. Le modèle utilise uniquement les brevets mécaniques et obtient une précision maximale de 72%. [Hu et al. \[2018\]](#) D'autres chercheurs font usage d'un modèle BERT préentraîné, leur permettant d'obtenir un score F1 de 66.83% et une précision de 84.26%. [Lee and Hsiang \[2020\]](#) Ces performances battent celles de DeepPatent, qui eux font usage d'une représentation vectorielle des mots puis d'un réseau de neurones convolutif comme classificateur. [Li et al. \[2018\]](#)

Une difficulté dans la comparaison des méthodes déjà employées sur le sujet est l'usage des données d'entraînement et des mesures de précision différentes rendant les comparaisons inéquitables.

Nous avons entraîné notre propre model. Pour cela nous utilisons PatentSBERTa pour le plongement lexical (sentence embedding) et nous avons comparé les performances de deux classifiers : une régression logistique ou un réseau de neurones convolutif. [Bekamiri et al. \[2021\]](#)

Score	PatentSBERTa + LR	PatentSBERTa + CNN
Précision (weighted avg.)	77%	75%
Recall (weighted avg.)	66%	62%
Micro F1-Score	72%	70%
Macro F1-Score	56%	48%
Weighted F1-Score	70%	67%

Les résultats "PatentSBERTa + LR" atteignent un bon weighted score F1, nous laissant espérer atteindre une classification correcte pour une grande partie des sous-classes. Le score macro, nous indique qu'il reste de lacunes dans les modèles, particulièrement pour les classes sous-représentées dans les données d'entraînement. Une voie d'amélioration pour ce modèle pourrait être l'ajout d'informations contextuelles sur les brevets, tel que développé dans certains projets similaires.

Nous avons fait usage du modèle basé sur une régression logistique pour classifier les brevets britannique sans classification CPC, nous permettant d'insérer ces brevets manquant à notre banque de donnée.

### 2.3.3 Labellisation

Dans l'optique d'extraire des informations des brevets à notre disposition, nous souhaitons appliquer un algorithme de reconnaissance optique de caractères (OCR). Pour ce faire, il est nécessaire de déterminer au préalable quelles pages contiennent du texte exploitable. Après avoir considéré des techniques de clustering pour identifier le contenu de chaque page de brevet, nous avons opté pour l'implémentation d'un réseau de neurones servant de classificateur multiclass. Cette approche permet non seulement de distinguer les pages contenant du texte de celles qui en sont dépourvues, mais également, dans l'optique d'un éventuel travail sur les schémas, de déterminer si ces derniers sont entièrement contenus sur une page ou seulement partiellement. Cette partie du projet n'a pas été développée d'avantage mais nous avons tout de même fait usage de ce classificateur entraîné par nos soins.

		Prédit		
		schéma_partiel	schéma	texte
Réel	schéma_partiel	118	52	3
	schéma	32	234	0
	texte	2	2	469

FIGURE 9 – Matrice de confusion du classifieur

La particularité des pipelines que nous avons développés réside dans la gestion de la variation des formats d'entrée ainsi que dans l'optimisation des temps de traitement. En effet, les données fournies sont disponibles soit au format PDF, soit au format TIFF, et sont stockées dans des archives tarball compressées. Cette configuration, compte tenu du nombre très important de fichiers et de leur taille, ne permet pas une décompression globale suivie d'un traitement séquentiel. Il est nécessaire d'analyser les fichiers individuellement, tout en maintenant les temps d'exécution à un niveau optimal. Cette contrainte s'applique tant à l'extraction de la classe de chaque page qu'à l'exécution d'une méthode d'OCR sur le document.

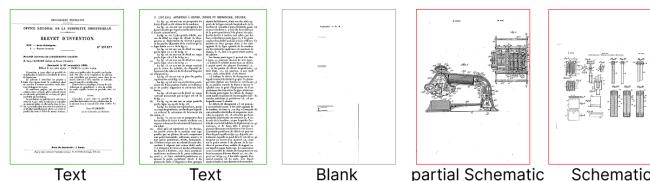


FIGURE 10 – Exemple d'image labelisé par un classifieur

Comme l'illustre la figure 10, nous avons étiqueté chaque page de nos brevets à l'aide de trois classes distinctes. Nous avons par ailleurs implémenté un test simple basé sur un seuil de pixels pour détecter automatiquement les pages blanches.

Une amélioration possible à cette méthode serait d'aussi être en mesure de détecter les pages de corrections (Erratum), ceux-ci peuvent jouer un rôle crucial dans la validité des données extraites. Cependant, leur nombre est si faible que nos données d'entraînement n'en contenant pas suffisamment pour entraîner cette nouvelle classe.

### 2.3.4 OCR

Afin de numériser le texte, nous avons cherché à déterminer quelle technique implémenter pour cette tâche. Notre première priorité fut le temps de traitement par brevet. Figure 13 démontre un temps d'exécution moindre que tous autres modèles.

Les performances de modèles plus récents tel que DocTR, bien que moins performant sur notre benchmark (Figure 12), reste un très bon modèle. [Mindee \[2021\]](#) Cette méthode gère adéquatement le texte en deux colonnes ainsi qu'une partie des champs inclinés, comme c'est le cas des dates de publication dans les brevets allemands.

Malgré ces alternatives et d'autres plus performantes en précision, nous avons choisi d'utiliser Tesseract pour l'extraction du texte des pages étiquetées.

A noter que Pero, entraîné sur des archives de journaux en tchèque, obtient de très bon scores de similarité et en temps d'exécution. [Kohút \[2021\]](#)

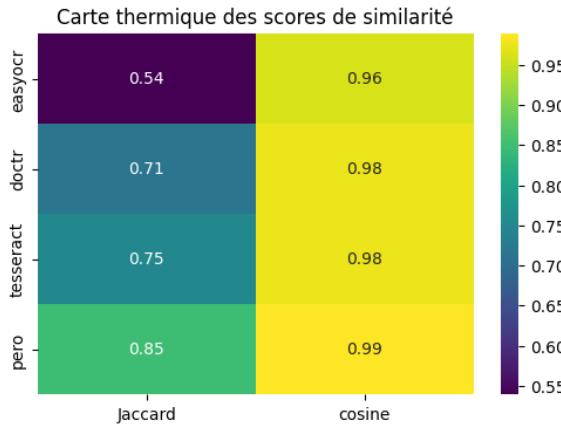


FIGURE 12 – Score de similarité avec une référence numérisé à la main

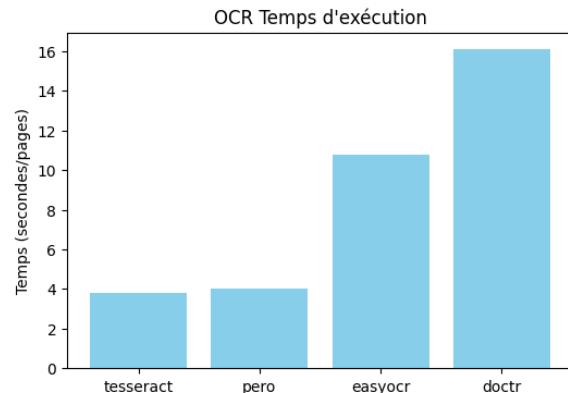


FIGURE 13 – Temps d'exécution par méthode d'OCR

Nous avons sélectionné deux mesures de similarité basées sur la proximité de l'OCR avec neuf brevets manuellement retranscrits (trois par pays). L'ordre des paragraphes n'étant pas particulièrement critique dans cette application et pouvant facilement varier en raison de la mise en page des brevets (alternant entre texte en deux colonnes et texte centré), nous avons choisi des mesures non sensibles à l'ordre des mots. La première est la mesure Jaccard, qui évalue la proximité du texte à travers l'intersection des mots communs rapportée au nombre total de mots distincts. La seconde, Cosine ou coefficient d'Ochiai, correspond au produit scalaire de la représentation vectorielle du corpus (le corpus est d'abord tokenisé selon la méthode "bag-of-words"). [SoftmaxAI \[2021\]](#) [Howson \[2020\]](#)

Comme nous le montre la figure 13, Tesseract demeure la méthode de numérisation la plus rapide. Par souci d'efficacité temporelle sur les millions de documents à traiter, nous avons donc opté pour cette solution sur l'ensemble des nouveaux documents. Les résultats de Tesseract, comme montré par la figure 12, sont également satisfaisants. Au total, nous disposons de numérisations provenant de Pero, notamment pour les brevets français. Une partie des brevets américains ont également été directement obtenus avec leur OCR déjà effectués, comme détaillé dans la partie 2.1.

### 2.3.5 Extraction

Une fois le texte de ces brevets obtenu, nous cherchons à en extraire les informations importantes pour notre analyse. Pour ce faire, nous avons décidé de comparer les performances des grands modèles de langage (Large Language Models) ainsi que des méthodes de traitement automatique du langage naturel plus classiques.



FIGURE 11 – Résultat avec DocTR

## Évaluation des grands modèles de langage (LLM)

Nous avons comparé plus de 22 modèles et mesuré leur performance dans l'extraction d'informations textuelles ou visuelles. Les résultats sont présentés par la figure 14<sup>2</sup> et la Figure 15 détaille le calcul des scores dont nous faisons usage.

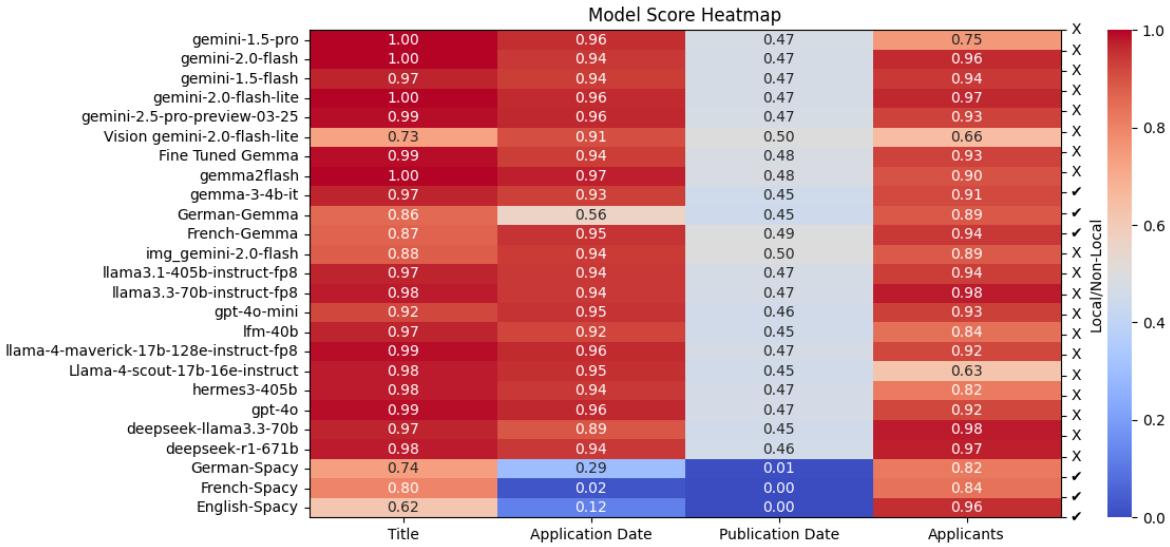


FIGURE 14 – Comparaison de modèles pour déterminer les valeurs manquantes

- 
2. En l'absence d'indication de langue spécifique, les prédictions ont été effectuées sur des brevets américains en langue anglaise.

— Titre :

Avec  $\text{sim}(t_1, t_2)$  le ratio de similarité des tokens du titre et  $\theta_T = 80$  le seuil minimal. Alors :

$$S_T = \begin{cases} 1 & \text{si } t_{\text{véritable}} = t_{\text{prédite}} = \text{None} \\ 1 & \text{si } t_{\text{véritable}} \neq \text{None}, t_{\text{prédite}} \neq \text{None}, \text{sim}(t_{\text{véritable}}, t_{\text{prédite}}) \geq \theta_T \\ 0 & \text{sinon} \end{cases}$$

— Date d'application :

Avec  $d_e$  la date d'application véritable et  $d_a$  la date d'application prédite :

$$S_{AD} = \begin{cases} 1 & \text{si } d_e = d_a \\ 0.5 & \text{si } \text{année}(d_e) = \text{année}(d_a) \text{ et } d_e \neq d_a \\ 1 & \text{si } d_e = d_a = \text{None} \\ 0 & \text{sinon} \end{cases}$$

— Date de publication :

Avec  $p_e$  la date de publication véritable et  $p_a$  la date de publication prédite :

$$S_{PD} = \begin{cases} 1 & \text{si } p_e = p_a \\ 0.5 & \text{si } \text{année}(p_e) = \text{année}(p_a) \text{ et } p_e \neq p_a \\ 1 & \text{si } p_e = p_a = \text{None} \\ 0 & \text{sinon} \end{cases}$$

— Inventeurs/Applicant :

Avec  $A_e$  la liste d'applicant/inventeurs unique, véritable, et  $A_a$  la liste d'applicant/inventeurs prédite. Avec  $M_A$  la taille de l'intersection de ces deux listes :

$$S_A = \begin{cases} 1 & \text{if } A_e = \emptyset \\ \frac{M_A}{|A_e|} & \text{if } |A_e| > 0 \end{cases}$$

FIGURE 15 – Calcul des scores de similarité

La figure 14 présente les scores obtenus par les différents modèles (grands modèles de langage ainsi qu'algorithmes de traitement du langage naturel plus classiques via Spacy). La colonne de droite indique si le modèle a été exécuté localement ou si une API a dû être utilisée, facteur important en cas de déploiement en raison des coûts associés. Les modèles Gemma sont des modèles développés par Google et disponibles pour une exécution locale. Nous obtenons de très bons résultats à partir de l'OCR des brevets en utilisant plusieurs grands modèles de langage. Nous avons également testé des modèles capables d'extraire les informations directement à partir des images des brevets (tel que Gemini-2.0-flash Vision), mais avec des résultats en moyenne moins performants que leurs versions basées sur le texte. Nous observons une différence très notable entre les LLM et les résultats obtenus avec Spacy. Cependant, les ressources computationnelles et les temps d'exécution diffèrent considérablement entre ces approches. La précision diminue particulièrement pour les dates (publication et dépôt de brevets), la détection de dates n'étant pas la spécialité de ce modèle. Cependant, particulièrement soucieux du temps d'exécution et des ressources nécessaires au traitement des données, nous avons choisi de faire usage de modèles Spacy entraînés par pays concerné.

Une amélioration de ce processus aurait pu être d'entraîner le modèle, non seulement une fois par pays d'application, mais aussi pour chaque changement dans les formulations de brevets, tout comme ce qui a été fait par dans la banque de donnée PatentCity. Bergeaud and Verluise [2024] Une autre piste d'amélioration serait de ne pas faire usage de l'OCR pour les éléments formatés de manière uniforme sur les brevets, tel que les dates de publication (voir en haut à droite de la figure 11) mais de faire usage d'un algorithme de reconnaissance d'images.

## 2.4 Validation de la consolidation des données

Nous pouvons faire usage des nombres publiés sur la concentration des brevets auprès des entreprises américaines pour comparer nos résultats. Ceux-ci proviennent du Congrès américain. [Watson and Holman \[1970\]](#).

Nombre d'entreprises	4 premières	8 premières	20 premières	40 premières	100 premières
Nombre de brevets (US Congress)	27'785	44'512	64'178	80'896	113'263
Nombre de brevets calculé	25'559	37'177	55'811	72'283	101'006

Nous observons une différence avec les chiffres de référence, mais celle-ci pourrait être expliquée par la méthodologie adoptée. En effet, notre méthode ne comptabilise que les brevets d'une entreprise portant le même nom, ne prenant donc pas en considération les filiales, contrairement aux données de la source.

Nous sommes aussi en mesure de vérifier que la consolidation des données effectuée nous permet de nous assurer de ne pas développer d'angle mort dans la future analyse en manquant une période historique ou un grand nombre de brevets à certaines périodes, en comparant les nombre de brevets à d'autres banques de données.

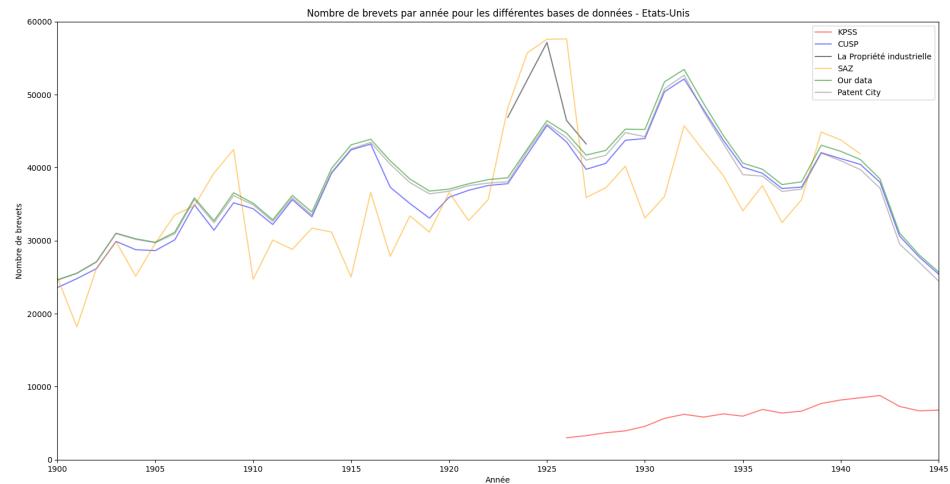


FIGURE 16 – Comparaison entre nos données et d'autres banques de données de brevets aux États-Unis d'Amérique

La figure 16 indique une très grande proximité avec le nombre de brevets d'autres sources. Nous avons ici fait usage des bases de données KPSS [Kogan et al. \[2017\]](#), USPC [Berkes \[2018\]](#), SAZ [Andrews et al. \[2020\]](#), PatentCity [Bergeaud and Verluise \[2024\]](#), et La Propriété Industrielle [UIPPI \[1885-1935\]](#).

Certaines années, la banque de données SAZ semble contenir plus de données, cependant c'est un constat à nuancer par le grand nombre de brevets manquants à d'autres années.

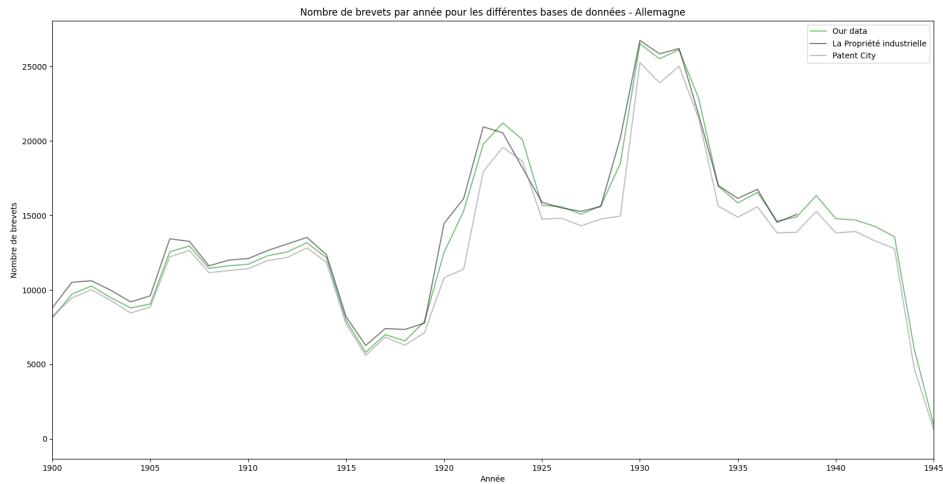


FIGURE 17 – Comparaison entre nos données et d'autres banques de données de brevets en Allemagne

La figure 17 nous permet de valider les brevets d'origine allemande. Nous sommes proches du nombre de brevets décrits par La Propriété Industrielle [UIPPI \[1885-1935\]](#) et nous avons légèrement plus de brevets que la base de données PatentCity possède pour cette période [Bergeaud and Verluise \[2024\]](#).

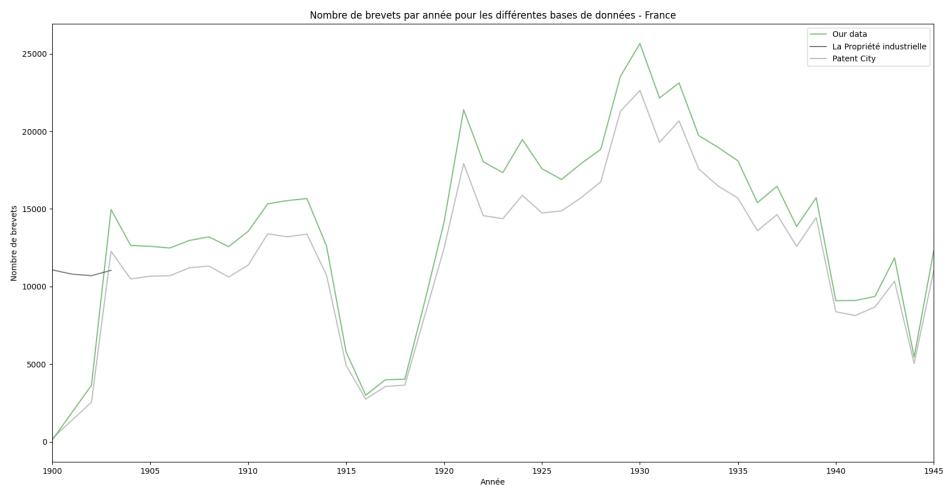


FIGURE 18 – Comparaison entre nos données et d'autres banques de données de brevets en France

À partir de la figure 18, nous observons avoir acquis plus de brevets pour une grande partie de la période sur laquelle nous nous concentrons. Cependant, nous observons un manque notoire de brevets pour les années 1900 à 1902 comparé aux chiffres disponibles dans La Propriété Industrielle [UIPPI \[1885-1935\]](#).

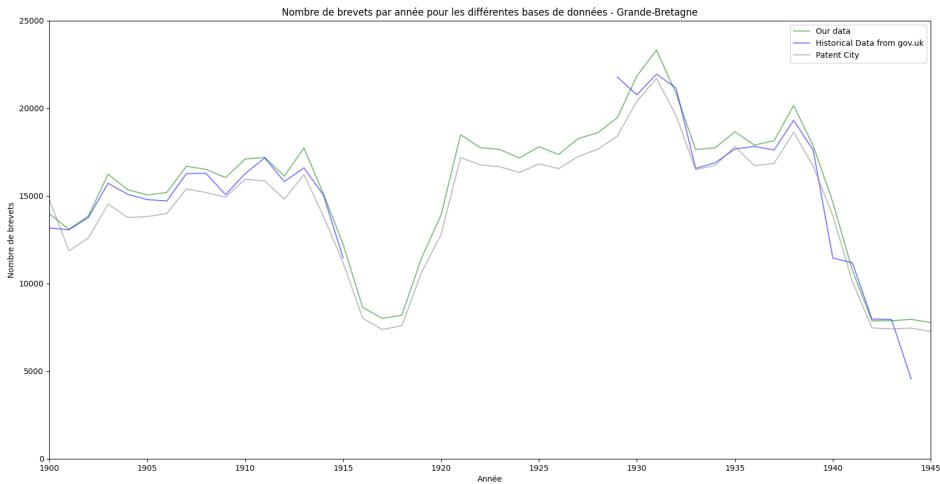


FIGURE 19 – Comparaison entre nos données et d'autres banques de données de brevets au Royaume-Uni

Enfin, la figure 19 démontre que notre base de données possède légèrement plus de brevets que PatentCity [Bergeaud and Verluise \[2024\]](#) et un nombre similaire aux données d'archives britanniques Go [\[2022\]](#).

## 2.5 Post-processing

La majeure partie du post-processing sur les données obtenues est faite afin de pouvoir mettre ensemble les entités nommées.

L'inventeur, la personne qui dépose un brevet et l'entité à laquelle un brevet est assigné peuvent être les mêmes individus. Cependant, ce n'est pas toujours le cas, cette différence dans la relation entre des entités nommées et le brevet peut être difficile à tracer, particulièrement parce qu'elle n'est pas la même dans les différents pays sur lesquels nous nous sommes concentrés dans ce projet et évolue au cours du temps.

Ces difficultés sont également liées aux pratiques légales. Par exemple, aux États-Unis, le brevet n'est accordé qu'au "first and true inventor" et pas à une entreprise. En Allemagne, la demande ne peut être faite par quelqu'un qui n'en est pas l'inventeur (*Anmelderprinzip*) [\[Chachereau, 2022, p. 182\]](#)

D'autres viennent de notre source de données : certaines métadonnées de l'office européen de brevets font la distinction entre l'inventeur et l'entité déposant le brevet. Une distinction qui n'est pas nécessairement faite sur certains brevets et dont nous ignorons comment l'entrée de la donnée a été réalisée. Nous ne faisons donc référence qu'aux entités nommées, celle-ci représentant tous les noms d'entités liés au brevet que nous avons pu identifier sans différencier leur relation au brevet. Par contre, nous ne considérons pas les noms de témoins, parfois mentionnés lors de la dépôse des brevets.

Souhaitant agréger le nombre de brevets possédés par entité nommée, nous devons implémenter un algorithme qui trouve les épellations communes d'une même entité.

Notre travail se concentrant sur une période assez courte (45 ans), dans un nombre de pays limité et principalement sur les entités possédant le plus de brevets, nous n'avons donc pas implémenté de critères géographiques tel que cela a été fait dans d'autres travaux similaires. [Li et al. \[2014\]](#) Nous n'avons pas extrait ces informations et pour des raisons de temps, nous n'avons pas cherché à les extraire ou à faire usage d'autres banques de données ayant fait ce travail d'extraction<sup>3</sup>

Une première étape consiste à standardiser l'écriture afin de maximiser nos chances de trouver des doublons. Nous retirons donc tous les caractères spéciaux ainsi que les termes généraux désignant des entreprises telles que "corporation", "ltd", "incorporated"... À cette fin, nous faisons usage d'une source compilant ces termes. [Networks \[2022\]](#)

Nous transformons ensuite des abréviations communes tel que "mfg" en "manufacturing". Une liste complète d'abréviations a dû être réalisée pour prendre en compte les différences linguistiques de nos données.

3. Tel que PatentCity. [Bergeaud and Verluise \[2024\]](#)

Le nombre d'entités distinctes étant très élevé (plus de quatre millions), il n'est pas judicieux d'appliquer un algorithme de comparaison entre toutes les entités (tel que Levenshtein).

Nous faisons à la place usage d'un algorithme permettant d'identifier des entités similaires avec un coût "sous-linéaire", une implémentation sous python de MinHash LSH. [Zhu \[2024\]](#)

Nous avons ensuite remarqué un nombre d'entités qui n'ont pas été agrégées dans un même cluster malgré leur désignation de la même entité.

Afin d'affiner les résultats, nous avons implémenté une comparaison supplémentaire. Cette fois-ci entre les entités nommées dans un même brevet. Nous calculons la distance Levenshtein ou un pourcentage de mots en commun<sup>4</sup> entre chaque entité nommée dans un même brevet et lions deux "clusters" d'entités si les entités sont suffisamment similaires.

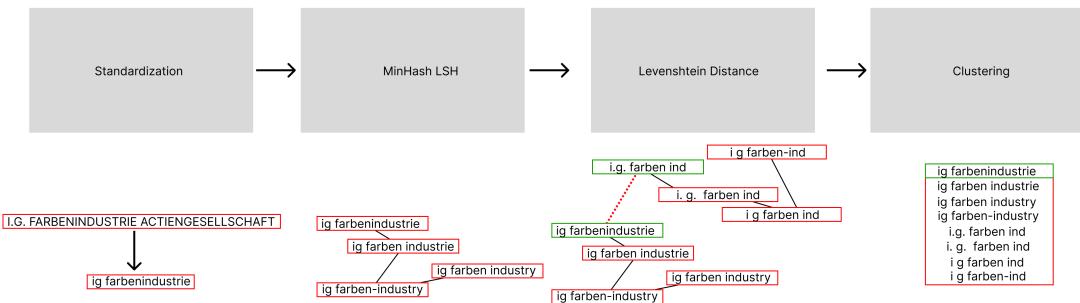


FIGURE 20 – Pipeline Clustering d'entité nommées

Enfin, une fois le graphe de correspondances construit, nous sommes en mesure d'exécuter un algorithme de clustering pour agréger toutes entités connectées en un seul cluster. L'algorithme identifie les clusters en utilisant une recherche en largeur d'abord (BFS). Chaque fois qu'un noeud (entité) non visité est rencontré, une nouvelle traversée BFS est lancée, explorant tous les noeuds connectés à partir de ce point de départ. L'ensemble de tous les noeuds découverts lors d'une même traversée BFS constitue un cluster distinct. Nous déterminons ensuite le nom de chaque noeud, l'algorithme travaillant avec les identifiants de notre banque de données pour des raisons d'optimisation, puis un nom principal pour chaque cluster.

Sur ce jeu de données, nous effectuons quelques nettoyages et clustering supplémentaires, principalement sur les grandes entreprises utilisant des variations dans leur écriture qui n'ont pas été mise en commun par l'algorithme.

Au total, nous obtenons 1.8 million de clusters, donc d'entité nommée théoriquement unique.

## 2.6 Limites

Nous pouvons noter plusieurs limites techniques dans la consolidation des données. Nous avons également déjà mentionné plusieurs pistes d'améliorations durant ce chapitre.

La précision dans la détection des entités liées aux brevets a malheureusement été limité par le coût computationnel liée à des méthodes plus avancées, tel que les LLM. Cependant, nous aurions dû entraîner plusieurs modèles Spacy par office de brevets pour gérer les changements de forme des brevets à travers l'histoire<sup>5</sup>.

Comme mentionné, nous ne faisons pas usage des erratum accompagnés de certains brevets dû à leur usage assez rare. Cependant, ces données pourraient être ajoutées à notre méthodologie.

Le modèle de classification employé ne prédit que les sous-classes dont nous avions quelques dizaines d'exemple durant l'entraînement. Ceci conduit notre modèle à ne pas prédire certaines classes, l'augmentation de la taille des données utilisées pour l'entraînement (tel que des données contemporaines) pourrait nous permettre de couvrir toutes les sous classes existantes.

Également concernant la classification, nous faisons uniquement usage du titre de l'invention. Comme l'ont prouvé plusieurs recherches sur ce sujet, il pourrait être bénéfique d'ajouter des informations supplémentaires sur chaque brevet pour obtenir une meilleure prédiction.

4. Grâce à Rapidfuzz [Bachmann \[2024\]](#)

5. Comme réalisé par PatentCity [Bergeaud and Verluse \[2024\]](#)

Enfin, malgré toutes les étapes de notre consolidation, nous n'avons pas réussi à acquérir toutes les données dans le temps imparti par ce projet. Le manque le plus marquant se situe en France entre 1900 et 1903.

### 2.6.1 Réflexion sur le clustering

A travers la partie expérimentation 3, nous sommes confrontés à une multitudes de cas nuancé où il aurait été bénéfique d'avoir une validation humaine lors du clustering.

C'est par exemple le cas pour l'entreprise allemande Siemens. Dans notre banque de données, les brevets de Siemens incluent ceux de Siemens & Halske, ainsi que des brevets de Siemens-Schuckertwerke. L'office européen de brevets, lie un certain nombre de brevets de "Siemens-Schuckertwerke AG" et de "Siemens & Halske AG" à "Siemens AG", liant ainsi dans nos données ces deux entités. Cependant, Siemens AG n'est formé qu'en 1966 de la fusion de "Siemens & Halske AG", "Siemens-Schuckertwerke AG" et "Siemens-Reiniger-Werke AG". Bien que Siemens-Schuckertwerke ait été une entité juridique distincte formée à partir des activités de courant fort de S&H et Schuckert, le contrôle stratégique et la direction étaient dirigées par Siemens & Halske et la famille Siemens, limitant probablement la concurrence entre ces deux entreprises, nous permettant de continuer notre analyse en cumulant leurs brevets ensemble. [Siemens \[1957\]](#)

La décision de si deux entités partageant des noms très proche doivent être liées peut cependant plus difficile et varie dans le temps, avec des acquisitions et des ventes d'entreprises rendant la mise en commun de brevets uniquement admissible sur une partie limitée de la période par exemple.

Il s'agit d'une limite assumée de notre méthodologie, car ces choix sont impossibles à réaliser manuellement sur une aussi grande période que celle de notre projet. Cependant, la transparence de notre algorithme, permettant de retracer les choix fait par le clustering, nous permet de corriger toute association erronée à posteriori.

## 3 Expérimentation

L'objectif de cette partie du projet est de démontrer le potentiel de ces données. Pour ce faire, nous chercherons à mieux comprendre les différentes données puis développerons une vue d'ensemble de la mesure de la concentration avant de rentrer dans deux études de cas spécifiques. Finalement, nous regarderons l'impact de changements d'usage et législatifs sur les données.

### 3.1 Classifications

Ce chapitre commence par une vue d'ensemble sur les classifications et leur usage. Avec ces données, nous sommes en mesure de calculer la proportion relative de brevets par classification. Mesurer le nombre de classifications par pays biaise les données dû au nombre supérieur de brevets américain, nous divisons le nombre de classifications par le nombre de brevets publié par chaque pays. Sur la figure 21, nous remarquons de grandes variations, notamment pour la section Y, particulièrement populaire pour les brevets américains. Celle-ci couvrant les nouvelles technologies sans classifications existantes, mais surtout les brevets classifiés sous le système américains USPC (Y10).

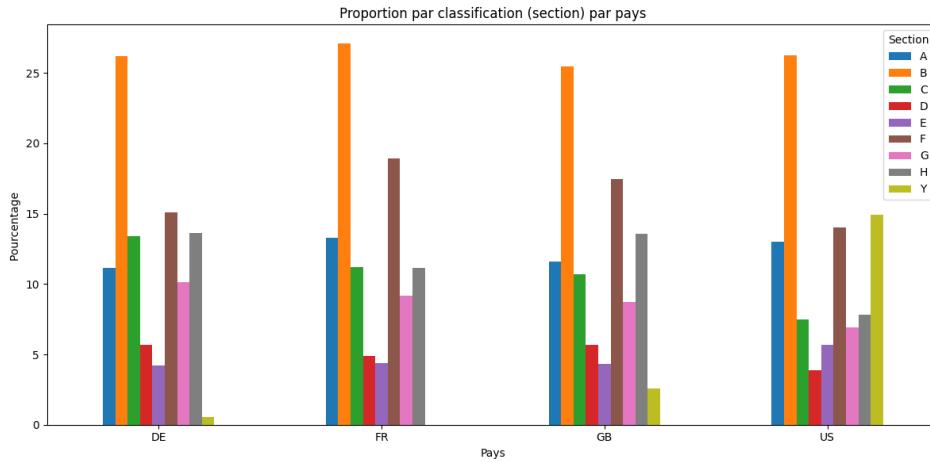


FIGURE 21 – Proportion de brevets appartenant à chaque classe par pays

Section	Description
A	Les besoins humains
B	Opérations et transport
C	Chimie et métallurgie
D	Textiles et papier
E	Constructions fixes
F	Mécanique, éclairage, chauffage...
G	Physique
H	Électricité
Y	Nouveau développements, anciennes collections

Les brevets sont liés à plusieurs classifications. C'est particulièrement le cas de la section Y, notamment la classe Y10. Les brevets liés à cette classification possèdent également d'autres classifications CPC. Nous ne nous intéressons donc pas à cette classe, car elle ne donne pas beaucoup d'informations sur l'invention. Cela pourrait être le cas d'autres classes, pour cette raison, nous allons analyser la cooccurrence des classifications.

A noter que la classification CPC a été créée pour aider les examinateurs de brevets dans les recherches. [Billington and Hanna, 2021, p. 8] CPC est basé sur la fonctionnalité du brevet et non sur l'industrie où le brevet sera utilisé. Cette limitation peut alors brouiller l'interprétation, rendant les analyses par marché plus complexe.

### 3.2 Cooccurrence

Nous pouvons essayer de mieux comprendre l'utilisation des classifications en étudiant la cooccurrence des sections et des classes dans les brevets. La figure 22 illustre, comme mentionné dans 3.1 que les brevets de section Y possèdent pour une très grande partie d'autres sections. Les autres sections partageant moins de cooccurrence.

Mieux comprendre la cooccurrence dans la classification CPC peut nous permettre par exemple à filtrer les sections ou classes qui ne sont pas spécifique à un domaine technologique et dont les brevets et l'activité de leurs inventeurs ne sont pertinentes à analyser.

La figure 23 illustre en échelle logarithmique les cooccurrences entre les classes.

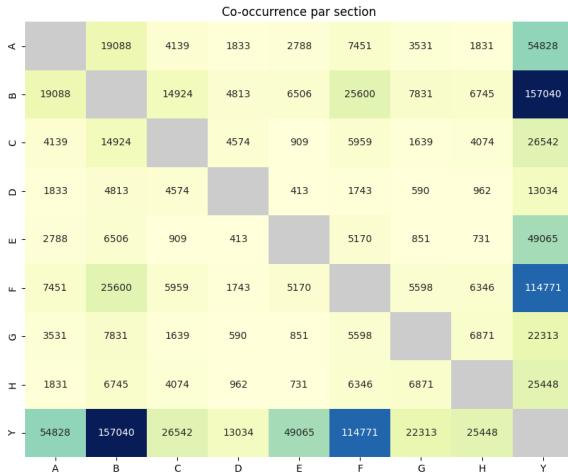


FIGURE 22 – Cooccurrence des sections

Un regard détaillé sur le nombre de sections par brevets nous permet de mieux comprendre la répartition de nos données. La figure 21 détaille qu'une majorité (73.47%) des brevets ne possèdent qu'une section unique. Un brevet comptant deux classifications au sein d'une même section est ici considéré comme appartenant à une seule section<sup>6</sup>.

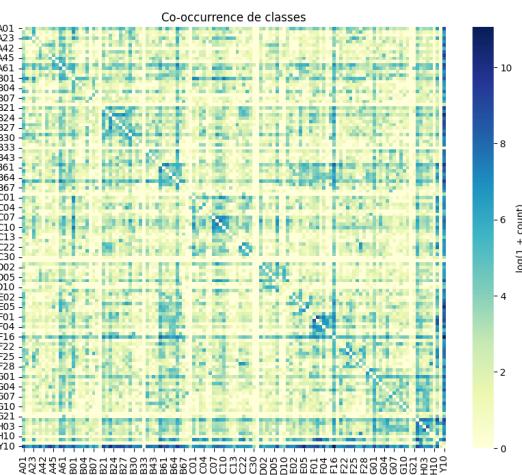


FIGURE 23 – Cooccurrence des classes (en log)

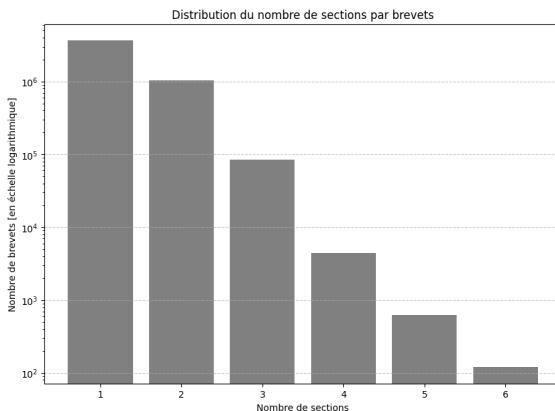


FIGURE 24 – Distribution du nombre de sections (en log)

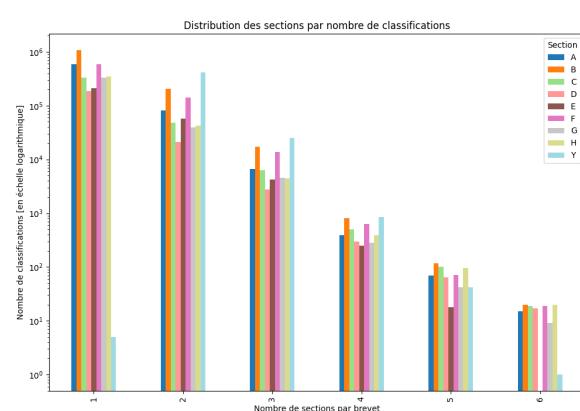


FIGURE 25 – Distribution du nombre de sections détaillées (en log)

La figure 21 nous permet de voir que la section Y est souvent la section accompagnatrice et très représentée dans les brevets avec plus d'une section. Elle n'est présente que 5 fois comme unique section (colonne 1) dans nos données. Nous observons aussi que la section E<sup>7</sup> n'est pas présente dans les brevets avec 6 sections, cependant cela peut être expliqué par le faible nombre de brevets possédant 6 sections : au total 120.

La classification CPC étant hiérarchisée, de nombreux brevets ne possèdent cependant une unique section, mais sont classifiés dans plusieurs classes dans cette section. Nous procédons de même pour les sous-sections.

On s'aperçoit sur la figure 27 que certaines classes ont beaucoup de cooccurrences (tel que B65<sup>8</sup> ou F16<sup>9</sup>), il s'agit de classes très généralistes. Dans l'objectif de trouver des domaines avec une forte concentration industrielle, il ne semble alors pas pertinent de développer des études de cas sur ces classes qui peuvent couvrir des inventions liées à plusieurs marchés distincts.

6. Par exemple lié à deux classes différentes dans un même section.
7. Construction fixe
8. Transport, emballage, stockage, manipulation de matériaux minces ou filamenteux
9. Éléments et unités d'ingénierie, mesures générales pour produire et maintenir un fonctionnement efficace des machines ou des installations, isolation thermique en général

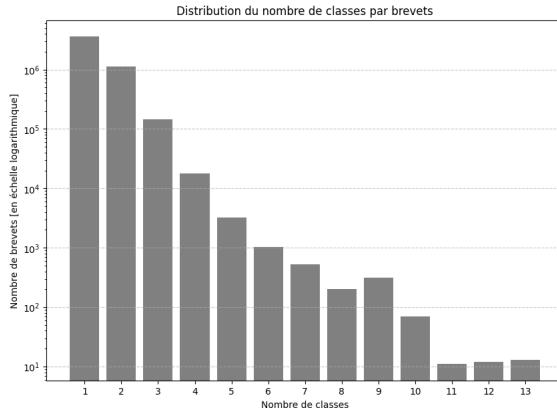


FIGURE 26 – Distribution du nombre de classes (en log)

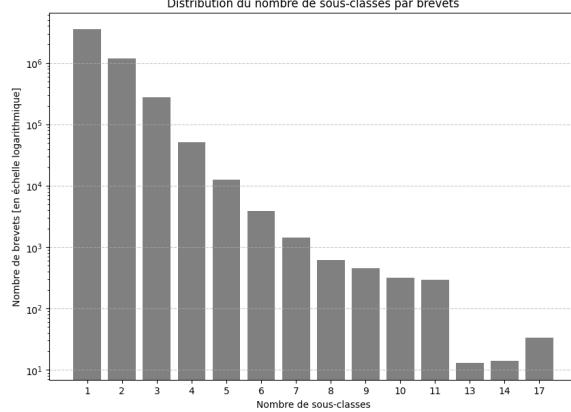


FIGURE 27 – Distribution du nombre de sous-classes détaillées (en log)

Nous pouvons observer sur la figure 26 que la distribution des sous-classes suit la même logique que la distribution des sections, avec une majorité d'éléments classés dans une seule classe (ou sous-classe, voir 27), mais une quantité néanmoins importante de brevets avec de multiple classifications.

L'usage des classifications influence l'analyse qu'on peut en faire. Nous devrons prendre en considération l'utilisation asymétrique de certaines classifications. Certaines étant utilisées en supplément à une classification primaire (comme F16), et pas nécessairement relative à l'industrie où l'invention trouve usage.

Pour déterminer si l'usage des brevets entre les grandes entreprises et le reste de notre base de donnée diffèrentes, nous avons mesuré le nombre de classifications que possèdent des brevets provenant des 5 premières entités nommées par pays et par année.

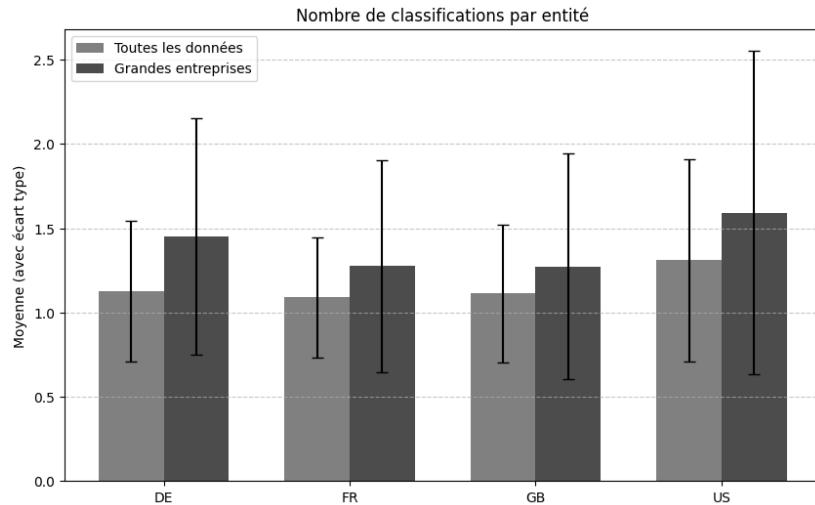


FIGURE 28 – Nombre de classifications par brevets

Nous observons une différence dans le nombre de classifications, avec les premières entreprises nommées (Grandes entreprises sur la figure 28) qui possèdent en moyenne plus de classifications que le reste de nos données. Cette différence pourrait par exemple venir de la généralisation accrue de certaines grandes entreprises, permettant des brevets liés à plusieurs domaines distincts, à l'intersection de leur compétences. Cette hypothèse est à nuancer en raison de la grande déviation standard observée dans nos données.

### 3.3 Concentration

Mettre en place des outils pour analyser la concentration industrielle est l'objectif principal de ce projet, dans ce but nous commençons par introduire des outils pour le calcul de la concentration, puis

les appliquons à quelques études de cas, où niveau des sections et des classes.

### 3.3.1 Mesure de la concentration

Notre première mesure de la concentration est le calcul de la proportion de brevets lié à une ou plusieurs entités nommées. Un ratio de concentration pour une entreprise de 20% veut ainsi dire que 20% des brevets sont liés à cette entreprise. Alors qu'un ratio de concentration de 50% pour les 10 premières entreprises veut dire que 50% des brevets sont liés à aux 10 entreprises possédant le plus de brevets.

$$C_i = \frac{E_i}{P_{total}} \times 100$$

- $C_i$  = concentration de l'entité  $i$  (en pourcentage)
- $E_i$  = nombre d'entrées liées à l'entité  $i$
- $P_{total}$  = nombre total de brevets dans la base de données

Si l'analyse porte sur les sections, classes ou sous-classes, nous ne comptabilisons qu'une entrée unique par niveau de classification considéré. Par exemple, si un brevet possède les classifications A01 et A02 mais que nous nous intéressons uniquement à la concentration au niveau de la section A, ce brevet ne sera comptabilisé qu'une seule fois.

Ainsi, en considérant la concentration des  $X$  premières entités :

$$S_X = \frac{\sum_{i=1}^X E_i}{P_{total}} \times 100$$

Avec  $S_X$  la concentration des  $X$  premières entités, avec les entités ordonnées par nombre d'entrées. La somme  $\sum_{i=1}^X E_i$  est la somme des entrées, filtrée pour ne pas comptabiliser les doublons, tel qu'un brevet lié à deux entités dans les  $X$  premières entités nommées. La mesure  $S_4$  correspond à la "Four-Firm Concentration Ratio". Cette métrique identifie la contribution des quatre premières entreprises. Cette mesure normalement appliquée aux ventes dans un domaine spécifique, qualifie le marché d'oligopole s'il atteint 60% de part de marché. [Parkin and Bade, 2006, p. 275]

Notre méthodologie est moins adaptée à l'indice de Herfindahl-Hirschman (HHI), celui-ci calculant la concentration d'un marché en considérant la part de marché de chaque entité. Rhoades [1993] Notre méthode de clustering se concentrant sur les grandes entreprises et risquant de regrouper les brevets de nombreux individus, les noms d'individus étant souvent plus similaires que les noms d'entreprises, elle leur attribuerait une part de marché disproportionnée, faussant ainsi cette mesure. De plus, les individus liés aux brevets d'une entreprise ne constituent pas nécessairement un concurrent sur un marché.

Cependant, nous essayerons ponctuellement de faire usage de cet indice en addition du ratio de concentration. Cette mesure ayant l'avantage, en mettant au carré la part de marché de chaque entité, d'incorporer la notion selon laquelle plus la concentration de la production dans un petit nombre d'entreprises est importante, plus la probabilité que la concurrence sur un marché soit faible est grande. Rhoades [1993] Pour le calcul de l'indice Herfindahl-Hirschman, nous limitons alors le nombre d'entreprises aux 50 premières, comme il peut en être l'usage. [Parkin and Bade, 2006, p. 275] Une mesure de 0.1 à 0.18 correspond à un marché en compétition. Au delà de 0.18, le marché est considéré comme peu concurrentiel. [Parkin and Bade, 2006, p. 276]

### Limite de la mesure

La mesure de la concentration du domaine de la propriété intellectuelle a des limites claires qui ont des conséquences sur la façon dont nous pouvons interpréter nos données.

La première est que deux brevets ne sont pas égaux. Contrairement à un marché, où deux entreprises peuvent rivaliser pour un service ou une marchandise auprès de la même clientèle, nous ne sommes qu'en mesure de comptabiliser les brevets et regarder à quelles entités ceux-ci sont liés. Cependant deux brevets n'ont pas la même valeur, une invention, et le monopole temporaire offert par un brevet sur cette invention peut donner une longueur d'avance à une entreprise pour capter un marché, d'autres brevets peuvent avoir très peu d'importance sans grande avancée technologique. L'usage réel des brevets, certains pouvant ne jamais être mis en production par exemple, ne peut pas être pris en compte dans

notre analyse. Des inventeurs peuvent décider de ne pas breveter leurs invention et de miser sur le secret industriel.

Enfin la politique de recherche d'une entreprise influe énormément, qu'elle acquiert des brevets auprès d'inventeurs indépendants (qu'on ne peut pas lier à l'entreprise, si le brevet n'est pas assigné au dépôt) ou possède un centre de recherche et développement interne, va grandement impacter le nombre de brevets lié à une entité.

Même dans le second cas, une entreprise peut décider de déposer uniquement les brevets économiquement viable ou pour chaque technologie développée par ses ingénieurs. Par exemple pour une question de morale au sein des employés de l'entreprise. [Lamoreaux and Sokoloff \[1999\]](#) Ces différences de politiques peuvent créer une sur-représentation ou un sous-représentation d'une entreprises. Les raisons peuvent également être fiscales, afin de diminuer son bénéfice imposable. [[Chachereau, 2022, p. 270](#)]

Il est donc important de noter que nous ne mesurons que la part de brevets lié à des entités, celle-ci n'étant pas un reflet exact de la part du marché que possède cette entreprise ou de son importance relative à cette période.

### 3.3.2 Vue globale

Nous commençons notre exploration par une vue statique de la mesure de concentration dans les différents pays.

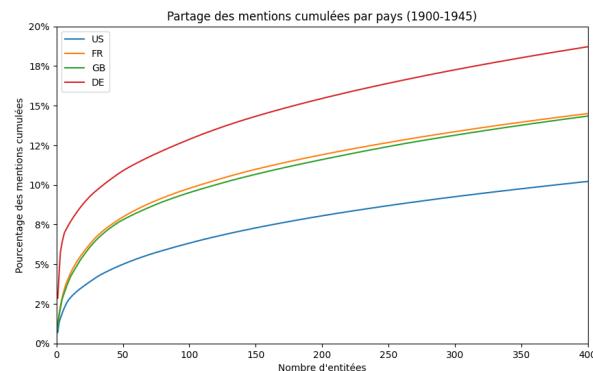


FIGURE 29 – Partage des mentions cumulées par pays jusqu'à 250

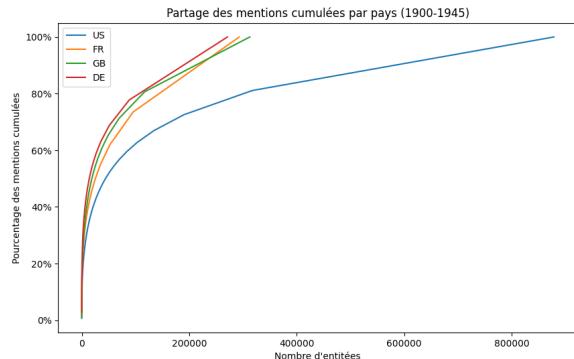


FIGURE 30 – Partage des mentions cumulées par pays

On s'aperçoit grâce aux figures 29 et 30 que l'Allemagne détient une plus forte concentration d'entités liées à un brevet, avec près de 15% des brevets liés à uniquement 200 entités nommées. Les États-Unis eux, semblent posséder le moins de concentration, avec environ 8% des brevets liés à 200 entités.

Plusieurs hypothèses peuvent expliquer la concentration supérieure observée en Allemagne. La première est que, dans un système de propriété intellectuelle plus difficile d'accès, avec une recherche d'antériorités (prior examination) contrairement à d'autres pays, le coût pour l'obtention d'un brevet est supérieur en Allemagne, exigeant également un renouvellement annuel croissant pour garder le brevet. [[Khan, 2005, p. 294](#)] Ceci pourrait alors restreindre l'accessibilité des brevets aux entreprises ayant les moyens, augmentant le ratio de concentration allemand.

De plus, les crises économiques, telles que les difficultés traversées après la Première Guerre mondiale, auraient pu complexifier l'accessibilité au dépôt de brevets, particulièrement pour les petits inventeurs. C'est en tout cas ce que semble

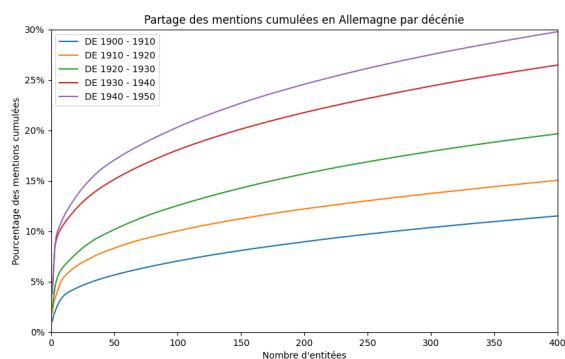


FIGURE 31 – Évolution de la concentration de brevets en Allemagne

indiquer la figure 31. Après tout les décisions de déposer une demande de brevets sont sensibles à la conjoncture économique. [Diebolt and Pellier, 2012, p. 619] On y voit que la concentration des brevets allemand par décennie a beaucoup augmenté, initialement (1900 à 1910) se situant à un niveau similaire à la moyenne américaine.

La structure de l'économie allemande pourrait aussi expliquer cette forte concentration : le pays possède de nombreux accords de cartels. Un historien allemand estime que le nombre d'accords a évolué de 4 en 1865, 70 en 1887, 300 en 1900, 600 en 1911, 1'000 en 1922, 2'100 en 1930 et 2'500 en 1943. [Crane, 2019, p. 1335] Cette cartellisation était encouragée par la politique de l'État allemand. [Crane, 2019, p. 1333] La pratique est légalisée en 1923 et l'influence des banquiers au sein des conseils d'entreprises<sup>10</sup>, favorable à la pratique, aurait joué un rôle dans la formation de nombreux cartels. [Barjot, 2013, p. 1045]

### 3.3.3 Étude temporelle

Comme développé dans la partie 3.3.2, nous sommes en mesure de calculer une vue globale statique permettant de juger de la concentration dans le domaine de la propriété intellectuelle dans un pays, mais nous souhaitons également développer une approche plus dynamique, nous permettant de comprendre l'évolution de cette mesure à travers le temps.

C'est donc l'objectif de cette partie. Nous souhaitons également voir quelles grandes variations sont visibles grâce à cette mesure et faire des hypothèses sur les causes de ces variations.

Nous calculons ce ratio par pays et par an afin d'obtenir la figure 32.

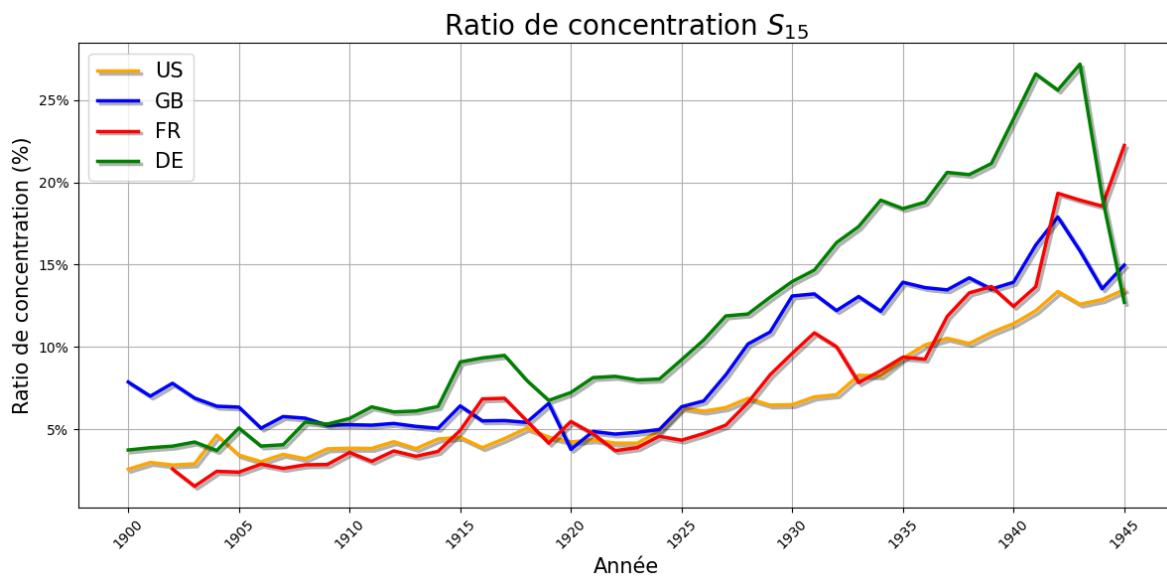


FIGURE 32 –  $S_{15}$  entre 1900 et 1945

Le graphique 32 permet de voir l'augmentation de la concentration entre 1900 et 1945.

Nous nous aidons également de la figure 33, pour comprendre l'effet des différentes crises traversées sur les plus grandes entreprises.

10. Pour certains, l'imbrication entre secteur bancaire et cartels industriels est inévitable. [Hilferding, 1981, p. 221-226] Cette proximité entre ces deux secteurs pourrait décourager la concurrence.

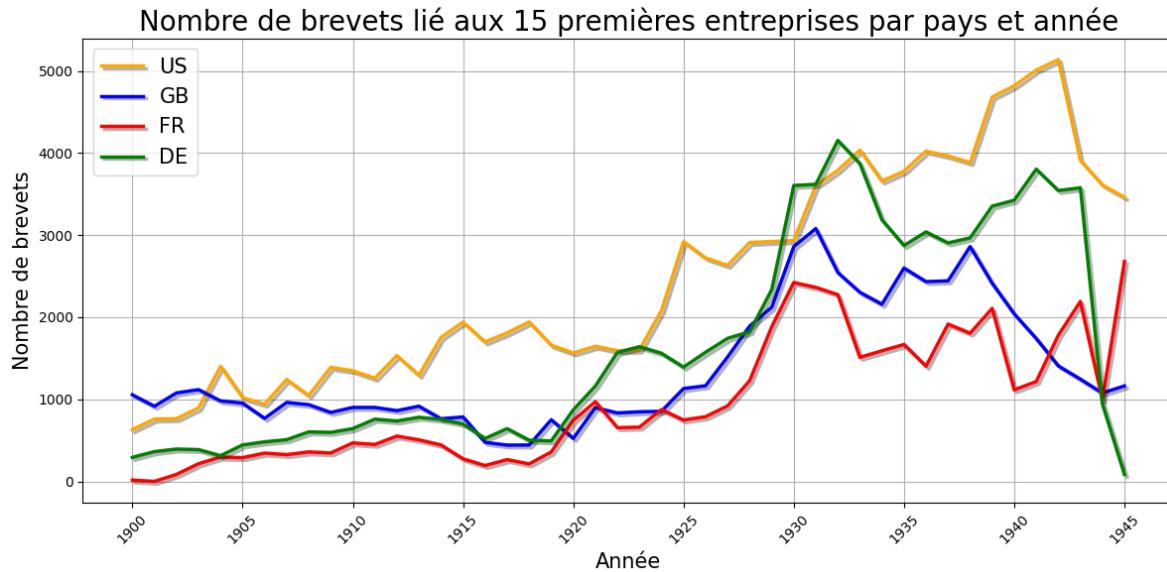


FIGURE 33 – Nombre de brevets pour les 15 plus grandes entreprises par année

Particulièrement visible est la forte augmentation en 1915 et 1918, correspondant à la Première Guerre mondiale. Inversement, les figures 19, 17, 18 nous indique une baisse dans le nombre de brevets délivré ces années-là. Les États-Unis semblent moins touchés par cette augmentation de concentration durant cette période, mais nous pouvons aussi constater une forte augmentation de brevets délivrés sur la figure 16. La figure 33 indique que cette période a aussi été un moment de baisse du nombre de brevets délivrés en Europe pour les entreprises en obtenant habituellement le plus. L'augmentation temporaire de la concentration durant la Première Guerre mondiale semble donc être le résultat de la chute du nombre de brevets obtenus durant cette période.

Les effets de la guerre sur la concentration de la propriété intellectuelle sont difficile à cerner. Certains gouvernements ont par exemple suspendu les brevets si une technologie pouvait avoir une application militaire. Et pour certaines entreprises, tel que Marconi (voir partie 3.3.5), la guerre fut une opportunité d'étendre leur "agenda monopolistique". [Christophers, 2016, p. 134] Il semblerait cependant durant la Première Guerre mondiale, les grandes entreprises jouissent plus de la propriété intellectuelle que le reste des acteurs du domaine.

Les conséquences de la grande dépression semblent également visibles sur la figure 33. Le krach boursier débute en 1929 et dure jusqu'à la fin des années 1930. Deux à trois ans plus tard<sup>11</sup>, le nombre de brevets publié par les plus grandes entreprises de chaque pays chute considérablement, démontrant l'impact de la grande dépression à travers l'Europe.

Nous observons également une chute dans la croissance de la concentration  $S_{15}$  observée durant la dernières décennie (1920-1930). C'est en France, que la chute semble avoir le plus d'impact, avec un recul de la concentration  $S_{15}$  d'environ 11% à 7%. La concentration au Royaume-Uni stagne durant près d'une décennie. Alors que l'Allemagne et les États-Unis ne semblent pas être touché. Une hypothèse peut être que les grandes entreprises, malgré avoir publié moins de brevets durant cette crise, restent nettement moins impacté que le reste des acteurs de la propriété intellectuelle.

Enfin, en dehors variations suite à des crises économiques et politiques, la figure 32 illustre l'augmentation conséquence dans la propriété intellectuelle dans la première moitié du XX<sup>e</sup> siècle. Déceler cette tendance dans nos données, nous donne bon espoir sur le potentiel descriptif de notre méthode pour comprendre l'usage de la propriété intellectuelle.

Nous observons que malgré une tendance positive tout au long de la première partie du XX<sup>e</sup> siècle, il existe des différences notoire entre les pays. Nous allons donc chercher à détailler cette évolution dans le ratio de concentration par pays.

## Allemagne

11. Correspondant à l'attente moyen entre la demande et la publication d'un brevet

Comme mentionné, la concentration augmente en Allemagne entre 1918 et 1914. En regardant en détail, en 1918, les entreprises en Allemagne déposent 51% de brevets en moins qu'en 1914. Pour les 15 premières entreprises, cette différence n'est que de 35.5% de brevets. Nous observons également que la concentration chute directement après la guerre pour retrouver un niveau similaire à celui de 1914.

En sortie de la Première Guerre mondiale, alors que la concentration diminue et se stabilise, nous observons tout de même une forte augmentation du nombre de brevets déposés par les premières entreprises du pays (Figure 33). Cette figure nous permet d'analyser qu'il y a une reprise de l'activité lié à la propriété intellectuelle et pas uniquement dans les grandes entreprises du pays. La concentration  $S_{15}$  en Allemagne dépasse celle des autres pays considérés ici à partir de 1920. En effet, comme vu dans la partie 3.3.2, le ratio de concentration en Allemagne est particulièrement haut à partir de cette période.

En regardant le figures 34 et 33, nous pouvons identifier d'autres moments clés pour le pays :

- Entre 1921 et 1924, alors que la république de Weimar connaît une crise d'hyperinflation. Nous observons une grande augmentation du nombre de brevets délivrés à cette période, éventuellement une conséquence de la chute du coût nécessaire à la publication d'une invention.
- Le nombre de brevets est en chute entre 1924 et 1928, une chute qui n'a pas impacté les plus grandes entreprises du pays. Une baisse de 21.7% des brevets déposés entre ces deux années alors que les 15 premières entreprises du pays voient leur nombre de brevets augmenter de 17%.
- En 1935, l'Allemagne publie 38.6% moins de brevets qu'en 1932. Cette fois les 15 premières entreprises, sont touché par une chute de 30.9% de brevets. Néanmoins, cela n'empêche pas le ratio de concentration d'augmenter comme nous pouvons le voir sur la figure 34. Comme mentionné, cette chute pourrait être la conséquence en Allemagne de la crise de 29.

Cette différence dans la capacité de grandes entreprises à mieux résister aux crises traversées par les autres acteurs de la propriété intellectuelle paraît naturel vu que durant les périodes de choc économique, les gouvernements ont souvent permis, voire encouragé, les ententes et la consolidation, estimant que l'excès de concurrence était alors préjudiciable. [Christophers, 2016, p. 139]

Entreprise	Nombre de brevets
Siemens	23460
AEG	12554
IG FarbenIndustrie	11306
Fried Krupp	4128
Telefunken	3202
Brown Boveri & Cie	2672
C Lorenz AG	1398
United Shoe Machinery	1273

TABLE 1 – Nombre de brevets possédé par les 8 premières entreprises en Allemagne

Les premières entreprises du pays sont majoritairement allemandes, à l'exception de United Shoe Machinery (États-Unis) et Brown Boveri & Cie (Suisse).

## États-Unis

La période est aussi clé dans le développement de nombreuses grandes entreprises américaines.

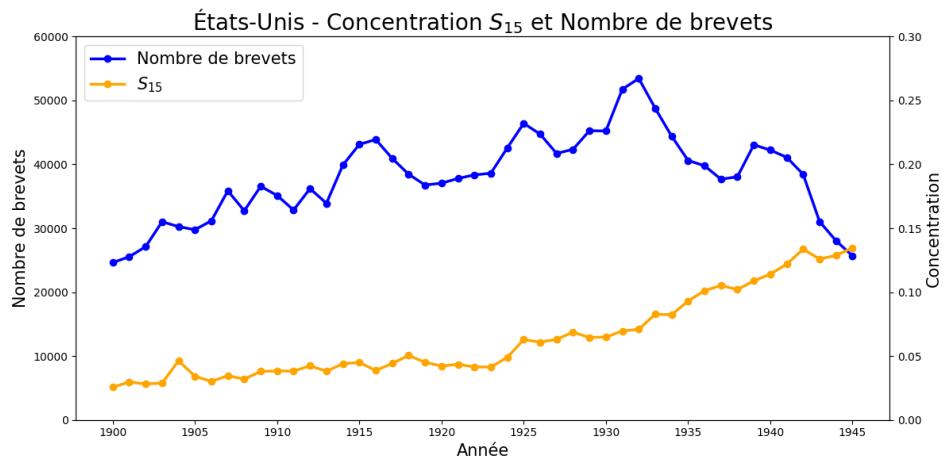


FIGURE 35 – Nombre de brevets publié  
et concentration aux États-Unis

Comme nous permet de le voir la figure 35, les brevets américains sont logiquement moins impacté par la Première Guerre mondiale que ceux du vieux continent. Nous pouvons également voir que les grandes entreprises actives aux États-Unis sont américaines, limitant ainsi l'impact de la guerre sur le marché de la propriété intellectuelle américain.

Entreprise	Nombre de brevets
General Electric	20750
Westinghouse electric & manufacturing company	18325
United Shoe Machinery Company	8071
Western Electric	7285
Radio Corporation of America	6787
Du Pont	5741
Bell Telephone Labor Inc	5436
General Motors Corporation	4906
Kodak	3941
AT&T	2885

TABLE 2 – Nombre de brevets possédé par les 10 premières entreprises aux États-Unis

Ce n'est cependant pas le cas de la crise de 29, qui a pour conséquence une chute historique dans le nombre de brevets déposés. Malgré un rebond en 1940, la tendance perdure jusqu'au moins en 1945. Cependant la figure 35 illustre que les grandes entreprises américaines ne semblent pas avoir souffert, gagnant près de 7 points de concentration entre 1930 et 1945.

### Grande-Bretagne

Tout comme en Allemagne, la propriété intellectuelle au Royaume-Uni est particulièrement impactée par la Première Guerre mondiale, comme nous pouvons le voir à travers le nombre de demande de brevets durant cette période.

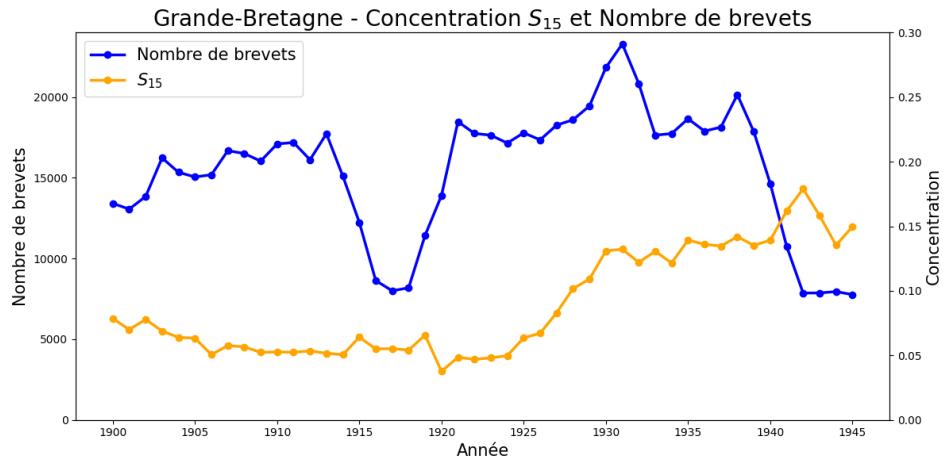


FIGURE 36 – Nombre de brevets publié et concentration au Royaume-Uni

Cependant, nous pouvons déceler les effets de la Seconde Guerre mondiale plusieurs années avant de voir les effets en Allemagne sur la figure 36, avec une baisse conséquence à partir de 1934. Contrairement à l'office allemand et américain des brevets, le temps d'attente pour la publication d'un brevet semble être stable et d'environ une année avec une faible déviation standard. Cette particularité pourrait expliquer la chute abrupte observée.

La figure 36 nous permet également de voir que contrairement à la croissance constante dans la concentration en France, Allemagne ou aux États-Unis, le Royaume-Uni a une concentration plutôt stable avant 1925 mais connaît une augmentation de 7 points en seulement 5 ans avant de se stabiliser.

Nous trouvons en Allemagne et aux États-Unis que les premières entreprises par nombre de brevets dans le pays sont pour la grande majorité des géants nationaux. L'industrie des brevets anglais semble cependant compter de nombreux acteurs étrangers auprès des entreprises les plus actives.

Entreprise	Nombre de brevets
The British Thomson Houston Co Ltd	10336
I.G. FarbenIndustrie	7466
Siemens	4845
General Electric	4607
Imperial Chemical Industries (ICI LTD)	2417
Standard Telephones and Cables Limited	2306
Marconi Wireless Telegraph Company	2085
Philips	2081

TABLE 3 – Nombre de brevets possédé par les 8 premières entreprises au Royaume-Uni

The British Thomson Houston Co Ltd été majoritairement possédé par l'américain General Electric. [Science Museum Group \[2024\]](#) Standard Telephones and Cables Limited est une filiale de Western Electric jusqu'en 1925 avant d'être racheté par l'américain ITT Inc. [Lewis \[2019\]](#)

D'autres entreprises restent anglaise, tel que Marconi et Imperial Chemical Industries.

## France

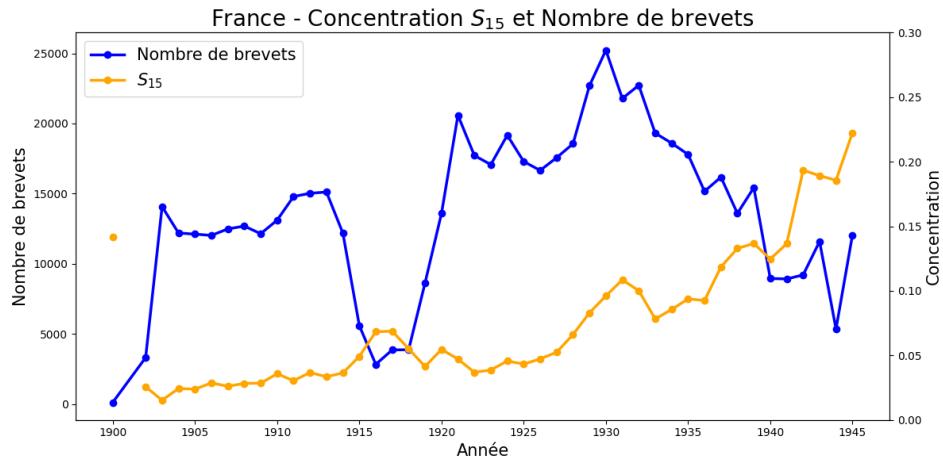


FIGURE 37 – Nombre de brevets publié  
et concentration en France

Comme nous pouvons le constater grâce à la figure 37, le ratio de concentration  $S_{15}$  en France augmente considérablement en 1900 et 1945, les 15 premières entreprises française en 1945, étant lié à plus de 20% des brevets du pays à la fin de cette période. La Première Guerre Mondiale et la grande dépression sont deux évènements qui semblent temporairement renforcer cette concentration mais ne durent pas plus de 2 à 4 ans. Nous constatons également que la concentration augmente particulièrement avec l'arrivée de Seconde Guerre Mondiale et la chute dans le nombre de brevets publié qui l'accompagne.

Alors que le nombre de brevets chute en 1939, la tendance était déjà à la baisse depuis 1932. Le nombre de brevets publié par les grandes entreprises durant cette période reste assez stable, comme l'illustre la figure 33. Cette augmentation de concentration semble alors être le résultat de la chute dans le nombre de brevets publié par les autres acteurs de la propriété intellectuelle.

Similairement au Royaume-Uni, en France, peu d'entreprises dans les plus grands acteurs de la propriété intellectuelle sont d'origine française.

Entreprise	Nombre de brevets
I.G. FarbenIndustrie	8890
Compagnie Française Thomson-Houston	4914
Philips	2991
Brown Boveri	2267
Le Matériel Téléphonique	2148
Westinghouse	1627
Fried Krupp	1422
Bosch	1342

TABLE 4 – Nombre de brevets possédé par les 8 premières entreprises en France

Em effet, I.G. Farbenindustrie est une entreprise allemande, nous discutons plus de cette entreprise dans l'étude de cas développée en partie 3.3.5. Philips est une entreprise néerlandaise, Brown Boveri suisse, Westinghouse américaine, Fried Krupp et Bosh allemand. Le Matériel Téléphonique est une filiale américaine de ITT Inc dès 1925. Godi [2023] La Compagnie Française Thomson-Houston ou Compagnie française pour l'exploitation des procédés Thomson-Houston ont été cluster correctement dans une entrée unique de notre banque de donnée. Il s'agit d'une filiale de l'américain General Electrics. Archives

Dans ces quatre pays, une tendance se désigne : les premières entreprises du pays en terme de brevets sont principalement actives dans le domaine de l'électricité ou de la chimie. Nous analyserons des classes liées à ces deux domaines dans la partie 3.3.5.

### 3.3.4 Par Section

Cette partie à pour objectif de démontrer qu'il également possible de détailler l'analyse par section dans le système de classification CPC. Ceci devrait nous permettre de voir l'évolution des différentes classifications et de leurs domaines technologiques principal.

La première chose visible sur la figure 38 est la nombre supérieur de brevets américains. Cependant cette différence est moins visible dans certaines sections que d'autres.

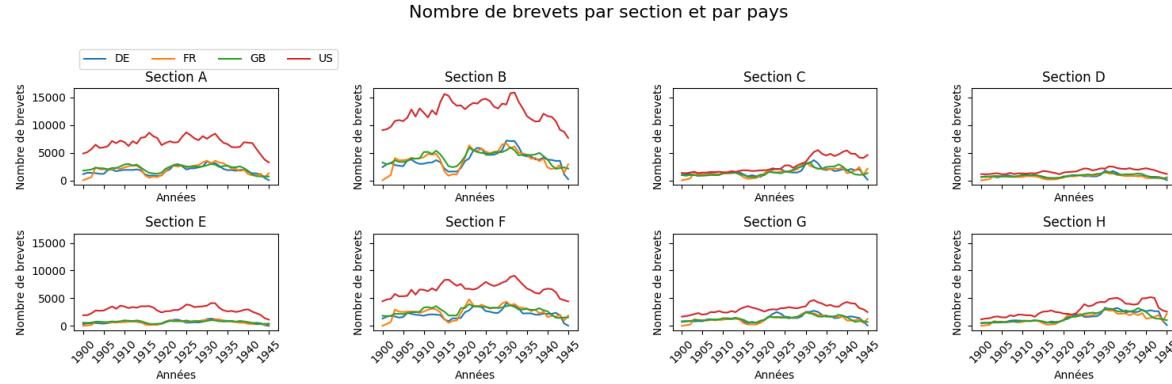


FIGURE 38 – Évolution du nombre de brevets par pays et par section

Dans la section C (Chimie et Métallurgie), nous trouvons un nombre de brevets américains, anglais, français et britanniques similaires, alors que dans la section B (Transport et opérations<sup>12</sup>), une différence notable entre le nombre de brevets américains et du continent européen peut être observé.

Nous sommes également en mesure de mesurer les variations dans la concentration de brevets par secteurs dans la classification CPC.

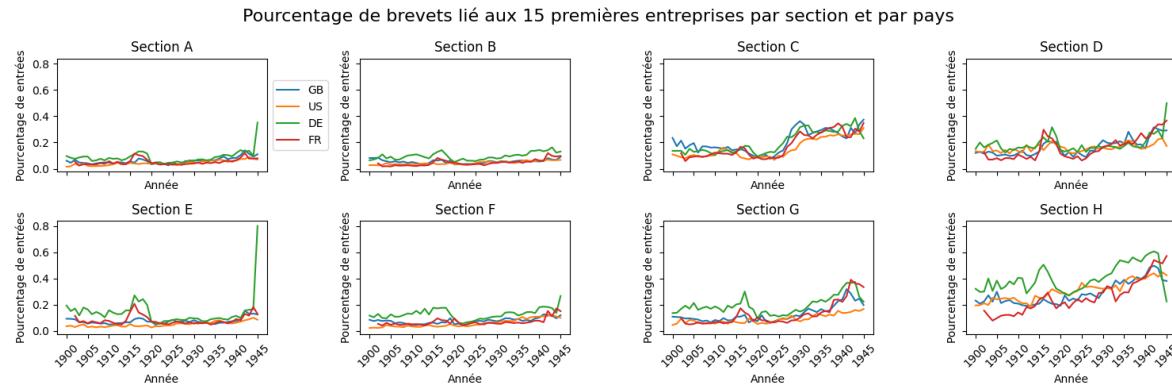


FIGURE 39 – Concentration  $S_{15}$  par Section

Nous observons une augmentation de la concentration au fil du temps, particulièrement dans le secteur C (Chimie et Métallurgie) et D (Textiles et papiers) à partir de années 1925. Enfin, nous sommes en capacité de mesurer la forte concentration des brevets dans le secteur H (correspondant au domaine de l'électricité).

Malgré une augmentation de la concentration dans une majorité des sections, nous pouvons voir que toutes les sections n'évoluent pas de manière identique. Une différence logique car les différentes sections CPC, liées à différents domaines technologiques n'ont pas évolué de la même façon durant la première moitié du XX<sup>e</sup> siècle. Cependant le lien entre classification CPC et domaine technologique reste flou, les sections CPC pouvant comprendre plusieurs domaines très différents brouillant les pistes.

12. Cette section inclus les processus de mélange et de fabrication mais aussi les technologies liées aux méthodes de transport tel que le rail et l'automobile

Nous nous sommes alors demandé si une analyse par classe permettrait de mieux lier une classification CPC à un domaine technologique et si les données permettraient également une analyse de la concentration à niveau.

### 3.3.5 Par Classe

La multiplication des classes rend en effet cette analyse compliquée. Le nombre de brevets n'étant pas uniformément réparti sur l'ensemble des classes, certaines montrent un haut degré de concentration, mais n'est le résultat que d'un faible nombre de brevets assigné à cette classe.

Pour cette raison, nous devons tout d'abord retirer les classes avec moins de 10'000 brevets.

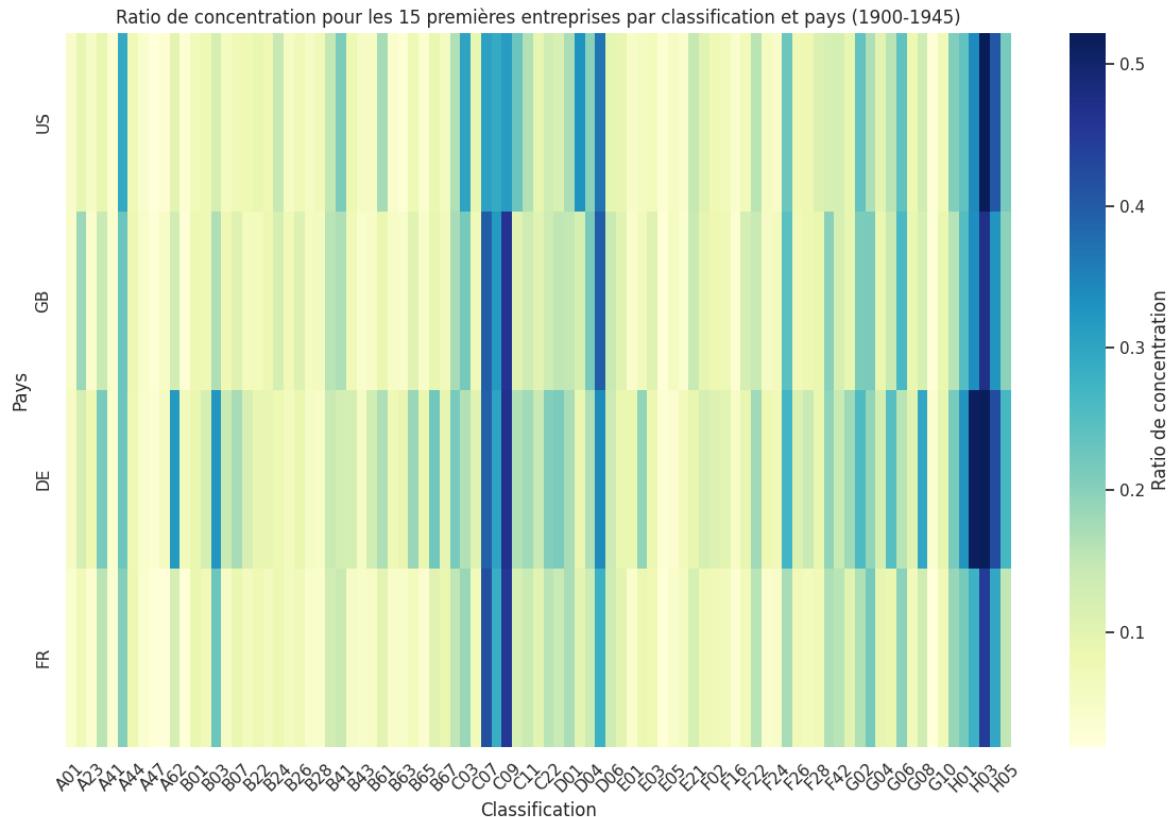


FIGURE 40 – Ratio de concentration entre 1900 et 1945 par classe

La figure 40 présente le ratio de concentration par classe pour les catégories restantes. L'internationalisation de la propriété intellectuelle est visible à cette échelle. Nous observons immédiatement qu'une majorité des classes possédant un plus haut degré de concentration sont consistantes à travers les pays, avec quelques particularités locales.

Nous pouvons développer cette analyse à travers le temps. En sélectionnant un critère, tel qu'un ratio  $S_4 \geq 60\%$  sur une moyenne mobile de 3 ans dans des secteurs très actifs, avec par exemple plus de 100 brevets déposés par ans. Ce critère nous permet de sélectionner un nombre limité de classes potentiellement intéressantes.

Ces classes correspondent à des secteurs et périodes avec une grande concentration industrielle.

Pays	Année(s)	Classe	Définition	Concentration $S_4$
GB	1942	H03	Circuiterie électronique	60%
GB	1929	C09	Teintures, peintures, résines, adhésifs...	61%
DE	1928-1931	C09	Teintures, peintures, résines, adhésifs...	68%
DE	1935-1937	C09	Teintures, peintures, résines, adhésifs...	65%
DE	1932-1942	H02	Production, conversion ou distribution électrique	67%

TABLE 5 – Classes avec de nombreux brevets et un ratio de concentration  $S_4$  élevé

La section H semble posséder au moins deux classes avec un haut degré de concentration. Nous pouvons également constater sur la figure 40, que la section H présente un haut degré de concentration dans tous les pays et dans de multiples classes. Nous pouvons regarder en détail comme les classes de cette section évolue au fil du temps.

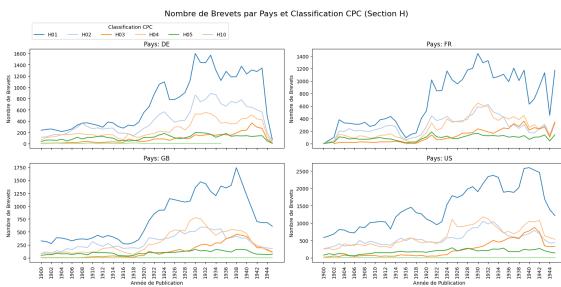


FIGURE 41 – Évolution du nombre de brevets par classes et par pays (Section H)

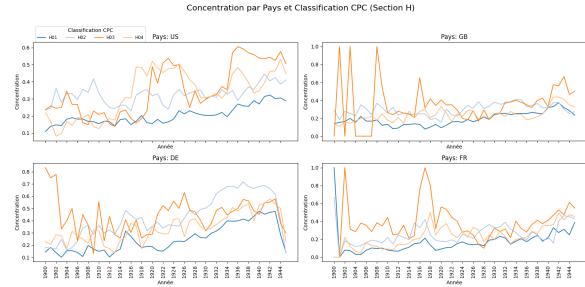


FIGURE 42 – Évolution de la concentration  $S_4$  par classes et par pays

La figure 41 montre que la classe H01, H02 et H04 possède un grand nombre de brevets avec des variations significatives. Nous pouvons l'analyser en regardant la concentration  $S_4$  (sans H05 et H10 qui possèdent un trop faible nombre de brevets) sur la figure 42 et observons que :

- H01 possède un très grand nombre de brevets mais une concentration limitée par rapport aux autres classes. Probablement lié à la nature très généraliste de cette classe définie comme "éléments d'électricité".
- H02 possède plus d'entrées aux États-Unis, alors qu'en France, Royaume-Unis et Allemagne, H04 et H03 montrent un niveau similaire.
- H02, H03 et H04 possèdent un haut ratio de concentration  $S_4$ , notamment dans des périodes où un nombre conséquent de brevets sont publiés.

Nous détaillons ci-dessous une étude de cas sur l'évolution de la concentration de H03.

### Évolution de la classe H03

Cette étude de l'évolution de la classe H03 a pour but de démontrer le type de recherche faisable avec nos données. Nous essayerons également de faire le lien avec des éléments historiques de cette époque ainsi que l'histoire de ces grandes entreprises.

La classe H03, correspondant à la circuiterie électronique, est particulièrement pertinente. Les plus grandes entreprises actives dans cette période étaient des précurseurs dans le domaine de l'électricité.

Nous y retrouvons de nombreux brevets, notamment de technologies liées à la radio et au télégraphe. Cette classification comprenant les inventions d'amplification, de modulation, les oscillateurs.

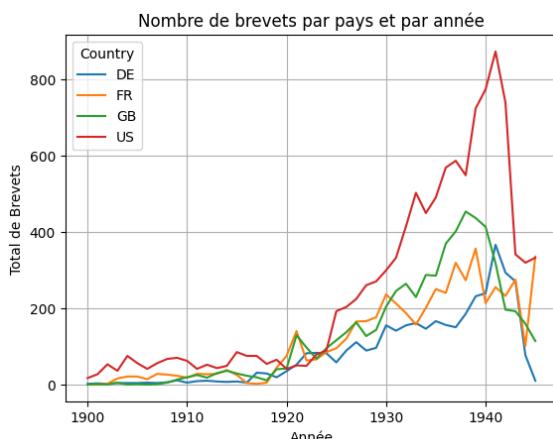


FIGURE 43 – Nombre de brevets dans classe H03

Comme l'illustre la figure 43, le nombre de brevets augmente fortement à partir de 1920. Nous commençons notre analyse à ce moment. Les brevets déposés avant, étant majoritairement liés à des individus.

Ces industries font usage de leurs nombreux brevets pour écraser la compétition. Certaines, tel que Westinghouse et General Electric ont mené une politique de brevet agressive. [Griset, 2004, p. 48]

D'autres, comme l'américain Marconi, obtiennent originellement une grande part du marché en rachetant la compétition tel que United Wireless Telegraph Company. [Griset, 2004, p.49] Cette politique est si agressive qu'aux États-Unis, les géants AT&T, General Electric, ainsi que Westinghouse créeront Radio Corporation of America (RCA) en 1919 à partir de l'américain Marconi, que General Electric rachète<sup>13</sup>[Archer, 1938, p. 173-174]

En 1921, AT&T rentrera dans un contrat de cross-licensing des brevets avec RCA. En 1921, c'est Westinghouse qui obtiendra de nombreux droits de brevets de RCA. Aitken [1985] RCA permettait alors le partage de brevets (patent pool), une stratégie de partage de brevets entre concurrent pour éviter l'arrivée de nouveaux acteurs dans l'industrie. [Christophers, 2016, p. 155]

1930 marque l'intervention étatique américain dans les activités de RCA une forme de monopole et invoque les lois anti-trusts. Times [1930] Dès 1932, General Electric et Westinghouse se séparent de RCA qui devient une entreprise indépendante. [Archer, 1939, p. 379]

Nous constatons sur la figure 44 que RCA tire avantage de cette situation, son nombre de brevets explose à partir de 1932.

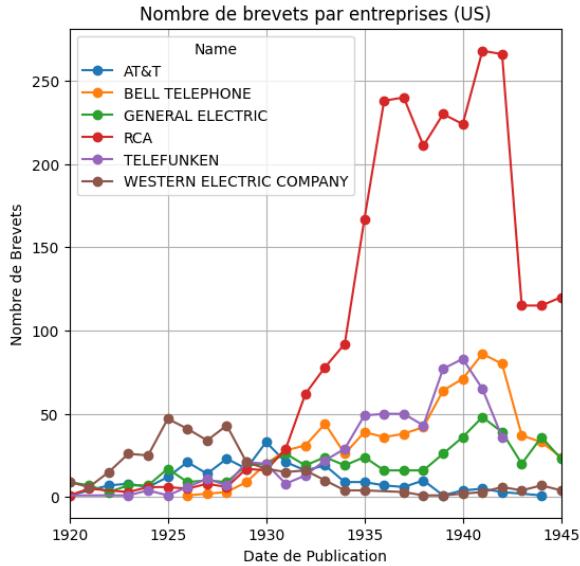


FIGURE 44 – Nombre de brevets aux États-Unis des plus grandes entreprises

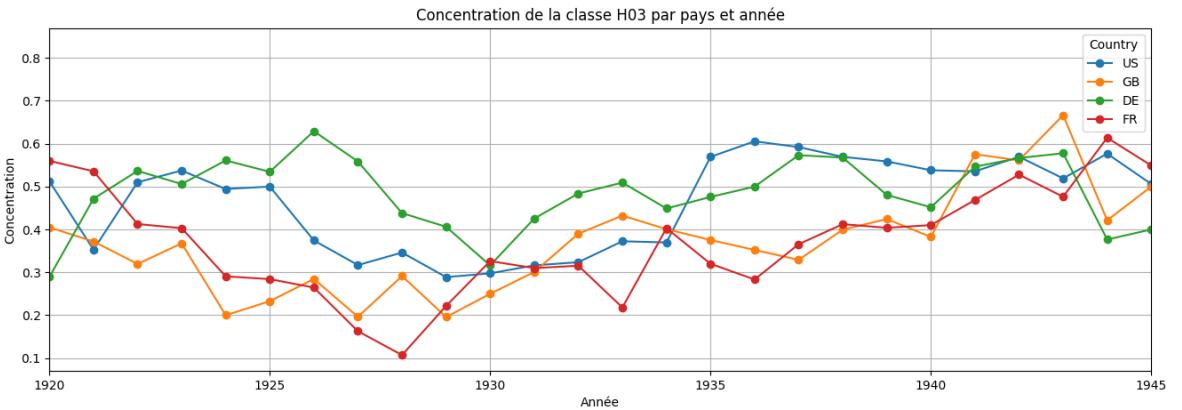


FIGURE 45 – Evolution du ration de concentration  $S_4$  pour H03

En regardant  $S_4$ , c'est à dire le ratio de concentration pour les cinq plus grandes entreprises par année et par pays, nous pouvons voir que :

- La mesure de la concentration dans les années 1920-1925 semble similaire à celle entre les années 1935-1940. Cependant la figure 44 nous permet de comprendre que le paysage industriel est complètement différent entre ces deux périodes. Nous faisons penser que le ratio de

13. Le rachat est poussé par l'armée américaine qui avait pris contrôle de la radio à l'entrée des États-Unis dans la première guerre mondiale en 1918.

la concentration sans détails n'est pas un indicateur suffisant pour juger de la concentration industrielle dans ce secteur.

- Malgré l'intervention des lois américains anti-trusts, c'est après la création de RCA comme une entité indépendante que la concentration aux états-unis augmente et semble uniquement endiguée par Bell Labs.

Nous avons sélectionné le ratio de concentration pour les deux plus grandes entités nommées.

Nous pouvons chercher à mieux comprendre la concentration dans la classe H03 en regardant la proportion de brevets liés à RCA et Western Electric. La figure 46 nous permet de déterminer l'origine des deux pics de concentration observés dans les données. Nous observons par exemple que la part de marché de Western Electric, bien que correspondant à son maximum à peu de brevets (visible sur la figure 44, représente une part de marché conséquente dans les années 1920 à 1925. L'indice Herfindahl-Hirschman atteignant 0.2 dans entre 1936 et 1937. Indicateur d'un manque de compétition sur ce marché. [United States Department of Justice \[2010\]](#).

Nous pouvons tirer des parallèles avec la situation en Allemagne. Comme le montre 45, le pays suit un ratio de concentration similaire haut à celui aux états-Unis.

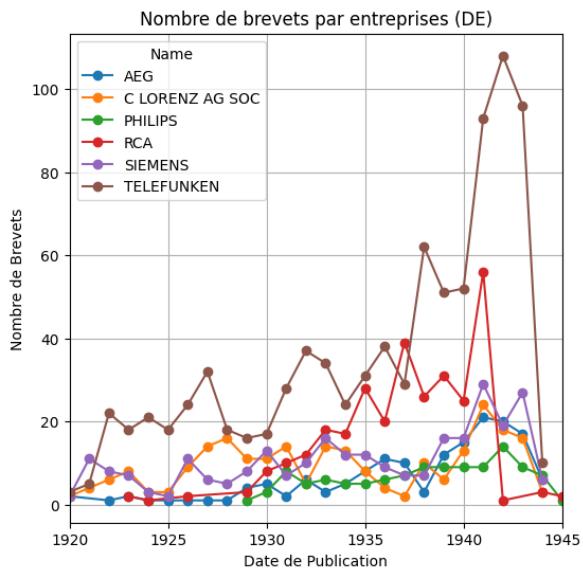


FIGURE 47 – Nombre de brevets en Allemagne des plus grandes entreprises

la société opérait. Les différents gouvernements, voyant une importance dans la création d'un géant national de la télécommunication ont aidé son développement. Pour Telefunken, c'est l'empereur Guillaume II qui finance, et devient actionnaire de l'entreprise qu'il souhaite mettre en compétition contre Marconi, eux financé par le gouvernement britannique. [Griset \[1983\]](#)

Cette compétition qui n'a pas duré, Marconi et Telefunk finissent par mettre en commun leur brevets. [Aitken \[1985\]](#)

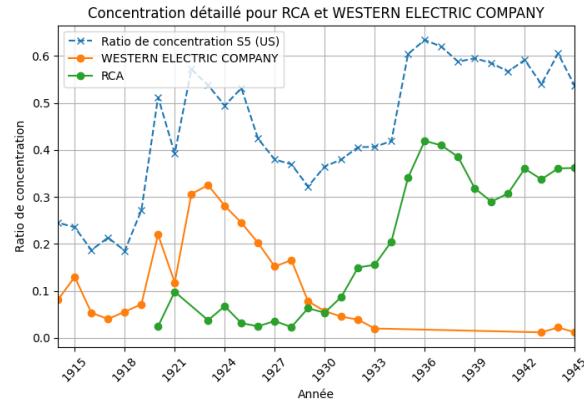


FIGURE 46 – Concentration détaillé pour Western Electric et RCA

Nous trouvons un schéma similaire entre dans le développement des brevets de la classe H02 aux États-Unis (figure 44) et en Allemagne (figure 47) avec un explosion de brevets dans les années 1930.

Cette fois Telefunken (alors encore Telefunken Gesellschaft für drahtlose Telegraphie m.b.H.) possède la majorité des brevets, un nombre qui atteint jusqu'à 100 brevets publiés par an, encore loin de maximum de RCA de plus de 250.

L'Allemagne connaît aussi ses alliances, avec des grands acteurs du milieu tel que "Deutsche Betriebsgesellschaft für drahtlose Telegrafie mbH (Debeg)", fondé par AEG Aktiengesellschaft, Siemens & Halske, Telefunken et l'entreprise belge Compagnie Télégraphie sans fil) en 1911 [Goetzeler \[1975\]](#). Cependant l'entreprise est nommée dans un nombre très restreint de brevets, inférieur à 10.

C'est aussi le cas de Telefunken, fondé en 1903, par la mise en commun des divisions radiographiques de AEG Aktiengesellschaft et Siemens. L'entreprise est d'une importance majeure, notamment dû au secteur critique à l'époque où

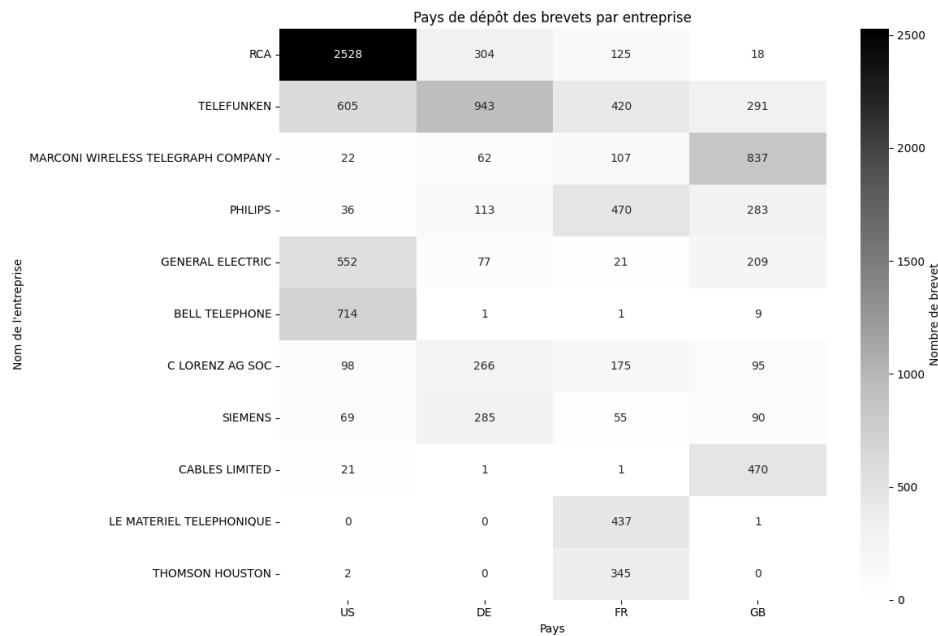


FIGURE 48 – Pays de dépôt de brevets par entreprise pour la classe H02

Nous sommes également en mesure d'observer quelles entreprises opèrent au-delà de leurs frontières.

Les principales entreprises : RCA (US), Telefunken (Allemagne) et Marconi (GB) ont déposé de nombreux brevets dans d'autres jurisdictions. Cette pratique n'est pas partagée par l'ensemble des acteurs du milieu, par exemple Bell Telephone possède très peu de brevets à l'étranger.

### Évolution de la classe C09

Tout comme pour la partie 3.3.5, nous cherchons à faire une étude de cas supplémentaire. Nous effectuons cette étude dans la section C, correspondant à la chimie et la métallurgie.

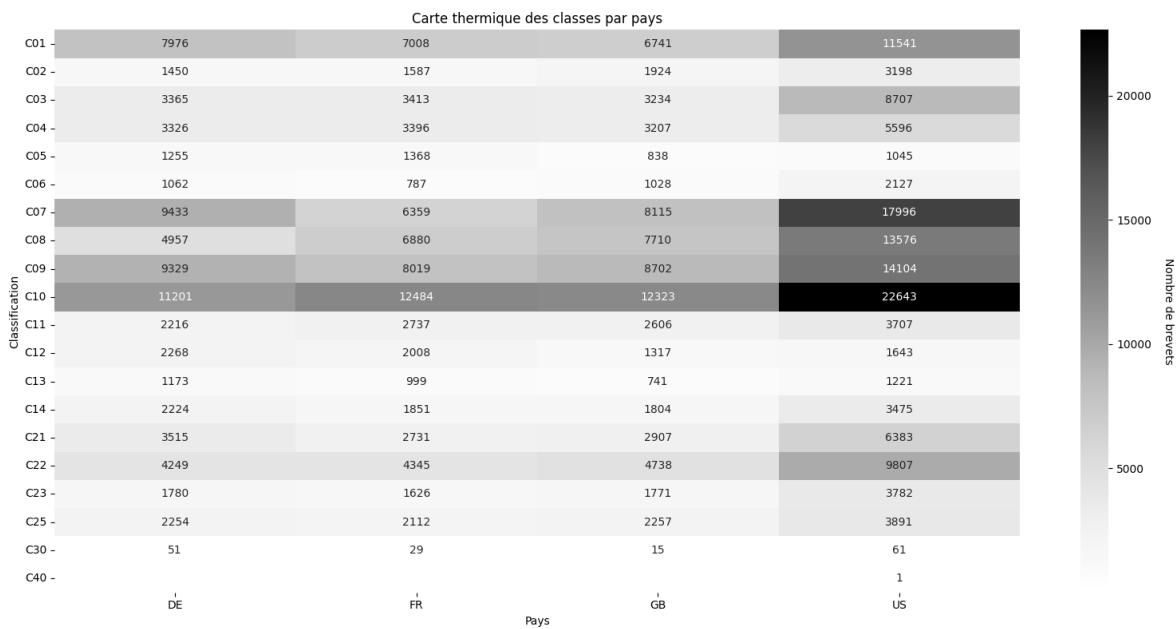


FIGURE 49 – Nombre de brevets par classe dans la section C

Un regard sur le nombre de brevets par classe dans cette section nous permet d'établir que C01, C07, C08, C09, C10 sont les classes les plus fréquentes de cette section. Ces sections correspondent à :

Classe	Nom
C01	Chimie inorganique
C07	Chimie organique
C08	Composés macromoléculaires organiques
C09	Teintures, peintures, résines, adhésifs, etc...
C10	Industrie du pétrole, du gaz, Combustibles

Nous pouvons analyser comment le nombre de brevets dans ces quatre classes évolue dans le temps.

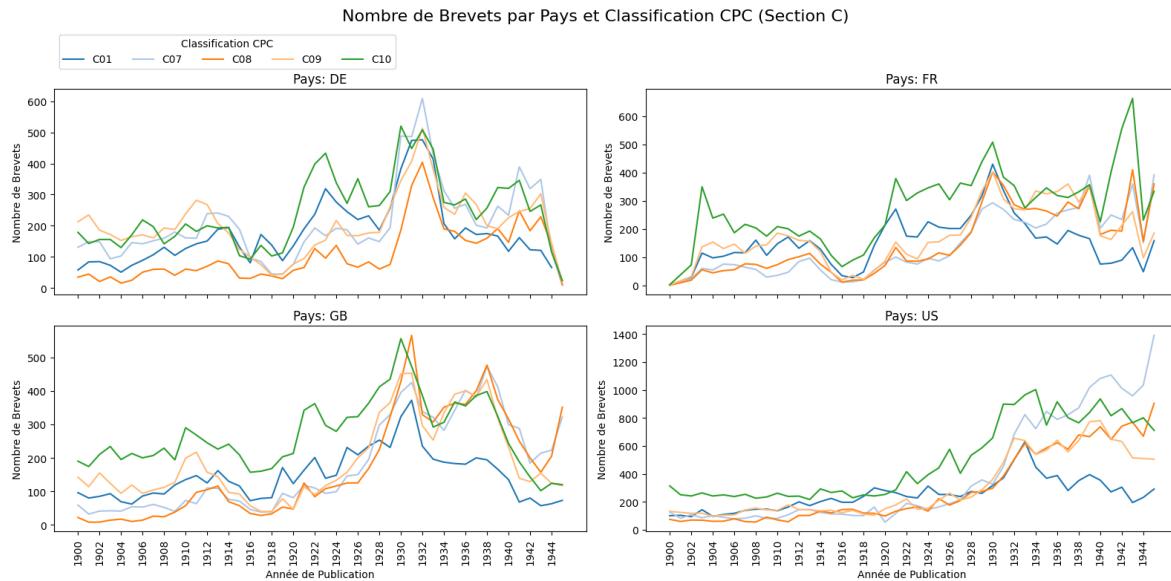


FIGURE 50 – Évolution du nombre de brevets dans les classes C07, C08, C09 et C10

La figure 50 nous permet de constater que le développement de l'industrie chimique américaine se situe particulièrement après les années 1920.

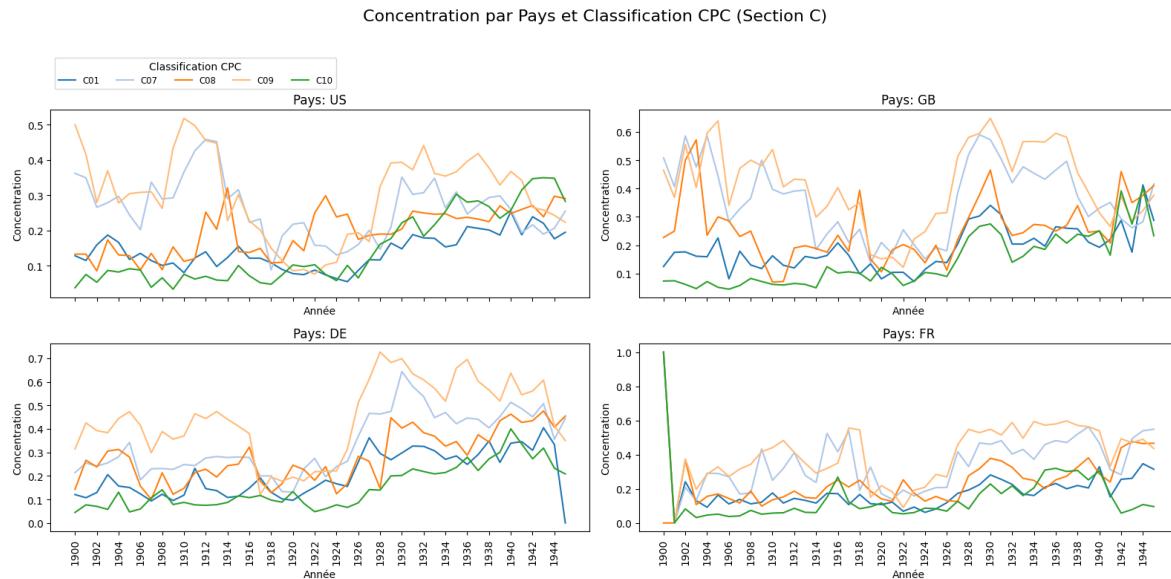


FIGURE 51 – Évolution de la concentration  $S_4$  par classe pour les classes C07, C08, C09 et C10

Alors que l'industrie de la chimie semble atteindre un haut degré de concentration, tel qu'en Allemagne dans les années 1930, nous observons que l'industrie pétrolière est en général moins concentrée

que d'autres industries chimiques. La classe C09 restant généralement la classe ayant le plus haut ratio de concentration  $S_4$ .

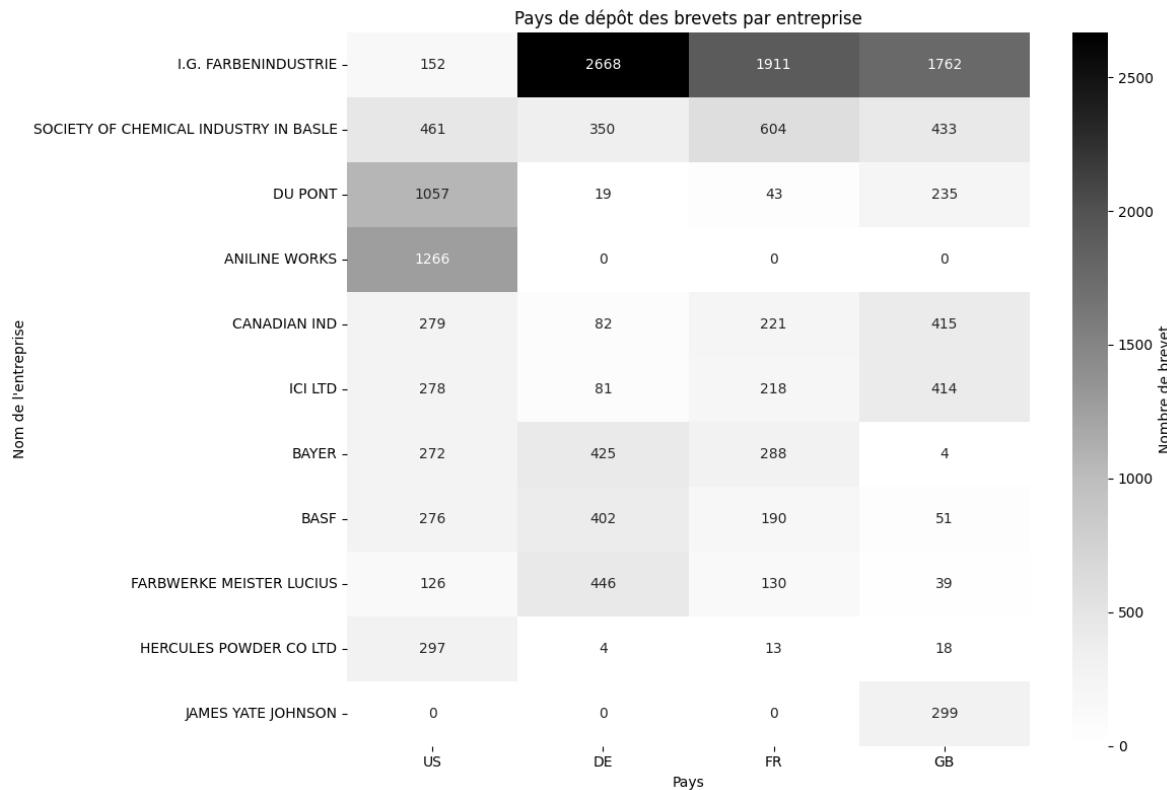


FIGURE 52 – Premières 5 entreprises par pays (Classe C09)

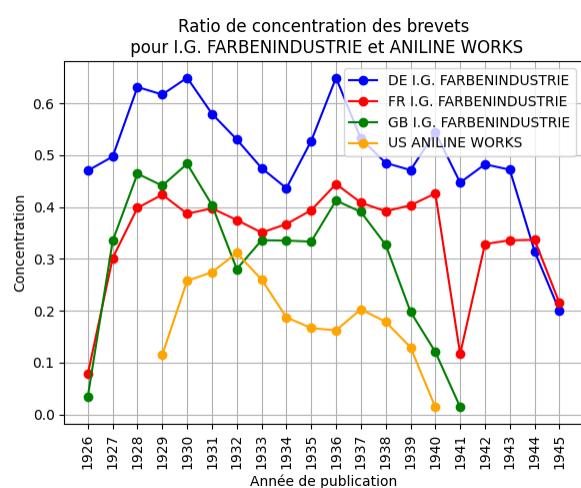


FIGURE 53 – Ratio de concentration de I.G. Farben et GAW

La figure 53 illustre l'importance colossale de I.G. Farbenindustrie dans l'ensemble des pays étudiés <sup>14</sup>. L'Allemagne apparaît alors comme pays très important dans le développement de brevets surtout

Après la Première Guerre mondiale, plusieurs grandes entreprises allemandes fusionnent : Agfa, BASF, Bayer, Griesheim-Elektron, Hoechst, et Weiler-ter-Meer pour former I.G. Farbenindustrie. [Tamm, 1978, p. 195] L'entreprise sera également d'une grande importance de le développement du régime Nazi. López-Muñoz et al. [2009]

Au total, les 5 premières entreprises actives par pays représentent 11 entreprises uniques, démontrant plus d'internationalisation dans ce domaine que dans la classe H09. L'entreprise General Aniline Works (GAW) est possédée en partie au bras américain de l'entreprise allemande I.G. Farbenindustrie : American I.G. Chemical Corporation. L'entreprise est rachetée entièrement par American I.G. Chemical Corporation en 1939 mais est saisie par le gouvernement américain en 1942, en vertu de la loi sur le commerce avec l'ennemi (Trading with the Enemy Act). Encyclopedia.com

14. En comptabilisant General Aniline Works avec I.G. FarbenIndustrie, l'entreprise possède le plus grand nombre de brevets dans la classe C09

lié au teintures, peintures, résines et adhésifs et plus généralement à la chimie, alors que le pays exclu originellement dans la loi allemande de 1877 les substances chimiques de la brevetabilité. [Chachereau, 2022, p. 135]

La Première Guerre mondiale a non seulement entraîné la saisie et la séquestration des actifs de brevets allemands à l'étranger, mais elle a aussi forcé l'industrie chimique allemande à se réorienter massivement vers la production de guerre pour soutenir l'effort national. [Travis et al., 1998, p. 52]

Nous remarquons un grand nombre de brevets de James Yate Johnson déposés en Angleterre sur la figure 52. Cet exemple illustre une limite des données de l'office européen des brevets. En effet, James Yate Johnson est un avocat londonien qui durant sa carrière déposa des centaines de brevets au nom de différentes entreprises. Les métadonnées disponibles pour ces instances auprès de l'office européen de brevets ne mentionnent que James Yate Johnson et non les entreprises derrière ces demandes de brevets.

Alors que I.G. Farbenindustrie fusionne en 1926, le Royaume-uni et la France connaissent le même phénomène de fusion.

- Au Royaume-uni, Brunner Mond, Nobel Industries, United Alkali Company et British Dyestuffs forment Imperial Chemical Industries (ICI LTD dans la figure 52) également en 1926. [Aftalion and Benfey, 1991, p. 140] Le Royaume-uni connaît de nombreux cartels dans cette période. Une étude place un tiers de la production industrielle aux mains des cartels. La pratique n'était d'ailleurs pas régulée, le gouvernement britannique encourageant même la pratique durant la Première et Seconde Guerre Mondiale. [Christophers, 2016, p. 138]
- En France, Établissements Poulenc Frères et Société Chimique des Usines du Rhône forment Rhône-Poulenc en 1928. Celle-ci ne comptabilise que 148 brevets liés à cette entreprise en France, et 286 au total, dans notre banque de données. La Société Industrielle des Matières Plastiques et la Société Générale de Dynamite forment la Société Nobel Française en 1927. Celle-ci n'est liée qu'à 74 brevets durant cette période. [Aftalion and Benfey, 1991, p. 143-144]

Nous pouvons regarder de plus près la concentration dans la classe C09 à l'aide de l'indice de Herfindahl-Hirschman.

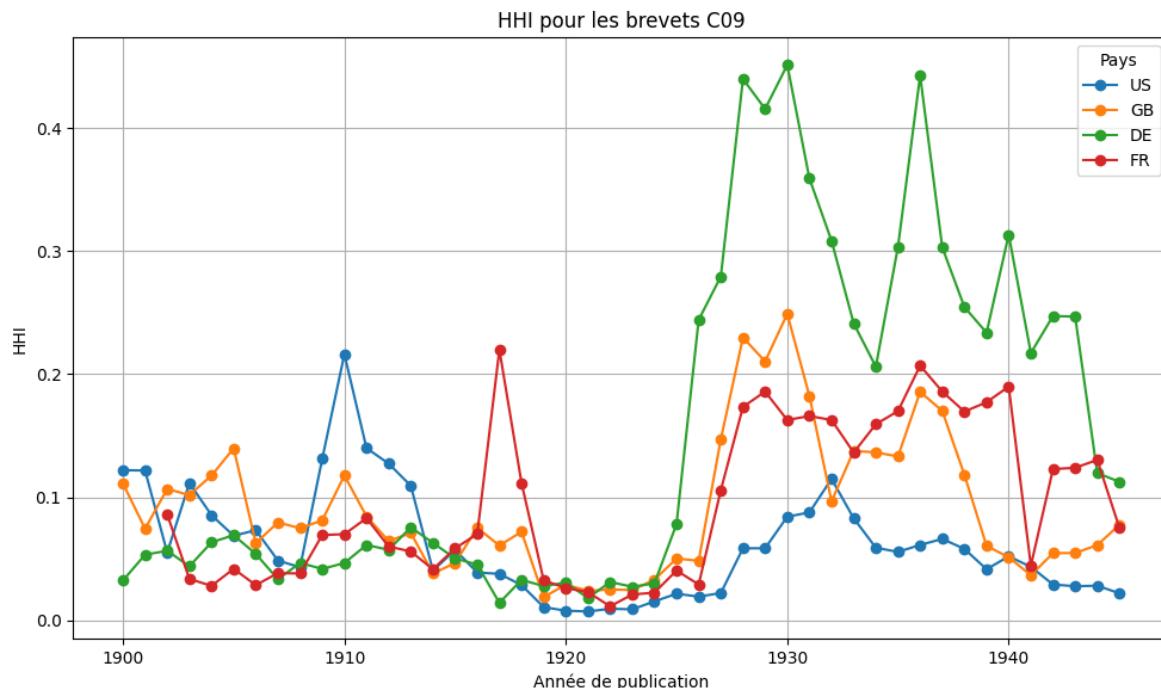


FIGURE 54 – HHI pour la classe C09

Comme apparaît dans la figure 54, plusieurs périodes atteignent un très haut indice. Nous pouvons décortiquer les entreprises possédant une très grande part de marché :

- Entre 1909 et 1911, les États-Unis atteignent un indice de 0.2. Le principal acteur est alors Bayer qui ira jusqu'à posséder 42% des brevets publiés en 1910 aux États-Unis.

- En 1917, la France atteint également un indice de 0.2. A cette période, la société de l'industrie chimique de Bâle dépose 44% de brevets, cependant cette concentration est à nuancer par le faible nombre de brevets, au total 36.
- En 1930, le Royaume-uni voit aussi une concentration supérieure à 0.2. I.G. Farbenindustrie est lié à 48% des brevets publiés dans la classe C06 dans le pays.
- Entre 1926 et 1940, I.G Farbenindustrie est nommé dans 54% des brevets publiés dans la classe C09 en Allemagne.
- Entre 1928 et 1940, c'est 40% des brevets de classe C09 en France qui sont liés à I.G. Farbenindustrie.

Ces deux études de cas démontrent l'usage de nos données, accompagnées par le ratio de concentration et l'indice de Herfindahl-Hirschmann dans la compréhension des dynamique économique de cette période.

### 3.4 Entités nommées

Nous observons dans la partie 3.3 qu'une partie importante des variations sont liés aux chocs économique ou politique. Une autre source de variation peut être les changements dans le fonctionnement des offices de brevets. [Diebolt and Pellier, 2012, p. 618]. Nous ne mesurons pas ces chocs additionnels dans la variation de la concentration industrielle mais nous pouvons interroger nos données pour comprendre quelles mesures peuvent refléter ces changements. Pour ce faire, nous pouvons chercher à mesurer l'évolution dans les pratiques d'attribution de brevets suite à des changement de politique d'entreprise et à des changement légaux.

Dans ce but, nous supprimons tout doublons dans les entités nommés sur un brevet en retirant les entrées similaire, puis retirons le nom de l'entreprise de la liste d'entrée pour limiter notre mesure aux nom d'individus lié au brevet.

#### 3.4.1 Changements législatifs

En 1936, l'Allemagne introduit davantage de droits des inventeurs et retire le concept de "Etablissementserfindung" (Invention d'entreprises). [Gispen \[2007\]](#)

Malgré l'adoption d'une politique plus favorable aux inventeurs, de nombreux brevets restent lié à des entreprises sur le papier, mais le brevet accorde une place plus importante à le reconnaissance de l'inventeur.

Werner Siemens, fondateur de Siemens & Halkse AG, avait comme politique d'assigner le droit exclusif des inventions uniquement à l'entreprise - entre 1877 et 1920 "the company can claim as its exclusive property inventions or improvements of any kind made by the employees.", stipulait leur "Patentrevers" (déclaration de brevet), une règle qui a fini par être abolie par un tribunal allemand. Siemens mise ensuite sur le concept de "Etablissementserfindung" (Invention d'entreprise), pour attribuer les inventions aux entreprises. [Gispen \[2007\]](#) C'est cette forme de brevets qui disparaît en 1936, introduit une compensation statutaire pour les inventions d'employés.

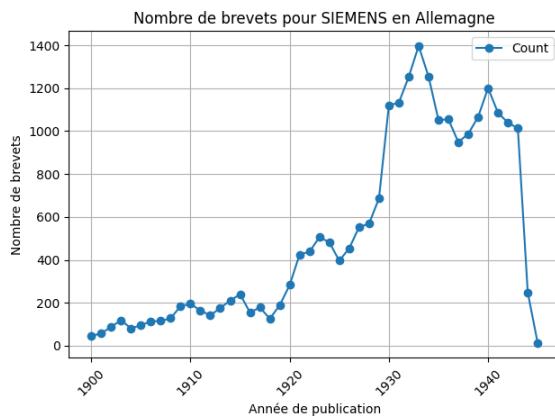


FIGURE 55 – Nombre de brevets lié à Siemens en Allemagne

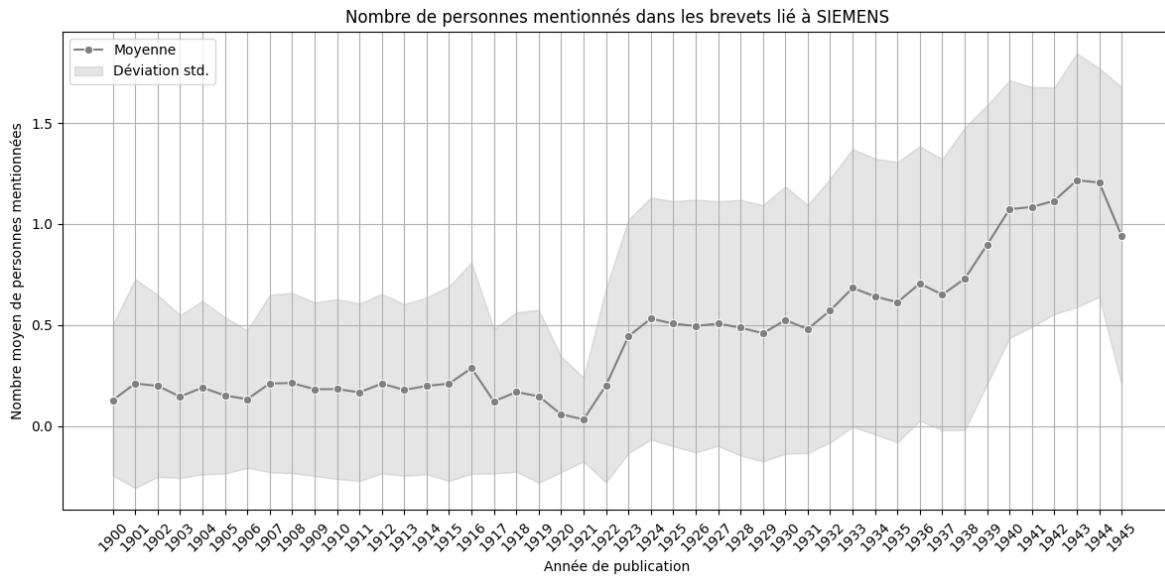


FIGURE 56 – Nombre d’inventeur nommé par brevet pour Siemens

Ce changement légal semble augmenter le nombre de personnes nommées par brevet quatre ans plus tard (attente moyenne entre la date de dépôt et la date de publication en 1936 en Allemagne).<sup>15</sup> Cependant Siemens continuera obtenir un grand nombre de brevets comme l’illustre la figure 55.

La figure montre aussi un premier palier vers 1922, dont l’origine reste à déterminer et constituerait un axe de recherche pertinent.

### 3.4.2 Centre de recherche et développement

Westinghouse et General Electric avaient adopté dans la fin du 19ème siècle des stratégies de rachat de brevets sans développer de centre de compétence en interne (R&D) Lamoreaux [1999]. Avant d’adopter dans la première partie du XX<sup>e</sup> siècle un modèle de recherche en interne. Nous pouvons chercher à comprendre comment ces changements sont quantifiable.

Nous n’avons pas été en mesure de mesurer ce changement de modèle d’innovation dans la concentration des brevets, ce que les données de Watson et Holman semblent également indiquer. Ils mesurent que la grande croissance dans le développement de centre de recherche et développement n’a que peu d’impact sur la concentration des brevets. [Watson and Holman, 1970, p. 115]

Nous allons donc chercher à voir si un impact sur le nombre d’entité nommées par brevet est mesurable.

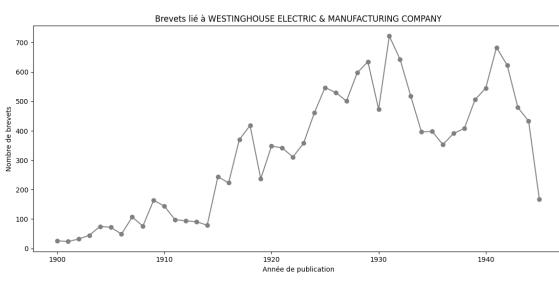


FIGURE 57 – Nombre de brevets de Westinghouse Electric & Manufacturing Company

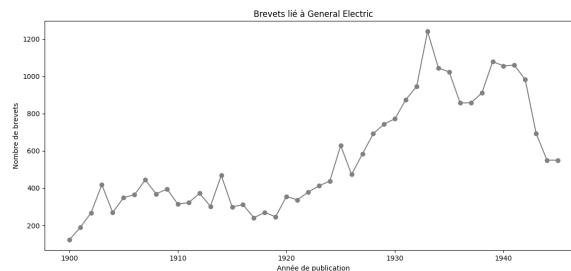


FIGURE 58 – Nombre de brevets de General Electric

En quelques années seulement, Westinghouse est mentionné sur de nombreux brevets, comme nous

15. Une tendance que nous retrouvons chez Fried Krupp alors qu’un inventeur unique est beaucoup plus systématiquement nommé dans les brevets liés à I.G. FarbenIndustrie.

pouvons le voir sur la figure 57. L'entreprise crée son centre de recherche et de développement à partir de 1906. [Brewer \[2012\]](#) A partir de 1900, General Electric commence à exiger la promesse des inventeurs d'assigner leurs inventions à l'entreprise. [Lamoreaux \[1999\]](#) Cette même année, l'entreprise crée un centre de recherche [Sheire \[1975\]](#) Malgré l'augmentation du nombre de brevets attribué à ces entreprises dans les années qui suivent, nous ne pouvons pas nécessairement lier ce résultats à la réussite de l'implémentation de centre de recherche et développement. En effet, nous ne mesurons pas de variations concrètes dans le nombre d'inventeurs liés aux brevets de General Electric et Westinghouse <sup>16</sup> comme il est visible sur la figure 59 et 60. Il est cependant simplement possible que malgré ces changements de structure, la politique d'assignation des brevets n'évolue pas à cette période.

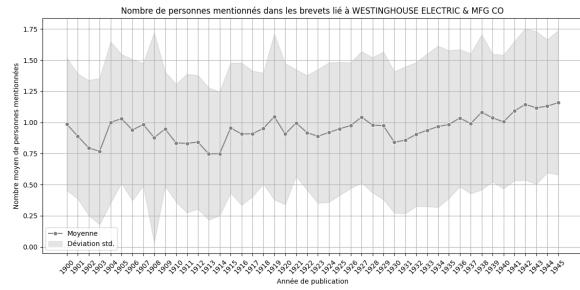


FIGURE 59 – Nombre d'inventeur lié aux brevets de Westinghouse Electric & Manufacturing Company

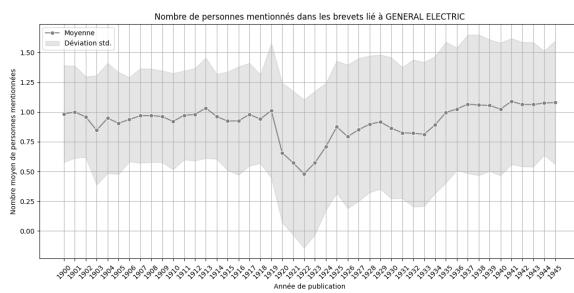


FIGURE 60 – Nombre d'inventeur lié aux brevets de General Electric

A noter qu'une limitation des données est que les brevets acquis après publication ne mentionnent pas les entreprises nouvellement liées à ces brevets, tout comme pour la figure 56. Ainsi, nous ne sommes capables que de comptabiliser les brevets liés avant le dépôt ou développés à l'interne avec accréditation de l'entreprise.

## 4 Conclusion

### 4.1 Apports et objectifs atteints

Cette recherche a rempli l'objectif de développer des outils robustes permettant l'analyse historique de la concentration industrielle à travers des brevets, tel que démontrée par la partie exploratoire de notre rapport.

L'étude de la propriété intellectuelle entre 1900 et 1945 en France, en Grande-Bretagne, en Allemagne et aux États-Unis révèle une augmentation considérable de la concentration dans la propriété intellectuelle au cours de cette période charnière, passant de 5% à 15-20% pour les brevets possédés par les 15 premières entreprises de chaque pays. Une augmentation que nous avons été en mesure d'analyser plus en détails, au travers de certains pays et de certaines disciplines.

Ces études de cas ont également démontré que l'augmentation de la concentration pouvait provenir de marchés divers avec des configurations économiques variées telles que des cartels nationaux ou par exemple des grandes entreprises internationales comme c'est le cas de I.G. Farbenindustrie.

Les données peuvent également être utilisées pour comprendre l'évolution dans l'usage des brevets liés à une entreprise spécifique dans un certain domaine comme le montre également nos deux études de cas.

Ces données permettent également d'observer l'effet des différentes crises traversées sur la concentration de brevets, offrant ainsi un prisme pour l'analyse des transformations économiques de la première moitié du XX<sup>e</sup> siècle.

Les données collectées et analysées peuvent servir aussi bien à une vue macro de l'économie de la propriété intellectuelle qu'à des analyses très ciblées sur des sous-classes d'inventions spécifiques.

16. Si moins d'un brevet est comptabilisé, alors les métadonnées du brevet ne mentionnent que l'entreprise en question.

## 4.2 Innovation méthodologique

Ces données issues de brevets de sources et d'époques variées, nous ont permis d'approcher la construction d'une banque de données avec la volonté de développer une approche robuste que nous pourrons également appliquer à des données plus récentes.

En effet, en réussissant à mettre en place des techniques d'extraction et de *clusterings* sur des données de différents pays, dans différentes langues et sur différentes décennies, tout semble indiquer que celles-ci peuvent être réutilisées avec encore plus de données, afin de par exemple mesurer la concentration industrielle dans l'ensemble du XX<sup>e</sup> siècle.

## 4.3 Implications théoriques

Nos analyses permettent de documenter les dynamiques de concentration empiriquement, ainsi que de mieux comprendre les marchés très concentrés de l'époque comme l'aéronautique, l'ingénierie à vapeur, la chimie, l'électricité, la téléphonie et la radio. [Christophers, 2016, p. 131] Ceci nous permet d'illustrer par des données comment la propriété intellectuelle devient un instrument stratégique dans la construction de cartels, d'oligopole et de monopole durant cette période.

## 4.4 Perspectives critiques

La création et l'utilisation d'une banque de données de brevets pour analyser la concentration industrielle présente plusieurs défis méthodologiques. Les notions d'inventeur et de déposants sont complexes, changent au cours du temps, les rendant difficilement capturables par un système automatisé. Traiter les fichiers d'archives vieux de 100 ans, venant de plusieurs sources, dans plusieurs langues et dont la forme varie durant la période d'étude a été le défi principal de ce projet. Nous avons fait usage de nombreuses techniques pour extraire les informations nécessaires au calcul de la concentration de brevets à partir de données d'archives, mais le choix de ces techniques a d'abord été guidé par l'abondance de données à traiter, dictant une attention particulière à l'optimisation et à limiter les coûts computationnels durant tout ce projet.

Notre méthodologie fait usage des classifications CPC mais celle-ci ne correspondent pas nécessairement à un domaine d'activité (par exemple : le tabac et les chaussures sont dans une même section alors que le textile rentre dans une autre).

Nous devons également porter un regard critique sur les données analysées. Il nous semble judicieux de rappeler que, premièrement, les brevets ne sont pas égaux. Ils dépendent notamment des nouveautés, des conditions conjoncturelles, de la capacité matérielle d'une entreprise à produire et mettre sur le marché et de leur validité aux yeux de la loi. Deuxièmement, notre analyse ne comptabilise pas les transmissions de brevets, ce qui affecte la mesure de la concentration réelle, laissant dans l'ombre les entreprises n'achetant que des brevets déjà déposés. Troisièmement, nous ne mesurons que les brevets publiés, donc pas une image complète de l'activité d'innovation et influencée par les critères de chaque pays sur l'éligibilité des inventions. Quatrièmement, il existe une différence de temps entre le dépôt et la publication de brevets, ce qui peut introduire des décalages temporels dans nos analyses. Cet effet est particulièrement visible en période de guerre quand les offices de brevets sont mis sous tension. L'Allemagne par exemple ne publie que très peu de brevets à la fin de Seconde Guerre mondiale. De plus, analyser l'évolution de la concentration à travers la propriété intellectuelle ne doit pas être sur-interprété. Nous ne sommes qu'en mesure d'essayer de comprendre la concentration dans l'attribution des brevets, celle-ci n'est pas nécessairement en lien avec d'autres formes de capital.

D'autres réflexions critiques sur le sujet méritent d'être mentionnées : Le brevet lui-même, en tant que mesure de l'innovation, est sujet à caution : sa délivrance ne garantit ni la qualité, ni la nouveauté du produit, ni même qu'il soit encore détenu par le déposant. Il est également crucial de reconnaître que le brevet n'est qu'un "des moyens" d'appropriation de l'innovation. Différentes entreprises peuvent adopter différentes stratégies. D'autres mécanismes, tels que le secret ou la publication uniquement partielle d'informations, peuvent être privilégiés par les entreprises. [Chachereau, 2022, p. 266-267] La question de l'utilité des brevets pour l'innovation est d'ailleurs un débat historique que ce projet de recherche n'a pas pour objectif de trancher.

De plus, la durée de validité d'un brevet est une variable importante que nous n'avons malheureusement pas dans nos données.

Enfin une grande limite de rapport concerne le manque d'expertise historique et économique ainsi que de temps, ne permettant pas de faire d'analyse approfondie. Nous nous sommes donc limités à l'exploration des données pour démontrer leur potentiel.

#### 4.5 Perspectives de recherche

Il serait intéressant de faire le lien entre cette mesure et d'autres données statistiques sur les grandes entreprises (profit, nombre d'employés, etc.) afin de mieux cerner les transformations structurelles des industries pendant cette période. Des travaux comparables ont été menés, mais se limitent à certaines périodes historiques aux États-Unis. [Watson and Holman \[1970\]](#) Cette nouvelle dimension permettrait d'enrichir notre compréhension de la concentration économique en général et du rôle spécifique joué par la propriété intellectuelle. L'approche développée dans cette thèse ouvre ainsi la voie à de nouvelles recherches sur les relations entre innovation, propriété intellectuelle et concentration économique, tant dans une perspective historique que contemporaine.

*Le code nécessaire à l'extraction des données, à l'analyse ainsi que la version la plus récente de la banque de données est disponible sur Github : <https://github.com/naelopode/PatentData>*

## Références

- V. Faubert. Quels enseignements tirer de la première mondialisation (1870-1914) ? *Économie & prévision*, 200-201(2) :217–226, 2012. doi : 10.3917/ecop.200.0217. URL <https://doi.org/10.3917/ecop.200.0217>.
- Serge Lapointe. L'histoire des brevets. *Les Cahiers de Propriété Intellectuelle*, 12(3) :1–23, 2000.
- Nicolas Chachereau. *Les débuts du système suisse des brevets d'invention (1873-1914)*. Éditions Alphil-Presses universitaires suisses, 2022.
- Claude Diebolt and Karine Pellier. 400 ans de protection par les brevets : Une contribution de cliométrie comparative 1. *Revue économique*, 633(3) :611–621, 2012.
- Brett Christophers. *The great leveler : Capitalism and competition in the court of law*. Harvard University Press, 2016.
- Naomi R Lamoreaux. *Learning by doing in markets, firms, and countries*. University of Chicago Press, 1999.
- Kees Gispen. Industrialists, inventors and the german patent system, 1877–1957 : A comparative perspective. In *The History of Information Security*, pages 53–81. Elsevier, 2007.
- Horst-Peter Götting, Michal du Vall, and Heike Röder-Hitschke, editors. *Enforcing Intellectual Property Rights*, volume 34 of *Schriften zum geistigen Eigentum und zum Wettbewerbsrecht*. Nomos, Baden-Baden, 1 edition, 2012.
- J. Baudry. Propriété industrielle. In P. Chauvin, M. Grossetti, and P. Zalio, editors, *Dictionnaire sociologique de l'entrepreneuriat*, pages 462–473. Presses de Sciences Po, 2014. doi : 10.3917/scpo.chauv.2015.01.0462. URL <https://doi.org/10.3917/scpo.chauv.2015.01.0462>.
- France. Loi sur les brevets d'invention. WIPO Lex Database, 7 1844. URL [https://wipolex-resources-eu-central-1-358922420655.s3.amazonaws.com/edocs/lexdocs/laws/fr/km/km008fr\\_1.pdf](https://wipolex-resources-eu-central-1-358922420655.s3.amazonaws.com/edocs/lexdocs/laws/fr/km/km008fr_1.pdf). Loi du 5 juillet 1844.
- B. Khan. An economic history of patent institutions. Online encyclopedia, 3 2008. URL <https://eh.net/encyclopedia/an-economic-history-of-patent-institutions/>.
- B. Zorina Khan. *The Democratization of Invention : Patents and Copyrights in American Economic Development, 1790–1920*. Long-Term Factors in Economic Development. Cambridge University Press, Cambridge, 2005. ISBN 0-521-81135-X.
- William F. Shughart. Industrial concentration, 2002. URL <https://www.econlib.org/library/Enc/IndustrialConcentration.html>.
- Donald Stevenson Watson and Mary A Holman. The concentration of patent ownership in corporations. *The Journal of Industrial Economics*, pages 112–117, 1970.
- Clair Wilcox. *Public Policies Toward Business*. R. D. Irwin, Homewood, Illinois, 1966.
- F. Machlup. *The Political Economy of Monopoly*. Johns Hopkins Press, Baltimore, 1952.
- Carl Kaysen and Donald F. Turner. *Antitrust Policy*. Harvard University Press, Cambridge, Mass., 1959.
- Time. Government : Defeat for United Shoe. *Time Magazine*, Mars 1953. URL <https://time.com/archive/6887391/government-defeat-for-united-shoe/>. Archive article.
- Naomi R Lamoreaux and Kenneth L Sokoloff. Inventors, firms, and the market for technology in the late nineteenth and early twentieth centuries. In *Learning by doing in markets, firms, and countries*, pages 19–60. University of Chicago Press, 1999.

Naomi R Lamoreaux, Kenneth L Sokoloff, and Dhanoos Sutthiphisal. The reorganization of inventive activity in the united states during the early twentieth century. Technical report, National Bureau of Economic Research, 2009.

Tom Nicholas. Independent invention during the rise of the corporate economy in britain and japan  
1. *The Economic History Review*, 64(3) :995–1023, 2011.

Michael J Andrews. Historical patent data : A practitioner’s guide. *Journal of Economics & Management Strategy*, 30(2) :368–397, 2021.

Leonid Kogan, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman. Technological innovation, resource allocation, and growth\*. *The Quarterly Journal of Economics*, 132(2) :665–712, 03 2017. ISSN 0033-5533. doi : 10.1093/qje/qjw040. URL <https://doi.org/10.1093/qje/qjw040>.

Enrico Berkes. Comprehensive universe of us patents (cusp) : data and facts. *Unpublished, Ohio State University*, 2018.

Michael J Andrews, Nicolas L Ziebarth, et al. Annual reports of the commissioner of patents data. *UMBC Economics Department Collection*, 2020.

Antonin Bergeaud and Cyril Verluse. A new dataset to study a century of innovation in europe and in the us. *Research Policy*, 53(1) :104903, 2024. ISSN 0048-7333. doi : <https://doi.org/10.1016/j.respol.2023.104903>. URL <https://www.sciencedirect.com/science/article/pii/S0048733323001877>.

Ashish Ghosh, Rajarshi Pal, and Rajendra Prasath, editors. *Mining intelligence and knowledge exploration*. Lecture notes in computer science. Springer International Publishing, Cham, Switzerland, 1 edition, November 2017.

Jieh-Sheng Lee and Jieh Hsiang. Patent classification by fine-tuning bert language model. *World Patent Information*, 61 :101965, 2020. ISSN 0172-2190. doi : <https://doi.org/10.1016/j.wpi.2020.101965>. URL <https://www.sciencedirect.com/science/article/pii/S0172219019300742>.

Jinyan Li, Xue Li, Shuliang Wang, and Quan Z Sheng, editors. *Advanced data mining and applications*. Lecture notes in computer science. Springer International Publishing, Cham, Switzerland, 1 edition, November 2016.

Jie Hu, Shaobo Li, Jianjun Hu, and Guanci Yang. A hierarchical feature extraction model for multi-label mechanical patent classification. *Sustainability*, 10(1), 2018. ISSN 2071-1050. doi : 10.3390/su10010219. URL <https://www.mdpi.com/2071-1050/10/1/219>.

Shaobo Li, Jie Hu, Yuxin Cui, and Jianjun Hu. DeepPatent : patent classification with convolutional neural networks and word embedding. *Scientometrics*, 117(2) :721–744, November 2018.

Hamid Bekamiri, Daniel S Hain, and Roman Jurowetzki. Patentsberta : A deep nlp based hybrid model for patent distance and classification using augmented sbert. *arXiv preprint arXiv:2103.11933*, 2021.

Mindee. doctr : Document text recognition. <https://github.com/mindee/doctr>, 2021.

Michal Kohút, Jakub et Hradíš. Ts-net : Ocr trained to switch between text transcription styles. In *International Conference on Document Analysis and Recognition (ICDAR)*. Springer, 2021.

SoftmaxAI. Jaccard Similarity; Text Similarity Metric in NLP; Study Machine Learning — studymachinelearning.com. <https://studymachinelearning.com/jaccard-similarity-text-similarity-metric-in-nlp/>, 2021. [Accessed 16-05-2025].

Ian Howson. Cosine\_similarity : Function for calculating the cosine similarity in GSgalgoR : An Evolutionary Framework for the Identification and Study of Prognostic Gene Expression Signatures in Cancer — rdrr.io. [https://rdrr.io/bioc/GSgalgoR/man/cosine\\_similarity.html](https://rdrr.io/bioc/GSgalgoR/man/cosine_similarity.html), 2020. [Accessed 16-05-2025].

Union internationale pour la protection de la propriété industrielle UIPPI. La propriété industrielle. *La propriété industrielle*, 1885-1935.

UK Go. Facts and figures : patents, trade marks, designs and hearings : 2022 — gov.uk. <https://www.gov.uk/government/statistics/facts-and-figures-patents-trade-marks-designs-and-hearings-2022/facts-and-figures-patents-trade-marks-designs-and-hearings-2022>, 2022. [Accessed 25-05-2025].

Guan-Cheng Li, Ronald Lai, Alexander D'Amour, David M Doolin, Ye Sun, Vetle I Torvik, Amy Z Yu, and Lee Fleming. Disambiguation and co-authorship networks of the us patent inventor database (1975–2010). *Research Policy*, 43(6) :941–955, 2014.

Profound Networks. Github - company designators : a dataset of international corporate designators, 2022. URL [https://github.com/ProfoundNetworks/company\\_designator](https://github.com/ProfoundNetworks/company_designator).

Eric Zhu. Github - datasketch : Big data looks small, 2024. URL <https://github.com/ekzhu/datasketch>.

Max Bachmann. rapidfuzz/rapidfuzz : Release 3.8.1, April 2024. URL <https://doi.org/10.5281/zenodo.10938887>.

Georg Siemens. *History of the House of Siemens*, volume I. Karl Alber, Freiburg/Munich, 1957. Traduction de A.F. Rodger.

Stephen D Billington and Alan J Hanna. That's classified ! inventing a new patent taxonomy. *Industrial and Corporate Change*, 30(3) :678–705, 2021.

Michael Parkin and Robin Bade. *Economics : Canada in the Global Environment – Solutions to Problems*. Pearson Education Canada, 6 edition, 2006. ISBN 978-0321312686. via St. Francis Xavier University.

Stephen A. Rhoades. The herfindahl-hirschman index. *Federal Reserve Bulletin*, 79(March) :188–189, March 1993.

Daniel A Crane. Fascism and monopoly. *Mich. L. Rev.*, 118 :1337–1370, 2019.

Rudolf Hilferding. *Finance Capital : A Study of the Latest Phase of Capitalist Development*. Routledge & Kegan Paul, London, 1981. Originally published as *Das Finanzkapital. Eine Studie über die jüngste Entwicklung des Kapitalismus*, Vienna : Wiener Volksbuchhandlung, 1910 (Marx-Studien, vol. III).

Dominique Barjot. Les cartels, une voie vers l'intégration européenne ? le rôle de louis loucheur (1872–1931). *Revue économique*, 64(6) :1043–1066, 2013. doi : 10.3917/reco.646.1043. URL <https://doi.org/10.3917/reco.646.1043>.

Science Museum Group. British thomson-houston co ltd. <https://collection.science museumgroup.org.uk/people/ap27016/british-thomson-houston-co-ltd>, 2024. Science Museum Group Collection.

Scott Lewis. Stc plc. <https://www.encyclopedia.com/books/politics-and-business-magazines/stc-plc>, 2019. Encyclopedia.com.

Jean Godi. Histoire de téléphone, 2023.

France Archives. Compagnie française thomson-houston (et autres compagnies réunies en 1968 dans thomson csf). <https://francearchives.gouv.fr/fr/findingaid/02ac762b4fa5efdc35108c90d2d1e2d2af0fe0e>.

Pascal Griset. Pionniers et grandes entreprises aux premiers temps de l'industrie radioélectrique américaine : Les réseaux hiérarchisés de l'innovation. *Ann. hist. électricité*, 2(1) :39, 2004.

Gleason L. Archer. *History of Radio to 1926*. Arno Press / New York Times, 1938.

Hugh G. J. Aitken. *The Continuous Wave : Technology and American Radio, 1900-1932*. Princeton University Press, 1985. URL <http://www.jstor.org/stable/j.ctt7zthqr>.

- Gettysburg Times. Government starts anti-trust suits. *Gettysburg Times*, page 2, May 1930.
- Gleason L. Archer. *Big Business and Radio*. 1939.
- United States Department of Justice. Horizontal merger guidelines. <https://www.justice.gov>, August 2010. URL <https://www.justice.gov/atr/file/810276/dl?inline>. Retrieved June 10, 2025.
- Herbert Goetzeler. Ferdinand braun und die drahtlose telegraphie an der nordsee. *Deutsches Schiffahrt-sarchiv*, 1 :8, 1975. URL [https://web.archive.org/web/20210709184832/https://www.ssoar.info/ssoar/bitstream/handle/document/49592/ssoar-dsa-1975-goetzeler-Ferdinand\\_Braun\\_und\\_die\\_drahtlose.pdf?sequence=1](https://web.archive.org/web/20210709184832/https://www.ssoar.info/ssoar/bitstream/handle/document/49592/ssoar-dsa-1975-goetzeler-Ferdinand_Braun_und_die_drahtlose.pdf?sequence=1). Retrieved June 01, 2025.
- Pascal Griset. La société radio-france dans l'entre-deux-guerres. *Histoire, économie et société*, 2(1) : 85, 1983. ISSN 0752-5702. doi : 10.3406/hes.1983.1317. URL [https://www.persee.fr/doc/hes\\_0752-5702\\_1983\\_num\\_2\\_1\\_1317](https://www.persee.fr/doc/hes_0752-5702_1983_num_2_1_1317). Le changement technique contemporain : approches historiques.
- Helmut Tammen. *Die I.G. Farbenindustrie Aktiengesellschaft (1925–1933) : Ein Chemiekonzern in der Weimarer Republik*. H. Tammen, Berlin, 1978. ISBN 978-3-88344-001-9.
- F López-Muñoz et al. The pharmaceutical industry and the german national socialist regime : I.g. farben and pharmacological research. *Journal of clinical pharmacy and therapeutics*, 34(1) :67–77, 2009. doi : 10.1111/j.1365-2710.2008.00972.x.
- Encyclopedia.com. "g.a.f. ." international directory of company histories. <https://www.encyclopedia.com/books/politics-and-business-magazines/gaf>. [Accessed 17-06-2025].
- Anthony S Travis, Harm G Schröter, Ernst Homburg, and Peter JT Morris. *Determinants in the Evolution of the European Chemical Industry, 1900–1939 : New Technologies, Political Frameworks, Markets and Companies*, volume 16. Springer Science & Business Media, 1998.
- Fred Aftalion and Otto Theodor Benfey. *A History of the International Chemical Industry*. University of Pennsylvania Press, Philadelphia, PA, 1991. ISBN 13978-0-585-19669-5.
- Nathan Brewer. Your engineering heritage : Westinghouse electric corporation. <https://insight.ieeeusa.org/articles/your-engineering-heritage-westinghouse-electric-corporation/>, 2012. [Accessed 05-06-2025].
- James Sheire. General electric research laboratory, March 1975. URL [https://npgallery.nps.gov/NRHP/GetAsset/NHLS/75001227\\_text](https://npgallery.nps.gov/NRHP/GetAsset/NHLS/75001227_text). Form No. 10-300.