N° 1 | 2020
Faire dialoguer les disciplines via l'indexation des connaissances : la recherche interdisciplinaire en débat

La CIB comme pivot de classement interdisciplinaire. La classification internationale des brevets au profit de l'interdisciplinarité

David REYMOND *Maître de conférences HDR*
'' Université de Toulon
Clara GALLIANO *ATER*
UFR Ingémédia
IMSIC
Université de Toulon, La Garde
Luc QUONIAM *Professeur des universités*
IMSIC EA792
Université de Toulon, La Garde

Édition électronique :
URL :
https://intelligibilite-numerique.numerev.com/index.php/articles/revue-1/2166-la-cib-comme-pivot-de-classement-interdisciplinaire-la-classification-internationale-des-brevets-au-profit-de-l-interdisciplinarite
DOI : numerev_1695
Date de publication : 13/05/2020

Cette publication est sous licence **CC BY-NC-ND** (Attribution - No commercial - No derivatives).

Pour **citer cette publication** : REYMOND, D., GALLIANO, C., QUONIAM, L. (2020) La CIB comme pivot de classement interdisciplinaire. La classification internationale des brevets au profit de l'interdisciplinarité. *Intelligibilité du numérique*, (1). https://doi.org/10.34745/numerev_1695

Si l’on parle aujourd’hui d’indexation pour désigner l’enregistrement d’un site internet dans la base de données d’un moteur de recherche, la notion d’indexation des connaissances remonte quant à elle à l’Encyclopédie de Diderot et d’Alembert ainsi que les projets qui s’en sont suivis (projet d’Otlet, la classification de Dewey…).
L'indexation des connaissances tire son origine du champ disciplinaire de la documentation, elle reste néanmoins au cours des questionnements et des débats liés à l’interdisciplinarité. L’indexation des connaissances peut-elle s’exercer au profit de l’interdisciplinarité ? Nous tentons d’y répondre dans cet article. Nous proposons une expérimentation visant à utiliser la classification internationale des brevets (CIB) comme pivot de classement interdisciplinaire. Le terrain étudié concerne la « science en train de se faire », ici les thèses de doctorat, mais peut être calqué à d’autres sources de données. Nous détaillons les résultats par un exemple lié directement aux sciences dures, mais qui peut être également appliqué aux sciences humaines et sociales. Cette étude originale est au croisement du monde académique et du monde technologique.

If we speak today of indexing to designate the registration of a website in the database of a search engine, the notion of indexing knowledge goes back to the Encyclopedia of Diderot and Alembert or the projects that followed (Otlet project, Dewey classification …).

The indexation of knowledge originates from the disciplinary field of the documentation, it remains nevertheless during the questions and the debates related to interdisciplinary. Can the indexation of knowledge be applied in favor of interdisciplinarity? We try to answer it in this article. We propose an experiment to use the International Patent Classification (IPC) as an interdisciplinary pivot. The field studied concerns "science in the making", here doctoral theses, but can be modeled on other sources of data. We detail the results by an example directly related to the hard sciences, but which can also be applied to the human and social sciences. This original study is at the crossroads of the academic world and the technological world.

Mots-clés :
Interdisciplinarité, CIB, Brevet, Indexation, Classification, Visualisation, Science, Thèses, Interdisciplinarity, Indexation, Classification, Patent, IPC, Thesis, Visualization, Science

---

1. Introduction

Si l'on reprend la terminologie proposée par Basarab Nicolescu, « la pluridisciplinarité concerne l'objet d'une seule et même discipline par plusieurs disciplines à la fois » (Nicolescu, 1996, p.64). L'interdisciplinarité permet de faire franchir un cran supplémentaire dans l'entrelacement des disciplines, puisqu'elle « concerne le transfert des méthodes d'une discipline à une autre [...] mais sa finalité reste aussi inscrite dans la recherche disciplinaire » (Nicolescu, 1996, p. 65-66), avec des degrés de bouleversements induits variables, du simple transfert-applicatif à l'engendrement de nouvelles disciplines, en passant par les modifications de nature épistémologique. Enfin, stade ultime, la transdisciplinarité, qui transcende les disciplines du fait de sa finalité, à savoir la compréhension du monde, « dont un des impératifs est l'unité de la connaissance ».

Cette classification a été appréhendée de façon légèrement différente par J-L Le Moigne (Kourilsky, 2002, p. 26-27) qui distingue l'interdisciplinarité de type pluri, soit l'emprunt par une discipline de méthodes validées issues d'autres disciplines de l'interdisciplinarité de type trans qui « en affichant son postulat de « dépendance au contexte » assume le primat de l'intelligibilité systémique sur la prévisibilité analytique ». Quel que soit le degré d'interdisciplinarité in fine, nous partons du postulat de l'existence d'objets de recherche qui ont une ascendance issue de différentes disciplines et, en ce sens, ces objets pourraient constituer un terrain propice aux croisements disciplinaires, au sein duquel il sera nécessaire de favoriser le dialogue. Afin d'identifier ces objets pluri/inter disciplinaires, nous proposons d'interroger un système d'indexation de connaissances (la base des thèses françaises) croisé avec un système d'indexation des objets ou procédés artificiels (la classification internationale des brevets). Dans ce qui suit nous contextualisons cette expérimentation, présentons la méthodologie puis les résultats obtenus, sous forme de visualisation, par le traitement d'un cas particulier.

2. Contextualisation et problématisation

Les nombreux débats sur le sens ou encore l'utilité de la recherche interdisciplinaire se confrontent souvent dans la façon de définir et de distinguer les termes tels que la « pluridisciplinarité », la « multidisciplinarité », la « transdisciplinarité » de Le Moigne (Kourilsky, 2002), ou même la « métadisciplinarité » (Prud'homme & Gingras, 2015). Le terme « polydisciplinaire » sera également rajouté par Morin pour désigner « l'association de disciplines en vertu d'un projet ou d'un objet qui leur est commun ; tantôt les disciplines y sont appelées comme techniciennes spécialistes pour résoudre

tel ou tel problème tantôt au contraire elles sont en profonde interaction pour essayer de concevoir cet objet et ce projet, comme dans l'exemple de l'hominisation » (Morin, 1994).
L'interdisciplinarité, quant à elle, est vue par les politiques publiques et les dynamiques scientifiques comme un instrument de lutte contre l'académisme de la recherche publique (Weingart et Stehr, 2000) et de promotion de ses retombées socio-économiques (Louvel, 2015). L'interdisciplinarité consiste en « la mise en relation d'au moins deux disciplines en vue d'élaborer une représentation originale d'une notion, d'une situation, d'une problématique » (Maingain, Dufour & Fourez, 2002). Selon Lenoir (2003), l'interdisciplinarité doit avoir un ancrage dans le réel. Pour Morin, elle signifie plusieurs choses : « que différentes disciplines se mettent à une même table, à une même assemblée, comme les différentes nations se rassemblent à l'ONU sans pouvoir faire autre chose que d'affirmer chacune ses propres droits nationaux et ses propres souverainetés par rapport aux empiétements du voisin. Mais l'interdisciplinarité peut vouloir dire aussi échange et coopération, ce qui fait que l'interdisciplinarité peut devenir quelque chose d'organique ».

Il a souvent été question de fractionnement des disciplines, de "division arbitraire du savoir" (Reisse, 1999), de "découpage des connaissances" ou encore "d'unité de savoir" dans certains domaines scientifiques. La spécialisation de certaines disciplines a créé de ce fait une caricature (utilisation d'un jargon scientifique propre par exemple) rendant les champs étroits et les scientifiques enfermés dans leur discipline (Guthleben, 2014). L'approche de l'interdisciplinarité apparaît alors comme une évidence, selon une logique de compétences. Les disciplines communiquent entre elles et sont parfois complémentaires les unes aux autres, mais pourtant : supprimer les frontières disciplinaires dans la production des connaissances parait irréel à l'heure actuelle. C'est pourquoi nous préférons penser que les procédés selon lesquels l'interdisciplinarité transforme et réorganise les disciplines, en tant qu'unités essentielles d'organisation et de différenciation des sciences contemporaines (Dubois, 2014) sont susceptibles d'exister.

Le travail sur la question de l'indexation et la classification ne date pas du XXème siècle bien au contraire, elle a été le sujet principal des travaux de Leibniz au travers des méthodes d'organisation des bibliothèques (Le Deuff, 2015). Nous pouvons également évoquer les projets de Otlet et Lafontaine, l'encyclopédie de Diderot et d'Alembert, la classification de Melvil Dewey (1851-1931) pour illustrer la volonté des scientifiques à vouloir rassembler, archiver et diffuser les savoirs en un seul lieu, en un seul point, une référence qui s'adapterait à tous les points de vue (disciplinaires). Cette notion même de référentiel est, de longue date, accomplie au niveau de la classification internationale des brevets (CIB) dont l'objet est de fournir un système de classification dynamique (s'adapte aux évolutions), univoque (chaque terme utilisé dans la description est défini et contextualisé –lingua franca) et international (indépendant de la langue) de tout ce qui est brevetable (objet ou procédé). Notre objectif est de mettre en lumière des objets « interdisciplinaires » par essence en mettant en exergue dans un système documentaire académique, ceux qui ont été abordés par différentes

disciplines. Le sujet/contenu de chaque document extrait à un sujet donné donnera les points de vue disciplinaires connexes. Avant de discuter de la méthodologie élaborée, nous présentons les référentiels utilisés pour construire ce pivot documentaire artificiel répondant à cette interrogation.

3. Les outils et ressources convoquées

3.1. La classification internationale des brevets

L'Organisation Mondiale de la Propriété Intellectuelle (OMPI, traduit de l'anglais WIPO « World Intellectual Property Organisation »)[1] a été créée le 14 juillet 1967 à la suite de la signature de la convention de Stockholm par ses États membres. A l'origine, l'organisation ne comptait que 51 États, qui sont aujourd'hui au nombre de 191 (dont la France depuis 1974).

C'est la source de données la plus complète au monde sur le système de propriété intellectuelle, ainsi que d'études empiriques, de rapports et d'informations factuelles sur la propriété intellectuelle. Sa mission principale est de promouvoir l'innovation et la créativité aux fins du développement économique, social et culturel de tous les pays au moyen d'un système international de propriété intellectuelle équilibré et efficace. Ses services encouragent les personnes et les entreprises à innover et à créer.

Aujourd'hui, plus de 110 millions de demandes de brevets ont été déposées dans le monde. Cette masse informationnelle est accessible grâce à l'Organisme Mondial de la Propriété Intellectuelle et l'interface d'accès à sa base mondiale EspaceNet[2].

Afin de classer et hiérarchiser les brevets, l'OMPI a mis en place l'IPC (de l'anglais « International Patent Classification »). La classification internationale des brevets (CIB)[3] a été créée par l'Arrangement de Strasbourg en 1971. C'est un système hiérarchique de symboles indépendants de la langue pour le classement des brevets et des modèles d'utilité selon les différents domaines technologiques auxquels ils appartiennent.

La CIB propose huit sections principales (cf. le Tableau 1 à gauche) avec plus de 70 000 subdivisions (cf. le Tableau 2 qui présente section par section les nombre de subdivisions qui les composent et permet d'apprécier la fine granularité de ce schéma de classement). Chaque subdivision comprend un symbole, composé de chiffres arabes et de lettres de l'alphabet latin, qui constitue un code hiérarchique (visible sur la colonne 3 du tableau suivant). Nous retrouvons ensuite ces symboles sur chaque document de brevet, attribués ou vérifiés par l'expertise de l'office national ou régional de propriété intellectuelle qui publie ce document.

La CIB répartit la technologie en huit sections principales :

- A Nécessités courantes de la vie
- B Techniques industrielles ; transports
- C Chimie ; métallurgie
- D Textiles ; papier
- E Constructions fixes
- F Mécanique ; éclairage ; chauffage ; armement ; sautage
- G Physique
- H Électricité

Section	A	Nécessités courantes de la vie
Classe	A21	Cuisson au four ; matériel pour la préparation ou le traitement de la pâte ; pâte à cuire
Sous-classe	A21C	Machines ou matériel pour la préparation ou le traitement de la pâte
Groupe	A21C1	Machines à mélanger ou à pétrir pour la préparation de la pâte
Sous-groupe	A21C1/06	à outils mélangeurs ou pétrisseurs montés horizontalement

Tableau 1 : Système de hiérarchisation proposée par la CIB (les sections et les modalités de leur subdivisions hiérarchique en classe, sous-classe… et l'exemple de leur description pour la classe A jusqu'au sous-groupe A21C1/96. Source OMPI

Cette classification[4] est très utile pour la recherche de documents brevet et pour effectuer une recherche sur l'état de la technique. Elle constitue un point d'entrée principal car il s'agit d'un champ d'indexation obligatoire et incontournable (Czajkowski, 2011; Durand-Barthez, 2013). Elle est très utilisée par les administrations chargées de la délivrance des brevets, les inventeurs potentiels, les unités de recherche-développement ainsi que tous ceux intéressés par les applications ou le développement de la technologie. La CIB est mise à jour régulièrement (au 1er janvier de chaque année). La version actuelle et les versions précédentes sont disponibles en anglais et en français. Cette révision est coordonnée par le Comité d'experts de l'Union de l'IPC.

Section	Nbre de classes	Nbre de sous-classes	Nbre de groupes principaux	Nbre de sous-groupes	Nbre total de groupes
A	16	84	1136	8048	9184
B	38	169	1983	15259	17242
C	21	87	1322	13387	14709
D	9	39	350	2726	3076
E	8	31	323	3122	3445
F	18	99	1099	8121	9220
G	15	85	723	7831	8554
H	6	51	547	8526	9073
Total	131	645	7483	67020	74503

Tableau 2 : tableau synthétique des subdivisions de chaque section et des tailles de ces subdivisions. Version de janvier 2019, source OMPI: https://www.wipo.int/classifications/ipc/fr/ITsupport/Version20190101/transformations/stats.html

En plus de ce schéma de classification, l'OMPI a développé des outils pour faciliter l'utilisation de cette classification :

- **IPCCAT**[5] : Outil d'aide au classement pour le système de la CIB, créé principalement pour faciliter le classement des brevets aux niveaux des classes,

sous-classes, groupes principaux ou sous-groupes de la CIB à l'aide des résumés ou d'un texte court

- **STATS** : Outil qui offre des prédictions de classement dans la CIB fondées sur une analyse statistique des documents de brevet contenant les termes qui font l'objet de la recherche
- **Inventaire vert** : Facilite la recherche d'information en matière de brevets relative aux technologies respectueuses de l'environnement
- IPCCAT-neural est un ensemble de réseaux de neurones (intelligence artificielle) accessible par une API qui, à un texte donné, retourne un code de la CIB. Initialement mis au point en 2003, l'outil est formé à la prévision du classement sous les 4 niveaux supérieurs de la CIB (section, classe, sous-classe et groupe principal). Il s'utilise de cette façon :
- Le service web de l'IPCCAT reçoit un fragment XML comprenant : texte du document, langue, niveau attendu de la CIB et le nombre de codes IPC espérés ;
- Le service retourne un nouveau fragment XML qui inclut le nombre demandé de catégories prédites au niveau IPC ainsi qu'un score de confiance pour chaque prédiction. Ce score de confiance est calculé en valeur absolue.

L'IPCCAT fonctionne aujourd'hui avec les textes écrits en anglais ou en français (Fiévet & Guyot, 2018). Nous utiliserons cet outil pour réaliser un système de pivot sur un système d'indexation de connaissances utilisant les disciplines en tant qu'entrée d'index desquels nous extrairons le résumé (contenant hypothétiquement une description de connaissance) et la discipline afférente. IPCCAT construira un pivot vers le monde « tangible » décrit par la classification brevet.

3.2. Le référentiel des disciplines

Il existe plusieurs nomenclatures pour hiérarchiser les domaines, sous domaines ou encore les disciplines et sous disciplines scientifiques. Cette différenciation est variable selon les institutions et organismes.

Nous en avons recensé quelques-unes et le Tableau 3 rend compte de la variété des nomenclatures utilisées selon les organismes (Domaines vs champ disciplinaires ou méta-regroupements) et du degré des subdivisions selon :

1- ERC (European Research Council)[6]
2- AERES/ HCERES[7] (Haut Conseil de l'Evaluation de la Recherche et de l'Enseignement Supérieur)
3- OST (Observatoire des Sciences et des Techniques)[8]
4- CNU (Conseil National des Universités)
5- INRA (Institut National de la Recherche Agronomique)
6- Ministère de l'enseignement supérieur et de la recherche et les Ecoles Doctorales

7- UNESCO (Organisation des Nations Unies pour l'Education, la Science et la Culture)
8- Wikipédia[9]

		Structures							
		ERC	AERES/HCERES	OST	CNU	INRA	Ministère et Ecoles doctorales	UNESCO	Wikipédia
Nomenclatures	*Domaines disciplinaires*	3	14			3	X		
	Disciplines	25		9	X	120	11		4
	Domaines scientifiques		3						
	Sous-domaines		37						
	Sous-disciplines et plus			X			X		
	Champs				6				
	Groupes				11				
	Sections				52				
	Sous-sections				X				
	Macrodisciplines					28			
	Métaregroupements					11			
	Concepts de niveau supérieur							24	
	Concepts génériques							X	
	Concepts spécifiques							X	
	Concepts relatifs							X	
	Branches								X
	Sous-branches								X

Tableau 3 : tableau des nomenclatures et hiérarchisation des disciplines selon les différents organismes et structures compétentes

Ce tableau n'est pas une analyse approfondie mais une liste exhaustive des différents référentiels qui montre la diversité des nomenclatures des disciplines (plus ou moins prédominantes selon les organismes avec les "sciences exactes", les "sciences de la vie", les "sciences technologiques" ou encore les "sciences humaines et sociales") et qui rend compte également de la complexité du dialogue interdisciplinaire à ce niveau de notre étude.

3.3. La base des thèses françaises

Le site français « theses.fr »[10] est une base de données ouverte en juillet 2011 recensant toutes les thèses soutenues depuis 2006, dans les établissements ayant choisi le dépôt sous forme numérique (et laissant de côté le dépôt papier). Historiquement, c'est une des applications du projet « Portail des thèses » confié à l'ABES (Agence Bibliographique de l'Enseignement Supérieur) en 2009, par le Ministère de l'Enseignement Supérieur et de la Recherche.

Les données sont issues des personnels des établissements de soutenance et s'agrègent depuis les applications sources : STEP (Signalement des Thèses En Préparation), STAR (Signalement des Thèses Archivage Recherche) et le catalogue collectif SUDOC (Système Universitaire de Documentation).
Depuis son lancement, la base de données a connu plusieurs évolutions majeures :

- Septembre 2011 : versement des données issues du Fichier Central des Thèses vers theses.fr
- Mai 2013 : enrichissement de la bibliographie nationale des thèses (thèses soutenues en France depuis 1985).

Les données sont accessibles via l'interface du site (facilitée par un moteur de recherche s'appuyant sur les différents champs d'indexation). Les données bibliographiques relevant du domaine public sont disponibles via une API. L'exportation sous six formats différents est également permise.
Pour mener à bien les opérations et organiser le comité de décision, l'ABES s'est entourée de plusieurs partenaires : MISTRD (Mission de l'information scientifique et technique et du réseau documentaire), DGESIP (Direction générale pour l'enseignement supérieur et l'insertion professionnelle), DGRI (Direction générale pour la recherche et l'innovation), Conférence des Présidents d'Université, Conférence des Grandes Ecoles, Formations doctorales dans les établissements habilités, Recherche privée ou monde économique, CNRS-INIST (Centre National de la Recherche Scientifique - Institut de l'Information Scientifique et Technique), Atelier national de reproduction des thèses de Lille, CINES (Centre Informatique National de l'Enseignement Supérieur), Intelli'Agence, ADBU (Association des Directeurs de Bibliothèques Universitaires), AURA (Association du Réseau des Utilisateurs des produits de l'ABES), CCSD (Centre pour la Communication Scientifique Directe) et TEL (Thèses En Ligne). Bien que ces organismes se soient entendus sur les opérations d'indexation des thèses, dans la pratique nous avons pu constater une immense variété[11] des disciplines.

En septembre 2019, la base nationale des thèses françaises compte 458 964 thèses (dont 385 902 thèses soutenues et 90 560 thèses accessibles). A chaque entrée bibliographique de thèse les champs recueillis par la plateforme sont :

- Discipline
- Établissement
- Date de soutenance
- Etat : « en préparation depuis le »
- Écoles Doctorales
- Langues
- Directeurs de thèse
- Domaines
- Description (français et anglais)
- Mots clés
- Jury de thèse (président, rapporteurs).

Pour rappel, l'accès au texte intégral de la thèse est soumis à l'autorisation accordée

par le docteur en question (et/ou les ayant-droits).

4. Méthode de recherche

Pour donner à voir les documents académiques qui abordent des objets selon différents points de vue disciplinaires et, en ce sens, répondent à une nécessité potentielle de dialogue interdisciplinaire, nous proposons une expérimentation qui vise à tester la CIB comme pivot de classement des résumés des publications.

Le terrain académique choisi se concentre ici sur la « science en train de de se faire », soit : les thèses de doctorat. Le sous-objectif de ce projet est de proposer des instruments de lecture et d'exploration des données collectées et traitées. Le schéma (Image 1) présente la chaîne de traitement globale de l'expérimentation : la collecte des données s'effectue sur l'API de la base des thèses françaises. Nous normalisons les disciplines recueillies en utilisant le référentiel CNU[12] et envoyons le résumé à l'API IPCCAT pour identifier un code CIB associé au texte. Le score nous permet de filtrer les résultats afin d'obtenir un niveau de fiabilité satisfaisant du procédé d'identification des codes CIB. Enfin, les données sont préparées pour différents outils de visualisation afin de mettre en exergue les éventuels rapprochements interdisciplinaires potentiels. Cette procédure est détaillée ci-après par la description des scripts de traitement.

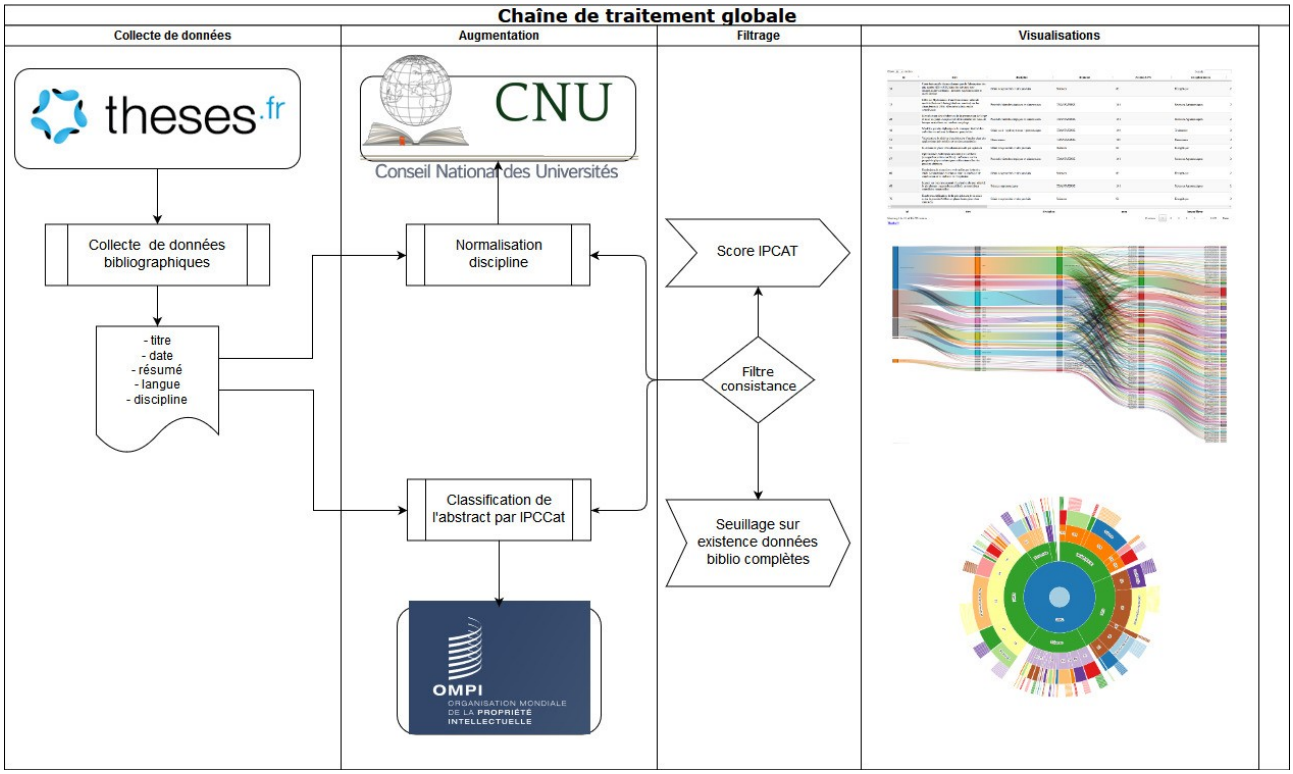


Image 1: chaîne de traitement globale de l'expérimentation : de la collecte (gauche) à la visualisation (droite) les différentes opérations de traitement, filtrage et seuillage des données bibliographiques

4.1. Les scripts

De nombreux scripts ont été réalisés pour la mise en œuvre de cette expérimentation. Ils sont complémentaires et sont facilement exploitables, voire modifiables, pour étendre les fonctionnalités (rajout de champs tels les établissements ou les directeurs de recherche) ou le terrain d'application (revues académiques, résumés des productions de laboratoire). Les scripts sont disponibles en Open Source sur Github[13].

- Collecte de données bibliométriques de la base des thèses françaises (theses.fr)

CollecteThese.pyest le script de départ. Il suffit d'adapter la ligne 12 valeur de 'requete' à vos besoins. Afin de rester respectueux des serveurs il ne faut pas changer la valeur du time.sleep. La ligne 13 correspond au nom de fichier dans lequel seront écrites les données.

- Préparation des données récoltées

Le script **LectureTheses.py** s'appuie sur un fichier JSON préparé par le script précédent. Il va récupérer (si possible) le résumé de chaque entrée bibliographique de thèse collectée par le script précédent. Les entrées qui n'ont pas de résumé seront exclues de la suite. Le résumé est alors catégorisé par IPCCAT[14]. Une série de CIB et de scores de classement est alors ajoutée aux données bibliographiques de la thèse. Le fichier JSON produit (ligne 20) contient les données bibliographiques augmentées de ces deux éléments (résumé et liste CIB).

- Nettoyage des disciplines

Les données ont nécessité la mise en place d'un nettoyage du champ "disciplines" tant la variété sur le plan lexical strict était importante (3307 disciplines différentes pour 16790 thèses recueillies). Sans prétention[15] le script **TraiteDisicipline.py** s'appuie sur un dictionnaire créé (le 08/06/2019) à partir de la description des sections disciplinaires du site du CNU[16]. Le dictionnaire utilisé est au format csv selon la nomenclature (Domaine; Numéro de section;[liste lexicale;]) cf. 'DisciplinesCNU.csv). Le script utilise une distance de Levenshtein améliorée pour pouvoir rapprocher indifféremment des tailles de chaînes et les positions de mots au mieux par rapport aux entrées du dictionnaire initial. Cf. fonction MatchSection. Le script a été lancé plusieurs fois pour converger au mieux et le dictionnaire récursivement adapté en rajoutant des termes pour qu'ils soient associés à la "bonne" section. Nous avons créé un domaine transverse et des numéros associés à ce domaine pour séparer ce que nous ne pouvions classer... Reste aussi que l'algorithme de classement et de rapprochement des unités lexicales serait à améliorer. Le script construit un fichier JSON reprenant les données bibliographiques de thèse précédentes et rajoute, pour chaque thèse, les champs :

1- "domaine" : le domaine disciplinaire (DEG, SCIENCES, LSH1,2, Pharmacie ; Transverse)

2- "section" : le numéro de section CNU (1 à 77 + quelques entrées au-dessus de 100)

3- "DiscipNorm" : la discipline "normalisée", première entrée dans le dictionnaire csv après le numéro de section.

4.2. Les visualisations

Une fois cette étape terminée, il ne reste plus qu'à préparer les données pour les visualisations. Les scripts suivants transforment les données structurées précédemment traitées dans des formats compatibles avec les différents instruments de visualisation sélectionnés pour certains à partir des exemples de la librairie D3JS[17] pour aider la lecture :

1- **FiltresJsonDataPivot.py** : export des données pour Datable[18] et PivotTable[19].

2- **GraphHierarchie.py** : pour chaque thèse, sélectionne la valeur du score maximal et le numéro CIB associé (si présent). Les scores nuls sont ignorés. Trois fichiers de type json sont produits dont la dénomination suit la nomenclature suivante :

> 1- HierarchieDiscipline-score-Titre : un tableau nommé imbriquant la hiérarchie des disciplines (corpus, domaine, section, discipline, etc. jusqu'au titre de la thèse ; des tableaux de tableaux) le score est le niveau de seuillage sur le score obtenu par IPCCAT ;
>
> 2- ValHierarchieDiscipline-score-Titre (même que précédemment mais la valeur de chaque nœud est donnée et fixée au nombre d'enfants de chaque niveau de branche ;
>
> 3- GraphHierarchie-score-Titre suit une autre représentation en graphe : la liste des nœuds (sections, disciplines, etc.) et des liens (relation d'appartenance) associés.

Chacun de ces fichiers est utile à une ou plusieurs des représentations différentes. Voici les différents diagrammes retenus pour l'exploration (d'autres non décrits ici sont disponibles sur le dépôt) :

- **PivotTable** : tableau croisé dynamique permettant de générer des graphiques à partir de croisement de données que l'utilisateur choisi via l'interface (histogramme des productions, cartes de chaleur par domaine, etc.).

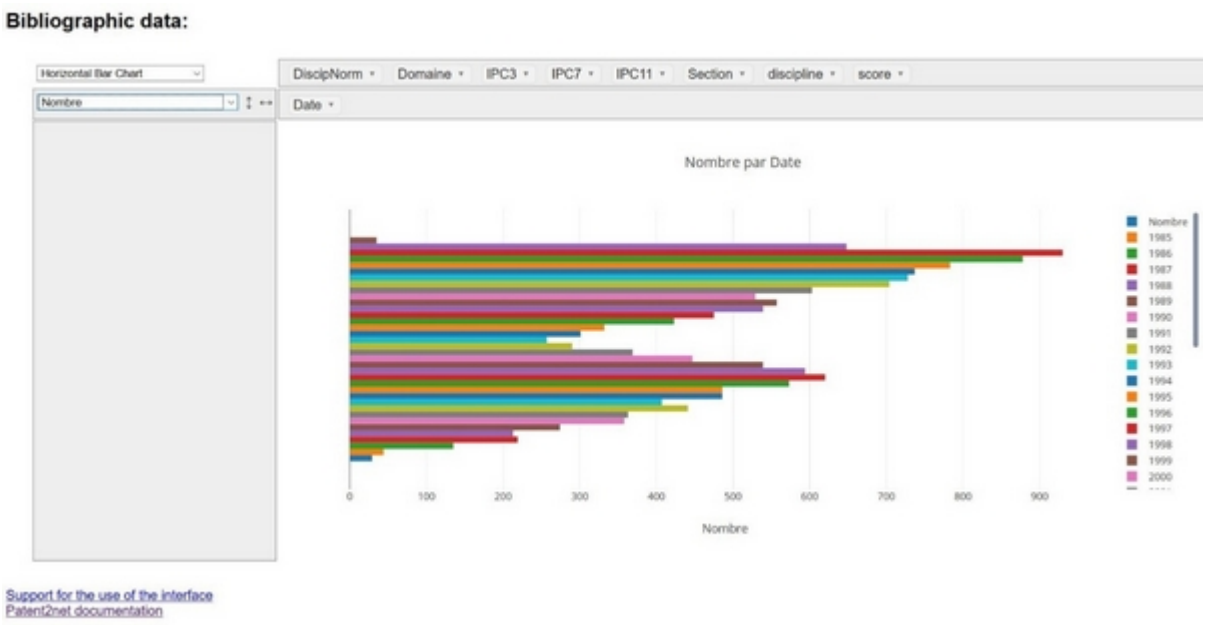
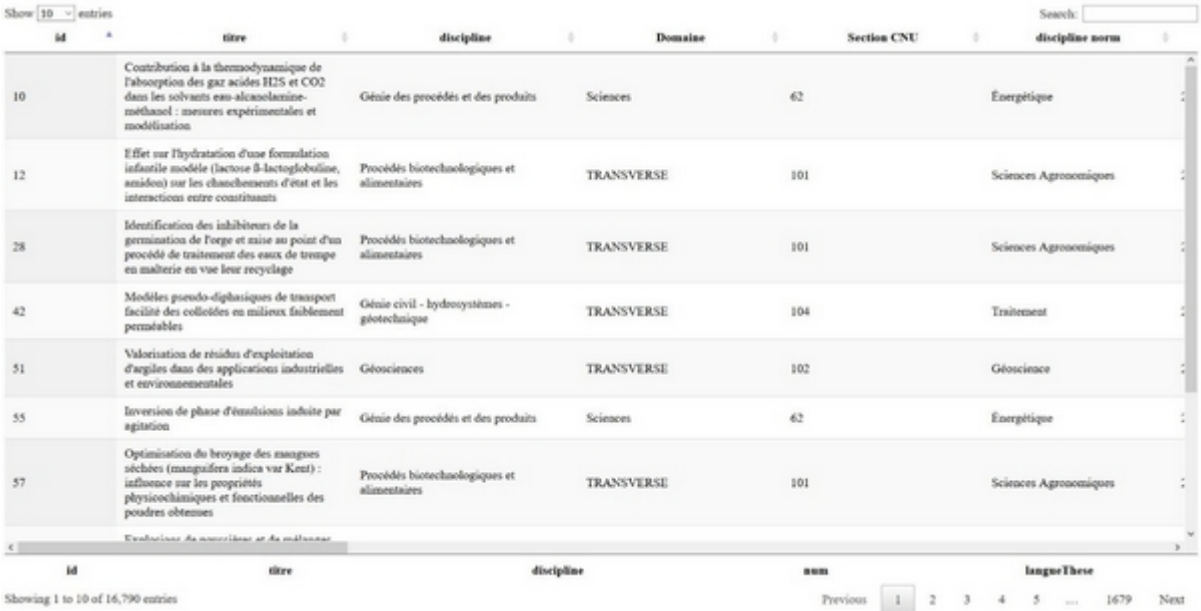


Image 2 : Tableau croisé dynamique avec, ci-contre, un histogramme modulé avec les différentes données pour obtenir "nombre par date" pour la requête "eau"

- **DataTable** : tableur de données permettant de visualiser et d'explorer (tri, recherche dans les différentes entrées bibliographiques du corpus).



Show 10 entries — Search:

id	titre	discipline	Domaine	Section CNU	discipline norm
10	Contribution à la thermodynamique de l'absorption des gaz acides H2S et CO2 dans les solvants eau-alcanolamine-méthanol : mesures expérimentales et modélisation	Génie des procédés et des produits	Sciences	62	Énergétique
12	Effet sur l'hydratation d'une formulation infantile modèle (lactose ß-lactoglobuline, amidon) sur les chanchements d'état et les interactions entre constituants	Procédés biotechnologiques et alimentaires	TRANSVERSE	101	Sciences Agronomiques
28	Identification des inhibiteurs de la germination de l'orge et mise au point d'un procédé de traitement des eaux de trempe en malterie en vue leur recyclage	Procédés biotechnologiques et alimentaires	TRANSVERSE	101	Sciences Agronomiques
42	Modèles pseudo-diphasiques de transport facilité des colloïdes en milieux faiblement permeables	Génie civil - hydrosystèmes - géotechnique	TRANSVERSE	104	Traitement
51	Valorisation de résidus d'exploitation d'argiles dans des applications industrielles et environnementales	Géosciences	TRANSVERSE	102	Géoscience
55	Inversion de phase d'émulsions induite par agitation	Génie des procédés et des produits	Sciences	62	Énergétique
57	Optimisation du broyage des mangues séchées (mangufera indica var Kent) : influence sur les propriétés physicochimiques et fonctionnelles des poudres obtenues	Procédés biotechnologiques et alimentaires	TRANSVERSE	101	Sciences Agronomiques

id | titre | discipline | num | langueThese

Showing 1 to 10 of 16,790 entries — Previous 1 2 3 4 5 … 1679 Next

Image 3 : Tableau de données des différentes informations appartenant à la requête "eau" (identifiant, titre de la thèse, discipline, domaine, section CNU, directeur de thèse...)

- **Collapsible Tree** : arborescence pliable/ modulable. Utilitaire pour visualiser l'arborescence disciplinaire reconstruite à partir des données du corpus. Chaque branche (domaine) s'ouvre sur les numéros de section, elle-même sur l'entrée choisie comme « clé » pour les différentes variations lexicales (ou autres entrées décrivant la section) identifiées (Image2). La version TidyTree.html se veut présenter l'ensemble de l'arbre (sans interactions) des domaines aux titres des thèses.

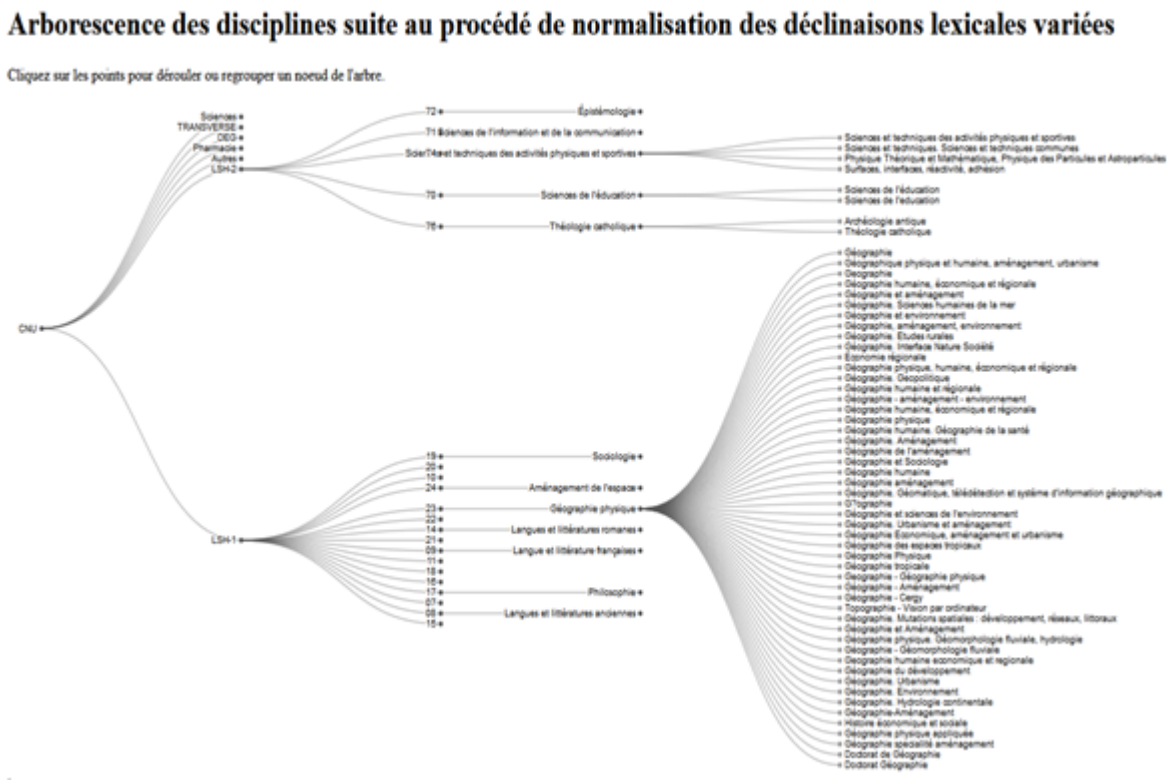


Image 4 : La cartographie des domaines disciplinaires, les disciplines du CNU (numéro de section) et l'imbrication des différentes écritures de ces dernières. Les branches sont repliables et dépliables pour pouvoir naviguer. Ici quatre disciplines sont « déroulées » : de haut en bas, les sections 74, 70, 76 et la section 23 qui permet de lire les nombreux déclinaisons de la géographie dans les disciplines extraites des entrées bibliographiques du corpus.

- **Nested Treemap** : carte proportionnelle imbriquée qui révèle le degré de représentation des domaines, sections et disciplines "normalisées" dans le corpus d'étude en rendant chaque élément proportionnel au nombre d'entrée (les code CIB). Le nombre de thèses associées est aussi inscrit (cf. Image 3).

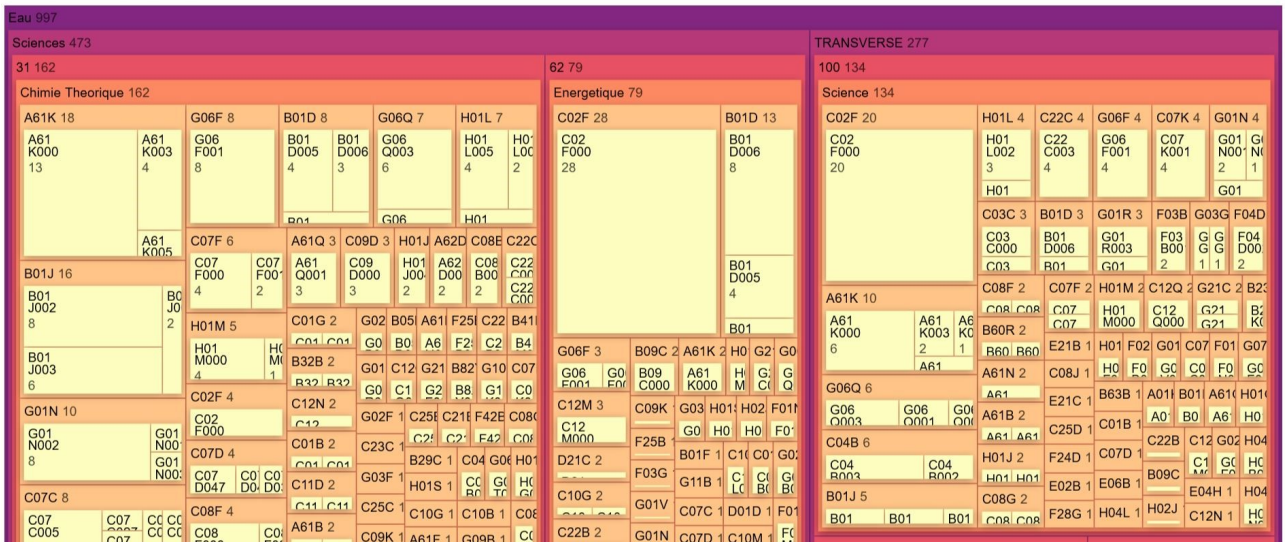


Image 5 : Carte proportionnelle imbriquée. Permet de visualiser rapidement les codes CIB prédominants par section disciplinaire et leur nombre de thèse afférentes.

- **Sankey** : diagramme de Sankey (ou diagramme de flux) met en lumière les croisements disciplinaires les plus marqués. Soit en partant des domaines vers les codes CIB, soit l'inverse (cf. Image 4).

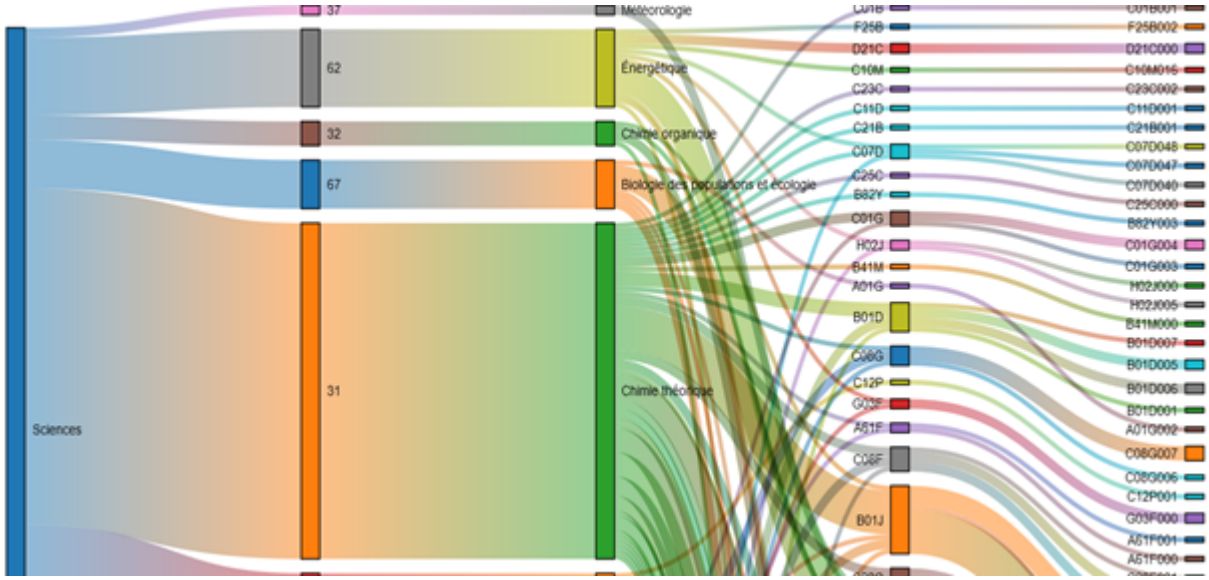


Image 6 : Extrait du diagramme de flux (Sankey) représentant les imbrications depuis les domaines disciplinaires aux codes CIB. Les relations et constructions (points de vue) interdisciplinaires sont manifestes par la lecture des croisements et codes couleurs associés (en milieu de figure, les codes CIB d'ordre 4).

- **Sunburst** : un cercle intérieur entouré d'anneaux de niveaux hiérarchiques plus

profonds qui permet une exploration interactive de la hiérarchie Corpus > domaines > section > discipline > codes CIB.

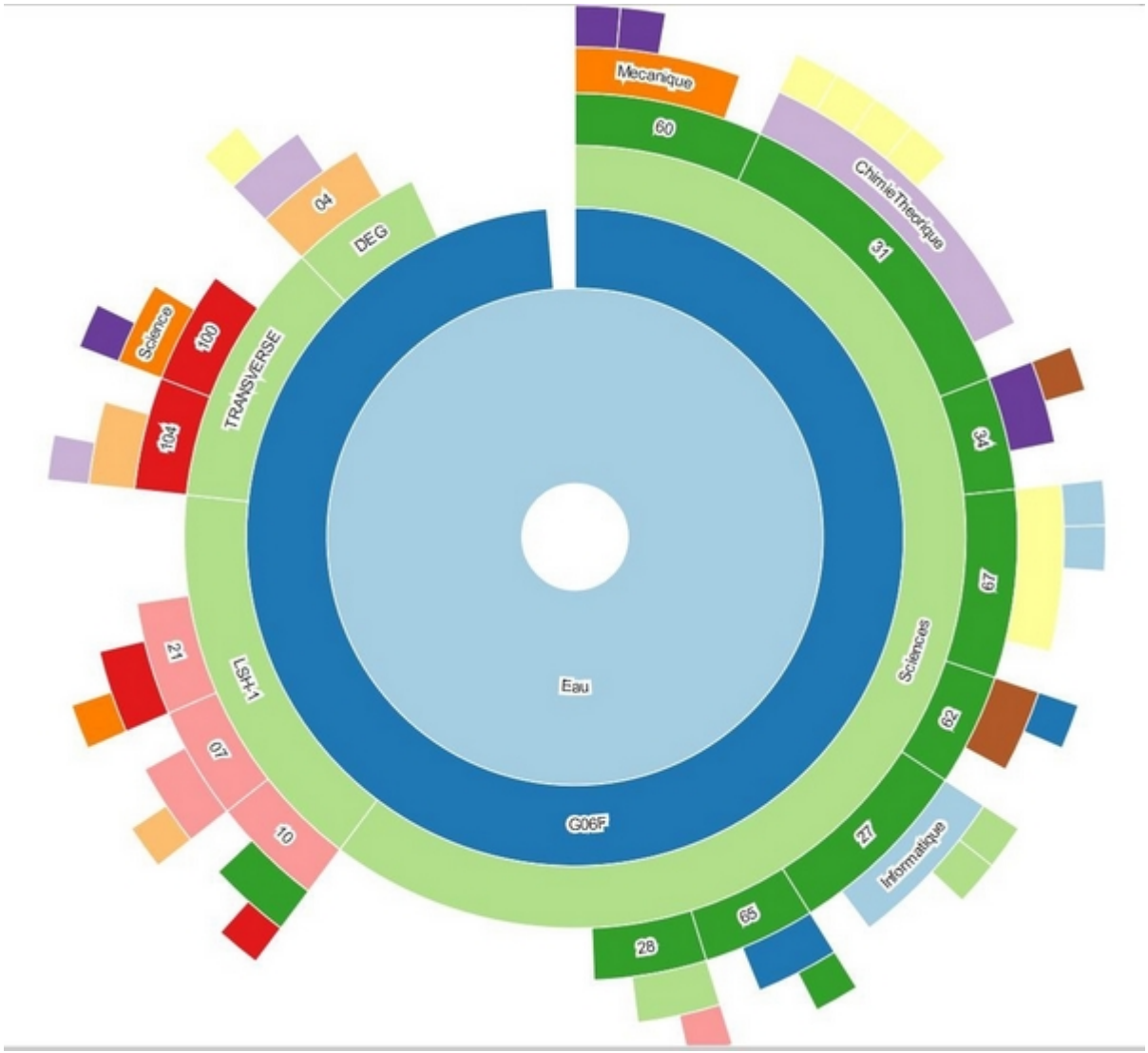


Image 7 : Exemple de diagramme (Sunburst) pour la requête "eau" avec les différentes hiérarchies interactives selon les disciplines, domaines et code CIB en référence aux thèses obtenues (corpus)

5. Collecte et exploitations des données

Nous avons choisi, pour ce test, d'utiliser un terme générique que l'on sait relevant forcément de différentes disciplines afin d'illustrer notre expérimentation : la requête ici est « eau ». Ce terme nous a permis de générer (au 08/06/2019) un corpus de 16790 entrées bibliographiques au terme de 15 heures de collecte. Un total de 444 entrées bibliographiques ne possédait pas de résumé.
Le Tableau 4 représente les résultats du procédé de traitement et de filtrage (troisième

colonne du schéma général (Image 1)) qui consiste à classer les résumés dans la CIB. Le score du classement obtenu nous sert à seuiller les résultats et éliminer les entrées. Ce tableau montre les évolutions du nombre d'entrées bibliographiques (ligne 3) en fonction du score (ligne 1) qui induit l'élimination des entrées dont le score est en dessous. La ligne 2 représente le nombre d'entrées éliminées. La dernière ligne du tableau révèle le nombre de disciplines présente au sein de chaque sous corpus.

Seuil score	0	600	800	1 000	1 200
Entrées éliminées	444	450	11 666	15 793	16 432
Thèses traitées	16 346	16 340	5 124	997	358
Disciplines distinctes	96	96	89	63	44

Tableau 4 : génération de sous corpus par seuil sur le score obtenu par IPCCat pour les résumés de thèse

Comme dit précédemment, nous avons produit les outils d'exploration des résultats sous forme visuelle qui permettent d'explorer ce corpus.

Grâce à ces illustrations, nous pouvons voir que « l'eau » (terme très générique mais emprunté) et les travaux utilisant ce terme se situent quelquefois au carrefour de plusieurs disciplines. Par l'exploration apparaît que certains codes CIB sont adressés par de multiples disciplines (par exemple, avec la requête "eau", nous obtenons : G06T – traitement de l'image, A01N – préservation des plantes, G06F - traitement électrique de données numériques, etc…). Ce qui permet de relier les textes et les résumés des thèses selon un système de hiérarchisation/classification précis et tangible.

6. Discussion (résultats et limites)

Les résultats obtenus par les visualisations nous permettent de poser clairement les avantages et les limites de cette expérimentation.
Selon le cas, nous obtenons des résultats pertinents et ce grâce à l'outil de classification utilisé (la CIB). Mais nous pouvons à ce stade dégager déjà deux limites :

- La classification des inventions : des objets techniques qui ne sont pas des concepts qui implique des biais évidents liés à la sémantique des termes (par simple opposition concept/concret). L'utilisation d'outil d'extraction sémantique avancé permettrait potentiellement de filtrer en amont les textes dont le contenu ne relève pas du concret ce qui augmenterait l'efficacité de l'instrumentation proposée
- Le classement grossier et imparfait des disciplines (mais ajustable) pose aussi quelques écueils : certains croisements disciplinaires affichés par le diagramme de

Sankey (par exemple Image 6) peuvent venir d'une erreur d'affectation d'une chaîne lexicale à une section disciplinaire...

Toutefois, à partir de la requête de départ, l'expérimentation peut être adaptée à toutes les disciplines à partir de n'importe quel mot clé. Par choix, nous avons privilégié une requête nous conduisant à des domaines de recherche d'applications tangibles (ingénierie, géographie, pharmacie, chimie sont les plus représentés). Par construction, nous avons évité de nombreux sujets de recherche qui ne se prêtent pas à ce procédé de pivot telle la philosophie ou la politique par exemple. Mais nous avons pensé à des requêtes qui pourront faire l'objet d'une nouvelle étude pour cette expérimentation, comme : "communauté", "information" ou encore "réseaux". A l'aide des résultats obtenus, nous pouvons voir que notre proposition est au croisement de deux mondes : le monde académique et le monde technologique (ce qui nous renvoie à la notion même d'interdisciplinarité et à la problématique de départ de notre étude).

Cette expérimentation peut être également modulée à partir d'autres sources de données. Nous avons choisi par avance un système d'indexation issu de la production scientifique à partir des résumés des thèses françaises, mais celui-ci peut s'appliquer à d'autres corpus (résumés d'article de recherche sur HAL, PubMed, Google Scholar, ArchiveSIC, ISTEX...).

Nous pensons également utiliser d'autres référentiels de classement documentaire (comme la classification de Dewey ou Mesh) mais les outils pour automatiser sont manquants.

7. Conclusion

L'instrumentation documentaire produite permet (à la limite de son perfectionnement) d'extraire à un sujet donné les positionnements et interrogations disciplinaires potentielles. En s'affranchissant des erreurs introduites par la projection sur le schéma de classement de la CIB des résumés de thèses, et par le classement approximatif des identification lexicales des disciplines, le résultat est interprétable par la considération de points de vue « disciplinaires ». De fait, les échanges et la communication peuvent alors se fonder sur ces points de vue singuliers afin de coconstruire le dialogue, s'entendre (ou identifier) sur les terminologies et aborder alors la problématique initiale sous un angle potentiellement nouveau, résolument interdisciplinaire.

Des travaux sont en cours pour améliorer le dispositif et circonscrire les limites précisément sur le plan lexical pour tenter d'éliminer les écueils de classification. En extension, en termes d'application, nous projetons de réaliser une approche comparative de différentes bases d'indexation documentaire afin de cartographier, au travers de la codification du schéma de la CIB, les différences et ressemblances de ces bases documentaires afin d'en révéler des éventuelles spécificités.

Cette expérimentation a été grandement facilitée par l'ouverture des données et la diffusion des connaissances rendues possibles à leur tour par de nombreux facteurs aujourd'hui indispensables, tels que : l'Open Science, l'interopérabilité des données, les vocabulaires contrôlés (Linked Open Vocabulary) et les principes FAIR pour la gestion des données.

8. Bibliographie

Czajkowski, A. (2011). *Using patent classification for searching, in particular the International Patent Classification (IPC)*. Consulté à l'adresse : https://www.wipo.int/export/sites/www/tisc/en/ppt/Philippines/patent_classification.pdf

Dubois, M. (2014). « *Private Knowledge* et "programme disciplinaire" en sciences sociales : étude de cas à partir de la correspondance de Robert K. Merton », *L'Année sociologique*, *64*, 1, p. 79-119.

Durand-Barthez, M. (2013). Former à l'information brevets dans l'enseignement supérieur. *Revue Internationale D'Intelligence Économique*, *5*(1), 25-38.

Fiévet, P., & Guyot, F. (2018). *Automatic Categorization of Patent Documents in the International Patent Classification (IPCCAT)*. Présenté à The International Conference on Search, Data and Text Mining and Visualization. (IC-SDV), Nice. Consulté à l'adresse https://haxel.com/ii-sdv/2018/Programme/monday-23-april-2018

Fourez, G. (dir.), Maingain, A. & Dufour, B. (2002). *Approches didactiques de l'interdisciplinarité*. Bruxelles : De Boeck.

Grant, J. L. (2011). *Searching Using Different Classification Systems*. Consulté à l'adresse https://www.wipo.int/export/sites/www/tisc/en/ppt/Morocco/Classification_Systems_EN.pdf

Guthleben, D. (2014). De l'indispensable interdisciplinarité. CNRS Le journal. Consulté sur : https://lejournal.cnrs.fr/billets/de-lindispensable-interdisciplinarite

Kourilsky, F. (2002). *Ingénierie de l'interdisciplinarité. Un nouvel esprit scientifique*. L'Harmattan, Paris, 2002.

Le Deuff, O. (2015). Utopies documentaires : de l'indexation des connaissances à l'indexation des existences. *Communication & Organisation*, 48(2), 93-106. https://www.cairn.info/revue-communication-et-organisation-2015-2-page-93.htm.

Lenoir, Y.(2003). La pratique de l'interdisciplinarité dans l'enseignement: pour construire des savoirs transversaux et intégrés dans le cadre d'une approche par compétences, CRIE

Faculté d'éducation, Université de Sherbrooke, P 76

Louvel, S. (2015). Ce que l'interdisciplinarité fait aux disciplines: Une enquête sur la nanomédecine en France et en Californie. *Revue française de sociologie*, vol. 56(1), 75-103. doi:10.3917/rfs.561.0075.

Morin, E. (1994). « Sur l'interdisciplinarité » Bulletin Interactif du Centre International de Recherches et Études transdisciplinaires n° 2 - Juin 1994 [En ligne, consulté en septembre 2018] http://ciret-transdisciplinarity.org/bulletin/b2c2.php

Nicolescu, B. (1996). *La transdisciplinarité*. Ed. du Rocher : Paris.
Prud'homme, J. & Gingras, Y. (2015). Les collaborations interdisciplinaires : raisons et obstacles. In : *Actes de la recherche en sciences sociales*, 210(5), 40-49. doi:10.3917/arss.210.0040.
Reisse, J. (1999). L'interdisciplinarité, conséquence d'une division arbitraire du savoir. *Revue de Géographie Alpine*, 87-1, pp 13-18. Consulté sur : https://www.persee.fr/doc/rga_0035-1121_1999_num_87_1_2913
Weingart, P. & Stehr, N. (eds.). 2000. *Practising Interdisciplinarity*, Toronto, University of Toronto Press.

9. Notes

[1] https://www.wipo.int/portal/fr/
[2] https://worldwide.espacenet.com/
[3] https://www.wipo.int/classifications/ipc/fr
[4] Il existe d'autres schéma de classification. La CIB est moins fine que l'ECLA par ex. mais est internationale (utilisée pour tous les brevets du monde). Un panorama des modalités de recherche à travers les différents systèmes de classification est proposé par Grant (2011).
[5] https://www.wipo.int/classifications/ipc/ipcpub/?notion=scheme&version=20190101&symbol=none&menulang=fr&lang=fr&viewmode=f&fipcpc=no&showdeleted=yes&indexes=no&headings=yes¬es=yes&direction=o2n&initial=A&cwid=none&tree=no&searchmode=smart
[6] https://cat.opidor.fr/index.php/Nomenclature_ERC
[7] https://esr-wikis.adc.education.fr/ca2co/index.php/Nomenclatures_:_2.1_Domaines_et_sous_domaines_scientifiques
[8] https://www.ird.fr › content › download › version › file › nomenclature OST
[9] https://fr.wikipedia.org/wiki/Discipline_scientifique
[10] http://www.theses.fr
[11] Sur le plan lexical le plus souvent (pluriels, coquilles) mais aussi au plan de la précision recherchée (histoire, histoire et archéologie…) ou inventive (Physicochimie, physicochimie, physique-chimie, chimie-physique…) et enfin sémantique (combustion, thermique, énergie thermique et combustion). De nombreux exemples liés à notre cas d'étude permettent d'apprécier cette variété cf. infra.
[12] Cette opération a dû être rajoutée face à la disparité des disciplines recueillies dans notre corpus de test. Le champ est en texte libre pour décrire une thèse et de fait de très nombreuses variantes lexicales sont présentes.
[13] Cf. https://github.com/ClaraGalliano/PivotCIB-NumeRev
[14] Voici le point d'entrée de l'Api IPCCAT : https://www.wipo.int/classifications/ipc/ipccat?&hierarchiclevel=
[15] Notre objectif n'est pas de normaliser les entrées de disciplines de la base des thèses

mais de produire un référentiel de lecture humainement lisible... Les auteurs ne s'engagent en rien sur la position éventuelle d'une (ou plusieurs) "sous"-discipline(s) dans une section erronée, ni sur les horribles choix potentiels que nous avons dû faire pour les besoins de la lecture !

[16] https://conseil-national-des-universites.fr/cnu/#/

[17] Cf. https://github.com/d3/d3/wiki/Gallery

[18] Cf. https://datatables.net/

[19] https://github.com/nicolaskruchten/pivottable/tree/master/dist