Médiations intellectives

David Reymond

Université de Toulon — EA 3820

Médiations intellectives

Habilitation à Diriger des Recherches
(Sciences de l'Information et de la Communication)

Mémoire présenté et soutenu publiquement le 23 novembre 2017

par

David Reymond

Devant le jury composé de :

- Pr. L. QUONIAM (Garant), CNU 71 - Université de Toulon
- Pr. B. JUANALS, CNU 71 - Aix-Marseille Université
- Pr. J. M. NOYER, CNU 71, I3M - Université de Toulon

Et des rapporteurs :

- Pr. L.FAVIER, CNU 71 - Université Charles-de-Gaulle (Lille - SHS)
- Pr. I. SALEH, CNU 71 - Université de Paris 8
- Pr. I. ROXIN, CNU 71/27 - Université de Franche Comté

Version hypertexte, amendée suite aux bienveillants commentaires du jury, en particulier le Pr. ROXIN.

À ma femme Séverine, qui me rappelle au quotidien les enjeux de la communication et mes propres limites, m'a soutenu et encouragé tout au long de ce travail. Elle a réussi par sa douce présence et sa persévérance à combler mes manquements à la vie de famille et à dynamiser mes efforts.

À mes enfants, forcément un peu délaissés dans cette entreprise, ils m'auront cependant témoigné patience et reconnaissance à leur manière.

À mes frères, depuis l'expert en opposition systématique à toute certitude, prenant le contre-pied de la moindre déclamation aussi référencée puisse-t-elle être au second, docile sur ce plan, qui réussi à calmer les discussions houleuses en conséquence en usant d'une nonchalance inébranlable.

À mes parents et ma famille, depuis toujours.

À mes beaux-parents, par amour, pour complexifier encore l'existence.

À mes amis, dont Jack, qui depuis le début de mes études marque de sa présence mon évolution.

À mon garant Luc, dont la personnalité n'a d'égal que la certitude de sa voie. Il m'a accueilli, ouvert à ses propres recherches que nous avons dynamisées. Le résultat de trois années de collaborations est synthétisé ici dans les chapitres portant sur le brevet. Je lui témoigne toute ma reconnaissance, ma gratitude et mon amitié.

À mon mentor Jean-Max, qui m'a ouvert les yeux sur nombre de travaux des Sciences Humaines et Sociales, par son recul et son envergure il m'aura soufflé les pistes essentielles à cette construction au cours d'échanges les plus variés et des plus joyeux.

À tous les membres du jury qui ont accepté de relire ce travail et de participer à sa soutenance.

Avant-propos

J'AI écrit ce mémoire à l'aide du processeur de texte LaTeX un langage de mise en page, qui m'a permis de maîtriser les aspects de présentation en tentant de faciliter sa lecture, notamment numérique. J'ai ainsi tenté de respecter au possible les règles des *petites leçons de typographie* de Jacques ANDRÉ (ANDRÉ, 2003) pour éviter les fautes du domaine.

J'appose quelques conventions de couleurs : les hyperliens internes au document sont en rouge (liaisons vers la bibliographie et depuis celle-ci pour revenir aux pages des citations, idem pour l'index et les notes de bas de page), les liens externes (URL) sont en vert. Figures et tableaux sont numérotés par chapitre suivent la convention de numérotation Chapitre.Numéro, également hyper reliés lorsque référence est faite.

Les citations sont soit marquées dans le texte par les traditionnels guillemets « », soit mises en valeur, en général pour les plus longues, par la mise en forme particulière suivante :

> *Ceci est une citation.*

La coloration est aussi utilisée au niveau syntaxique pour l'explicitation de quelques lignes de code (en annexe). Les noms propres sont dénotés en PETITES MAJUSCULES. Onglets et lettrines interviennent pour le plaisir des yeux. Les épigraphes sont des citations généralement historiques en phase avec le chapitre qu'elles introduisent. En espérant modestement compenser par cette voie esthétique quelques passages qui pourraient être difficiles à lire.

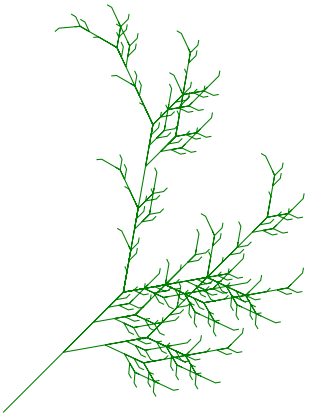
Enfin, en faisant abstraction des hyperliens, le document est réalisé pour être imprimé

en recto-verso ce que le lecteur aura deviné à la vue des onglets : ainsi, dans la version pour impression, les onglets sont symétriquement apposés au recto et verso de chaque page de sorte à se superposer par transparence et s'accentuer par sur-impression. Il y a deux versions de ce document, l'une destinée à la lecture sur écran, l'autre est compilée pour pouvoir être imprimée. Ainsi, dans cette dernière version les URLs des documents bibliographiques (url, DOI, ISBN) accessibles en ligne ne figurent pas, ni les hyperliens internes au document.

Encadrés

Ce type d'encadré dénote des éléments saillants pour mettre en valeur l'essentiel de mon propos. Une convention de couleur dont la légende reste à écrire, me permet de nuancer la portée ou le registre des éléments ainsi dénotés.

LaTeX est complété par BibTeX pour la gestion de ma bibliographie. Hormis la partie CV qui est spécifique à la présentation de mes travaux par ordre chronologique, le lecteur trouvera une bibliographie par chapitre présentée par ordre alphabétique, puis une bibliographie générale segmentée selon les grandes rubriques académiques (livres, articles de revue et de conférence, etc.) et par ordre alphabétique.



Sommaire

Résumé

L'INTERNET étend notre monde, l'autonomie du réseau le rend comparable à un organisme vivant qui associe humains et non humains, associations que l'hypertexte retranscrit, associations que le web mémorise en données. L'hypertexte se pose depuis en métaphore de la Communication, mais aussi des processus sociotechniques, partiellement un reflet et une extension de la réalité. Les notions centrales des SIC, telle par exemple l'identité (marque, organisationnelle ou individuelle), la mobilité (diaspora, culturelle, professionnelle...) ou le savoir (académique, technique) sont représentatives d'un ensemble de questions vives de caractérisation des formes sociales, culturelles, organisationnelles et info-communicationnelles actuelles dont l'écosphère est le témoin et quelquefois la manifestation. Les mettre en regard, les assembler permet de réfléchir sur des continuités ou des discontinuités conceptuelles mais aussi empiriques qui sont au centre des réflexions des chercheurs en communication. En conséquence, l'enjeu est décisif quant à l'élaboration d'instruments de reconstitution de faits établis sur ces questions.

Une extension des Sciences de l'Information et de la Communication (SIC), en appui sur des données numériques, se pose au sein des humanités digitales par l'exploration de faits observables dont les fonctions de médiations convoquées quantifient l'épaisseur symbolique et médiatique. Le couplage incontournable (information, fonction de médiation) réunit en premier lieu information/communication. Que ce soient le filtrage, l'agrégation, la transposition... tout élément de construction de l'émergence informationnelle se retrouve dans ce couple (information, fonction de médiation) qui transporte les éléments de son herméneutique.

Pour les SIC, les artefacts se posent ainsi en médiateurs offrant la documentation de faits et libère un niveau informationnel nouveau. Artefacts extensibles, ouverts, partageables et adaptables à souhait sur des phénomènes informationnels et communicationnels pas forcément compliqués mais d'envergures. Ces dispositifs permettent le

passage à l'échelle des mégadonnées (big data), et ouvrent enfin la perspective d'études scientifiques inédites en adressant la complexité humaine non plus par une limitation à une perspective circonstanciée d'un cadre spécifique mais par le « statistiquement vrai » ou « fortement probable » des sciences naturelles. L'exploitation des données informationnelles et de leur fonction de médiation permet de relativiser la portée et l'incidence, laissant espérer la généralisation (sous réserves) que l'on pourra alors cerner... en degré d'intellectivité.

L'anthropocène devient alors une source de captas potentiels propres à alimenter une nouvelle forme de documentation dévènements informationnels et communicationnels. L'enjeu scientifique en est l'inscription des pratiques des humanités digitales au sein des SIC, pratiques qui dépassent la simple donnée informative pour s'attacher aux modalités de production : les médiations intellectives adressent les mégadonnées en général (ouvertes ou pas). Réciproquement, la captation dévènements ouvre à la reconstruction d'unités libérant la perspective d'analyse comparée sur des élaborations de références normalisées.

En application directe de cette voie de recherche, la documentation brevet est une source majeure de l'intelligence compétitive. Apprendre à l'utiliser est un enjeu de taille sociétale, dont la complexité impose une médiation nécessaire qui s'inscrit clairement en SIC.

Modifiables et adaptables, les outils de médiation se positionnent comme une instrumentation de construction de moyens de pilotage sémiotiques, des artefacts délivrant des herméneutiques spécifiques et rendant concrètes les relations qu'entretiennent open data, datamining et algorithmique. La base européenne des brevets alimente le phénomène croissant de l'open-data : plus de 150000 publications mensuelles nouvelles sont recensées à la date de rédaction de ce document. Cette immense source informationnelle est négligée bien que d'intérêt aussi en intelligence compétitive, marketing, innovation, et l'ensemble des sciences de l'ingénieur et de la santé en général. La base regorge d'innovations, de sources potentielles à l'inspiration, de réponse à des problèmes concrêts, mais bien évidement aussi d'informations stratégiques et concurrentielles. Par son volume et sa dynamique, cette documentation constitue un flux de données structurées, faciles à appréhender. Pourtant, fouiller cette documentation, extraire de l'information pertinente en regard de ces problématiques très variées demeure encore d'une difficulté majeure.

Nous avons construit une gamme d'outils modulaires et open-source, facilement dérivables et exploitables : des artefacts de médiation pour l'analyse des documents brevets. La modularité permet la mise en œuvre de chaînes de traitements spécifiques (d'autres fonctions de médiation) à un contexte particulier (organisation, entreprise, collectivité territoriale...). Ainsi, au-delà de la documentation d'un domaine technologique particulier, de l'analyse compétitive (qui travaille sur quoi, avec qui et pour qui et avec quelle expertise), P2N peut être utilisé pour la production de corpus documentaires, ainsi que les chaînes de traitement spécifiques pour leur analyse. Ce peut être, par exemple, pour la construction de perspectives historiques de l'étude de l'évo-

lution technologique d'un domaine, ou bien l'élaboration d'une biographie, etc. Les différents traitements sont ouverts depuis la captation de la source brevet, leur exploration jusqu'à la présentation des résultats obtenus. Chacune de ces chaînes est ré-exploitable afin de pouvoir procéder à une contextualisation propre à des questionnements d'intelligence économique, en phase avec la dynamique informationnelle générale et la dynamique organisationnelle interne s'adaptant aux multiples usages de la veille.

Ces travaux ouvrent directement la documentation brevet pour des usages nouveaux (résolution de problèmes, créativité, innovation, intelligence économique) et des voies de recherches interdisciplinaires associées, mais aussi adresse de façon inédite le plan de la formation et l'éducation STEM et SHS en général en associant d'une part une source de documentation négligée et, d'autre part, la pratique de l'exploitation des données documentaires raisonnablement massives ce qui dépasse largement le cadre de la documentation brevet.

Introduction

Le développement extraordinaire du numérique catalyse la transformation actuelle du monde. Celle-ci affecte de manière forte et durable les pratiques communicationnelles et les pratiques de recherche mais aussi les conditions de production, de circulation et d'exploitation des savoirs. De ce point de vue nous avons constaté que la séparation entre information science (IS) et approche communicationnelle est de moins en moins fondée. Le web produit les données, et des outils pour leur exploitation, et à partir de ce problème, les bases de mon travail sont de développer des méthodes d'appréhension de l'information numérique, méthodes en général fondées sur une approche documentaire, méthodes d'agencements d'instruments existant pour produire des artefacts de « lecture », des médiations intellectives. Ainsi, que ce soient des sites web, leurs usages ou des banques de données documentaires, mon apport dirige l'agencement d'instruments permettant de porter un regard nouveau facilitant la compréhension de grandes quantités de données. La singularité se situe dans la mise en œuvre qui est portée, d'une part et à priori, par une contextualisation humaine des traitements opérés et, d'autre part, qui articule la croisée de différents niveaux informationnels pour que le cyberespace se donne à voir.

Dans l'écriture de ce mémoire je poursuis deux objectifs : le premier est de synthétiser mon parcours et mes différents travaux en les structurant autour d'un fil directeur apparu au gré de mes avancées et collaborations. Le second, plus étayé, ancrera ces travaux au plan disciplinaire et positionnera une direction actuellement développée de recherches. En ce sens, j'ai choisi de construire un complément à mes travaux antérieurs pour situer en premier lieu la nouveauté et l'intérêt de ces recherches mais également le potentiel catalyseur de ces travaux inscrits au sein des humanités numériques.

L'objet est de faciliter l'identification d'informations pertinentes à une problématique posée : la détermination ou le dévoilement d'informations utiles relativement à sys-

tème d'interprétation donné. Cette tâche passe par l'utilisation d'outils de traitement de grandes tailles de données pour explorer, classer, résumer et visualiser des informations (re)construites. Le développement en mode *open source* laisse libre l'extrapolation de l'existant : chacune des briques élémentaires est réutilisable pour déployer d'autres chaînes de traitement, et dériver l'outil pour le contextualiser [1] à une problématique particulière. Ainsi, par exemple libre aux usagers d'adapter si besoin l'outil à la production d'indicateurs technométriques de mesure de la production « brevet » sur des échelles variées (régionale, nationale, d'entreprise ou sur un sujet particulier) peu importe encore que ces indicateurs soient dynamiques ou statiques. Sur un autre niveau informationnel, l'ouverture se prête encore à l'introduction de dictionnaires terminologiques spécifiques pour affiner une analyse particulière. Ce peut-être encore l'exploitation d'autres modes de visualisation pour appréhender les données massives que constituent potentiellement l'espace des réponses de la base brevet. Le terrain est aujourd'hui préparé pour déployer des recherches variées de facilitation d'accès à cette très vaste documentation et, réciproquement, leur utilisation au travers de jeux de données constitue un terrain de préparation aux humanités digitales (histoire, droit, marketing...).

Le lecteur remarquera d'une part que la notion d'information se pose de manière ouverte à des niveaux d'échelles différents (simple texte, sites, menus de navigation puis document, les métadonnées) et, d'autre part, que la notion d'information sous entendue entre dans le cadre de pratiques intellectuelles qui renvoient à des milieux cognitifs différents. Cette séparation rend possible l'introduction de variables de contextualisation pour adapter les opérations d'herméneutique aux données massives. C'est dans la tension de ces deux mouvements que je développe un nouveau paradigme de recherche.

Ce sont ainsi pour l'essentiel ces deux éléments qui guident le développement qui suit avec un objectif de généralisation des résultats. Ma démarche de recherche est de type spagyriste qui consiste à extraire, décomposer et assembler, en vue de construire autre chose. J'accumule ainsi et bénéficie de ma propre activité de recherche pour continuer à apprendre, et pas simplement à produire.

La seconde partie [2] de ce document sera dédiée à l'inscription épistémologique de cette généralisation : à partir de l'infosphère de DE ROSNAY, le médium algorithmique de Lévy est dérivé comme instrument d'explicitation et de compréhension. Je souligne d'abord l'expansion massive du numérique au plan du quotidien : en particulier la communication, les échanges, et la documentation qui sont au cœur de mes préoccupations (cf. chapitre 1 p. 25–41). Dans ce cadre là, la notion d'information est déconstruite pour la reconstituer en l'abordant dans le cadre de la sphère sémantique en général : la granularité, informationnelle couplée à la pertinence des chaînes de leur mise en lumière vont permettre d'aborder autrement l'émergence sémantique. À

1. Constituant fondamental d'une pédagogie novatrice pour l'accompagnement et l'initiation aux humanités digitales centrés sur l'apprentissage de la capacité à la réutilisation de briques de code élémentaires.

2. Mon autobiographie, parcours et réalisations détaillées sont absentes ce cette

partir de l'information, la question de déconstruction/reconstruction de l'information est reprise à des échelles variées pour montrer que le processus même masque une sémantique propre dépassant la précédente. Ainsi, à partir d'entités primaires (dont le niveau peut pré-définir un degré sémantique) seul le couple entité informationnelle / fonction de médiation construit le sens donné par le chercheur (cf. chapitre 2 p. 43–74). Cette reconstruction s'opère de facto dans le domaine du numérique à l'aide dartefacts (collecte, agrégation, complétion, complémentation etc.) qu'ils soient primitifs ou élaborés. Ces unités atomiques composent alors les outils desservant la médiation du phénomène observé (oligopticons de LATOUR, le « voir » du cyberespace de BALTZ, des **médiations intellectives**). Enfin je montrerai que cette série d'opérations instruit une classe d'activité de recherche inscrite au sein des humanités digitales (cf. chapitre 3 p. 75–95).

La troisième partie développe l'espace d'une activité de médiation de l'information : celle-ci est documentaire (au niveau des brevets) mais s'opère à des échelles très variées. Les plans informationnels (niveau de pratiques, niveaux cognitifs) utilisés s'étendent depuis les textes (résumés, revendications, descriptions), les métadonnées (inventeurs, mandataires, etc.) et le schéma de classement des inventions. Par construction le document brevet est un document vérifié et co-construit pour lequel les résultats des traitements produisent de nombreuses informations factuelles : qui travaille sur quoi, pour qui, etc. Différentes chaînes de traitement sont produites par l'outil P2N en vue de produire des herméneutiques adaptées et différenciées : depuis le simple traitement d'appréhension d'un univers brevet (la réponse à une requête) aux constructions relationnelles que cela implique. Que ce soit au niveau des métadonnées (réseaux d'inventeurs par exemple) ou au niveau des contenus lexicaux (mots associés) (cf. chapitre 4 p. 99–131) les relations masquées sont dévoilées par un formalisme ou des structures de données adéquates. Enfin, le dernier chapitre met en scène l'ensemble des éléments abordés par la construction progressive et argumentée d'un procédé de *médiation intellective* opéré par hybridation de texte et de données documentaires suivi d'un traitement particulier.
Une requête, même fine, peut produire en réponse un très grand [3] ensemble de brevets qui sont en soit fastidieux à trier, à explorer pour en déterminer le substrat à un problème donné. Alors que, on dispose pour chaque brevet de :

- titre
- résumé
- classement dans la Classification Internationale des Brevets [4]

Ces trois segments de texte alimentent des instruments de traitement automatique qui participent chacun séparément à l'éclairage des contenus. C'est ainsi au plan sta-

3. Des ensembles de 200 à 2000 documents sont des tailles raisonnables.
4. La description des registres de classement de la CIB qui constitue un niveau méta. Ces descriptions sont des données textuelles décrivant chaque registre (plus de 70000) inscrites dans une autre base, une ressource documentaire complémentaire qui est écrite dans un langage normalisé. Ce texte décrit les inventions que l'on range dans ce registre.

tistique (via les nuages de mots par ex.) ou encore au plan des associations (par la méthode des mots associés (TURNER, COURTIAL, MICHELET)) que s'initie l'exploration. Ce sont ensuite des techniques de classification automatique des documents (par ex. selon la méthode REINERT) qui pourront faciliter l'appréhension de ces gros volumes de documents en les classant selon leur ressemblance terminologique ou les associations terminologiques qu'ils réalisent. Actuellement ce traitement potentiel se fait isolément dans chaque ensemble de segments de texte (cf. chapitre 5 p. 133–154).
Constat est fait par ailleurs, que pris isolément, le bruit provenant des termes génériques [5] masque le plus souvent des signaux plus faibles de ces ensembles. D'autant plus qu'à l'extrême, ces bruits sont de deux ordres : soit une erreur de transcription véritable issue des OCR ou bien, à l'opposé, un terme très spécifique grandement informatif. En conséquence les instruments de classification automatique ne sont pas toujours pertinents. Si l'on juxtapose les trois segments de texte, les fonctions différentes de chacun des segments introduisent-elles un bruit qui nuise à la qualité des classifications ? Quel gain opérationnel en recherche d'information peut-on espérer ?

Partant de là, le traitement développé dans le dernier chapitre (cf. chapitre 6 p. 155–178), sera de reconstruire le texte de description des registres de classification des documents pour l'injecter au sein de chaque document brevet afin de reconstituer un document hybride. Ce nouveau document additionne résumés et titres originaux puis le texte normalisé qui les décrivent. La mise en relation des contenus textuels avec leurs descripteurs reproduit alors des relations jusqu'alors masquées (implicites dans la base de données et, de surcroît au plan qualitatif, ce sont des relations qualifiées par l'expertise humaine « normalisée »). Afin d'estimer les effets, des mesures des résultats obtenus sont opérées et analysées : depuis les fondements infométriques des choix opérationnels de seuillage, de sélections de termes (mots-associés) qui composeront l'interface, aux algorithmes de classification automatique. Le résultat est une nouvelle chaîne de traitement autonome, qui fournit un éclairage singulier des univers brevet et, comme la plupart des autres chaînes de P2N, ouvre aussi à l'accompagnement à l'utilisation de ses sous composantes en tant qu'instruments des humanités digitales et, en ce sens, une direction de recherche dépassant la documentation brevet en général.

5. Il suffit de prendre les mots clés servant à la construction de la requête qui par définition se retrouvent dans tous les résumés ou titre des résultats. Ce sont d'évidence des termes très fréquents mais triviaux des corpus.

L'information envisagée en tant que phénomène en émergence n'est jamais stabilisée, alors que la conception technique a besoin de figer la réalité des objets et des relations en vue de formaliser les règles qui seront appliquées à leur gestion.

Les pratiques documentaires engendrent ce type de tension et se caractérisent par une médiation visant à la réduire. Cette médiation doit être prise en compte dans le cadre de l'analyse des conditions d'adéquation entre l'offre et la demande des produits d'information. Elle aboutit à engendrer un coût : celui que crée l'adoption de solutions techniques permettant de rigidifier la gestion des flux d'informations par l'adoption des représentations de besoins définis a priori.

Comment réduire ces coûts d'origine organisationnelle ? La réponse réside dans une diminution de la rigidité des représentations : c'est-à-dire dans l'introduction d'une grande souplesse au niveau de la conception des chaînes de traitement technique, afin de pouvoir les faire évoluer de manière à tenir compte de la rapidité d'évolution dans les modes d'usage de l'information.

— Turner William LA FORMATION DES PROFESSIONNELS DE L'INFORMATION, *Bulletin des Bibliothèques de France* (1995)

I

Inscription épistémologique

CHAPITRE 1

Écosystème informationnel : l'infosphère

Il existe un évènement dense, qui s'appelle l'Anthropocène.
Ce n'est pas un évènement court et brutal, soudain et rapide.
Ce n'est pas un objet spectaculaire ou médiatique.
C'est cependant un fait qu'aucune science actuelle ne sait penser et qui, en profondeur, interroge toute distinction actuelle entre sciences dites sociales et sciences dites naturelles.
C'est l'évènement épistémologique fondamental d'aujourd'hui.

— Hervé REGNAULD, L'évènement anthropocène - EspacesTemps.net (2017)

١

DANS CETTE PARTIE, je pose quelques constats sur l'avènement du numérique par lequel Internet est devenu l'entrée et le passage du monde vers une écologie cognitive. Il en constitue à la fois une extension mais aussi sa mémoire. L'ensemble pouvant desservir de façons inédites des desseins de recherches comportementales, sociétales et, par extension, en information et communication.

S'approprier l'Anthropocène pose nécessaire de repenser une épistémologie adéquate, en phase avec l'événementiel numérique : des faits, des événements, des phénomènes constituent des données, des témoins d'actions, des traces et empreintes d'échanges ou de réalités qui s'accumulent dans l'écosphère, peuvent être repérées ou construites à dessein pour identifier des modes opératoires, réactions, actions, ou relations [par exemple]. Les capter, les dissocier, les agréger pour alimenter la réflexion, ou reconstituer la Scène est l'enjeu du cheminement opératoire pendant de cette épistémologie.

L'internet est construit depuis sa création en un miroir de la société, en devient aujourd'hui une extension et se pose ici en fournisseur de données.

1.1 De l'Internet à l'écosystème informationnel

En quelques années, depuis son ouverture au public, le réseau s'immisce dans nos vies, agrège tous les médias et transporte des quantités extraordinaires de données. Au plan machinique, il relie des unités de traitement et de stockage proportionnelles dont la quantité croît au quotidien [1]. Le réseau interconnecte ainsi différentes sphères informationnelles dédiées à la communication (écrite, orale, multimodale) en support aux échanges humains ou mixtes (Web, réseaux sociaux, blogs, ...) ou encore en support à l'information, la mémorisation, la structuration documentaire des contenus. Ces sphères sont seulement distinctes par les logiciels d'arrière plan, pour leur majorité bâtis sur les mêmes protocoles de transport et d'adressage de dispositifs et de

1. Cf. http://internetlivestats.com

l'hypertexte.

Ces mêmes protocoles sont utilisés pour l'écriture et la lecture de documents potentiellement recomposables.

1.1.1 Support pervasif de transport informationnel

Sans en revenir aux origines, que de nombreux travaux retracent (HUITEMA, 1995; HAFNER, 1996; LEINER et al., 1997; GUICE, 1998; DROMARD et SERET, 2006; SCHAFER et THIERRY, 2013b; A. SERRES, 2000; TÉTU et RENZETTI, 1995), le média est global. Internet mondialement répandu [2], s'étend dans des cités connectées depuis les habitations individuelles jusqu'aux infrastructures critiques de la ville (DOUKAS et al., 2011).

Le réseau étend encore sa disponibilité quasi planétaire, y compris dans les pays émergents, en nous suivant dans les moyens de transport au moyen de smartphones (terminaux internet) : les exemples sont innombrables et les progrès technologiques laissent à penser que le phénomène n'en est qu'à ses débuts. L'internet suit une croissance exponentielle (ADAMIC et HUBERMAN, 1999), nous suit où que nous allions, nous précède par ses objets connectés qui seront probablement, à terme, part du quotidien.

Pour DE ROSNAY (2007, p. 57), Internet serait :

> *une sorte de système nerveux dont les internautes seraient les neurones... Il possède une structure de base, fractale (telle que décrite par* Mandelbrot*), bâtie sur le modèle des capillaires sanguins de l'être vivant.*

L'auteur décrit ainsi l'objet par une métaphore : un réseau de liens entre le canal de transmission et les individus pensants, dont la complexité intrinsèque est reconstruite par mimétisme de la complexité naturelle.

En sus de cette fonction *support de transmission*, la connexion de matériels lui octroie une capacité de mémorisation numérique que l'on peut segmenter en différentes sphères informationnelles non totalement disjointes pour observer les objets des SIC sous l'angle des médias (Web, tv, radio), de la communication interpersonnelle (mail, chat, etc.), des réseaux sociaux (Facebook, Twitter, etc.) ou encore, par exemple, sous les angles des sphères de la médiologie (DEBRAY, 2000) (graphosphère, vidéosphère,

2. Ce dont on pourrait, par ailleurs, s'inquiéter (BADILLO et N. PÉLISSIER, 2015) remettant en scène les risques liés à la surveillance et le Big Brother de G. ORWELL.

logosphère). Le lecteur notera ici que la variété de ces sphères informationnelles, porteuses de sens par leur dénomination, masque la réalité de leur appréhension dès lors que l'on souhaitera les observer, les identifier, ne serait-ce que les échantillonner. Aucune de ces sphères n'est immédiate, il n'est pas d'instrument capable de délivrer un quelconque corpus représentatif ou encore un échantillon contrôlé.

Plongeons dans la dimension technique pour en extraire la substance nécessaire à une reconstitution utilisable à des fins de recherche en SIC.

1.1.2 Interconnexion de réseaux et d'espaces informationnels

La sphère la plus connue, médiatique (ROUQUETTE, 2010), nous pourrions dire la plus interdépendante de notre réalité est le web. Si l'ouverture de l'internet au grand public a conduit à amalgamer au sens commun, Web et internet, c'est qu'il s'agissait dès lors, du vecteur de popularisation le plus enthousiasmant. C'est ainsi que le premier navigateur « grand public » (Mosaic) par ses richesses graphiques inédites, contribue à partir de 1993, à l'adoption sociétale du Web (SCHAFER et THIERRY, 2013b).

1.1.2.1 L'espace documentaire

Avant d'aborder les sphères informationnelles et les usages, il nous faut rappeler les briques fondamentales. Les premières publications par ses fondateurs le projettent comme une construction technique (BERNERS-LEE, 1989; BERNERS-LEE, CAILLIAU et al., 1992) visant à améliorer l'organisation documentaire. Le schéma 1.1 de la proposition original de BERNERS-LEE (1989) oppose aux systèmes centralisés d'organisation hiérarchique des documents une implémentation décentralisée pour la mise en relation de documents composés (hypermédia).

Le projet conceptuel des auteurs est ainsi de construire un espace informationnel réalisant[3] le rêve de BUSH (1945) par la mise en relation effective des documents telle que la projetait OTLET (1934). Visionnaire, la proposition s'appuie sur les limites de l'orga-

3. Cf. une comparaison de BERNERS-LEE à ce sujet : (1995).

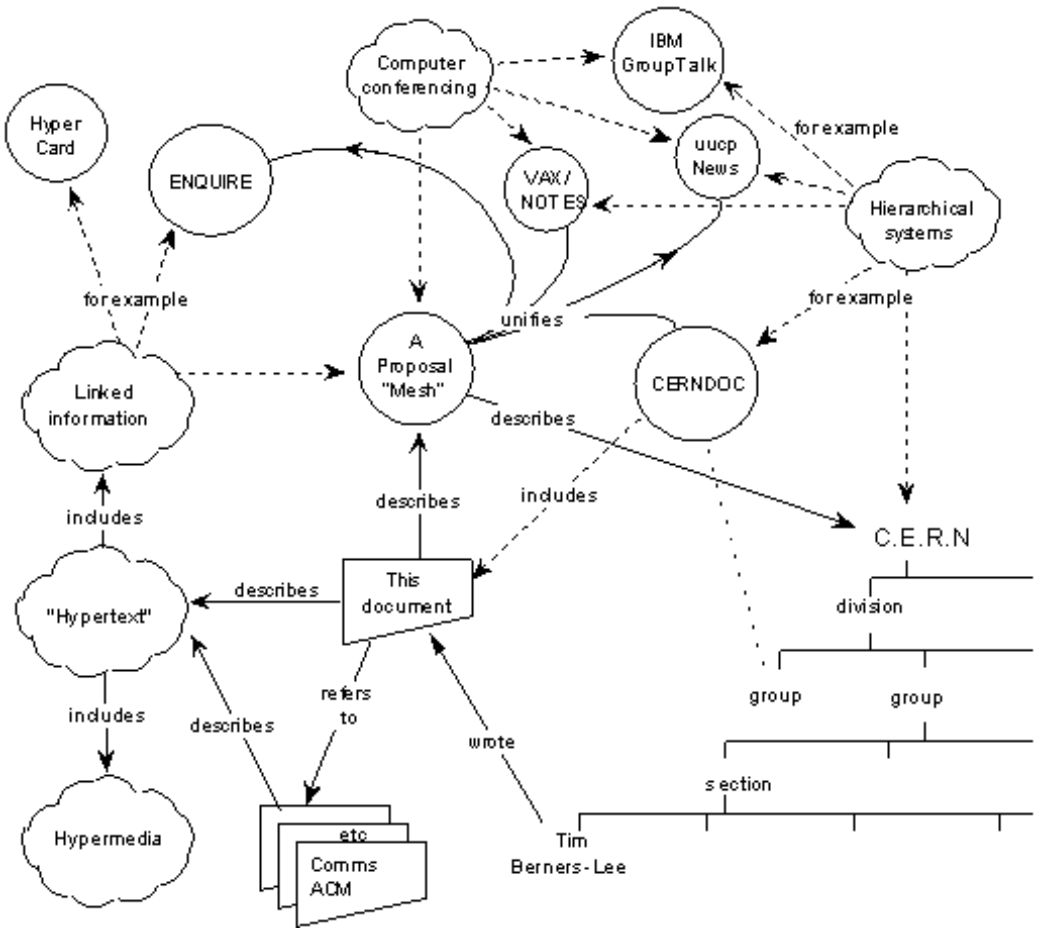


FIGURE 1.1 : Schéma original de BERNERS-LEE montrant l'hypertexte en réponse aux limites de l'organisation de l'information au CERN (1989)

nisation hiérarchique des documents et des mots-clés pour indexer afin de mettre en avant la notion d'hypertexte décentralisé de NELSON (1967). L'espace documentaire du web est depuis fondé techniquement sur un système d'adressage de ressources (URL - *Uniform Resource Locator*), un protocole de communication (HTTP - *HyperText Transport Protocol*), une gamme de métalangages de description des contenus (HTML - *HyperText Markup Language*) dont les balises peuvent être utilisées pour leur présentation. Á l'aide d'un navigateur, l'auteur promeut une série d'applications variées : de documentation (de la création à leur recherche) mais aussi d'inventaire de compétences personnelles supposant une contribution généralisable.

Depuis, ces éléments technologiques ont simplement évolué : pour n'en retenir que l'essentiel, les URL sont devenues des URI[4] (*Uniform Resource Identifier*), puis se sont internationalisés en IRI[5] (*International Resource Identifier*). Le protocole de communication s'est étendu au niveau sécuritaire, et le langage HTML muni d'une grammaire extensible formellement réservée à la description des contenus séparée de la présentation par les feuilles de style CSS. Les navigateurs ont depuis acquis des capacités de traitement dynamique via la structure du document (WOOD, 1999) et, à l'aide du standard Ecmascript (ECMA, 1999) qui régule le langage javascript, ces technologies sont dynamiques et portées sur la majorité des terminaux (ordinateurs, téléphones...). Les serveurs offrent l'accès à des documents mais aussi à des services à l'aide des mêmes technologies.

Interconnecté et ouvert

Appuyé sur une infrastructure ouverte, en perpétuelle extension, le canal de communication s'étend. Constitué d'éléments reliés et décrits, le Web évolue en un espace documentaire s'enrichissant au quotidien. Les formats de données et protocoles de communication sont ouverts et évolutifs.

4. Les ressources sont non seulement localisables mais aussi identifiables par un mécanisme s'appuyant sur les métadonnées descriptives de celles-ci. Principe fondamental du web sémantique qui dépasse le cadre de ce document car résolument dédié à une instrumentation machinique visant à construire un niveau d'informations et de services produit par raisonnement sur les données et services du web (BERNERS-LEE, HENDLER et LASSILA, 2001 ; SHADBOLT, W. HALL et BERNERS-LEE, 2006).

5. Les URLs sont limités à l'utilisation du jeu de caractère ASCII, alors que les IRI utilisent la norme unicode (UTF8).

1.1.2.2 Les usages et imprévus

Si ses premiers pas dans la société l'ont fait paraître obscur et dédié à des érudits ou savants informaticiens, les avancées fonctionnelles et ergonomiques par constructions cumulatrices de couches de haut niveau d'abstraction des briques élémentaires précédentes ont permis de promouvoir lectures ou créations de contenus à la portée des enfants des écoles élémentaires. Les outils de gestion de contenu (CMS), de classement, notation ou annotation dont les fonctions offertes vont de la lecture, écriture (multimédia et hypertexte) au classement et la description des URL. On crée des contenus et des contenus sur les contenus. L'utilisation peut aussi être collaborative, par ex. pour un jugement de valeur sur un contenu, le partage pour offrir ses découvertes ou encore l'annotation des contenus propulse quelquefois les internautes du rang de lecteur ou auteur ou collaborateur, voire à celui de « documentaliste ». La conséquence est en effet déroutante et l'on peut encore s'étonner d'un tel phénomène social comme le rappelait CHAZELLE (2013a) au Collège de France :

> *Des heures à peaufiner des pages web... c'est un phénomène sociologique déroutant. C'est très difficile de prévoir que de tels concepts, formant un cocktail d'altruisme et de narcissisme, puissent marcher. De nos jours, le développement de Facebook c'est assez prévisible, pas très innovant, alors que le web est révolutionnaire, pas au sens technique mais au niveau social.*

Ainsi, la place qu'occupe le web, en tant que dispositif sociotechnique, dans les sociétés du XXIe siècle, n'est plus à démontrer et s'accompagne d'une demande nouvelle de savoirs et de savoir-faire (BARATS, 2013). Facilitée, la production informationnelle enrichit encore et toujours le web ou le pollue. De nombreux travaux font état d'un « territoire du web [...] comme un entrelacement de liens hypertextes à n dimensions[6], maillant réel et virtuel, social et technique, ancrage et flux, superposant des régimes d'espaces, à différentes échelles (individuelles/ collectives) »(PINÈDE, 2014) qu'il faudra peut être tenter de démêler pour les « lire ». Nous reviendrons sur ce fait. Le web dépassant l'internet et vice-versa par l'immersion du réseau dans les sphères sociales (ou réciproquement l'immersion des sphères sociales dans le réseau) qui provoque des changements sollicitant la « renégociation des frontières ». La première frontière

6. L'auteure entend ici les multiples dimensions issues de la sémantique et de la syntaxe de la relation inscrite par le lien hypertexte.

est par évidence celle de notre monde physique dont l'internet devient une extension.

1.2 Bouleversement des frontières

L'ampleur du phénomène est telle que les recherches reconsidèrent la notion de frontière, classiquement notion de notre monde physique et aujourd'hui abordée, par exemple en géographie (SOULAGES, 2013), sous l'angle du numérique et de l'information du domaine en particulier (SZONIECKY, 2013).

1.2.1 Des territoires

La notion de « territoire » s'est répandue lorsque la « montée des réseaux et la mondialisation de la communication semblaient "gommer" les frontières » et doit tenir compte des réseaux physiques ou sociaux (TÉTU, 1992). TÉTU a montré que la notion de territoire, ses frontières sont construites par les moyens de communication qui structurent les échanges et construisent ses représentations anticipant sans commune mesure un bouleversement BOUHAÏ, HACHOUR et SALEH (2014, p. 6) :

> *la reconfiguration des activités humaines sous l'impact du numérique aurait la particularité de provoquer de nouvelles négociations, en tout cas la remise en cause des* statu quo *ataviques limités par des espaces consensuels qu'ils soient territoriaux ou conceptuels.*

Les frontières ne sont pas abolies mais déplacées, « transcendées » au point qu'il convient de les reconsidérer sur de nombreux plans. Réciproquement ce sont ainsi les nouvelles frontières de l'Internet géopolitique, les sociétales (KRASTEVA, 2013), les frontières sociales ou politiques (ROUET et SOULAGES, 2013) que l'on tenter de cerner. On considère même la notion de « citoyen » (KRASTEVA, 2013) comme nouvelle. En parallèle, l'internet ouvre, évidemment, à des usages inattendus (DUFOULON, 2012).

1.2.2 À l'humain

Ainsi, en reprenant l'expression de l'écosystème d'information de DE ROSNAY«, les dispositifs numériques s'inscrivent dans cet écosystème qui se transforme avec les technologies numériques, leurs acteurs et leurs usages qui imposent des transformations tant cognitives, qu'organisationnelles ou culturelles » (BOUHAÏ, HACHOUR et SALEH, 2014) pour les appréhender. Sans aller jusqu'au propos de WEINBERGER (2007) dans (PISANI et PIOTET, 2011, p. 161), qui ne promet rien de moins que :

> *le passage de la connaissance - quête dominante, en Occident, du moins depuis la Renaissance - à la sagesse, quête asiatique millénaire. Une sagesse qui – travers occidental – se trouverait dans les données et les métadonnées.*

Cette écosystème possède (SALEH et al., 2013) enfin

> *[...] une influence considérable sur nos habitus sociaux, nos modalités relationnelles et interactionnelles ainsi que nos activités intellectuelles, cognitives et interprétatives. Ces agencements sociotechniques numériques – au premier chef, les dispositifs hypertextuels et hypermédiatiques – ont donc pour une grande partie reconfiguré nos écosystèmes privés (familiaux, amicaux) et publics ou professionnels et entraîné une réorganisation des usages tout autant que des pratiques.*

La définition de l'humain dans ses rapports actuels entretenus avec les environnements numériques, la nature et le traitement des big data »(BEN AMOR, RENUCCI et ZÉNOUDA, 2013, p. 346) constitue ainsi un **courant de recherche porteur de questions fondamentales des SIC**.

1.2.3 Et ses artefacts

Les développements technologiques ont vu récemment l'avènement d'objets facilitant la connexion du monde physique au réseau internet et vice et versa. Les URI vues précédemment permettent d'adresser les éléments du réseau indépendamment du protocole de transport [7], mais aussi des éléments physiques du monde pas forcément connecté. L'internet des objets est constitué d'une foison de capteurs et automates offrant le signal qu'ils captent ou construisent (et peuvent transformer) à leur lecture [8] au réseau.

7. Cf. la RFC 3986 de janvier 2005 signée BERNERS LEE et al.
8. Une lecture qui n'est pas forcément ouverte à tous.

Pour BENGHOZI et al. (2012) proposent (p. 15) la définition suivante de l'internet des objets (IdO ou en anglais IoT - *Internet of Things*) qui :

> *permet, via des systèmes d'identification électronique normalisés et unifiés, et des dispositifs mobiles sans fil, d'identifier directement et sans ambiguïté des entités numériques et des objets physiques et ainsi de pouvoir récupérer, stocker, transférer et traiter, sans discontinuité entre les mondes physiques et virtuels, les données s'y rattachant.*

Par ces dispositifs, l'infrastructure mémorielle et de transport informationnel précédente sert, en sus de l'organisation documentaire et des échanges informationnels, à la numérisation en temps réel de notre environnement. L'application de cette instrumentation semble sans limite dans la vie quotidienne : depuis l'individu (santé : de l'échelle micro à macro), son habitat (des échelles personnelles autour de bébé, à l'échelle d'un quartier ou d'une ville), aux moyens de transports. Les applications dans le monde économique suivent (ou précèdent) en conséquence.

Un nouveau rapport au monde

L'IoT promet ainsi une foison de capteurs que l'on peut convoquer à la description de situations. Ils constituent une extension de notre capacité de perception : si les signaux sont habilement traités, assemblés et utilisés à des fins explicatives complémentaires les unes des autres.

Le phénomène est appelé la « datafication » (LYCETT, 2013 ; VAN DIJCK, 2014 ; BASTIN et FRANCONY, 2016). MAYER-SCHÖNBERGER et CUKIER, p. 72 réservent cependant cette appellation aux données « tabulaires » pouvant être analysées :

> *To datafy a phenomenon is to put it in a quantified format it can be tabulated and analysed.*

Cette restriction peut s'entendre comme une forme de filtre « utilitaire » : une fonction de sélection de données extraites du « big data ». Je note avec (COLIN, 2015, p. 14-21) que cette discrétisation de phénomène traverse cependant nos vies dans une variété inédite : les sentiments et émotions, les interactions et relations, la parole, mais aussi la culture, la connaissance, la technologie...

1.3 Un écosystème d'informations

L'imbrication virtuel-réel, les mélanges sphères privées-publiques, professionnelles ou amatrices, l'augmentation de l'homme (CLAVERIE, 2010 ; FERRY, 2016) par ces technologies ont des incidences sur notre capacité à échanger par delà le temps, à informer et communiquer, ou encore accéder à un savoir immense interculturel. S'impose de suivre ces évolutions et de s'en saisir : cet écosystème informationnel porte des signes factuels de l'activité humaine, une réplique discrète (au sens mathématique) d'un système complexe au sens de MORIN et LE MOIGNE (1999). C'est cette notion de complexité qui, par construction, se retrouve dans l'écosystème informationnel : la sphère mémorielle est une transcription, certes partielle (seulement une partie de la réalité humaine est enregistrée), et en ce sens un phénomène complexe. Le résultat est, par définition, de nature complexe car la transcription n'insère ni ordre, ni sémantique. Les instruments de la discrétisation tendent cependant à dépasser nos propres sens impliquant un potentiel de traduction supérieur à celui délivré par chacun de nos sens. La difficulté primaire tenant à leur utilisation, la reconstitution d'une réalité, reconstitution qu'il convient d'expliciter.

1.3.1 Descripteurs des données informationnelles

Les données complémentaires qualifiées par les machines sont une aide pouvant s'avérer utile. En effet, l'ensemble des données de l'internet est taxé de métadonnées de catégorisation et ce, de façon automatique plus ou moins implicite : description temporelle (tous les dispositifs serveurs sont connectés aux serveurs de temps universels), description spatiale (avec la notion de lieu dans le virtuel par ex. Facebook, Wikipédia, et le positionnement GPS ou IP de nos appareils), auxquelles se rajoutent les métadonnées spécifiques du transcripteur informationnel (les outils de gestion de contenus en tout genre depuis les instruments d'édition de site aux applications de réseaux sociaux ou Wikipédia).

Encore plus données que celles produites par nos sens, qui sont heureusement aug-

mentées de métadonnées. Ainsi, DE ROSNAY (2007, p. 117) poursuivra :

> *le phénomène Internet nous fait entrer dans un nouveau paradigme : il nous oblige à tenter de comprendre, par la synthèse plutôt que par l'analyse, comment les éléments se combinent dans des ensembles plus complexes qui rétroagissent sur leurs éléments. Cette démarche, qui fonde et légitime toute action consciente au cœur de la dynamique des réseaux, devrait nous rapprocher de la nature et de notre rôle au sein de l'écosystème informationnel Internet dont nous sommes désormais partie intégrante.*

1.3.2 Extension du monde

En conséquence, le réseau et son infrastructure constitue une extension de l'homme, au-delà de la convergence de l'ensemble de nos médias, au-delà d'un gigantesque réservoir documentaire (MORGAN, 1994 ; DROMARD et SERET, 2006 ; OTLET, 1934), une extension de son territoire (BEAUDE, 2008 ; LÉVY, 1997), de son environnement, de son activité, de sa vie. Selon G. BERRY (2008), le Monde « devient numérique » .

Nous retrouvons ainsi dans l'écosystème informationnel, l'infosphère de FLORIDI cité par RICHMOND (2016, p. 218) ou *l'Hominescence* de Serres (2004, p. 18), pour lequel le philosophe suggère d'en construire une épistémologie pour la rendre utile à la compréhension de la société :

> *la communication et ses technologies ouvrirent d'autres voies dans l'espace et l'instant, amenant de nouveaux liens et une expansion inattendue des connaissances. Lorsque des millions de messages deviennent source d'information, la société devient pédagogique en son entier. Reste encore à écrire la nouvelle épistémologie de ce savoir manipulé.*

Sous cette vue, le constat établit me permet de formuler quelques postulats et corollaires suivants utiles par la suite :

1. Le réseau internet, ses utilisateurs et ses diverses branches rallient entre eux une partie de la population humaine devenue non négligeable. En ce sens, il devient le terrain d'étude des activités humaines restreintes au numérique.
2. L'analyse des transcriptions traduira une réalité au mieux discrétisée du continuum réel.
3. La capacité mémorielle du réseau lui confère une dimension de reflet sociétal.

4. L'analyse des phénomènes sociétaux appuyée sur des données factuelles issues de l'internet peut porter sur des phénomènes en cours, ou sur une reconstruction de phénomènes passés.
5. Le réseau est historiquement un support de la connaissance humaine, qu'elle soit sociétale (postulats précédents) ou académiques (colons historiques du web).
6. Le support de la connaissance scientifique de l'Humain peut s'approcher par agrégation de contenus à partir du réseau internet.

Une documentation du monde à reconstruire

Par cet état de faits, le web peut constituer une matière première à des recherches en sciences humaines : transporteur et transcripteur de messages, d'échanges, de manifestes, d'opinions, de communication directe ou pas. L'intérêt scientifique se pose en évidence sur cet immense réservoir de témoins factuels. La complexité inhérente à sa construction nous impose avant tout de lui porter un regard herméneutique de (re)construction documentaire pour opérer à son analyse par des artéfacts propres à l'instrumenter.

1.4 Un incommensurable fouillis ?

1.4.1 Les six principes de l'hypertexte

Prenant le contexte comme l'enjeu de l'acte de communication, Pierre LÉVY modélise celui-ci en s'inspirant des hypertextes informatiques [9] (1991). Application de la notion d'association, l'hypertexte s'identifie comme une matérialisation du savoir commun. Cela inaugure pour LÉVY une « nouvelle géométrie de la communication » :

> *L'objet principal d'une théorie herméneutique de la communication n'est donc ni le message, ni l'émetteur, ni le récepteur, mais l'hypertexte qui est comme la niche écologique, le système toujours mouvant des rapports de sens qu'entretiennent les précédents. Et les opérateurs principaux de cette théorie ne sont ni le codage, ni le décodage, ni la lutte contre le bruit par la redondance, mais ces actions moléculaires d'association et de dissociation qui réalisent la métamorphose perpétuelle du sens.*

9. Précurseur, LÉVY projetait ce que deviendrait le web et le réseau internet.

LÉVY propose six principes abstraits pour « préserver les chances de multiples interprétations du modèle de l'hypertexte » qui se retrouvent dans/par le web actuel.

Métamorphose : Le réseau hypertextuel est sans cesse en construction et en renégociation [...] Son extension, sa composition et son dessin sont un enjeu permanent pour les acteurs concernés, que ceux-ci soient des humains, des mots, des images, des traits d'images ou de contexte, des objets techniques, des composants de ces objets, etc.

Hétérogénéité : Les nœuds et les liens d'un réseau hypertextuel sont hétérogènes. Dans la mémoire on trouvera des images, des sons, des mots, des sensations diverses, des modèles, etc., et les liens seront logiques, affectifs... Dans la communication, les messages seront multi-modaux, multimédia, analogiques, digitaux, etc. Le processus socio-technique mettra en jeu des personnes, des groupes, des artefacts, des forces naturelles de toutes tailles, avec tous les types d'association que l'on peut imaginer entre ces éléments.

Multiplicité et d'emboîtement des échelles : L'hypertexte s'organise sur un mode « fractal », (Multifractal) c'est-à-dire que n'importe quel nœud ou n'importe quel lien, à l'analyse, peut lui-même se révéler composé de tout un réseau, et ainsi de suite, indéfiniment, le long de l'échelle des degrés de précision [...]

Extériorité : Le réseau ne possède pas d'unité organique ni de moteur interne. Sa croissance, et sa diminution, sa composition et sa recomposition permanente dépendent d'un extérieur indéterminé : adjonction de nouveaux éléments, branchements sur d'autres réseaux, excitations des éléments terminaux (capteurs), etc. [...]

Topologie : Dans les hypertextes, tout fonctionne à la proximité, au voisinage. Le cours des phénomènes y est affaire de topologie, de chemins. Il n'y a pas d'espace universel homogène où les forces

de liaison et de déliaison, où les messages pourraient circuler librement. Tout ce qui se déplace doit emprunter le réseau hypertextuel tel qu'il est, ou est obligé de le modifier. Le réseau n'est pas dans l'espace, il est l'espace.

Mobilité des centres : Le réseau n'a pas de centre, ou plutôt, il possède en permanence plusieurs centres [...]

1.4.2 Organisation ou agencements complexes

Le lecteur reconnaîtra aisément dans ces principes (posés en 1991) leur réalisation dans l'internet et le web. L'auteur pose ainsi des préceptes que l'on peut associer à une méthodologie opérationnelle pour aborder l'infosphère :

Autoadaptation : pour considérer la métamorphose perpétuelle le regard construit doit pouvoir s'adapter aux mutations et changements, en reconstituer la dynamique pour délivrer un niveau informationnel subséquent.

Homogénéisation : L'hétérogénéité du réseau est une barrière à son entendement et, en même temps, une source informationnelle multimodale. La mise en relation d'événements factuels reconstituant une scène conduira probablement à une meilleure caractérisation de cette dernière

Distance : Élément central de la topologie, la notion de proximité présuppose de l'existence de distances [10]. Des métriques qui permettent de produire des analyses mesurées des faits, de les rapprocher ou encore d'estimer les biais et erreurs de mesure. La notion de voisinage permet de s'abstraire et de dépasser la contrainte des nombres réels.

Les autres principes impliquent la modestie de toute reconstitution objective. Celle-ci

10. Je prends l'élément central le plus concret de cette notion. Les propos de LÉVY seraient à rapprocher plus génériquement avec les topos d'Alexandre GROTHENDIECK, (GROTHENDIECK et VERDIER, 1972) pour mathématiser l'ensemble de ces principes, pour couvrir leurs aspects empiriques d'une formalisation catégorielle ou ensembliste par exemple.

ne peut être que partielle [11], reconstituer prudemment un éclairage focalisé [12]. Enfin, la généralisation est délicate par la multiplicité et l'emboîtement des échelles [13].

Ces éléments se retrouvent dans la notion de systèmes complexes (MORIN et LE MOIGNE, 1999) qui donne un cadre opérationnel pour aborder de tels phénomènes. Un phénomène perçu complexe, n'est pas réductible à un modèle déterminant la prévision certaine de ses comportements, il se représente par un système complexe (LE MOIGNE, 2003, p. 140). La modélisation systémique fonde son originalité sur sa capacité à respecter cette *dialectique constitutive de toute complexité : devenir en fonctionnant et fonctionner en devenant, en maintenant son identité*. Ainsi, il ne s'agit plus de (ibid., p. 42) :

> *« tout » expliquer de l'objet considéré (avec quelques risques d'échouer dans l'entreprise !) mais, plus modestement, interpréter ce à quoi nous nous intéressons, sans nous assurer de la totalité de cette interprétation.*

1.5 Conclusion préliminaire

Le lien avec la théorie de l'acteur réseau de LATOUR (2006) se pose ici en évidence. L'internet étend notre monde, l'autonomie observée du réseau le rend comparable à un organisme vivant qui associe humains et non humains, associations que l'hypertexte retranscrit, voire mémorise. Posons alors, comme LÉVY, l'hypertexte en métaphore de la Communication, des processus sociotechniques, et de « toutes les sphères de la réalité où les significations sont en jeu ».

Multi-échelle, fractal, attracteur, petit-monde, constituent quelques termes clés du réseau. De nombreux travaux issus de disciplines variées montrent que l'on peut/doit l'observer sous des angles très différents. Aucun instrument ne peut rendre compte sous les angles multiples que constituent les points de vue à poser ou les faits à discerner. LATOUR et LÉVY invitent à produire une herméneutique de cet écosystème, un mode de lecture qui devra se situer à un niveau conceptuellement adaptable par tracé

11. Le principe d'extériorité implique que ce que l'on observe est régi par des causes extérieures et, en conséquence, il ne peut être absolu.
12. Le principe de « mobilité des centres » implique le risque de ne pas observer ce qui se doit.
13. Une caractéristique déterminée qui aurait la propriété d'invariance d'échelle pourrait être généralisée. La détermination de cette invariance d'échelle est un verrou scientifique d'actualité.

des observables, par des (re)constructions synthétiques en appelant à de multiples sens (dont la vision). C'est ainsi que, par essence, cette herméneutique s'appuiera sur une instrumentation apte à extraire des flux infimes de l'écosystème pour rendre possible la lecture de faits. LATOUR et VENTURINI (ibid., p. 255) définissent ainsi des *oligoptiques*, des instruments capables de saisir une réalité (précise) du monde au service d'un infra-langage dont les termes « ne désignent pas ce qui est cartographié mais la façon de cartographier quelque chose à partir d'une nouvelle définition du territoire ».

Oligoptiques

Les postulats précédents, posent l'existence de faits (captables) au sein de l'écosystème informationnel. Celui-ci est d'une telle complexité (au sens de MORIN et LE MOIGNE) qu'il faut utiliser et créer des instruments (eux-mêmes complexes) pour rendre lisibles ces faits (devenus captas). Ces conditions viennent en support des prémisses d'une lisibilité étendue du monde numérique : propre à identifier des morphodynamiques elle en référera à un cadre d'interprétation et s'inscrit dans une dynamique interprétative temporelle.

Les oligoptiques font partie de l'instrumentation constitutive de la documentarisation d'un fait ou d'un ensemble de faits pour réaliser un phénomène observable. En ce sens, ils opèrent en médiateurs des événements (numériques) au service de la production d'informations factuelles, pour desservir la documentation d'une scène.

Ici le lien avec les SIC se pose en évidence. Malgré la rupture à laquelle l'on peut s'attendre par les constats et avancées précédents, je prends la voie d'inscription épistémologique d'en montrer une continuité, le point de convergence des recherches en information et communication [14]. Le raisonnement poursuivi va repositionner la notion d'information de l'infosphère et des SIC pour construire une notion apte à desservir l'élaboration d'oligotiques tant au plan conceptuel qu'opérationnel.

14. Quitte à abandonner l'un des tiraillements probablement le plus fréquent de la discipline : info ou com?

CHAPITRE 2

De l'information à un signe (ou l'inverse ?)

Nous ne trouvons pas convaincante n'importe quelle explication des sciences qui parlent d'inscription, de reliure, de physiographe, d'instrument, de diagrammes ; mais seulement celles qui rattachent ces pratiques au mouvement de mobilisation.
Inversement, nous ne trouvons pas également convaincantes toutes les explications — et Dieu sait s'il y en a — en terme de groupes, d'intérêts, de classes, de cycle économique ; mais seulement celles qui proposent en même temps un mécanisme précis pour que ces groupes, intérêts, classes et cycles soient additionnés quelque part grâce à certaines techniques nouvelles d'inscription. [...]
Pour résumer, il faut que vous inventiez des objets qui soient mobiles, immuables, présentables, lisibles et combinables.

— Bruno LATOUR, Textes fondateurs de la Sociologie de la Traduction (2006)

La mathématique est l'art de donner le même nom à des choses différentes.

— Henri POINCARÉ, Science et méthode, Flammarion (1908)

DE NOMBREUX TRAVAUX s'accordent sur la polysémie du concept « information ». Évoquer cette polysémie est presque un cliché (CASE et GIVEN, 2012, p. 45-76), et souvent, dans la littérature, d'autres termes sont utilisés pour le désigner.

Cette altérité pourrait provenir d'une quête de précision par les auteurs mais, curieusement, ces autres termes empruntés s'avèrent quelquefois, à leur tour, polysémiques. On pourrait y voir une forme de mathématiques, récursive, comme le souligne l'épigraphe de ce chapitre d'Henri POINCARÉ.

L'évolution même du terme « information » et les disciplines qu'il traverse le montrent toujours d'une variété immense et la notion est présente dans de nombreuses sciences (SEGAL, 2003). D'après SAVOLAINEN (2016), SCHRADER ((1986, p. 179)) identifie 134 nuances de caractérisations de l'information en science de l'information seulement. Constats qui conduiront (BURGIN, 2010) à positionner les sciences de l'information comme des méta sciences dans son développement encyclopédique des disciplines de l'information.

Dans les sciences exactes, il s'emploie en veillant à ne pas lui donner de sens particulier : c'est une grandeur mesurable (BATTAIL, 1997) qui permet d'établir la négation de l'entropie (BRILLOUIN, 1959) ou *néguentropie* en s'opposant au chaos. Certains (BASKERVILLE et MYERS, 2002) posent alors la Science de l'Information comme centrale aux autres disciplines (l'auteur oppose le positionnement conventionnel (avant 2000) tel que schématisé par la figure 2.1 à une restructuration qui fait occuper une place centrale tel que schématisé sur la figure 2.2) ou en devenir de l'être (WADE, BIEHL et KIM, 2006) en tant que discipline singulière, unique. C'est ainsi une sur-discipline pour certains (SIDOROVA et al., 2008). Le concept est actuellement au cœur d'une véritable révision scientifique (LELEU-MERVIEL et USEILLE, 2008 ; FURNER, 2015) et l'on peut noter encore un développement récent du concept au plan philosophique (FLORIDI, 2011 ; CAPURRO et HJØRLAND, 2003), développement qui dépasse le cadre de ce document. Le concept de l'information est posé comme un concept théorique (DOUCETTE et al., 2007, p. 198), un super concept qui lui permet d'englober des sous-concepts.

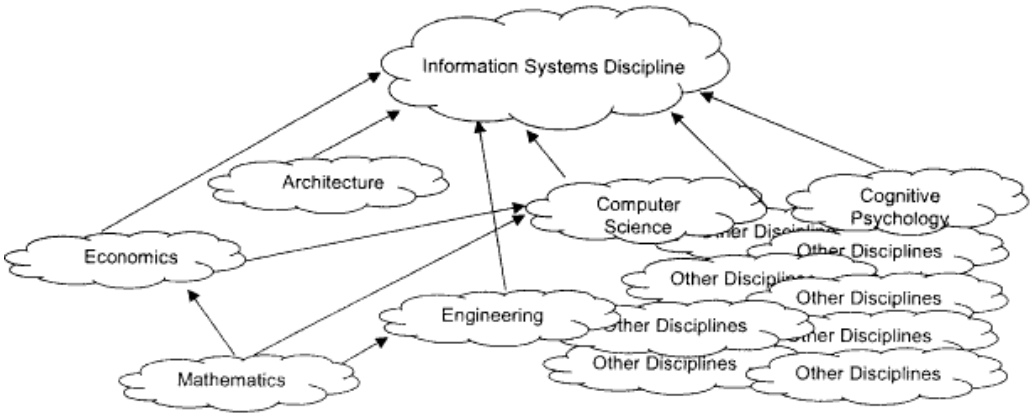


FIGURE 2.1 : Positionnement conventionnel des sciences de l'information d'après BASKERVILLE et MYERS(2002)

Plus localement, l'inscription disciplinaire du concept dans les SIC est claire mais reste toujours tiraillée (IBEKWE-SANJUAN, 2012) : l'emploi du terme « information » est généralement esquivé et s'oppose le plus souvent à la communication.

Par la suite, je puise dans une littérature variée pour m'inspirer de ces courants de pensée et mettre en exergue une caractéristique commune des différentes constructions de « l'information » pour m'abstraire des choix des auteurs afin de déterminer une notion satisfaisante du concept. J'en resterais forcément non exhaustif sur le plan des travaux car je ne souhaite pas établir une vue complète des concepts. Mon objectif n'est pas non plus de prendre parti sur ces questionnements mais de construire une représentation suffisamment flexible qui donnera un cadre pour assembler les données factuelles du chapitre précédent : l'intérêt en est une abstraction des diverses prises de position qui permette de les articuler, les étayer, et d'initier une voie pragmatique de la recherche en sciences de l'information et de la communication.

2.1 Quelle information ? Tour d'horizon

2.1.1 Sans définition

Information est un terme n'ayant pas, à ce jour, fait l'unanimité sur sa signification, c'est un concept hybride à terminologie variée (JUANALS, 2003, p. 16-24). L'information

selon MORIN (1977, p. 310) a une

> *emprise sur toutes choses physiques, biologiques humaines. Elle entend désormais régner de l'entropie à l'entropos, de la matière à l'esprit.*

Sans diminuer son rôle, prenons d'abord sa définition quantitative, débarrassée de toute référence directe subjective et signifiante, celle qui fait d'elle l'instrument privilégié pour comprendre l'ordre thermodynamique du monde de DE ROSNAY (1975, p. 171-172) dans (E. COSINSCHI et M. COSINSCHI, 2009, p. 87). Cette notion est cependant insuffisante pour y associer l'information des SIC. Par exemple, BALTZ (2013) raccroche le social et l'économie : « l'information est la matière première du numérique » puis précise que celle-ci « est toujours le fruit d'élaborations complexes et d'âpres conflits socio-économiques et commerciaux », annonçant clairement une dimension journalistique, sémantique et économique (tout au moins) à cette notion. Il rajoutera, au plan de l'éducation, qu'il convient d'englober autour du numérique (et par conséquent de l'information) toute la complexité de la nature environnante : « les formations au numérique n'ont que plus d'intérêt si l'on y réserve une place de choix à l'information qui porte avec elle toute la complexité de notre environnement quotidien ». BALTZ souligne alors qu'autour de la notion même il est beaucoup plus, et que ce plus porte « la nature environnante »...

Ce qui nous permet de passer d'un composant atomique, privé de sens, à une notion « divine » avec le même terme. IBEKWE-SANJUAN (2012, p. 57) postulera ainsi qu'il est « impossible de dériver une grande théorie de l'information » de par cette polysémie...

En d'autres termes, l'information est complexe (au sens de la systémique cette fois) et se pose en premier rang à la fois dans la dualité matériel/immatériel par transposition du virtuel-réel, mais est aussi conceptuelle en opposant signifiant-signifié.

2.1.2 Sciences de l'Information et de la Communication

Traditionnellement est évoquée « LA » science de l'information et de la communication pour LE COADIC (2004), DESCHAMPS (2010), IBEKWE-SANJUAN (2012). Mais sont

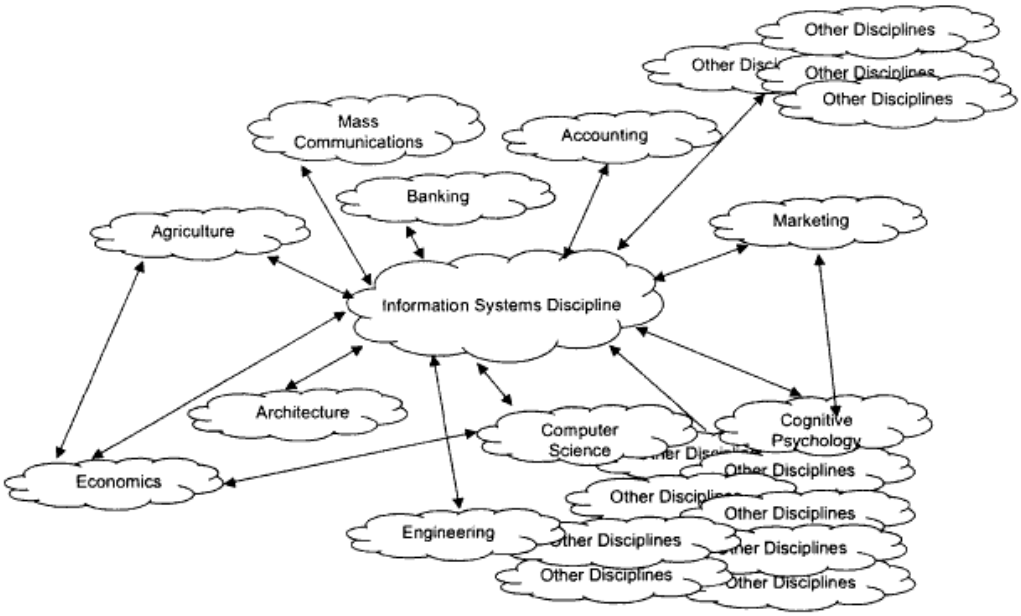


FIGURE 2.2 : Positionnement révisé des sciences de l'information d'après BASKERVILLE et MYERS(2002)

évoquées aussi « LES » sciences de l'information et de la communication pour BENOIT (1995), MUCCHIELLI (2006), OLIVESI (2014), JEANNERET et OLLIVIER (2004)... tout en cherchant à se distancier de l'information neutre, mathématique, cette polysémie recréée est peut-être une conséquence du propre de la cognition (DEANE, 1988 ; JACQUET, VENANT et BERNARD, 2005) dans l'étude des propriétés de l'information ? Cette distanciation est d'autant plus étonnante que SHANNON aura marqué le développement scientifique de nombreuses disciplines historiques, et en créer des nouvelles telles que les SIC en inspirant les fondateurs (VITALIS, 2007 ; PROULX, 2007).

Le modèle trouve ainsi une bonne place dans *la théorie générale de l'information de la communication* d'ESCARPIT (1991) qui spécifie toutefois, aux prémisses des SIC, nombre de réserves et de critiques, exacerbées depuis pour maintes raisons. BALTZ (2007a) souligne aussi l'injustice de cette situation en proposant de ramener ce modèle où il se doit : d'une part au cœur de la transmission d'information et d'autre part, pour établir potentiellement une mesure quantitative objective de celle-ci. Le but original du modèle n'est « que » de réaliser une transmission idéale [1] entre un émetteur et un récepteur pour un message donné, indépendamment de ce que les interlocuteurs en

1. Au sens où le message qui arrive s'approche du mieux qu'il peut de celui qui est parti et ce indépendamment du bruit que peut produire l'environnement ou le canal.

produisent ou peuvent en produire ou pas (BALTZ, 2007b). BALTZ poursuivra (2009) en développant une mesure de l'information fondée sur le nombre de questions que l'on peut poser pour déterminer une information. Cette notion ouvre à la notion de points de vue et à la sémantique des données à contrario de l'information de SHANNON.

La séparation

La position en SIC relève d'une opposition traditionnelle et systématique avec l'informatique en posant un principe de séparation : les premiers s'occupant du sens du message, les seconds de son transport dans le modèle de Shannon. La seconde séparation implicite est celle de l'Information Et de la Communication. Ces deux coupures sont bloquantes pour la prise en compte de l'infosphère pour les recherches en SIC.

2.1.2.1 La limite du seul message

Pour IBEKWE-SANJUAN (IBEKWE-SANJUAN, 2012, p. 44) citant Moles, l'information est une composante de la communication :

> *la théorie de l'information de* Shannon *est devenue à partir des années 1960, l'une des pierres angulaires d'une science plus générale, la science des communications, dont la théorie de l'information au sens strict n'est qu'une petite partie.*

Mais cette inclusion est peut-être simplement un jeu d'écriture comme le note ABITEBOUL (2013) :

> *d'une certaine façon, déterminer l'information mise en jeu dans un objet quelconque, depuis une bactérie jusqu'à un phénomène comme le cours des actions, ou le mouvement des planètes, est une étape essentielle pour comprendre cet objet. Mais cela tient d'autres sciences que l'informatique.*

La détermination de l'information dépend, pour lui, d'autres sciences (ce qui ne nuit pas à son utilisation).

Ainsi, conduire l'étude de la communication par l'explicitation des informations en mise en jeu, en se limitant au contenu seul des échanges est d'évidence limité car relève d'un niveau descriptif trop particulier pour la considération de certains modèles de la communication non mécanistes. En d'autres termes, l'étude du message seul se relève d'un plan insuffisant pour répondre à la pragmatique d'une situation (DE AGUILERA, 2006). MOLES propose d'établir la qualité d'un message reçu dont la valeur s'éta-

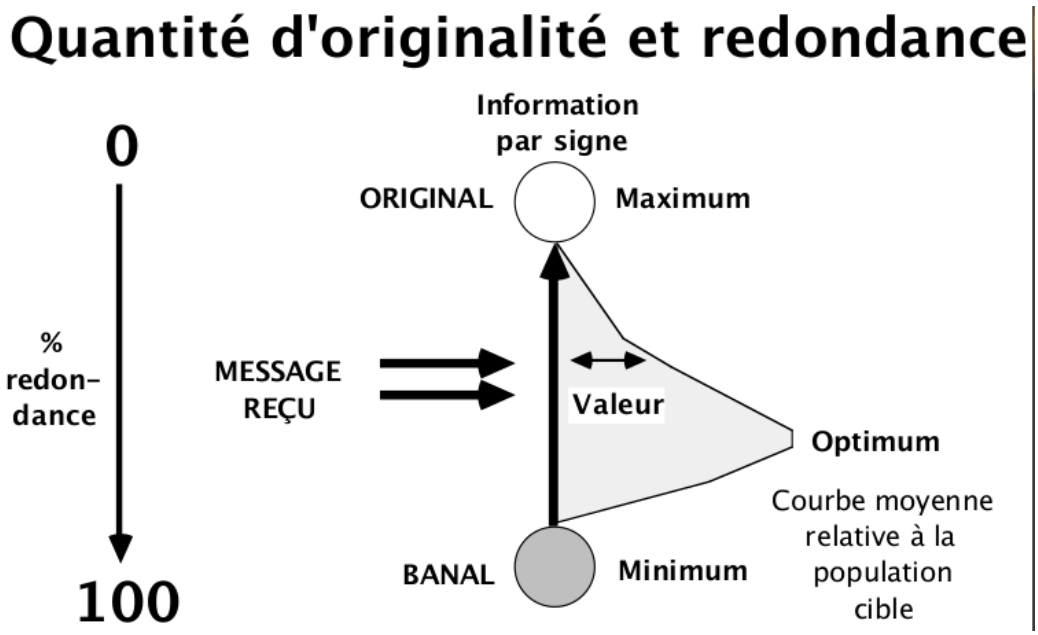


FIGURE 2.3 : L'originalité d'un message d'après MOLES (1972, p. 93)

blit entre la banalité (redondance totale) et l'originalité sur un optimum relatif à la population cible, par son schéma 2.3. Son modèle, bien que ne relevant pas de toutes les situations, introduit à l'issue du modèle de SHANNON (le message est reçu), la Communication pour laquelle l'efficacité se mesure via l'information « originale » pour les récepteurs (MOLES, 1972, p. 299) :

> *la mesure de l'information ne tient pas plus dans le nombre de symboles transportés que dans le retentissement efficace de ces symboles, mais dans l'originalité du groupement de ceux-ci opposée à la banalité du prévisible.*

2.1.2.2 De la Communication aux signes de la communication

Pour commencer, les réflexions du domaine font remarquer que circonscrire la définition de l'information par construction à partir de son dual que pourrait être la communication est risqué. Le postulat de JEANNERET (2002) en témoigne : « la communication a affaire avec elle-même, avec les relations qu'elle établit et le jeu qu'elle institue, avant de concerner quelque information que ce soit ». Pourquoi pas mais dès lors qu'il s'agit de la penser, d'en parler ou de l'écrire, il y a transcription de ces relations et jeu et en ce sens, sans doute, production d'information. Ne serait-on pas sur le paradoxe de la poule et de l'œuf à séparer ainsi? Ou alors, l'auteur suppose, avec nous, que la réflexion sur la communication doit porter sur quelque chose à nouveau, qui est exté-

rieur au message (QUÉRÉ, 2000). Voyons, selon le même procédé que précédemment, quelques éléments et définitions du domaine :

Pour MAIGRET (2015, p. 16) :

> *La communication n'est pas tant un donné (celui de la nature) ni un flux de données (celui de l'information, au sens mathématique) qu'un rapport permanent de sens et de pouvoir dont les cristallisations sont les contenus et les formes des médias. Quelque chose de complémentaire, aussi complexe soit-il à un message dénué de sens en général.*

Cette assertion est implicite dans le modèle de WATZLAWICK, dont l'axiomatique définit dans l'acte de la communication le transport d'un message qui impose un comportement.

L'école de Palo-Alto, a proposé le concept de métacommunication qui décrit le discours sur la communication pour combler ce manque. Selon ce courant, la communication se décrit par une transmission d'« indices » associée à une transmission d'« ordres ». Ces derniers relevant de la méta-information puisque définis comme des « informations sur une information ». L'on retrouve cette distinction (PICARD, 1992) de façon plus ou moins implicite [2] dans les modèles de SHANNON, à celui de RIDLEY et RIDLEY ou encore de JACOBSON dont la notion de contexte intervient dans le modèle linguistique ou encore dans les modèles incluant les conditions sociales comme variable de l'interprétation. Ainsi, pour WATZLAWICK et al. (1972, p. 52) « toute communication présente deux aspects : le contenu et la relation, tels que le second englobe le premier et par suite est une métacommunication ».

Dominique WOLTON inscrit (WOLTON, 2012, p. 12) la science de la communication sur le second aspect : « tout se complique quand on passe de l'information, le message, à la communication, la relation » que l'auteur situe autour de trois sens : le premier vise au partage et à la volonté de mieux se comprendre, le second à transmettre et le dernier à négocier.

2. SHANNON s'exempte dès l'origine de toute interprétation du message dans son modèle des télécommunications et se défend clairement de distinguer la nature du message. Cf. section 2.1.2 p. 46

2.1.2.3 Composite, morphodynamique, variable

Des extraits de modèles précédents, apparaissent les bribes d'une essence complexe d'une imbrication récursive que d'autres ont déjà notée (MIEGE, 2004) et associée au paradigme de la complexité de MORIN telle « l'approche communicationnelle » de MUCCHIELLI (2000). Ce dernier pose une réflexion sur les fondements scientifiques des sciences de l'information et de la communication et montre comment le phénomène communicationnel et la problématique dans laquelle on l'insère sont dépendants d'un positionnement épistémologique, du choix d'une théorie de référence et de ses concepts, ainsi que d'une méthodologie d'étude.

Ceci en conséquence, abonde à nouveau sur la plasticité des concepts convoqués, des concepts à qui l'auteur/chercheur donnera un sens partagé seulement si cadre, contexte, et procédé(s) en œuvre utilisés pour les définir sont exposés. En d'autres termes, la communication est composée de signes (pour ne pas dire informations) in-sensés pris séparément et dont l'assemblage (juxtaposition, agencement, mise en relation, procédés...) fabrique le message, permet son décodage et achève la communication par le partage et la négociation bilatérale de sens. On retrouve une nécessaire complémentarité entre des signes privés de sens accompagnés d'un ensemble support à leur interprétation.

L'information

Ainsi, que l'on étudie la communication, les processus d'information, ou l'information elle-même, tout reviendrait, pour le chercheur, à établir des constituants informationnels élémentaires qui exposent, expliquent et argumentent les processus en jeu.

2.1.3 Tentative : le terme *information* a-t-il un sens ?

Convenons que proposer une définition de la notion d'information revient à lui donner un sens. Au plan Saussurien du terme, cela revient à élaborer et formaliser le contexte de l'interprétation de ce terme qui est lui-même sujet à un « horizon de pertinence focalisé » (LELEU-MERVIEL, 2010a)... à partager. Élaborer les conditions d'une interpré-

tation unique de ce concept composite ne se fait donc qu'à travers un contexte pour un des sous concepts au détriment des autres. Ainsi, prétendre une théorie globale de l'information ne peut conduire qu'à des paradoxes entre les contextes. Pour s'en convaincre, il suffit de prendre une définition d'une discipline de cette notion comme point de départ puis les controverses et querelles scientifiques qui s'en sont suivies, le plus souvent au sein même de la discipline[3].

Un autre voie est celle du théorème d'incomplétude de GÖDEL. Donner sens conduit à axiomatiser dans le langage, définir des postulats dans un espace dont les règles permettent d'inférer, déduire, conclure. Le langage nous autorisant à élaborer ce type de construction au plan que l'on souhaite (le contenu des échanges entre les bactéries, le contenu d'un journal, le contenu d'un texte, le contenu d'un image, etc.) chacune des axiomatiques prises avec une définition trop générale du concept entrera en conflit avec l'une ou l'autre des axiomatiques. GÖDEL, sur la base du langage formel, nous donnera même le caractère indémontrable de cette définition dans le langage humain courant. Ceci permet d'interdire formellement cette universalité du concept information (GIRARD, 1971). Ainsi, le concept ne peut se satisfaire d'une définition générale.

2.1.4 Passons aux textes : de l'information ?

Il est commun dans notre section de prendre le texte comme source d'information. Pour JEANNERET (2001), le texte est un objet :

> *matériel, singulier, complexe, hétérogène ; cet objet repose sur une union intime entre le support et le message ; il repose sur des codes stricts et d'autres plus flous en matière d'assemblage de signes (le texte alphabétique étant un cas particulier) ; il peut être doté de sens par la confrontation à des modèles acquis ; il propose des marques pour une relation énonciative (implication de communication) et des représentations du monde ; tout en définissant ses propres frontières, le texte est ouvert, car il entre en relation, explicite ou non, avec d'autres textes.*

Par sa définition, l'auteur confère au texte une puissance telle qu'à lui seul il peut fournir de nombreuses informations différentes que seule une analyse permettra de dis-

3. Sans doute car elle est le lieu même de la communication et des échanges. Par essence même, pas de querelles si l'on ne partage pas directement ou indirectement des points de vues.

cerner que ce soit directement ou en utilisant d'autres sources pour faire sens (modèles, codes, relation par détection des marques, représentation et enfin les relations). Le texte peut se « découper » sur de nombreux plans. L'on peut distinguer ainsi pour les études des textes, l'hypertexte (BALPE, LELU et PAPY, 1996), le paratexte (G. VARET et M.-M. VARET, 1995, p. 640), le metatexte ou encore le code, notion (en littérature (DUBOIS, 1973)) qui « apparait comme la mieux à même de désigner les contraintes et normes qui régissent le fonctionnement textuel ». Pour DUBOIS le code renvoie cependant à un « carrefour de perspectives bien plus qu'à une définition ferme », à l'inverse du code informatique traduisant le plus souvent un algorithme dans un langage nécessairement rigoureux pour une interprétation sans ambiguïté pour les machines. Dans la même lignée, MOLES va introduire dans la théorie de l'esthétique informationnelle le concept de « supersigne » qui va permettre de prendre en compte les éléments extra-linguistiques dans la modélisation du texte (BOOTZ, 2001, p. 28).

L'on retrouvera cette fragmentation encore dans les études de la production de sens à partir des élémentaires (signes, textes ou données), qui se fait encore selon les cognitivistes (LELEU-MERVIEL, 2003) en hiérarchisant en niveaux de représentations de données, le niveau humain en constituant (à ce jour), le degré ultime : « le récepteur doit être capable d'inférer, à partir du message littéralement transmis et des circonstances de l'énonciation, l'intention avec laquelle le locuteur produit ce message dans le contexte particulier où il est émis ». On retrouve ainsi, en complément du message littéral : circonstance, message, intention et contexte pour compléter l'« information » totale associée au processus de communication.

En résumé

Ce tour d'horizon a montré que l'on ne peut se dispenser d'une définition de la notion d'information, que le caractère polyscopique et polysémique impose quelques prudences quant à sa formalisation. En SIC il est fréquent d'une part de se démarquer des études sur « l'information » et d'autre part, pour les chercheurs d'employer d'autres termes relevant d'une sphère sémantique proche. Parcourons quelques-uns de ces termes pour en montrer la « quasi équivalence » car relevant de notions construites pour enfin entrer dans le concept de données (ou obtenues), terme (trop) connoté aux sciences de l'informatique.

2.2 Décomposition cyclique

S

La notion d'information se décompose en « sous-concepts » primaires, élémentaires. Voyons comment d'autres concepts, fréquents objets de recherche des SIC, partagent cette propriété. En les parcourant de façon non exhaustive, m'en pardonnent les lecteurs, j'userai de simples exemples de nature différentes qui suffiront pour étayer mon propos. Il me fallait un point d'entrée, j'ai choisi la notion de trace.

2.2.1 La trace, l'empreinte, le signal

Daniel PÉLISSIER (2015) offre un historique de la notion de trace, posant Alexandre SERRES (2002) en précurseur. Ce dernier la caractérise au sein des SIC comme une mémoire, une empreinte, un indice et un écrit. PÉLISSIER dénote en sus de cette polysémie une notion complexe. SERRES soulignait déjà en 2002 les enjeux que cette notion de trace porte, deux aspects indissociables : **un caractère numérique** puis, en conséquent, **des outils de traitement associés**. Alain MILLE (2013) va distinguer de son côté la trace de l'empreinte par un **nécessaire processus cognitif**. Il propose, en suivant une définition opérationnelle (rattachée aux usagers) : « la trace est **constituée à partir d'empreintes** numériques laissées volontairement [4] dans l'environnement informatique à l'occasion de processus informatiques ». Dans la même lignée, MERZEAU [5] (2013) les considère aussi ainsi, sous l'angle de « personnelles », et leur confère une « dimension informationnelle » chère au yeux des géants du Web impliquant un risque (2011, p. 92) pour les internautes et l'individu en général [6]. La définition de MERZEAU est encore une définition opérationnelle de la notion de trace, qui, pour l'auteure, trouve son utilité dans le cadre des recherches sur les données personnelles et les dépassent. La trace est ainsi une information **reconstruite** dans sa définition à l'aide de données numériques captées çà et là.

4. Rajouter « ou non » ne devrait pas nuire à la définition et l'élargit à la plupart des activités sur le Web.

5. Mme le Pr. MERZEAU s'est éteinte subitement le 15 juillet 2017. Bien que non central dans mes recherches, son apport conséquent est très présent dans ce document, un modeste et involontaire (au départ lors de la rédaction.) hommage, témoignage à ses travaux.

6. Ce qui en montre tout au moins la puissance potentielle révélatrice tout au moins du comportement des individus.

FLON et al. (2009) propose une autre conception de la trace comme constituée de données :

> *Les données ainsi constituées, qui peuvent être de diverses natures (logs, fichiers structurés XML, capture d'écran, enregistrement audio ou vidéos, etc.) nécessitent ensuite d'être analysées et interprétées à partir d'informations externes sans lesquelles il est difficile de leur donner un sens.*

DAVALLON, NOEL-CADET et BROCHU (2003) prendront une position un peu différente, restreinte aussi au domaine du web, celle de la trace comme une notion qui considère la relation entre les sites web pour contextualiser les activités des usagers, la relation hypertexte est la base de la trace des usagers et des usages [7]. Poursuivons.

JEANNERET (2011) dénote en plus la complexité de la notion de trace. Il souligne les confusions potentielles entre signe, indice et trace et que la trace se décompose pour lui en « inscriptions ». Son raisonnement rappelle notamment que la transformation par MERZEAU (2011, p. 92) de l'empreinte en indice puis en trace est une « construction de l'esprit ».

Cette construction, qu'elle soit réfléchie (et/ou), calculée (et/ou), inférée ou déduite ou interprétée, imaginée (et/ou) projetée, peu importe, est **une transformation**... mais alors de quoi ? Cet élément ne serait-il pas en soi un constituant de l'information elle-même ?

De Ω à la trace

Je note dans ces travaux, que face à la complexité de la notion chacun a recours d'une part à une contextualisation un cadre, puis définissent le concept comme résultant d'une transformation de « sous-concepts » ou de composants atomiques par des opérations particulières. Les sous-concepts ne sont pas stabilisés ce sont tantôt des empreintes, tantôt des signes, tantôt des logs, des données. Notons les Ω.

7. La relation existante entre les sites est plus de l'ordre du « chemin » que de la trace, de notre point de vue et, au plan opérationnel, la reconstruction des traces laissées par les usagers sur l'ensemble des chemins du web est à ce jour impossible ou ne peut être que partielle, en partant des dispositifs des individus. Les outils tels que Alexa permettent de reconstituer ces traces sur des panels anonymisés, issus d'usagers qui ont installé les outils Alexa sur leur navigateur. Cf. http://www.alexa.com/

Ainsi, JEANNERET (2011) déclame :

> *Mettre la main sur la trace, ce serait à la fois tenir un modèle performant de la communication et pouvoir manipuler le réel.*

En réintégrant l'Hominescence de l'écosystème informationnel, ces traces que BOULLIER nomme « répliques » pour les « désigner » sont constitutives d'évènements. Bien que les notions précédentes de trace se rattachent déjà au domaine de la communication quelquefois, voyons plus particulièrement un mode de décomposition en constituants élémentaires : les signes.

2.2.2 Des signes de communication à la « donnée »

Dans la tradition de PIERCE, par extension de tout ce qui se pense, tout communique avec tout, et tout ce qui communique le fait avec des **Signes**. Pour les sémioticiens, le **signe**, à son tour, est polysémique et polyscopique. Il s'étudie en langage naturel, et est une prémisse « démonstrative, nécessaire ou probable » chez ARISTOTE. En posant la question de la modalité de son étude VERHAEGEN (2010), propose la construction d'un regard sémiotique qui sera élaboré, réfléchi, argumenté et circonstancié. Au-delà de la sémiotique, ce procédé d'appréhension se retrouve aussi dans la linguistique et la sémantique en général pour qui le Signe est un constituant élémentaire. La multiplicité des « substances » de HJELMSLEV lorsqu'il étend le signe/texte en objet de la linguistique (BADIR, 2005) en sont une preuve. Ainsi le signe devient aussi « concept » (KYHENG, 2005 ; BONDÍ, 2008). En conséquence, la notion de signe est prise (RASTIER, 2005) tant comme **objet matériel** (par ex. celui de l'écrit) et **objet logique** (celui du formalisme des calculs). Et l'on retrouve aussi cette distinction dans les travaux fondateurs sur le concept de l'information chez CAPURRO et HJØRLAND (2003). RASTIER dénote la présence de cette notion dans les travaux de HILBERT qui la délivre par définition de tout sens :

> *en mathématiques [...] l'objet de notre examen ce sont les signes concrets eux-mêmes dont la forme nous apparaît immédiatement avec évidence.*

Pour les littéraires, qui s'adonnent à des études de communication particulières, le

« code » est introduit pour discuter et considérer l'information textuelle. Pour l'esthétique informationnelle de MOLES (MATHIEN, 2003) le « supersigne » est défini. Ainsi, à nouveau, les sciences de l'information et de la communication ont besoin de considérer plus que l'information-message. Ceci se manifeste dans les travaux non seulement par une variété lexicale très ample accompagnée de qualificatifs variés sur les états des messages. L'ensemble produit des nuances différentes de l'information, nuances nécessaires pour construire et exposer un regard particulier sur l'information ou la communication. Ainsi, la plasticité de la notion informationnelle est transportée (ou bloquée selon le point de vue) d'une discipline à une autre, d'un regard à l'autre, d'une perspective à une autre par une contextualisation complémentaire au signe élémentaire. Le choix du terme signe, à nouveau arbitraire, n'est pas sans risque non plus comme esquissé précédemment. Apparaît ainsi, dans les divers positionnements des diverses disciplines, que le concept requiert l'existence d'entités atomiques privées de sens (dorénavant les signes élémentaires ou diaphories) associés à une opération herméneutique que l'on retrouve dans la fonction sémiotique de HJELMSLEY (BONDÍ, 2008). Cette construction lui confère ainsi une flexibilité adaptable aux différents points de vue qu'il conviendra de cerner plus précisément.

On peut voir deux caractéristiques communes de ces travaux : d'abord celle qui établit une précision sémantique du concept « information » employé pour cadrer les analyses et développements par des « sous-concepts » puis, celle qui reste simplement focalisée sur une dimension pratique et opératoire pour mesurer et établir des faits.

Fonction de composants atomiques

- trace, inscription, empreinte, témoin, ou signes[a], sont tous des composants atomiques **en contexte**, des Ω . Ces composants, dans les propos des auteurs, rappellent les « données » primaires, brutes des informaticiens, comme celles qui alimentent le *Big Data*, des faits ou des évènements,
- chacune de ces entités résulte d'une **opération de transformation** d'une entité primaire...

a. Hors sens particulier de l'écrit (sauf à prendre les lettres isolément) car celui-ci dispose à son tour d'une telle profondeur en pouvant, in fine, les englober.

Notons aussi la nature discrète de ces composants atomiques, séparés d'une dynamique temporelle que l'on pourrait associer à des flux (flux de visites sur un site, fil des contributions, météo, température, sourires des auditeurs, etc.). Certains de ces flux se manifestent à différentes fréquences. Rares sont les considérations de cette dimension temporelle qui confèrerait cependant à ces entités informationnelles une forte ressemblance avec la notion de signal.

Ce sont ces traits caractéristiques que l'on retrouve dans la définition de la « donnée » reprise ci après de LELEU-MERVIEL (1996, p. 115). L'auteure rajoutera des éléments caractérisant le contenu élémentaire, des critères assurant sa véracité, son authenticité en condition préalable au faisant-sens :

> *une donnée est en principe un fait objectif, le plus souvent quantifiable, encore faut-il préciser pour un groupe social qui a établi un consensus sur les attributs choisis, sur la méthode de mesure, sur le code utilisé, et qui a confiance en l'honnêteté du processus ou qui peut le contrôler. Hors de ces conditions, une donnée n'est pas une « donnée », mais peut être suspectée de n'être ni neutre, ni objective.*

Pas étonnant que pour l'analyse de leur signification, GALINON-MÉLÉNEC (2011) propose aussi la prise en compte du contexte de leur production et de leur interprétation.

Les S.I-C

Par extension (tout en restant interne à la discipline des SIC) ce sont les conjonctions objective/subjective, concret/abstrait, formel/informel qui permettent d'opposer[a] l'Information de la Communication. Ce qui me permet de souligner que **les contenus faisant sens pour les uns peuvent devenir les données pour les autres...**

a. Entre autres formes d'exclusions mutuelles ou de prédominance de l'un ou de l'autre, peu importe.

2.2.3 Le sens de la donnée

Dans la conception cognitiviste, pour conceptualiser l'information il est nécessaire de construire un regard : « une hypothèse sur la construction de schèmes de compréhension signifiants, structurants et organisants, qu'élabore l'intelligence à partir d'aspects qualifiants discrets reliés par des liens. La mise en *liction* des différents aspects

permet de combiner une représentation élaborée porteuse d'indices de compréhension novateurs (LELEU-MERVIEL et USEILLE, 2008, p. 30). Les auteurs poursuivront : « cette mise en relation de différents aspects, des propriétés qualifiantes que des unités informationnelles partagent (ou entrent en résonance) permettent d'aboutir à leur conceptualisation ».

En conséquent, LELEU-MERVIEL (2010b, p. 18-19) propose la définition suivante :

> *L'information traduit un point de vue singulier déterminé notamment par des savoirs, des expériences antérieures, une culture, une perspective propre gouvernée par un questionnement qui révèle et rend pertinentes certaines diaphories et pas d'autres.*

En conséquence, on ne peut accéder à ce point de vue singulier à partir des diaphories primaires sans connaissances ou l'explicitation de tout ou partie des éléments constitutifs des points de vue.

Les travaux de LELEU-MERVIEL (ibid.) instruisent la précédente déconstruction du concept information. L'auteure s'appuyant sur BARTHES, montre que tout est trace (2013) et, bien qu'en dehors du texte et des données numériques (qui sont plutôt mon point d'entrée), la conclusion est faite que « ce n'est plus la marque résiduelle qui caractérise la trace, mais la traque du processus producteur à travers les rétentions diaphoriques spectrales qu'il a engendrées ». L'on retrouve ainsi à la fois la notion de différence qui donne sens et aussi le concept dual souligné précédemment : celui d'une unité atomique privée de sens, et un processus associé. Processus qui s'identifie au travers du schéma 2.4 de l'auteure (2010b). Je confèrerai à cette représentation une dimension supplémentaire d'auto référentielle, propriété de récursivité esquissée précédemment, (la « boucle » de LÉVY (2015), traditionnelle en informatique, mais qui réside ici dans l'espace abstrait de la relation entre « informations »). En effet, quel que soit le niveau informationnel (ceux esquissés précédemment par exemple tels information, signe, trace, texte...) *Noumène* et *Pattern* sont différents pour chacun de ces niveaux. Par exemple, le « noumène » de la sonde thermométrique pour l'électronicien est différent de celui de l'informaticien et encore différent de l'utilisateur pour un usage identique de lecture de surcroît.

Ainsi, l'information est aussi associée à une opération de transduction ou de « tra-

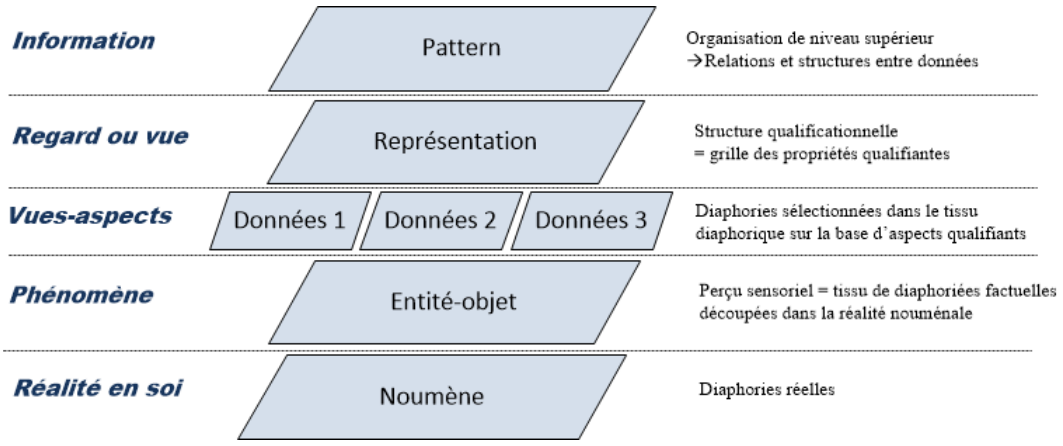


FIGURE 2.4 : Du noumène à l'information via la conceptualisation : synthèse (d'après Leleu-Merviel, 2010)

duction » qui, dès le niveau des données sélectionnées comme aspects qualifiants par le prélèvement sur le tissu diaphorique complexe découpé dans la réalité nouménale, laisse « diverger les choix et les représentations d'un individu à l'autre ».

En citant NEUMANN (2006), LELEU-MERVIEL (2010b) propose la définition suivante de cette action de communication séparant encore contenu et processus :

> *le faire-sens comme un processus actif qui extrait des éléments opérationnels convoyés par le message, tandis que l'information est la valeur que l'interprétant vient assigner à un signal* a priori *indéterminé*

Ainsi, le faire sens se lie (ou se lit) dans le schéma 2.4 par la relation (non représentée encore, cf. p. 62 pour une instanciation restrictive du concept) entre les différents niveaux, cette relation qui construit leur passage au niveau supérieur et qui pourrait se décrire comme une série d'opérations ou algorithme. Ce que nous prendrons comme postulat :

Le faire-sens s'élabore à partir de constructions qui peuvent être des algorithmes.

2.3 Herméneutique de la mise en relation ou algorithmes ?

En introduction de la leçon inaugurale au Collège de France CHAZELLE exclamait : « Oui, depuis plus d'un milliard d'années dans la vie il y a des algorithmes » (SCHA-

FER et THIERRY, 2013a). Les êtres vivants, leurs cellules élaborent et appliquent des algorithmes. N'en déplaise à chacun, nous opérons tous des algorithmes. Un algorithme peut par exemple décrire une méthode de construction de corpus, d'enquêtes, un ensemble de questions visant à évaluer un produit, ou encore un raisonnement un tant soit peu construit. Bien évidemment, on ne peut tout formaliser, ce n'est pas cette question qui est abordée ici, il s'agit de considérer ce qui peut l'être et ce à différents degrés. Sans entrer dans la réflexion sur la cognition, nombres d'activités du chercheur sont des procédures établies, des séries d'opérations qu'il aura à expliciter pour les rendre reproductibles. Que ces opérations soient issues d'une intuition, qu'elles soient déterministes ou pas, il s'agit de constructions qui s'apparentent (car pas forcément formelles) à des algorithmes.

Dans notre cadre, les diverses opérations du faire sens précédemment esquissées s'expriment par des algorithmes et participent de la fonction de médiation des données de masses que l'on ne peut appréhender sans instrument. Leur variété fait que l'instrument est unique et le plus souvent doit être construit « à la demande ».

2.3.1 Médiation de la donnée

Pierre LÉVY pousse le raisonnement plus loin (LEVY, 2014) en situant les algorithmes dans une sphère sémantique au dessus des éléments informationnels primaires vus précédemment (cf. figure 2.5 en p. 62). Le schéma 2.4 en p. 60 établit un opérateur entre le niveau *World Wide Web* et la sphère sémantique suggérant la mise en place d'un niveau cognitif à cette étape n'opérant non plus sur les données elles-mêmes mais sur les algorithmes.

La figure 2.6 reprend le processus de conceptualisation de LELEU-MERVIEL et représente des fonctions de passage entre les différentes couches constitutives. La première (en partant du bas) est celle du filtrage (tel la discrétisation opérée par exemple par le microphone en transformant le signal analogique en signal numérique, mais ce peut être aussi le filtrage, qu'il soit non intentionnel ou bien réfléchi, qui exclut l'inutile, le procédé de sélection de l'information de Moles. La seconde fonction, de sélection et d'agrégation, reporte le processus d'élimination de la quantité informationnelle (un

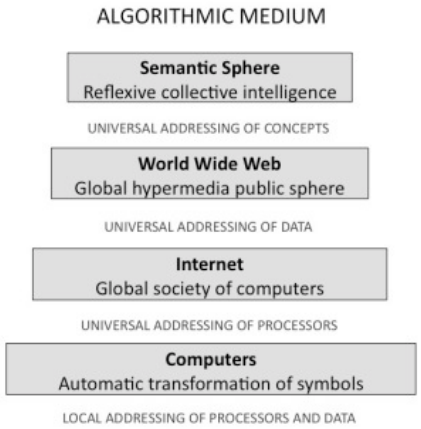


FIGURE 2.5 : Le médium algorithmique de Pierre LÉVY

choix réfléchi comme par exemple l'extraction dans les bases de données par une requête) pour produire de nouvelles données (on retrouve sur ce plan les « obtenues » de LATOUR). Le troisième niveau consiste en l'augmentation de données qui revient tout au moins, à compléter, situer ou complémenter les données des niveaux précédents afin d'élaborer un point de vue. Enfin, le dernier par la visualisation (au sens du Voir) conduit enfin à l'information, le *pattern* emprunté à BATES.

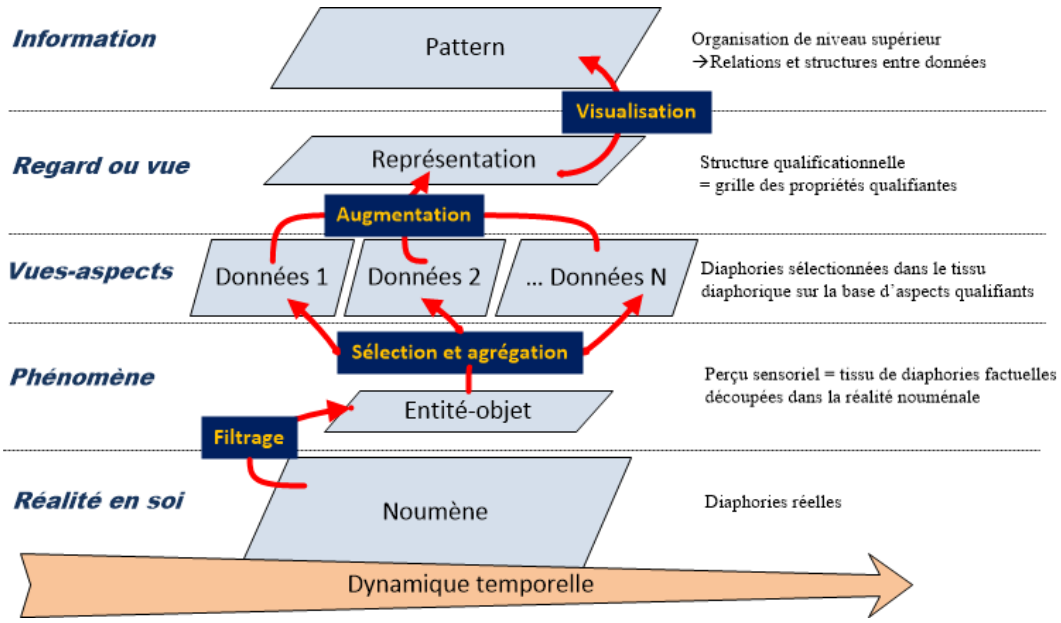


FIGURE 2.6 : Herméneutique de la donnée

2.3.2 Inscription épistémologique

La considération de la donnée constitutive d'éléments informationnels élémentaires revient au calculable et à l'opérationnel, le fondement d'une science qui situe clairement mon positionnement dans la famille des épistémologies positivistes à laquelle de nombreux noms (GELLNER, 1984, p. 611) ont contribué : depuis la méthode cartésienne, Auguste COMTE et son discours sur l'esprit positif, puis les étapes religieuses, métaphysiques et positives, CONDORCET, la méthodologie hypothético-déductive de Claude BERNARD, le positivisme logique de HILBERT, J. MONOD (naturalisme désordonné et matérialisme ordonné), K. POPPER et le réalisme de B. D'ESPAGNAT, B. RUSSELL, T. KUHN, R. THOM avec la querelle du déterminisme, ou encore RUELLE avec le chaos déterministe.

L'extension de données factuelles mémorisées ou transportées par l'univers numérique que nous alimentons, qui nous accompagne, situe la position de l'Hominescence de SERRES au plan d'informations élémentaires, **des faits numériques**.

Herméneutique du fait numérique

Capter ces données, les assembler, les augmenter, les transformer, les explorer, développe une nécessaire herméneutique élaborée. L'Hominescence ouvre un terrain immense de données des activités humaines et au-delà de la constitution contrôlée de corpus, de données, de traces ou signes qu'elle autorise, de la réduction de celles-ci pour se rendre capables de leur entendement, il nous faut prévoir une certaine plasticité de la fonction herméneutique mais aussi l'expliciter. Cette explicitation passe par la considération de l'algorithmique, portée en tant que constituant de l'information cette dernière prise au sens non seulement des mathématiques mais aussi de la Communication.

Car, nous venons de le montrer, c'est non seulement une condition nécessaire à l'exploitation de l'immensité des données, nécessaire à délivrer les conditions de l'appropriation de faits nouveaux , mais aussi à l'entendement de la mise en relation par l'analyse de ces explicitations. Le concept d'information adjoint à ces fonctions de transformation se prédispose à l'élaboration d'un cadre interprétatif transposable.

Cette posture rejoint ainsi les propos de Pierre LÉVY (2015) qui complètent (voir la figure 2.5) en p. 62) ce positionnement épistémologique :

> *Quatre*[8] *échelles ont été ajoutées depuis le milieu du 20e siècle. Nous observons encore l'invention progressive de nouveaux systèmes de codage, principalement centrés sur l'adressage des processeurs, des données et des métadonnées. La construction du médium algorithmique s'en suit. Nous sommes maintenant prêts pour ajouter une quatrième couche d'adressage et, cette fois-ci, ce sera un adressage universel de métadonnées sémantiques. Pourquoi ? Premièrement, nous sommes encore incapables de résoudre le problème de l'interopérabilité sémantique entre les langues, les classifications et les ontologies. Et deuxièmement, sauf pour quelques statistiques d'approximation et des méthodes logiques, nous sommes encore incapables de calculer des relations sémantiques, incluant les distances et les différences. Ce nouveau système symbolique sera un élément clé d'une future révolution scientifique dans les humanités et les sciences sociales, nous amenant à un nouveau type d'intelligence collective réflexif pour notre espèce. Je voudrais fortement souligner ici que la catégorisation sémantique des données restera dans la main des gens. Nous pourrons catégoriser les données que nous voulons et ce à partir de différents points de vues.*

LÉVY postule de la portée d'une telle représentation, le médium algorithmique qui permet de construire des lectures des sphères informationnelles de l'infosphère reproduite sur la figure 2.8 en p. 67 (ibid.). Ainsi[9] :

> *Ce nouveau niveau de manipulation symbolique sera opéré et partagé dans un environnement mixte combinant les mondes virtuels et la réalité augmentée. Les deux niveaux les plus bas du schéma 2.8 p. 67 représentent l'internet actuel : une interaction entre « l'internet des objets » et les clouds dans lesquels toutes les données convergent dans une infosphère ubiquitaire. Les deux niveaux les plus haut, le « senseur sémantique et l'intelligence collective réflexive » dépeint la condition humain qui va libérer le futur.*

8. Traduction personnelle. Citation originale : « Four layers have been added since the middle of the 20th century. Again, we observe the progressive invention of new coding systems, mainly aimed at the addressing of processors, data and meta-data.The construction of the algorithmic medium is ongoing. We are now ready to add a fourth layer of addressing and, this time, it will be a universal addressing system for semantic metadata. Why? First, we are still unable to resolve the problem of semantic interoperability across languages, classifications and ontologies. And secondly, except for some approximative statistical and logical methods, we are still unable to compute semantic relations, including distances and differences. This new symbolic system will be a key element to a future scientific revolution in the humanities and social sciences, leading to a new kind of reflexive collective intelligence for our species. I want to strongly underline here that the semantic categorization of data will stay in the hands of people. We will be able to categorize the data as we want, from many different point of views. All that is required is that we use the same code. The description itself will be free. »

9. Traduction personnelle. Citation originale : « this new level of symbolic manipulation will be operated and shared in a mixed environment combining virtual worlds and augmented realities. This new level of symbolic manipulation will be operated and shared in a mixed environment combining virtual worlds and augmented realities. The two lower levels of the [above] (cf. 2.8 p. 67. slide represent the current internet : an interaction between the « internet of things » and the « clouds » where all the data converge in an ubiquitous infosphere... The two higher levels, the « semantic sensorium » and the « reflexive collective intelligence » depict the human condition that will unfold in the future. »

2.3.3 Apparté : méfiance et raison

La puissance de la considération statistique des faits humains (DESROSIÈRES, 1988) donne matière à inquiéter. Un principe de précaution s'impose d'évidence car on pourrait par exemple proposer de les utiliser à des fins de production de règles sociétales. Alors que l'on pourrait considérer de démocratique à l'extrême car la raison statistique représente finement (DESROSIÈRES, 2008) le reflet de nombreux aspects sociétaux, ce mode de construction de la gouvernance pose la mise en place d'une source informationnelle unique pour un gouvernement. BERNS (2013) en fournit la définition la plus générale :

> *gouverner réside dans le rapport qu'il établit avec le réel : il s'agit de gouverner à partir du réel, à partir des activités existantes, et non plus de gouverner le réel, ou le concret avec l'idée que le concret et son gouvernement seraient des objets de décision ; il s'agit donc de gouverner comme si l'on se contentait de recueillir ce qui est déjà là, en recueillant l'activité humaine, prise en considération et montrée comme vivante et consistant.*

FIGURE 2.7 : http ://geektionnerd.net/prism/

ROUVROY et STIEGLER (2015) dénoncent les risques liés à l'instrumentation décisionnelle par l'utilisation de données factuelles pour établir les règles, dans ce qu'ils appellent la gouvernementalité algorithmique.

> *Par gouvernementalité algorithmique, nous désignons dès lors globalement un certain type de rationalité (a)normative ou (a)politique reposant sur la récolte, l'agrégation et l'analyse automatisée de données en quantité massive de manière à modéliser, anticiper et affecter par avance les comportements possibles.*

Collecter, traiter et puiser ainsi les données est en conséquence la source d'inquiétudes sur de nombreux aspects sociétaux actuels (ROUVROY et BERNS, 2013 ; THOMAS, 2009 ;

BERNS, 2013 ; STIEGLER, 2006 ; STIEGLER, 2004). Ces inquiétudes pourraient retrouver en partie traduction dans un tel système d'écriture issu de l'écosphère pour caractériser populations, comportements, interdits, etc. C'est aussi un écho de travaux plus anciens liés à la mauvaise utilisation de la puissance de synthèse de considérations statistiques que DESROSIÈRES rappelle (2000) le risque classique dû au mauvais usage des statistiques, des nombres en général, mais qui trouve sens **en enfreignant la règle d'explication de l'information utilisée**.

Ainsi, pour ROUVROY et BERNS (2013) :

> *Une donnée n'est plus qu'un signal expurgé de toute signification propre mais c'est aussi ce qui semble assurer leur prétention à la plus parfaite objectivité : aussi hétérogènes, aussi peu intentionnées, tellement matérielles et si peu subjectives, de telles données ne peuvent mentir !*

C'est la construction, l'herméneutique que l'on opère sur ces données qui, elle, peut mentir !

2.3.4 Ouverture

La question n'est pas de donner ici crédit ou pas à de telles modalités d'usage des descriptions statistiques des individus dans un cadre général. Je situe mon propos pour établir cette compétence « algorithmique » au service de la recherche en sciences humaines. D'autant que les inquiétudes se fondent sur la « boîte noire », la transformation aveugle qui régirait les prises de décision uniquement sur les résultats de cette dernière. Par l'explicitation de cette boîte, des cadres d'usages ouvrent des perspectives nouvelles d'appropriation et de compréhension. Il en va de soi dans le cadre de la recherche.

Ainsi, par construction cette écriture permet de situer à la fois les approches de l'information de *bas niveau* s'appuyant strictement sur les données et leur transport pour décrire par exemple un ou des objets de l'IoT (BUYYA et DASTJERDI, 2016), caractériser les constituants élémentaires de la bibliométrie et de l'infométrie en général ou encore la construction de corpus textuels, historiques, de linguistique (RASTIER, 2011) et passer au *haut niveau* de la sémantique (sociale). L'explicitation algorithmique (aussi pauvre soit elle au plan de sa formalisation que de sa complexité) permet de situer la

critique sur un plan de la mise en relation opérée, un algorithme si explicité suffisamment. Explicitée encore suffisamment, formalisée pour ouvrir alors aux calculs formels de CHAZELLE, discursive des opérations réalisées dans tous les cas. C'est sous couvert de cette explicitation de la mise en relation que CARMES et NOYER (2014) offrent l'espoir d'atteindre par les pratiquants des facultés cognitives nouvelles :

> *L'automatisation d'un certain nombre des tâches intellectuelles doit elle aussi être pensée dans sa pleine et entière positivité c'est à dire pas seulement selon le point de vue statique et essentialiste de la « perte » de facultés cognitives, mais aussi selon le point de vue de l'émergence de productivités nouvelles, voire de libération des facultés pour de nouvelles danses créatrices...*

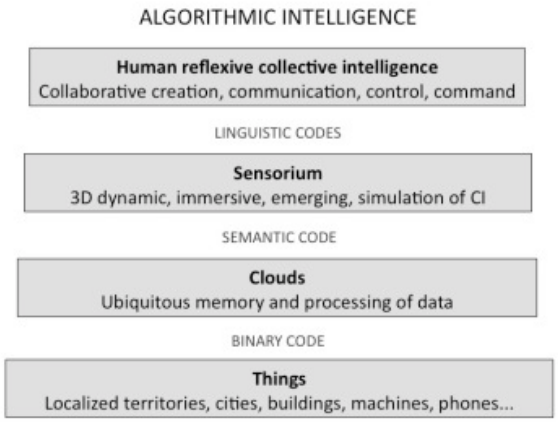


FIGURE 2.8 : L'émergence de la culture algorithmique de Pierre Lévy

Outil de la pensée

La réflexion de LÉVY revendique la création d'instruments pour l'accomplissement de tâches intellectuelles. L'utilisation d'automates qui in fine inspirent la créativité portant la puissance de la pensée au-delà de son niveau initial. Le potentiel informatif de l'infosphère ouvre pléthore de situations à analyser en combinant des médiations variées aux données.

2.3.5 Perspective d'une transdisciplinarité ?

La dimension interdisciplinaire, à la base de la posture de recherche décrite précédemment, fait écho à l'interdisciplinarité inhérente aux sciences de l'information et de la communication (OLLIVIER, 2000 ; BOUGNOUX, 2001). Les SIC présentent les ressources nécessaires pour prendre en compte les multiples dimensions de la place de la technique, mais également pour interroger le rôle des méthodes tout comme les

étapes d'acquisition et de traitement des données numériques (OUAKRAT et MÉSANGEAU, 2016).

S'il est clair que le regard à construire sera de nature pluridisciplinaire, un pas reste à produire pour l'interdisciplinarité au sens de NICOLESCU (1996) qui :

> *concerne le transfert des méthodes d'une discipline à l'autre... (mais dont) la finalité reste aussi inscrite dans la recherche disciplinaire.*

Cela sous-entend d'emblée une harmonisation des méthodes de recherche (et du langage de chacun). Par la construction opérée qui allie contenu et dynamique processuelle au sein du concept de l'information l'on se rapproche ainsi de l'interdisciplinarité de type «trans» de LE MOIGNE (2002, p. 25).

Dans le cadre des sciences de l'information et de la communication l'aboutissement de ces travaux s'inscrira dans une construction transdisciplinaire qui selon NICOLESCU, peut être définie :

> *ainsi que le préfixe « trans » l'indique, (comme) ce qui est à la fois entre les disciplines, à travers les différentes disciplines et au-delà de toute discipline. [...] (Et elle a comme) finalité la compréhension du monde présent, dont un des impératifs est l'unité de la connaissance.*

Le regard actuel posé se légitime par la question « comment est produite, transcrite et utilisée l'information? » et la stratégie de recherche devient une adaptation des concepts et des méthodes au sein d'un édifice pluridisplinaire (ibid.) que nous espérons pouvoir dépasser par ces recherches :

> *l'addition-juxtaposition des disciplines se révélant suffisant et appelant, tout au plus une interdisciplinarité de principe.*

Le propos s'ouvre ainsi à la mise en application effective au sein des humanités numériques, fondée sur des méthodes et une herméneutique instrumentée des données informationnelles au sein d'une « indiscipline » telle que présentée (BESNIER et PERRIAULT, 2013, p. 13) en « renouveau des sciences de la communication ». Mais ouvrons tout d'abord une explicitation de la notion d'information entendue à ce stade.

2.4 Unification conceptuelle

Nous avons vu que la plasticité de la notion « information » est transportée (ou bloquée selon le point de vue) d'une discipline à une autre, d'un regard à l'autre, d'une perspective à une autre par une contextualisation complémentaire au signe élémentaire. Le choix d'un terme information, signe, trace, texte n'est pas sans risque non plus comme esquissé précédemment. Apparaît ainsi, dans les divers positionnements des diverses disciplines, que le concept requiert l'existence d'entités atomiques privées de sens (un méta-descripteur générique) associées à une opération herméneutique que l'on retrouve dans la fonction sémiotique de HJELMSLEY (BONDÍ, 2008) qui en donne le sens. Cette construction lui confère ainsi une flexibilité adaptable aux différents points de vue qu'il faut maintenant cerner plus précisément.

2.4.1 Construction d'un point de vue

HOFKIRCHNER (2009) cherche à unifier le concept d'information et dénote aussi la superposition à différents niveaux et différents degrés de nombreux concepts avec le super concept « information >> : « structure, data, signal, message, signification, sens, signe, psyché, intelligence, perception... ». (SAVOLAINEN, 2016) propose en conséquent quatre points de vue du concept :

- l'information comme **processus d'être informé**, proche des travaux de BUCKLAND (1991) ;
- l'information comme **une revendication sur le monde**, une proposition. L'information est alors un résumé « fourni de sens » représentant les assertions faites sur les objets ;
- l'information construite socialement en tant qu'Artefact humain au sein de situations sociales ;
- les définitions structurelles régissant l'organisation et la structure de l'information contiennent elles-mêmes leur propre information et sont en ce sens informatives. **La méta-information.**

2.4.2 Propriétés informatives

Citant GOGUEN (1997) qui revendiquait une définition sociale pour définir le concept d'information (BOWKER et al., 2014, p. 29), BURGIN retient (2010, p. 178-179) à l'information les cinq premières propriétés suivantes (parmi les sept initiales de GOGUEN reproduites ci-après) :

1. située (peut être comprise dans un contexte (temporel, lieu, et de groupe) seulement),
2. locale (le contexte induit un niveau empirique de définition),
3. émergente (se révèle seulement subséquemment après traitement définit comme une interaction avec le message),
4. contingente (l'extraction de l'information dépend de plusieurs paramètres qui sont changeants et modifient l'état de l'information),
5. représentée ou « incarnée » dans un médium,
6. vague car elle n'est pas pré-déterminée et peut varier d'une situation à une autre,
7. puis ouverte car son extraction est un processus non fini.

Ces propriétés, associées aux différents points de vue précédents paraissent satisfaisants pour cerner le concept. Remarquons-en toutefois le caractère limitée au plan opérationnel : qualifier l'« information », qualifier de la sorte un message transmis semble difficile à opérationnaliser. Il suffit de tenter mentalement de l'effectuer.

2.4.3 Une approche paramétrique

PERVEZ qui recense la diversité des conceptions de l'information au sein de travaux très variés (physique, économie, sociologie, ...) depuis les années 1950 et identifie quatre dimensions (données, processus, canal ou technologie/utilisation et fonctions ou applications). L'auteur propose d'utiliser en sciences de l'information ces dimensions pour construire un support d'aide à l'interprétation constructive de l'information fondé sur les quatre formes de SPENCER BROWN (BROWN SPENCER, 1969).

De son côté, BURGIN (2010, p. 104-119), avec pour objectif d'unifier la compréhension du terme information, démontre comment une théorie globale de l'information peut élaborer les relations d'une diversité de sens très vaste. Son approche mathématique s'appuie sur l'inscription d'un paramètre de distinction des types d'information (le paramètre *système infologique*[10]), pour distinguer social, personnel, chimique, biologique, cognitif... qui lui permet d'intégrer tous les types d'information en un seul concept (J. E. BRENNER, 2012).

2.4.4 Théorie du message de CAPURRO

CAPURRO considère la société actuelle (l'écosphère décrite au chapitre 1) comme celle de « messages » et de « messagers », des humains reliés par différents moyens de communication, en particulier numériques, activant synergies de plans (et par deçà) ethnique, économique, culturels... Il continue son œuvre de synthèse du domaine de l'information (KELLY et BIELBY, 2016, p. 220-267) en fournissant une approche à la fois philosophique et herméneutique de la compréhension de ce phénomène complémentaire aux travaux de BURGIN (J. BRENNER, 2014).

2.4.5 Recomposition

2.4.5.1 Herméneutique fonctionnelle

Ces lectures éparses, courant de pensées, fondamentaux théoriques disciplinaires quelquefois, montrent que des concepts d'information différents, désignant des objets différents se retrouveraient par la considération d'un signe neutre (l'information mathématique) et d'une gamme[11] de fonctions herméneutiques (reprenant *a minima* les propriétés précédentes de GOGUEN). L'on retrouve cette approche formulée différemment chez BATES (BATES, 1999) et FLORIDI en 2005 (LELEU-MERVIEL et USEILLE, 2008).

10. Traduction personnelle de : « infological system ».

11. Par la suite la gamme de fonctions est désignée par « fonction » au singulier, le lecteur pouvant considérer $\varphi = (\varphi_1, \varphi_2, \varphi_3, \cdots, \varphi_k)$ comme une seule et même fonction. La plupart du temps au cours des exemples utilisés, une seule des propriétés sera évoquée (contextualisation).

Si l'on admet que la fonction herméneutique associée peut être un algorithme (cf. postulat 2.2.3 p. 60), on retrouve ce constat au sein même de la science informatique, exprimé ci-après par CHAZELLE (2013) :

> *Au-delà du scepticisme ou de l'engouement du jour pour la dernière nouveauté informatique se cache un de ces changements de paradigme chers à* Kuhn. *Outil conceptuel subversif, l'algorithme ouvre la possibilité d'un regard nouveau sur les sciences et les technologies qui s'étend bien au-delà de ses applications pratiques.*

L'auteur poursuivra plus loin :

> *En particulier, elle [la notion de lien entre programmes et données] ouvre la voie à une approche algorithmique de ces processus plus riche et plus expressive qu'un traitement purement informationnel.*

C'est cette relation entre programme et données qui est adressée ici, supposée ici la même que celle entre signe et information quel que soit la profondeur de la « malle » associée au mot. De surcroît, l'on peut aussi convenir de la plasticité nécessaire à ces opérations sur le concept lui-même de l'opérateur de relation.

2.4.5.2 Notation plastique et conceptuelle

Le concept d'information doit être suffisamment plastique pour intégrer l'ensemble des éléments précédents.
Nous noterons [12] ce concept :

$$\phi_k = (\Omega, \varphi)$$

Avec Ω qui désigne un concept non formalisé – bien pratique pour représenter l'information ou la « donnée » privée de sens – et φ les fonctions associées à, par exemple, la contextualisation ou leur description. L'indice, noté k dans ϕ_k désigne le « plan du regard », le plan conceptuel d'une étape de construction du point de vue, le paramètre infologique.

12. Le choix des signes de représentation utilisés ici est totalement arbitraire, guidé par une recherche d'esthétique auto-complaisante du document, modifiables.

Une représentation hybride : le modèle des informations

Après cette déconstruction du concept, en retenant l'idée d'existence d'une information paramétrée de BURGIN, des propriétés singulières décrites par GOGUEN, et de la synthèse de dualisme de CAPURRO, l'ensemble peut se formaliser par la considération d'entités informationnelles neutres, le paramètre infologique et des « fonctions » herméneutiques associées.
Au plan formel, je noterais l'information par : $\phi_k = (\Omega, \varphi)$ en faisant abstraction, pour la simplification dans les propos suivants, du paramètre infologique, qui sera implicite dans l'utilisation du modèle.

2.4.5.3 Données, obtenus, captas et herméneutique associée

Que ce soient le filtrage, l'agrégation, la transposition... tout élément de construction (l'émergence) qui en cerne au possible ou au nécessaire jusqu'à, si possible, la limite de la subjectivité pour que l'ensemble ϕ_k transporte les éléments de son herméneutique (fiable ou pas). Cette notation pseudo-formelle est suffisamment flexible pour tantôt devenir formelle par instanciation dans les sciences du traitement de l'information en introduisant, par exemple, les instrumentations de mesure. Ou, suffisamment flexible encore, pour embarquer aussi les fonctions d'herméneutique telle une mesure de la fiabilité de l'ensemble Ω associé. φ agrège récursivement les diverses opérations de traduction et demeure ainsi suffisamment flexible enfin pour représenter une description discursive « littéraire ».
Ainsi, le donné primaire Ω peut être lui même défini par toute opération sur des obtenus latouriens (ou *captas* (DRUCKER, 2011 ; DRUCKER, 2010)) issues eux mêmes d'une construction différente : $(\Omega', \varphi')_k$ et ainsi de suite une combinaison d'autres opérateurs herméneutiques primaires récursivement à partir des signes bruts les plus élémentaires.

MUGUR-SCHÄCHTER statue (LELEU-MERVIEL et USEILLE, 2008) que l'information résulte de l'interaction des données avec la structure de réception, et qu'il est possible d'obtenir des informations différentes à partir des mêmes données. La séparation signe élémentaire et fonction herméneutique associée offre la possibilité de discuter de ces points de vue en considérant/formalisant $\phi_k = (\Omega, \varphi)_k$ et $\phi_k = (\Omega, \varphi')_k$ ou plus, si né-

cessaire l'interprétation des informations différentes s'opérant au niveau de la différentiation φ, φ' car par définition construits sur l'égalité des φ_i pour tout $i < k$. Ainsi, la discussion peut se faire sur deux plans : le premier formel dépendra du degré de formalisation et explicitation des $\phi_k = (\Omega, \varphi)$ associés, le second pour s'attacher à exprimer les différences de ces constructions qui sont les mises en relations, les herméneutiques. L'avantage en est de pouvoir passer à un plan méta en regard de ce qui est observé : les émergences et les opérateurs de leur construction.

La construction théorique de ce chapitre permet de continuer dans un domaine d'application, celui des humanités numériques et, pour une entrée en matière du chapitre suivant, MANOVICH (2013, p. 14) ouvre à l'intérêt de porter les questions des sciences humaines et sociales aux algorithmes que l'on aura reconnus en support du suivi des observables, reconstructeur des réalités, traceur des agencements, producteur de captas, une batterie d'artefacts au service d'une herméneutique appropriée aux mégadonnées :

> *Why should humanists, social scientists, media scholars, and cultural critics care about software ? Because outside of certain cultural areas such as crafts and fine art, software has replaced a diverse array of physical, mechanical, and electronic technologies used before the twenty-first century to create, store, distribute and access cultural artifacts.*

CHAPITRE 3

Les humanités numériques

Les sciences humaines ne détiennent certes aucun monopole, mais elles ont fortement incité au développement, parfois même à la création, de méthodes originales d'analyses de données, classification, classification automatique. Elle continueront à pousser au progrès, car, outre le prêt à porter qui est toujours utile, il leur faut souvent du « sur-mesures ».

— Lentin ANDRÉ, Rapport sur les applications des mathématiques aux sciences de l'homme, aux sciences de la société et à la linguistique, p. 16 (1984)

Tandis que l'informatique obtient sa reconnaissance officielle dans les instances académiques, associée ensuite dans le cadre du CNRS à l'automatique, à l'analyse des systèmes et au traitement du signal, les aspects « sciences humaines et sociales » de l'ancienne cybernétique refont surface sous deux formes nouvelles : les sciences cognitives et les sciences de l'information et de la communication. Seules l'intelligence artificielle et la « systémique » (elle-même avatar de la recherche opérationnelle) font passerelles entre ces champs scientifiques séparés.

— P.E. MOUNIER-KUHN, L'informatique en France de la seconde guerre mondiale au plan calcul : l'émergence d'une science, p. 571 (2010)

ε

RAPPORTANT récemment les propos de Dominique BOULLIER, la journaliste du journal le Monde, Marine MILLER (2016) évoque l'éducation et les métiers du numérique soulignant l'imbrication étroite de nouvelles méthodes algorithmiques allant de pair avec l'écosphère. Les propos du sociologue rejoignent ceux du mathématicien Henri POINCARÉ (1911, p. 26).

> *On s'accorde à dire que l'enseignement littéraire, bien compris, c'est-à-dire dépouillé de tout appareil inutile de pédantisme ou d'érudition, est le plus propre à développer en nous l'esprit de finesse. Et comme l'esprit de finesse est nécessaire à tout le monde, parce que tout le monde doit vivre, on conclura que la culture littéraire est nécessaire aux savants comme à tous les hommes. Seulement on croit généralement qu'ils en ont besoin pour devenir des hommes et non pour devenir des savants ; et c'est là qu'on se trompe.*

Je propose d'alimenter la discussion non pas sur ce que sont ou pas les humanités numériques (CITTON, 2015) ou digitales (LE DEUFF, 2014), débat richement instruit par ailleurs (WELGER-BARBOZA, 2012 ; MOUNIER et DACOS, 2015), mais poser des bases constitutives en continuité de ce qui précède pour en montrer l'affiliation et dénoter quelques instanciations. J'adhère au point de vue de CASILLI (2014b) pour qui la question des humanités numériques (HN)

> *est un domaine de recherche qui prend en compte les transformations introduites par les technologies de l'information et de la communication, et leur impact sur les sciences économiques et sociales ainsi que sur les sciences humaines.*

MCCARTY (2014) tente en effet de cerner des « communs méthodologiques » aux différentes disciplines des sciences humaines qui reposent sur une formalisation (MOUNIER et DACOS, 2015) du raisonnement scientifique : « ce n'est donc pas la capacité de traitement statistique de grandes masses de données qu'implique le recours à l'informatique, mais plutôt la contrainte de désambiguïsation et de formalisation inhérente à toute activité de programmation ». Cette position souligne l'apport algorithmique dans les HN.

De son côté, sans évoquer les HN, MERZEAU (2011, p. 92) proclame « dans ce processus mémoriel propre au numérique, l'externalisation ne porte alors plus seulement sur la

mnème (la trace elle-même), mais aussi sur l'anamnèsis : le recueil et le traitement des traces, désormais délégué aux algorithmes ». Cette dernière se retrouve comme une construction du point de vue (Ω, φ) dans le concept de la médiation abordé en p. 91.

C'est en effet cette notion de trace que BOULLIER (2015) promeut en source de la 3e génération des sciences sociales, dans le courant des humanités digitales (GOLD, 2012 ; D. BERRY, 2012 ; WELGER-BARBOZA, 2012 ; MOUNIER et DACOS, 2015).

3.1 Anamnèsis ciblé : empiries numériques, captas, obtenus, tout dépend du point de vue

La capacité mémorielle, l'infrastructure internet, enregistre **des faits**, noumènes ou diaphories, issus du réseau de ses contributeurs (malgré eux quelquefois), laissant des signaux des activités humaines (ou pas) sur tous les plans (professionnels, personnels, organisationnels), toutes les sphères qu'elles soient informationnelles ou communicationnelles. Les HN prennent appui sur la disparition de cette frontière qui séparait le chercheur de son corpus de recherche certains objets nécessitant des collectes de données ciblées selon des modes variés d'observation (désengagée, participante, sondage, entretiens, recueils...) en appelant le plus souvent à une méthode et un échantillonnage rigoureux, devenant sujet quelquefois à des interprétations subjectives. Cette phase de constitution de données d'appréhension des objets des humanités se franchit aujourd'hui et, sous couvert de reproduire une méthodologie équivalente (tout aussi rigoureuse, délimitée, discutable par évidence sur la base de ces limitations), le chercheur visera à (re)constituer un ou des corpus à partir du foisonnement de l'infosphère à condition de franchir la fracture (CASILLI, 2014a) imposée par le numérique. MERZEAU (2009) voit dans cet essor du numérique une transformation environnementale qui remet en question « les modèles conceptuels qui servent à les formaliser ». Des algorithmes de construction qui ne peuvent être des boites noires déjà prêtes car d'une part il est peu probable que la construction d'un regard portant sur un objet scientifique des humanités soit déjà toute faite [1] et, d'autre part, le chercheur ne peut accep-

1. Ce serait d'évidence peu informatif.

ter une quelconque « gouvernementalité algorithmique » méconnue de son objet (ou de la construction de celui-ci).

À titre d'exemple, SZONIECKY et LOUÂPRE étudient quelques outils de traitement et de cartographie afin d'élaborer une méthode générique (2015) pour « l'analyse des écosystèmes d'informations numériques qui composent un univers complexe ». Si, par construction sur les systèmes complexes, la méthode ne peut être générique, les auteurs identifient à leur tour une nécessaire formalisation pour élaborer des cadres interprétatifs comparables. Ils élaborent ainsi un cadre par application de la sémantique différentielle de BACHIMONT (2007) pour rendre interopérables les émergences proposées par les chercheurs. Par ailleurs, l'ouvrage de ROMELE et SEVERO ouvre de nombreuses pistes applicatives et méthodologiques quant à l'utilisation de **captas**.

J'explore par la suite des exemples de telles constructions informationnelles, des élaborations de points de vue plutôt génériques[2] sans me soucier ni de leur complexité (qui induirait indirectement avec son augmentation un temps de calcul pouvant devenir non négligeable), ni de leur faisabilité automatique (qui imposerait une activité manuelle), ou encore sans présumer de l'utilisation négative que l'on peut en faire. Des domaines très variés sont montrés pour apprécier l'étendue, chacune des activités du chercheur reposant a priori sur la documentation de son objet d'étude par la construction d'un recueil de données approprié, puis le traitement et enfin l'exploration avant d'élaborer les étapes d'analyse en application directe de l'artefact médiateur (cf. figure 3.2 p. 93).

3.1.1 Cas des instruments de bas niveau

À travers l'internet, les objets connectés fournissent pléthore de capteurs. Ces objets deviennent de potentielles sources informationnelles qui ouvrent la possibilité de dépasser[3] messages et contenus de la communication par d'autres éléments informationnels complémentaires que ce soit environnementaux (température, luminosité,

2. La partie suivante de ce document est dédiée à une série de constructions ciblées et opérationnelles.
3. En d'autres termes, il s'agit aussi d'étendre les sens en augmentant la perception que l'on peut avoir d'une situation. Une augmentation qui consiste à compléter le « regard » par d'autres angles de vue, d'autres médias, d'autres signaux. Par des *oligopticons* sur des captas variés.

...) ou comportementaux (langages de communication non verbale,...), afin de compléter la perception de l'objet d'étude initial à l'ensemble des signes de la communication (paroles, gestes, écrits, images, musiques) au travers des différentes machines à communiquer (VERHAEGEN, 2010). Les capteurs transforment en soi les « signes analogiques » (MEUNIER et PERAYA, 2010, p. 28) de la communication en signaux numériques que le chercheur peut collecter pour mémoriser son expérimentation, son cadre, et mettre en contexte. Au-delà de la mémorisation, s'agit aussi étendre sa conception de l'objet, cerner son regard et les interprétations possibles, le « documenter » (VAYRE, 2014). Les capteurs démultiplient certes les données mais surtout les capacités de retranscription du réel. Ces appareils opèrent en soi une fonction de discrétisation du continu. Pour utiliser ces sources informationnelles il convient d'en caractériser a minima certains éléments physiques. Ces éléments décrivent la transduction et sont généralement fonction de l'utilisation qui en est faite, en amenant sur un plan informationnel k. Pour les signaux vidéo, la fréquence de numérisation qui permet de cerner le tempo des signaux et d'informer sur la période de captation résultant de cette source. Par exemple, pour une caméra son nombre d'images par secondes définit cette période dont il faut tenir compte dans un contexte donné. Ainsi, dans le cas où cette caméra est utilisée pour capturer les signes de communication non verbale émanant des acteurs, l'échantillonnage devra se faire à haute fréquence pour pouvoir saisir ces signes en décomposant les mouvements mieux que ne peuvent le faire nos yeux, sauf par « chance » ou hasard. À l'inverse, pour montrer la perte de cheveux au cours de la rédaction d'une telle réflexion, ce sera une fréquence faible qu'il faudra utiliser[4]. De nombreux exemples simples peuvent être identifiés aussi quant aux couleurs...

Cette phase d'échantillonnage se complète d'évidence d'une phase de sélection. L'augmentation s'appliquant à la quantité informationnelle, la phase de sélection en saisit la partie utile au contexte. Dans cet exemple les signes gestuels, mimiques ou autres éléments informationnels tels la détection d'un cheveu dans sa chute sont déterminants.

Ces étapes franchies, le donner-sens doit se poursuivre par l'élaboration constitutive d'informations de plus en plus élaborées.

4. Heureusement!

3.1.2 Contenus (humains) du web

ε

De nombreux travaux décrivent les pages web organisationnelles (REYMOND et PINÈDE, 2010), les blogs, les documents académiques (AGUILLO, 2010 ; THELWALL, 2002 ; THELWALL et WILKINSON, 2004 ; THELWALL, BINNS et al., 2002) sont autant de sources informationnelles qui traduisent une réalité duale de l'organisation, de l'individu (blogs et réseaux sociaux), de groupes (BJÖRNEBORN et INGWERSEN, 2001 ; BJÖRNEBORN, 2004 ; THOMS et THELWALL, 2005). Une discipline de l'informétrie (NOYER, 1995) s'est développée pour élaborer des cadres interprétatifs, des modèles d'analyse et explorer ce vaste domaine essentiellement par transposition de la bibliométrie (THELWALL, 2008) par analogie entre le lien hypertextuel et la citation (SMITH, 2004). L'hyperlien est une source informationnelle à part entière, nouvelle et difficile à cerner. Selon les objets ciblés, la collecte de données, le filtrage et l'analyse produit sens en communication des organisations (BARATS, 2013) ou en communication sociale (MILLERAND, RUEFF et PROULX, 2010), etc.

À ce niveau informationnel (le texte) l'élaboration d'une « trace » pour un objet de recherche devient un document élaboré. Document qui se définirait comme un assemblage de captas : une reconstitution par la collecte d'une mémoire de l'activité d'un groupe, d'un individu ou d'un collectif. Le support textuel et hypertextuel autorisera dès lors des traitements très variés pour l'exploration de ce reconstruit : classification et classement automatique ou manuel, mots-associés, citations et co-citations, résumé automatique et l'ensemble des traitements que les techniques du traitement automatique des langues facilitent. Des procédés qui servent au-delà de la synthèse, à l'extraction d'informations particulières depuis la terminologie utilisée au style de rédaction et, bien entendu des entités nommées (dates, lieux, personnalités, etc.). Une panoplie d'instruments qui opèrent un traitement automatisé des contenus. Ces instruments sont en soi des φ de construction d'éléments informationnels de niveau « supérieur ». D'évidence les mêmes opérations sont combinables sur les données constitutives des corpus ou sur les métadonnées associées qui rendent possible l'entrecroisement des échelles informationnelles de niveaux différents (et ce, récursivement).

3.1.3 Activités humaines : utilisations ou usages

Sur un autre plan, les données d'activités de visites sur un site [5] (prenons simplement les « clicks ») capturés par les logs des serveurs. Ces derniers mémorisent le flux de l'activité des usagers sur le serveur, une numérisation du moment de leurs activités. Le flux de logs peut être considéré comme une donnée (série temporelle de signes) dont la fréquence d'échantillonnage est celle de la fréquence d'activité des humains et non humains autour du site (d'aucun y retrouvera une bribe pragmatique (partielle, j'en conviens) de la théorie de l'acteur réseau (VENTURINI et LATOUR, 2010)) : la « mémorisation » de l'entrée en résonance entre l'Humain et le non-Humain.

3.1.4 Exemple hors SIC : lettres, langues et données humaines

Un autre plan encore, celui de la littérature. On pourrait construire un corpus de textes d'un auteur X à l'aide d'un (Ω, φ). Puis un second corpus constitué par les évènements journalistiques issus d'extraits de publications contemporaines à X $(\Omega^{'}, \varphi^{'})$. L'hypothèse est faite que ce second corpus contient une base d'éléments informationnels ayant potentiellement inspiré X. Une opération d'identification et d'extraction de termes associés dans l'un et l'autre des corpus, leur mise en relation temporelle vont exprimer un nouvel ensemble $(\Omega^{''}, \varphi^{''})$ explicitant plus facilement l'évolution de l'auteur dans son univers. J'espère que le côté littéraire du lecteur pourra souscrire à cet exemple simple, construit pour évoquer la puissance d'une telle notation : les corpus sont caractérisés de manière à exprimer leurs limites, les fonctions de filtrage et d'agrégation caractérisées pour circonscrire leur portée et réduction éventuelle, afin de rendre l'ensemble interprétable dans un contexte raisonné et exposé. Cela ne donnerait probablement rien pour certains auteurs, mais l'explicitation précédente en portera probablement la cause ou la source informationnelle pour chercher ailleurs.
SCHÖCH montre (2012) que les recherches en littérature peuvent se reconfigurer en s'appuyant sur un triptyque opérationnel : l'identification de nouvelles configurations textuelles que les méthodes numériques permettent de découvrir. Il emploie un algorithme informel descriptif de la « motivation narrative ». Le second élément proposé

5. Transposable sur n'importe quel dispositif TIC

par l'auteur est l'introduction d'une base de données, son processus d'alimentation et d'enrichissement conférant au chercheur une puissance d'adaptation à des contextes littéraires variés. Ainsi, la figure 3.1 montre un processus itératif d'adaptation de la « lecture » instrumentée d'un corpus pour alimenter une base de données descriptive dudit corpus. Si le corpus d'étude se situe au plan d'un Ω pré-établi (les romans français entre 1760 et 1800), la base de données permet au chercheur de se situer sur un plan méta, complémentaire au corpus précédent. La construction proposée donne la possibilité de reconduire le processus à d'autres ensembles. Le troisième élément

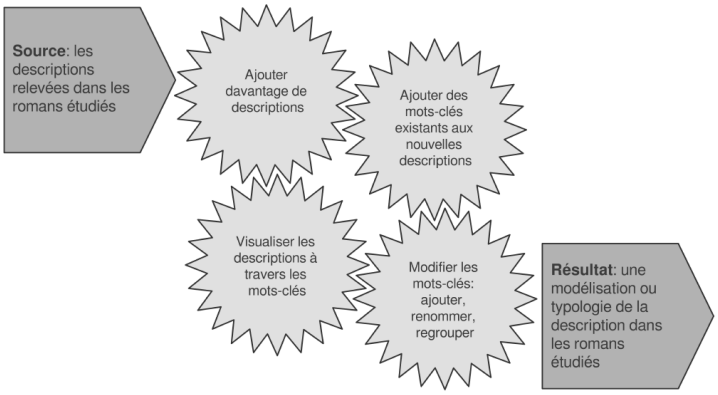


FIGURE 3.1 : Le processus itératif de l'établissement de la base de données selon SCHÖCH

souligne les tensions issues du tournant numérique induites par le renouveau méthodologique, les rapports entre le chercheur individuel et les autres disciplines, l'alliance de « l'approche herméneutique des sciences humaines aux formalisations et aux calculs de l'informatique et de la statistique ».

3.2 La difficile question de la Communication

J'ai inscrit dans le chapitre 2 p. 43 mon positionnement quant à la non séparation entre l'information et la communication. Je pose dans cette section une piste en accomplissement théorique de ce qui précède, pour l'élaboration d'études de communications qui convoque les travaux de QUETELET et permet d'instruire des recherches sur des bases inédites : la considération de l'altérité par un représentant prototype moyen en contexte.

3.2.1 La communication inter systèmes

Abraham MOLES en rapprochant psychologie sociale et communication (DEVÈZE, 2004) va introduire une fine analyse en extrapolation du schéma de la communication qu'il définit (MATHIEN, 2003) en « science de l'interaction des êtres et des choses est plus ou moins indépendante de ces êtres et de ces choses ». Il développe d'après PERRIAULT (2007) comme « l'action de faire participer un individu ou un organisme situé à une époque, en un point donné, aux expériences et stimuli de l'environnement d'un autre individu ou d'un autre système situé à une autre époque, en un autre lieu, en utilisant les éléments de connaissance qu'ils ont en commun. ».

Dans ce contexte, pour VITALIS (2007) « désigner par information l'échange qui se produit est erroné » : une confusion entre ce mot et les termes voisins de « donnée » et de « connaissance » au risque de « valoriser le rôle des technologies [6] et de minimiser voire d'ignorer le rôle des individus. L'auteur poursuit :

> *[...]Si les technologies peuvent aider à une meilleure mobilisation des données, ces dernières ne deviennent des informations, ou mieux des connaissances, qu'après un effort d'appropriation de la part des individus. Sans cette appropriation, les meilleures technologies du monde sont sans effet.*

Du système à l'individu

MOLES introduit ainsi l'acteur, l'environnement, le contexte en complément du modèle de SHANNON, et le complique par quelque chose qui n'est pas contrôlable, définissable et paramétrable : l'Autre, de qui dépend forcément la communication. On passe de la communication entre systèmes machiniques (que le modèle de SHANNON idéalise) à la Communication.

3.2.2 L'Autre

Dominique WOLTON approfondira le regard de VITALIS et cerne avec finesse l'évolution sociétale de la communication par sa démarcation de l'information au sein des SIC, afin de poser la problématique de l'échange essentiellement sur le récepteur « la

6. Sans doute l'auteur a voulu suggérer que le modèle de SHANNON ne s'applique pas à ces dimensions de l'information, de nature différente de celle du modèle et qu'il convient de regarder les acteurs de cette communication.

question de la communication, c'est-à-dire de « l'autre », avec l'obligation et la difficulté de la cohabitation est bien au cœur des défis nouveaux »(WOLTON, 2007). Dans son point de vue, WOLTON développe cinq caractéristiques de la communication en un problème qu'il qualifie d'« interdisciplinaire » par les constats suivants :

- la discontinuité entre l'homme et les outils qu'il a inventé pour communiquer ;
- la communication se construit en prenant en compte le point de vue de l'autre, alors que l'information se transmet ;
- par défaut, le récepteur est rarement en phase avec l'émetteur ;
- l'altérité est l'enjeu le plus difficile de la communication ;
- l'« incommunication » devient en quelque sorte l'*horizon de la communication.*

À cette position, je peux opposer que d'une part, dans tout rapport de communication l'incommunication guette mais ce n'est pas seulement l'Autre qui détient l'épée de Damoclès ou peut en porter la cause, par évidence. Ainsi, que l'on soit compris ou incompris par manque de phase du récepteur et de l'émetteur, que l'incommunication existe (par un fait ou un autre) sous-entend d'abord que le message a bien été transmis (que le modèle de SHANNON a bien fonctionné) sinon la déformation du message peut être en cause et, dans tous les cas, de devoir considérer le message pour établir ce fait. Ainsi à message correctement transmis, il faut par évidence se rendre extérieur à la Communication pour l'observer en tant que telle. Ce qui nous ramène au modèle des informations développé précédemment (cf. 2.4.5 p. 71).

D'autre part, situer le problème de la communication sur « l'Autre » conduirait à tenter de cerner ce dernier ? Ainsi, le constat est fait (« Introduction générale » 2014) que la « communication humaine se construit sur une incommunication manifestant la part de plus en plus active du récepteur et sa résistance à une communication réussie avec l'Autre [...] Il y a autant d'altérités que de singularités humaines, et par définition elles sont toutes absolues et radicales. ». Caractériser et déterminer l'Autre nous amène à dépasser les frontières de toutes les sciences, ne pouvant nous donner une définition individuelle de nous même suffisante à la caractérisation de toute situation de communication (libre arbitre, hasard, choix individuels obligent par évidence).

Complexité sociale

Cette voie d'approche de la Communication se heurte à l'immensité[a] des altérités par définition de l'unicité de l'individu.

a. Pas seulement en nombre.

Toutefois, VIAL (2013) pose une piste associant écosphère et perception :

> *Si l'Autre n'est pas en soi une donnée engendrée par la technologie, l'Autre comme phénomène nous est néanmoins donné par l'environnement techno-perceptif dans lequel nous vivons, dans des conditions historiques et techniques déterminées, qui sont aujourd'hui celles du « système technique numérique »*

La CNIL nous défendant de constituer des informations nominatives (ce qui dans ce cas de figure n'aurait scientifiquement aucun intérêt), je continuerai le raisonnement en m'attachant non plus à l'autre mais à l'ensemble des autres.

3.2.3 Les autres

Dans le premier tome de sa Méthode, MORIN (1977) décrit le fonctionnement des circuits de l'information dans un modèle biologique de la communication pour lequel l'ADN est une mémoire informationnelle pour les cellules se développant par construction à l'aide de cette dernière. La mémoire est alimentée par le milieu ambiant et sert de support au programme de construction et adaptation de l'embryon à son environnement. DAGENAIS (2007) suggère de transposer ce modèle biologique sur le plan social.

JUANALS et NOYER (2007) notaient dans les travaux de D. H. HYMES les fondements d'une anthropologie pragmatique et communicationnelle que les auteurs dérivent en une « ethnographie de la communication » pour « étudier le statut des processus communicationnels situés au cœur des collectifs d'énonciation complexes ». Les auteurs ouvrent par là une piste alternative à la précédente : il ne s'agit pas d'aborder la Communication en cernant le libre arbitre, mais de déterminer les variantes observables que l'on qualifierait/différencierait au plan ethnographique, linguistique et contextuel. Si la portée d'une telle approche est loin d'un déterminisme absolu quant au fonctionnement de la Communication, elle ouvre une perspective aujourd'hui abordable.

La perspective du numérique

Le tracé (volontaire ou reconstruit) de situations de communication variées (à une échelle impensable pour un individu disposant de plusieurs vies), peut couvrir nombres de cas de figure en étayant et instruisant une telle variation, qui dès lors qu'elle serait suffisante, devrait amener à pouvoir les typer.

Comme le suggèrent JUANALS et NOYER, les travaux de D.H. HYMES offrent une piste méthodologique à associer aux divers procédés de traitement du langage (pas forcément uniquement linguistique) tels les sémiotiques, les mots associés ou leurs combinaisons. Ainsi, tel l'ADN pour le vivant, l'internet se pose en mémoire informationnelle de la Communication.

3.2.4 Quetelet : documenter l'Autre (prototype)

Adolphe QUETELET (1796–1874) était un mathématicien statisticien et astronome belge qui se passionna pour le calcul des probabilités appliquées, entre autres, à l'étude des caractéristiques sociales et physiques des humains (EKNOYAN, 2007), ce qu'il appela la *Physique Sociale* (QUETELET, 1842). Son approche a été de construire un « homme moyen » selon des traits caractéristiques : ainsi, l'index de QUETELET, utilisé comme référence par les médecins, détermine l'indice de masse corporelle (IMC) pour estimer la corpulence corporelle. Une représentation synthétique, qui plus que le schéma de SHANNON, a été source d'incompréhensions et de critiques et, dans ce cas particulier, de mésusage.

Deux constats : le premier est que les propos de QUETELET ne sont pas de déterminer les caractéristiques justes et adéquates pour un individu (la moyenne) mais bel et bien de se donner un référentiel qui permette de constater un écart (une anomalie éventuelle) : l'utilisation de l'index est de poser en différenciation les caractéristiques d'un individu en regard de l'index. Le second se pose sur l'analogie fréquente, une confusion de l'approche de QUETELET, entre la recherche de la normalité (DAVIS, 1997) que les détracteurs opposent à QUETELET et la (simple) construction d'un réfé-

rentiel permettant d'interpréter de trop vastes dimensions caractérisant l'individu. La première conduit d'évidence à toutes les dérives possibles[7], la seconde instrumente une méthodologie d'analyse comparée pour qualifier avec un degré de certitude (jamais de « vrai » ou de « faux »), un état en regard de la norme. La confusion est aussi probablement étymologique : la notion de « normale » de QUETELET devant être entendue comme la fréquence moyenne mathématique ou, au plan médical, comme ce qui s'oppose au pathologique et non ce qui doit « être ». La citation suivante de l'auteur (QUÉTELET, 1829), malmené depuis le début de ses travaux, en témoigne :

> *Je suis loin sans doute de prétendre que quelques tableaux numériques isolés peuvent suffire pour déterminer complètement tous les éléments si compliqués de nos sociétés modernes. Il faudrait, pour remonter des effets aux causes, ou pour conclure de ce qui est à ce qui sera, avoir égard à un ensemble de circonstances qu'il n'est point donné à l'homme de pouvoir embrasser : de là, la nécessité de négliger toujours, dans toute espèce d'appréciation, un certain nombre de circonstances dont il aurait fallu tenir compte.De là aussi, l'absurdité des résultats auxquels conduit souvent cette énumération incomplète, ou le trop d'importance qu'on attache à un élément qui ne devrait être considéré que comme secondaire. La mauvaise foi pourra même porter à ne choisir dans une série de résultats, que ceux qui sont favorables au principe qu'on voudrait faire prévaloir, en passant sous silence ceux qui lui seraient contraires : et c'est ainsi, comme on l'a fort bien observé, que tout pourrait se prouver par les nombres de la statistique. Non sans doute : il faudrait rejeter aussi la physique, la chimie, l'astronomie, en un mot, toutes les sciences d'observation qui rendent les services les plus éminents et qui font le plus d'honneur à l'esprit humain. Pour l'ignorance, elle se montre toujours par assez de côtés, pour qu'on n'ait point à la redouter ; quant à la mauvaise foi, il faut s'attacher à la combattre, en prenant dans la statistique même les éléments qu'elle cherchait à cacher, afin de substituer avec plus d'assurance le mensonge à la vérité.*

Ainsi, en s'abstenant de la tentation de considérer l'Homme Moyen comme la référence à viser ou la façon normale d'opérer (NANNIPIERI, MURATORE et al., 2016), mais plutôt, par exemple, au sein d'une méthode d'exploration d'analyse de situations (PLANTIN et RUSSO, 2016), la construction de QUETELET est d'un apport colossal et doit être vue comme la création de caractérisations de référence par rapport à l'immense complexité dimensionnelle humaine. Le développement technologique permet aujourd'hui d'entrevoir la construction d'un (en fait plusieurs) « Homme Moyen » dynamique, en évolution perpétuelle, un prototype plus proche que ne l'était une construc-

7. Cf. (MATTELART, 2003, p. 21-24) pour la mise en application en normes de gouvernement.

tion sur des données statistiques statiques pour ne pas dire forcément vieillissantes, en phase plus étroite[8] avec la réalité. Ces caractérisations permettraient en outre de situer l'altérité précédente et de reconstruire un typage à partir de l'« Homme Moyen » dans un objectif ethnographique de la communication (JUANALS et NOYER, 2007).

En effet, les précédents chapitres se sont attachés à définir l'écosphère à la fois comme une extension du monde mais aussi sa mémoire. D'une part le travail de recherche s'attacherait à l'élaboration de situations de communication non pas de l'Autre mais des Autres et par la reconstitution de situations « moyennes » en vertu de caractéristiques sociales (le typage sociologique, linguistique ou culturel est un point de départ). Disposer de méthodes (φ de construction de capta) de reconstitution établies (discutables, performables, etc.) permet de le maintenir en phase avec la dynamique de ces situations (pour en saisir l'épigénétique) et de construire des représentations de la Communication à partir de caractérisations. L'application n'est pas de rendre compte de la Communication et la caractériser numériquement mais bien de situer et d'explorer les phénomènes communicationnels en regard de considérations moyennes (dynamiques par construction) qui ouvrent à la caractérisation des négociations, phénomène central de la communication (WOLTON, 2014). Ainsi, dans cette construction, le propos est d'élaborer des référentiels permettant de doter d'instruments de caractérisations intermédiaires entre la Communication (l'idéale, celle qui est effective) et l'incommunication (celle qui nous guette à chaque confrontation avec l'Autre) pour élaborer une continuité de vraisemblance entre ces deux opposés. Il s'agit de donner matière à l'élaboration de plusieurs « informations » en situant les messages, les prises de positions, et un moyen de considérer les échanges dans une analyse comparée des communications. Les référents et les fonctions d'herméneutique associées sont contextuels, dynamiques et auto-adaptatifs. Cette esquisse de proposition poursuit un objectif non pas déterministe mais combine une instrumentation analytique pour diagnostiquer les situations avec un potentiel pragmatique qui reste à développer[9].

8. Écart que l'on peut mesurer ou tout au moins estimer le plus souvent.

9. Il ne s'agit pas d'embrasser ici la Communication toute entière mais nombre de situations se prêtent d'emblée à une telle position (qui n'exclut évidemment pas l'éthique essentielle de l'anonymisation au risque de paraître insistant) : la communication des organisations via leur sites web, leur blogs, forums ou, dans leur intranet, les échanges avec les clients, les prospects, les partenaires. Ce peut être aussi en s'appuyant sur les réseaux sociaux par lesquels des constructions statistiques nombreuses sont déjà à l'œuvre. Un domaine très vaste et probablement moins biaisé (cf. mensonges, supercheries, illusions de

La dynamique de l'homme moyen

Si en aucun cas nous n'accepterions d'être catalogué comme homme moyen, nous faisons tous probablement partie d'un groupe qui (ré)agit ou pense de façon similaire à une situation dans un contexte donné. L'essor remarquable du marketing[a] (informations, services ou biens) appuyé par les systèmes de recommandations et leurs applications en sont la preuve : si les considérations ethniques, culturelles, sociales peuvent en être des variables explicatives, la prise en compte seule de ce typage des individus (fabriqué par construction d'un comportement moyen sur des caractéristiques établies[b]) suffit à qualifier le comportement moyen et en prédire une suite. Calquer sur ce modèle un mode raisonné d'élaboration de cadre d'analyse de la Communication pour en extraire les clés de sa performation probabiliste, des règles d'identifications des modalités de communication les plus effectives ou de partage des connaissances.

a. Notamment web mais pas seulement : les points de vente par exemple sont aussi en lice. Cf. l'exemple de l'expérimentation de la station Tesco qui capture le faciès des clients pour leur projeter sur un écran une publicité ciblée selon les caractéristiques physique d'âge et de sexe http://www.bbc.com/news/technology-24803378.

b. Évidement cette approche est très discutable dans certains cas notamment au plan de l'éthique de laquelle tout chercheur doit se faire le Gardien.

3.3 Une démarche non déterministe

La mise en œuvre abstraite précédente a permis de noter une potentielle récursivité du processus se situant à deux niveaux.

3.3.1 Mise en abyme : la vache hilare

En reprenant l'Ω de la section 3.1.3 p. 81 des usagers des sites web, cet ensemble peut à son tour être filtré pour isoler les humains des non-humains. Le flux de logs résultant devient alors la trace des interactions des usagers avec le site considéré. Prendre l'immensité des sites (y compris les réseaux sociaux) comme source d'information potentielle, conduit (CUKIER et MAYER-SCHÖNBERGER, 2013) à pouvoir extrapoler quant

LAMIZET (1995)) est celui de la médecine et des échanges avec les patients souvent transcrits, domaine dans lequel la communication (effective) est indispensable.

à son potentiel expressif :

> *Amitiés, pensées, échanges, déplacements : la plupart des activités humaines donnent désormais lieu à une production massive de données numérisées. Leur collecte et leur analyse ouvrent des perspectives parfois enthousiasmantes qui aiguisent l'appétit des entreprises.*

Sous un autre angle, dans l'hypersphère informationnelle, MERZEAU s'attache aussi à la dimension mémorielle (MERZEAU, 2011, p. 92) tout en la reliant à des algorithmes : « l'hypersphère est un nouveau régime de mémoire qui affecte toutes les activités, lieux et acteurs investis à des degrés divers dans l'agencement du collectif ». L'entrée des « indices » (Ω) dans l'ordre du traitement est la première étape de cette reconfiguration. Elle rapproche les deux pans de la culture que la *graphosphère* avait durablement écartés : celui de l'empreinte (enregistrement, contextualité, courts-circuits, données personnelles, affects... des Ω) et celui du code φ (calcul, normes, déliaison, indexation, monétisation... des φ, dispositifs sociotechniques, des algorithmes). L'opération d'extraction du flux de logs des activités humaines n'est autre qu'un algorithme, simple à opérer et dont la connaissance de son existence seule suffit à étendre la portée d'interprétation des données. Bien sûr ce n'est pas toute la complexité de la vie qui est visée et l'on peut trouver nombre de situations qui ne se prêtent pas à telle caractérisation. L'important est de disposer d'une flexibilité de représentation suffisante qui permette (au sens des DH) d'adresser des phénomènes humains, leur contexte et de les expliciter tel que le proposaient VENTURINI et LATOUR (2010) :

> *digital data are representative only if their processing chain (identification, extraction, integration, analysis, publication) remains close the work of social actors. To be sure, we are not saying that quali-quantitative methods will allow us to smooth out all the complexity of collective life. Quite the contrary, the advantage of these methods is that they are flexible enough to follow some social phenomena along each of their folds.* [10]

10. Traduction personnelle : Les données numériques sont représentatives seulement si leur chaîne de traitement (identification, extraction, intégration, analyse, publication) reste proche du travail des acteurs sociaux. Certes, nous ne disons pas que les méthodes quali-quantitatives nous permettront de limiter toute la complexité de la vie collective. Bien au contraire, l'avantage de ces méthodes est qu'elles sont assez souples pour suivre certains phénomènes sociaux le long de chacun de leurs plis.

3.3.2 Retour de la complexité

Le second niveau relève du fait que si les informations générées ne sont pas suffisantes (Ω, φ) pour la problématique posée, l'exploration peut conduire (souvent) à devoir ré-itérer le processus avec d'autres données, d'autres plans, d'autres filtres, autant de couples (Ω, φ) jusqu'à la conclusion.

Se reconnaît là un processus adaptatif à la complexité inhérente des objets de notre observation. Cette méthodologie de recherche retrouve ses fondements dans la méthode de MORIN et LE MOIGNE (1999), boucle traditionnelle de la rétroaction en cybernétique reprise ici en construction de l'information (HOFKIRCHNER, 2013). Non pas pour les modéliser mais pour les explorer, les expliciter et les comprendre, ouvrant ainsi la voie à décrire ces procédés (Ω, φ) comme des médiations de l'écosystème informationnel.

Le concept rejoint l'épistémologie des SIC (COURBET, 2004) qui « adhérant le plus souvent au postulat du déterminisme, selon lequel il existe des régularités derrière l'apparente complexité de la réalité » par des démarches holistiques, car le couple (Ω, φ) se construit sur le domaine continu dont il constitue un instrument d'analyse, différent d'un quelconque cloisonnement par disjonction, il relève en plus d'une analyse multi échelle opératoire sur de la donnée factuelle. Plus drôle, le même objet peut situer l'Ω au plan même de φ sous-jacents et s'attacher à décrire ou mettre en relation à l'aide d'un autre φ. On passe ainsi, avec le même pseudo-formalisme, à l'étude de **la mise en relation**.

3.4 Médiations ou algorithmes de la communication ?

Au sein des Sciences de l'information et de la communication, la médiation est devenue un champ de recherche important dans le sens où les types de médiation ordonnent la production, la diffusion et l'appropriation de l'information au sein de l'espace public (LAMIZET, 1995). Pour le chercheur, de toute discipline scientifique, son cadre théorique repose sur des « outils intellectuels pour interpréter, donner du sens, produire du savoir, mais aussi pour penser des actes et fonder la pratique »(GARDIÈS,

2012, p. 13).

La médiation est un concept phare des sciences de l'information et de la communication, comme « processus d'échange, de transmission et de traduction ». (RÉGIMBEAU, 2012, p. 76). Le terme désigne « l'espace dense des constructions qui sont nécessaires pour que les sujets, engagés dans la communication, déterminent, qualifient, transforment les objets qui les réunissent, et établissent ainsi leurs relations. » (JEANNERET, 2007). Pour LIQUÈTE, se définir comme le lien entre l'énonciateur et le récepteur (LIQUÈTE, FABRE et GARDIÈS, 2010) :

> *la médiation met en place, grâce à un tiers, des interfaces qui accompagnent l'usager et facilitent les usages. Elle permet de créer un lien et de concilier deux choses jusque là non rassemblées pour établir une communication et un accès à l'information. La médiation lorsqu'elle s'appuie sur des dispositifs matériels ou humains en capacité de lier information et communication, peut être qualifiée de médiation documentaire*

En conséquence, la définition du φ est une construction relevant des SIC, devant être explicitée pour contextualiser les Ω produits et sources. La médiation documentaire exprime directement ce point en prenant (RÉGIMBEAU, 2012)é : < appui sur le traitement documentaire basé sur l'usage de normes professionnelles universelles visant un usage collectif, elle s'oriente aujourd'hui vers la mise en place de dispositifs techniques et humains plus complexes qui incluent des réécritures de l'information, revisitant ainsi les formes médiatrices dans les pratiques professionnelles à partir des composants d'un processus de communication prenant pour objet l'usager, les idées, le contexte technique, les contenus et la pratique (sociale, économique, politique...). [Parmi ces visions], l'une focalise les pratiques professionnelles et celles-ci, dans la lignée de G. RÉGIMBEAU peuvent se définir par « ses intérêts sur le document, l'autre, sur une catégorie particulière d'information, et la dernière, sur le processus humain.

3.5 Artefacts de la médiation

La figure 3.2 p. 93 représente une catégorisation des φ supports à la construction d'information de niveau différents. À partir de noumènes ou de diaphories, la première étape consiste en leur collecte (production de captas) puis leur sélection (filtrage) ré-

sultant en un Ω_1 issu de cette première transformation. Un second type de traitement vise à agréger ces éléments informationnels, des traitements descriptifs élaborés sur des mesures informationnelles tel que les dénombrements statistiques. Le traitement est endogène en ce sens qu'il ne construit qu'un élément informationnel du même plan que les données en entrée, c'est simplement un processus qui permet de les appréhender. Par exemple, à l'issue de ces deux seuls traitements l'on retrouve les indicateurs classiques de scientométrie (sélection des productions d'un auteur ou d'un journal, dénombrement quantitatif de la production) dans l'ensemble donné des productions scientifiques.

Puis, une « augmentation » des données se construit en convoquant d'autres bases informationnelles qui viennent en complément. Elles se veulent explicatives (traitement exogène) ou participantes (complément endogène) d'un objet plus élaboré. Se retrouvent à ce niveau les ensembles terminologiques, les ontologies, ou plus simplement d'autres données issues d'autres collectes et sélections. À partir de deux ensembles Φ_k et Φ'_k de même niveau k, l'on retrouve les principes de corrélation. Dès lors que Φ_k et Φ_j pour $j \neq k$, il s'agit d'une mise en relation de données. Catégorisation, classification se retrouvent dans ce cadre faisant passer le niveau informationnel d'un niveau k (les données d'entrée) à un niveau $k+i$, celui des catégories.

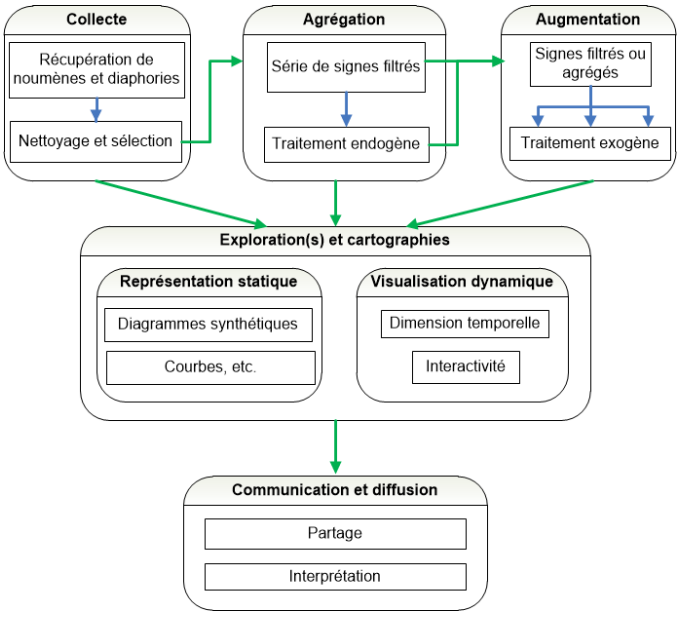


FIGURE 3.2 : L'artefact médiateur. Les opérateurs de construction de sens et les boucles de rétroaction sont omises.

3.6 Limites, portée et perspectives

3.6.1 Limites

J'ai précédemment énoncé quelques limites du concept des informations. La première se situe aux plans mathématique et informatique au niveau de l'insuffisance de formalisation du concept, des règles d'application et des espaces de conceptualisation. C'est une faiblesse majeure pour son utilité et un frein à son usage en mathématiques.
Réciproquement, nul besoin non plus de changer les pratiques en SIC. L'explicitation du φ ne doit pas forcément s'en tenir à un formalisme bourbakiste. La plupart des travaux rapportent une méthodologie rigoureuse de construction d'échantillon, de questionnaires et de leurs objectifs, de collecte et de construction de corpus soignés. Que ce soit dans le domaine de l'intelligence économique, de l'étude de la communication, de l'infométrie et la bibliométrie en général, les applications sont d'évidence, sans que soit posé a priori le besoin d'un autre formalisme que celui de la rigueur scientifique dans la construction des corpus, l'exposition des résultats et leur analyse.

3.6.2 Portée

Le pseudo formalisme introduit ouvre la perspective de formalisations plus abouties esquissées précédemment. L'enjeu en est de situer la réflexion sur les constructions humaines du faire sens des données, la mise en relation des données, des informations, et de leur usages. Une version plus aboutie de ce formalisme ouvre la perspective de nous faire guider par les machines sur la base d'un faire-sens en émanation de la sphère sémantique collective de Pierre LÉVY, s'inspirer des émergences construites par l'humain pour analyser les données humaines.

Les notions centrales des SIC, telle l'identité (marque, organisationnelle ou individuelle), de la mobilité (diaspora, culturelle, professionnelle, ...), par exemple, sont représentatives d'un ensemble de questions vives de caractérisation des formes sociales, culturelles, organisationnelles et info-communicationnelles actuelles dont l'écosphère est le témoin et quelquefois la manifestation. Les mettre en regard, les assembler per-

met de réfléchir sur des continuités et discontinuités conceptuelles mais aussi empiriques qui sont au centre des réflexions des chercheurs en communication. En conséquence, l'enjeu est décisif quant à l'élaboration d'instruments et artefacts de médiation alliant la collecte d'information pour la reconstruction de faits établis de ces notions à l'exploration des résultats d'un fait observable quantifiant l'épaisseur symbolique et médiatique de la notion, relativisant la portée et l'incidence, laissant espérer la généralisation sous réserves que l'on pourra alors cerner...

Dominique WOLTON propose un scission en parlant des sciences de la communication (2012, p. 14), je propose à l'inverse une fusion. Peut-être devrions nous dire « la science des informations » pour les SIC?

II

Application à la documentation brevet

CHAPITRE 4

Le document brevet

Les inventions appartiennent à la société pour le moins autant qu'aux inventeurs, et nul ne peut y prétendre un droit absolu et éternel

— C. COQUELIN, Dictionnaire de l'économie politique, p. 215 (1873)

La distinction qui est habituellement établie entre science et technique ne tient pas, car dans un cas comme dans l'autre on a fondamentalement affaire aux mêmes processus et aux mêmes opérations ; il s'agit d'un seul phénomène qu'on peut appeler « technoscience ».

— L. QUÉRÉ, Les boîtes noires de B. LATOUR ou le bien social dans la machine ((1989))

4

Par essence l'enregistrement de demandes d'inventions mémorise une diaphorie de la technoscience en inscrivant des traces de l'activité d'innovation (Griliches, 1998 ; de Kermadec, 2001). Les compter détermine une mesure de l'activité technologique (Barré et Laville, 1994 ; Hidalgo-Nuchera, Iglesias-Pradas et Hernández-García, 2009). Agréger ces dernières par domaine technologique permet de cerner des signes de la production de connaissance (Gibbons et al., 1994). En établir les origines géographiques et les évolutions temporelles reconstitue un témoin de la circulation des connaissances (Moed et Halevi, 2014). Au niveau du document, l'invention est une source d'informations utiles pour catalyser la créativité (Dou, Haudeville et Wolff, 2015) et penser « hors de la boîte ». Un corpus ciblé de ces documents leur confèrera le statut de ressource documentaire de l'information stratégique (Oubrich et Barzi, 2012). De plus, les brevets étant régulés notamment par les offices nationaux de brevets, les demandes portent des modalités de mise en œuvre de la propriété intellectuelle dans un système juridique international (Granstrand, 1999).

Le brevet est généralement perçu comme un document consignant la protection d'une invention. D'une part moins de 5% des brevets sont actifs et constituent une barrière à l'utilisation mercantile enpêchant notamment la fabrication de l'invention qui est décrite [1] et, d'autre part, les inventions sont décrites par les documents brevets, et rarement ailleurs. Ce dernier point implique que les documents brevets sont une source informationnelle technologique unique. L'épigraphe de Coquelin qui œuvra pour la construction du système brevet au 18e siècle rappelle ce point fondamental : le pendant pour obtenir le droit de la protection d'une invention est de fournir la description d'une invention, puis de l'offrir à l'humanité.

1. En aucun cas cette règle empêche de *composer avec*, bien qu'il faille en acquérir les droits pour produire, pas pour innover en une nouvelle invention.

Le contenu décrit par le brevet est généralement, par nature, inventif, nouveau et une réponse à un problème donné. En ce sens, il offre une gamme informationnelle non négligeable, esquissée précédemment, couvrante pour la panoplie des catégories d'informations de la partie précédente. Le second épigraphe de QUÉRÉ propose de situer la documentation brevet au même titre que la documentation académique notamment pour l'étude générique des sciences, la *technoscience* de LATOUR. J'ouvrirai en ce sens quelques perspectives de recherche en montrant que le brevet constitue un objet d'information pour les SIC (QUONIAM, 2013).

Au cours de ce chapitre, les aspects techniques et formels du document brevet, de sa création, son dépôt et sa vérification seront rappelés et présentés sous l'angle des informations qu'ils véhiculent et pour lesquelles je soulignerai notamment la qualité des informations consignées (cf. 4.1.3 p. 105). Les vérifications, amendements et constructions collaboratives impliquent un risque moindre d'extrapolation des contenus et des descripteurs ouvrant à des usages très variés et en pleine expansion depuis que la libération des bases sur internet promeut une encyclopédie technologique de l'Open-Data (cf.4.2 p. 111). Je montrerai que la captation des *empiries*, que ce soit au plan des contenus ou des métadonnées associées, intéresse non seulement de nombreux plans académiques tels l'infométrie, l'intelligence compétitive, l'économie et le marketing, mais aussi au plan opérationnel la veille informationnelle des petites et grandes entreprises. Ainsi, la documentation brevet se démarque par une prédisposition à une approche académique complète : depuis la recherche d'amélioration de techniques de traitement, d'application et d'extraction d'informations et d'accompagnement des usages et de médiation.

Paradoxalement cette documentation est sous-utilisée dans le monde académique (DURAND-BARTHEZ, 2013 ; QUONIAM, 2013 ; QUONIAM, KNIES et MAZIERI, 2014), alors que cette envergure utilitaire souligne l'intérêt d'une médiation attentive à cette documentation par les sciences de l'information et de la communication : l'ampleur applicative est balancée par une complexité notoire tant au niveau administratif, que technique ou terminologique (PARANJPE, 2012). L'ensemble porte l'enjeu de ces recherches en SIC à un niveau d'intérêt sociétal majeur (cf. note 118).

4.1 Le document brevet

4.1.1 Généralités

D'un point de vue contenu, le document brevet présente une invention sur le plan technique, ainsi que ses revendications (usages protégés) au plan juridique (OMPI, 2015, p. 3) en réponse à deux fonctions : la **protection des droits du déposant** et la **divulgation de l'invention**. Le dépôt d'une demande est réalisé en échange, pour le déposant, de la protection de la fabrication/diffusion/revente dans une zone géographique et pour une durée prédéterminée généralement à 20 ans, de l'invention sous couvert qu'elle en soit bien une.

4.1.2 Cycle de vie

Si l'invention est généralement l'aboutissement d'une idée, individuelle ou collective, le dépôt d'une demande de brevet est le point d'entrée d'une série d'étapes et d'interactions avec un office de dépôt et ses différentes instances qui instruisent, in fine, une **construction collaborative du document brevet**. Ce dernier se construit progressivement en une série de documents structurés et associés à des métadonnées descriptives (BONINO, CIARAMELLA et CORNO, 2010), série qui marque les étapes vérifiéés par l'office de dépôt.

Le tableau 4.1 représente la progression constitutive d'un document brevet dans le temps, et les différents éléments de sa composition au fur et à mesure de l'avancement de la procédure de publication.

TABLE 4.1 : Les différentes étapes de la publication de demandes

Phase	**Étape**	**Données précisées**
Demande	Date de dépôt (*filing date*)	Quelques métadonnées
Demande publiée	18 mois après la date de dépôt	Texte complet et métadonnées

Brevet obtenu	Deux ans après la date de dépôt	Texte complet et métadonnées amendés
Brevet expiré/cédé	20 ans après la date de dépôt (ou avant)	Texte complet et métadonnées amendés

4.1.2.1 Dépôt de demande et délivrance de brevet

Une demande peut suivre trois voies (OMPI, 2015, p. 4) : la première est nationale pour des effets uniquement dans le pays dans lequel la protection est demandée, la seconde régionale[2], et le traité de coopération (PCT) définit depuis 1978 une procédure internationale. La figure 4.1 représente la procédure type des différentes étapes génériques selon la voie de dépôt et (plus ou moins suivi) selon les offices (ibid., p. 5-7).

Document collaboratif

Le document brevet est issu d'un long processus d'échanges entre l'office de dépôt et le déposant, convoquant expertises, vérifications et validation.

La procédure synthétisée par la figure 4.1 souligne d'une part la durée relativement longue de l'examen d'une demande et, d'autre part, la mobilisation de ressources d'expertises et de vérification des contenus qui conduit à la collaboration constitutive du document final, son contenu ainsi que les étapes clés de cette construction.

La procédure de dépôt varie selon les offices nationaux ou régionaux, le lecteur intéressé suivra un approfondissement de ces questions relevant de subtilités selon les offices et les pays (cf. la procédure de dépôt à l'INPI : (*Le formulaire brevet. Comment remplir votre dossier de dépôt de brevet ?* 2015)), mais aussi selon les types d'inventions, que l'on protège différemment selon les pays (OMPI, 2015 ; QUONIAM et REYMOND, 2013 ; ADAMS, 2012 ; BOUGRINE, 2001). Ces éléments dépassant le cadre des besoins pour l'approche proposée, nous en restreignons la présentation à la procédure internationalisée, représentative à quelques détails près des procédures nationales et ré-

2. La notion de région dans le domaine correspond à un regroupement de pays d'une région du monde (Europe, Afrique, ...)

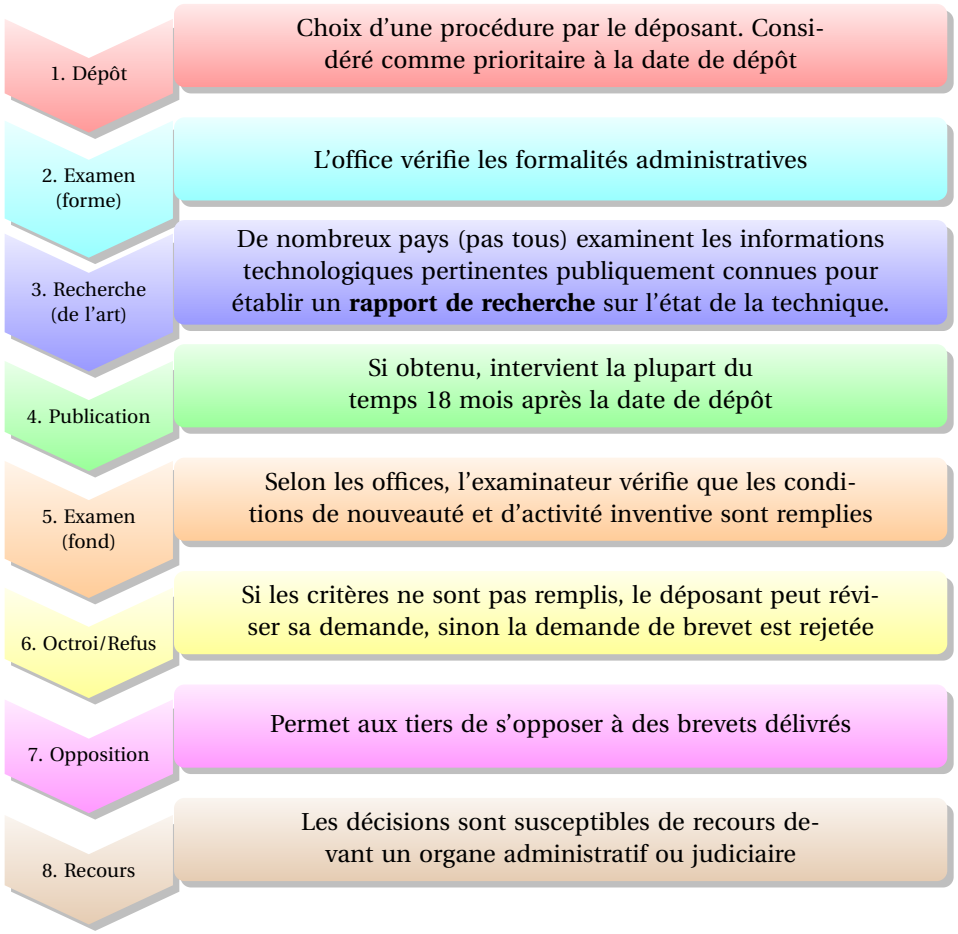


FIGURE 4.1 : Procédure de délivrance des brevets. Le schéma montre les différentes étapes à droite et leur description d'après (OMPI, 2015, p. 5-7).

gionales, qui suffit à décrire les éléments d'intérêt pour le raisonnement suivi.

4.1.2.2 Procédure de dépôt internationalisée

Depuis 1978, le traité de coopération établit une procédure (relativement suivie) au niveau international qui représente un consensus de méthode de dépôt (GURRY et FINK, 2016, p. 11) sur lequel je focaliserai pour cerner globalement cette riche procédure d'instruction des documents brevet. La figure 4.2 présente une vue d'ensemble du système du traité de coopération (*Patent Cooperation Treaty* - PCT). En règle générale les déposants suivent d'abord une procédure nationale (qui reprend le processus de la figure 4.1) puis l'étendent, dans un délai de 12 mois à compter de la date de priorité, à une demande internationale. Cette phase entend une procédure d'expertise

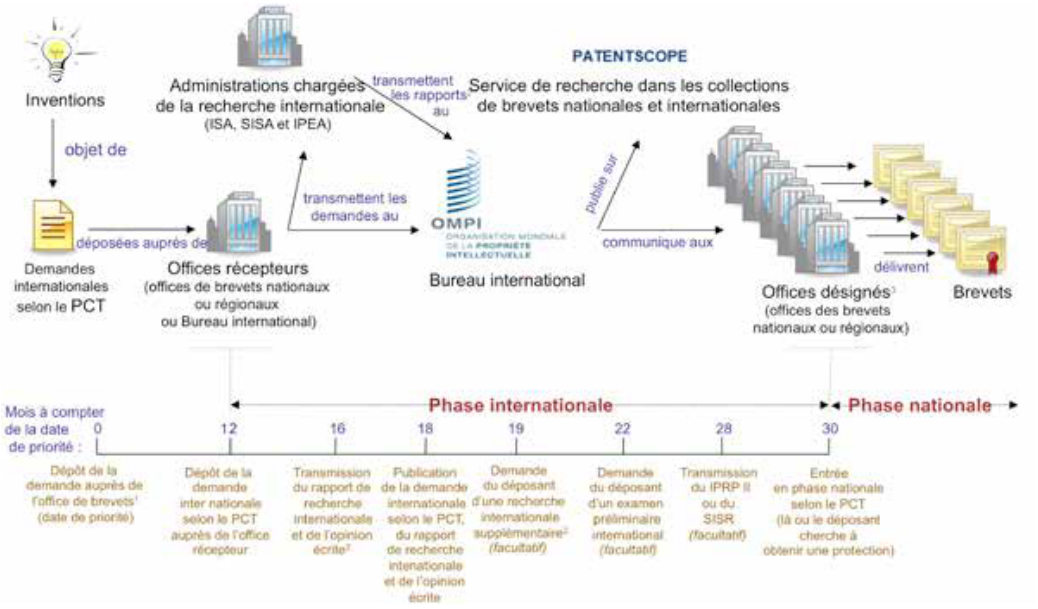


FIGURE 4.2 : Vue d'ensemble du traité de coopération internationale d'après GURRY et FINK (2016, p. 12)

et de recherche internationale[3] à l'issue de laquelle les administrations transmettent les rapports (un rapport de recherche ou de brevetabilité) aux offices désignés (ibid., p. 12) qui viendront compléter la demande initiale.

4.1.3 Document structuré

Les différentes étapes de la procédure de dépôt, de la collaboration avec les offices et l'enregistrement de données à chacune de ces étapes impliquent l'enrichissement progressif et collaboratif du document. Au niveau de son contenu, l'imposition d'une structure poursuit un effort de normalisation renforçant la richesse du document.

3. Processus couvert par trois administrations faisant autorité internationale ralliant trois différentes procédures d'expertise : l'examen préliminaire instruit par l'IPEA (*International Preliminary Examining Authorities*), la recherche internationale cautionnée par l'ISA *Authorities as International Searching*, puis la recherche supplémentaire couverte par le SISA (*Supplementary International Searching Authority*).

Richesse structurelle normalisée

Au-delà de l'appréciation qualitative du contenu, le processus collaboratif vise à enrichir le document brevet par une granularité descriptive des différentes parties du document :

- découpage des contenus en sous-documents spécifiques (résumés, revendications, description) ;
- recommandations de rédaction des contenus ;
- descriptions normalisées (inventeurs, déposants, etc.).

4.1.3.1 Contenu

Au plan de son contenu, un brevet standard est normalisé par les offices qui imposent un aspect formel de présentation de la page de couverture représenté par la figure 4.3 appuyé sur un « plan de rédaction ». Ce plan est structuré selon les éléments suivants : page d'accueil, description, revendications, dessins (optionnel) et rapport de recherche internationale (DE KERMADEC, 2001).

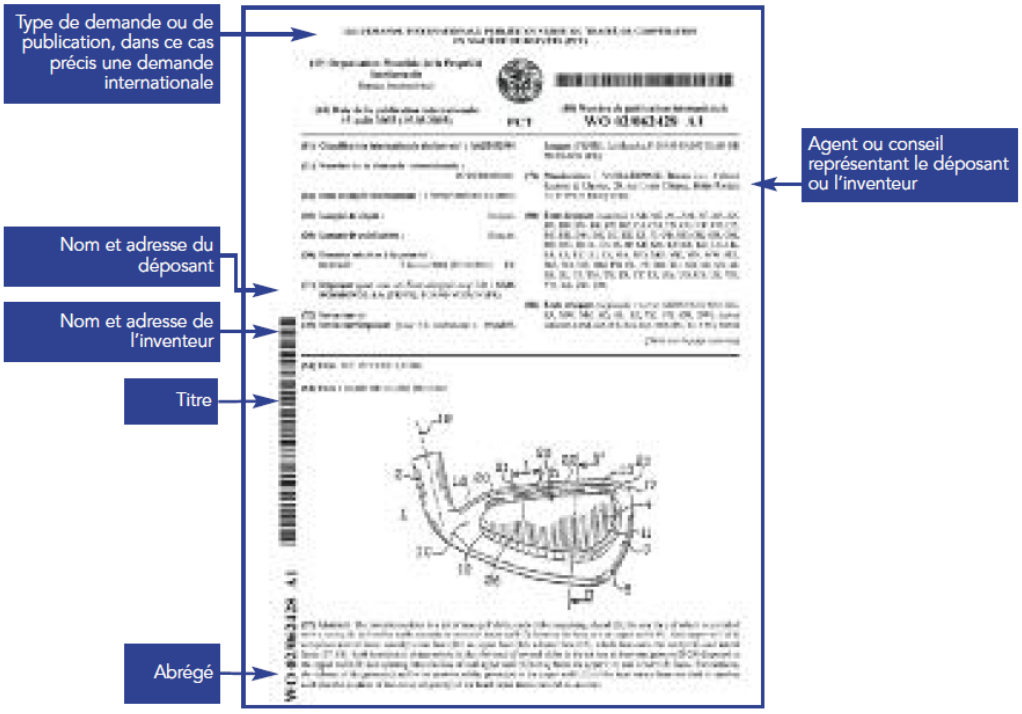


FIGURE 4.3 : Description de la page de couverture type d'un document brevet (OMPI, 2013, p. 3).

L'Institut National de la Propriété Intellectuelle (INPI) recommande un plan type de rédaction des contenus (*Le formulaire brevet. Comment remplir votre dossier de dépôt de brevet ?* 2015, p. 8) : le premier paragraphe rapporte le cadre général de l'invention (domaine technique), puis le second paragraphe l'état de la technique antérieur, puis vient l'exposé de l'invention... Les documents brevet peuvent comporter deux types d'information, à savoir « l'information d'invention » et « l'information additionnelle » qui complète l'innovation technologique.

Richesse technologique

Les inventions brevetées des nouveautés, la description d'une invention doit être suffisamment informative de sorte à pouvoir être reproduite par quelqu'un de l'art.

4.1.3.2 Description des documents et des étapes

De même, la normalisation est imposée sur les données descriptives de ces documents, des « métadonnées standards » : les titre, date, numéro de document, numéro de publication, liste de déposants et d'inventeurs, domaines (voire sous-domaines) technologiques, citations et références. La page de données bibliographiques d'une demande internationale voit aussi sa représentation standardisée (cf. figure 4.4) autour de ces contenus normalisés.

4.1.3.3 Métadescription : schéma de classification internationale

La classification internationale s'appuie sur un schéma de classement hiérarchique normalisé. Le schéma est révisé périodiquement tous les cinq ans et conçu pour permettre un classement uniforme à l'échelle internationale des inventions. Les libellés du schéma de classement sont décrits en français et en anglais (OMPI, 2016, 2, §2) et l'objectif de l'expertise de ce classement est celui de la recherche d'information technique (ibid., 23, §75). Ainsi, la règlementation internationale (l'Arrangement de Strasbourg de 1971) impose à toute demande d'invention d'être classée a minima dans un domaine (ibid.) qui ouvre la voie à des procédures de recherche indépendantes des

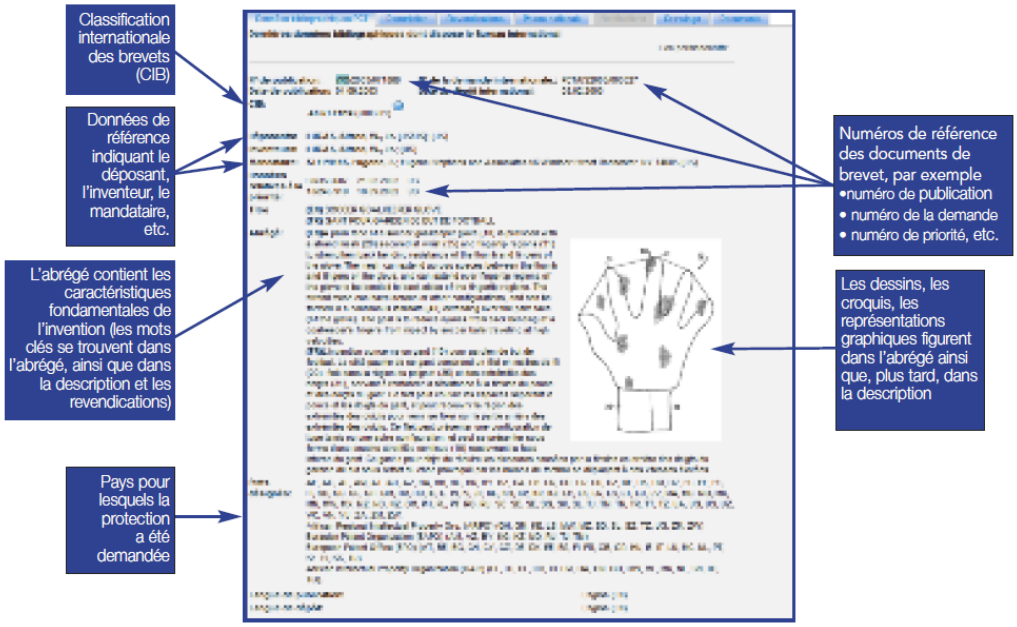


FIGURE 4.4 : Description de la page de données bibliographiques d'une demande de brevet internationale.

langages (CHEN et CHIU, 2013) en offrant un pivot consensuel. La Classification coopérative [4] (*Cooperative Patent Classification*-CPC) entend progressivement améliorer le système de classification de l'IPC depuis janvier 2013 en affinant la hiérarchie du schéma de classement tout en rassemblant le schéma de classement des deux principales bases (américaine - UPSTO et Européenne-EPO).

Classé dans un registre international

Le document brevet est classé par les auteurs et les experts des offices dans un système de classement international rigoureusement entretenu par un cycle performatif mobilisant une large communauté internationale. Ce classement permet une lecture par « facette » indépendante de la langue de description de l'invention.

4.1.3.4 Documents enrichis et vérifiés

L'enregistrement numérique de chacune de ces étapes en différentes versions du document, le découpage structuré des contenus eux-mêmes décrits par des métadonnées conduit à l'obtention de documents semi-structurés. En effet, en interne les conte-

4. Cf. http://www.cooperativepatentclassification.org/index.html.

nus sont amendés, révisés et progressivement enrichis par les experts et les auteurs. Leur description bibliographique qui intègre la dimension temporelle est vidée au plan administratif pour chacune des étapes les offices de dépôt. En « externe », le classement, méta-descripteur de l'invention, est expertisé et souvent complété manuellement par les spécialistes. Lors de l'examen de la demande, les experts construisent un rapport de recherche complémentaire qui veille à l'inscription de la demande dans l'état de l'art (classement et citations) (MICHEL et BETTELS, 2001) pour décider si la demande répond aux critères de brevetabilité.

ALCACER et GITTELMAN (2006) remarquaient qu'au sein de la base européenne des brevets, 40% des brevets comportent des citations ajoutées par les examinateurs et, en sus, qu'en moyenne, les deux tiers des citations sont insérées par les examinateurs comme l'avaient remarqué (A. B. JAFFE et TRAJTENBERG, 2002, p. 390). L'insertion de références à d'autres documents construit un positionnement relatif des documents et, en ce sens, des technologies sous-jacentes, processus comparable à l'architecture informationnelle de Wikipédia [5]. Bien que des différences statistiques importantes aient pu être constatées au sein des offices, pour lesquels les modalités législatives de dépôt sont mises en cause [6]. Les citations ajoutées par les examinateurs constituent l'élément informationnel primaire de nombreux travaux de recherche (MEYER, 2000). Les citations réalisées par les examinateurs de l'Office Européen des Brevets sont typées : X (référence de l'art antérieur posant le doute sur une revendication), Y (référence comme X qui demande la conjonction avec d'autres documents), A (références de l'état de l'art), D (références combinant les types A, X ou Y) (HARHOFF et WAGNER, 2009). La documentation de recherche de l'OEB englobe la littérature brevet ainsi que tout autre élément imprimé ou non imprimé pouvant présenter un intérêt technique pour la procédure de délivrance des brevets. La littérature non brevet (*non patent literature*) est typée XP. Récemment, l'inclusion de citations par URL vers des ressources du web marque une évolution de la référence « non brevet » bien qu'encore fragile en

5. À la différence de format et de qualité prêt : la relation entre les technologies n'est pas en format hypertexte natif, elle est cependant construite par des experts qualifiés.

6. Lors du dépôt d'une demande auprès de l'office américain UPSTO, les déposants doivent faire figurer un ensemble de références exhaustives sous peine que l'invention soit refusée (MICHEL et BETTELS, 2001, p. 190). Ceci entraîne un biais de la part des déposants qui préfèrent citer un brevet éloigné plutôt que de livrer une demande incomplète. À l'office européen, l'exhaustivité n'est pas exigée, les références complémentaires sont ajoutées par les examinateurs.

tant que source informationnelle (ORDUNA-MALEA, THELWALL et KOUSHA, 2017).

4.1.4 Document scientifique ?

Est-ce que le brevet constitue un document scientifique ? Dans l'absolu il est évident que non : si la condition de brevetabilité garantit l'unicité de la réponse à un problème et sa nouveauté, rien n'oblige à ce que la solution soit viable. Nombre d'exemples de demandes de brevet proposent ainsi des inventions farfelues s'opposant aux lois fondamentales de la physique. D'autres encore sont insignifiantes en apport technologique (cf. figure 4.8) mais peuvent peut-être déclencher des idées. Cependant, la marginalité de ces dernières demandes fait considérer les brevets comme une ressource incontournable de la documentation technique (JÉRÔME, 2006).

Inclure cette documentation en complément de la documentation académique tel que le proposent QUÉRÉ, LATOUR, CALLON, Mac Millan (2006) soulève, au-delà des évidentes oppositions recherches fondamentales / recherches appliquées, tout au moins la question de la représentativité. En effet, breveter une invention n'est pas la seule voie pour les industriels : l'on peut garder secret des inventions (COUTENCEAU et BARBARA, 2014, p. 70) pour notamment ne pas alerter les concurrents Ainsi, la description de certaines inventions échappent à leur publication. Soit. Pouvons-nous garantir l'exhaustivité en recherche académique ? Je prends pour preuve du contraire les publications « post mortem » qui, quelquefois, révèlent des travaux inédits et d'envergure.
Ensuite, l'on pourrait discuter de la séparation entre la recherche académique et le domaine brevet, une forme de réciproque de la question précédente. Le brevet comme voie de publication pour le monde académique a été posée depuis la mise en marche de politiques initiées par le *Bayh Dole Act* en 1980 aux USA qui donna alors le droit aux universités de faire breveter des inventions issues d'un financement public. L'État Français soutient également une politique de valorisation par l'intéressement financier[7] des chercheurs et des établissements : le décret du 13 février 2001 a modifié le décret du 2 octobre 1996 et porte de 25% à 50% le complément de rémunération versé à l'agent sur les sommes perçues chaque année par l'entité qui a déposé le brevet. Le

7. Cf. le rapport d'information du Sénat (ADNOT, 2006).

décret du 26 septembre 2005 ajoute une prime d'intéressement au brevet d'invention. Enfin les programmes d'investissement en cours et à venir visent à promouvoir la valorisation de la recherche publique et privée, notamment par l'exploitation de brevets regroupés de façon cohérente tel le fonds souverain *France Brevets* (« Rapport d'activité 2015 » 2016, p. 21).

Le débat a été posé (CAPART, 2006) en Europe et reste encore ouvert en soulevant des questions éthiques. Les universités sont invitées à pratiquer la protection du savoir académique et à le monnayer (VAN OVERWALLE, 2006). Position politique que l'on peut opposer à la mission fondamentale de la recherche qui, en tant constructeur du savoir, se doit de le diffuser à la société. De fait, certains établissements s'impliquent plus que d'autres (JÉRÔME, 2006) par les questions éthiques que ces choix soulèvent (MAY, 2006) : non seulement ce choix laisse planer l'ombre du financement de la recherche via les brevets (CAPART, 2006) et, d'autre part, protéger une technologie peut constituer un risque de poser des verrous à l'innovation (BELLEFLAMME, 2006).

D'un point de vue ressource documentaire, les écueils précédents peuvent être évités en considérant la documentation technique et la documentation académique comme complémentaires au plan scientifique,ce qui permet d'approcher l'exhaustivité : science et technologie se complétant mutuellement. D'autant plus que la corrélation entre l'activité d'invention et le degré de connaissance scientifique a été établi par NARIN, HAMILTON et OLIVASTRO (1997; 1991).

4.2 La base européenne des brevets, une source de l'open

Parmi les bases de données « brevet » librement accessibles sur le web (WIPO, 2009, p. 5-6), *EspaceNet* est un service de recherche offert par l'Office Européen des Brevets (OEB ou *European Patent Office* - EPO). *EspaceNet* permet d'effectuer des recherches textuelles sur les données bibliographiques et légales des brevets. La base de données mondiale centralise à ce jour plus de 89 bases de données issues de 50 organisations, affiliées à l'Office Mondial de la Propriété Industrielle (OMPI), faisant autorité.

4.2.1 Une base très complète

Pour être acquise, une invention ne devant pas être décrite par ailleurs, le contenu de cette base est en conséquence disjoint du reste du web et des autres sources de documentation. L'OEB concentre, selon une temporalité variée par la fréquence de mise à jour entre les bases des offices de dépôt adjoint, une majorité des demandes de publication d'invention.

Représentativité

Il s'agit de la plus grande collection libre d'accès de documents brevets complets (de la demande à leur acceptation et les rapports de recherche afférents) qui compte plus de 90 millions d'entrées datant de 1836 à nos jours.

EspaceNet a ouvert en 2005 un service web offrant l'accès des robots de collecte (*Application Programming Interface* (API)) à leurs bases de données (KALLAS, 2006).

Ouverte et centralisée

Cette base de données structurée peut se voir comme une gigantesque encyclopédie technologique, concentration des inventions de l'humanité en un ensemble documentaire gigantesque.

Ce service web, appelé *Open Patent Service*, offre la possibilité de télécharger gratuitement les notices et les contenus de la documentation brevet à des fins d'analyses locales jusqu'à 2,5 Go de données par semaine.

4.2.2 Du brevet aux « commons », qu'un pas !

Si la base contient évidement des technologies protégées, une majeure partie de celle-ci est libre de droits (HIRATA et al., 2015).

Constituant d'envergure de l'open-data

La taille, son unicité et l'accès libre aux données descriptives de la documentation brevet en font une source non négligeable du phénomène de l'Open Data, et plus précisément, de l'open-knowledge.

La taille de la base, son unicité et l'accès libre aux données descriptives de la documentation brevet en font une source non négligeable du phénomène de l'Open Data. De surcroît, l'origine fondatrice du dispositif d'acquisition de brevet souligné par l'épigraphe de COQUELIN, par cette phase qui transforme les descriptions des inventions en biens publics au bout d'un temps variable selon des conditions (acquittement des coûts, présence territoriale, durée maximale) en un bien commun de l'humanité sont **a contrario des dictionnaires, encyclopédie et autres jusqu'à ce jour négligés.** La description des inventions, enrichie de relations internes aux autres documents brevets, et aussi externe en référence à la littérature académique, est ainsi partie prenante du volet connaissance des *commons* au sens d'Ostrom (HESS et OSTROM, 2007). Les auteurs proposent un modèle d'analyse pour comprendre la connaissance comme un système écologique social partagé pour lequel la base de données en accès libre libère un potentiel inédit. Libre d'accès, la base des brevets fait donc partie des *Commons*. L'OEB dispense la base la plus vaste à ce jour et la plus représentative du domaine de la documentation brevet.

4.2.3 Une base documentaire inexploitée

Les travaux portant sur la documentation brevet ciblent pour l'essentiel les professionnels de cette documentation laissant de côté les programmes d'éducation (DURAND-BARTHEZ, 2013). Le contexte historique de la recherche d'information au sein des brevets (ADAMS, 2012) marquée par des bases de données commerciales d'accès très coûteux et le manque d'outils pour les utiliser sont une des principales raisons de ce manque. Il convient de développer l'instrumentation d'une herméneutique associée à sa médiation pour viser à lui convoyer une portée sociale à la hauteur de l'origine même du processus d'acquisition de brevets. La section suivante montre les aspects

usages pour en inférer les fonctionnalités nécessaires.

4.3 Usages du document brevet

Le document brevet est, par définition, une transcription détaillée d'une invention pour protection de revendications et issue de résultats d'un processus de recherches. Au-delà des activités professionnelles autour des brevets, pour valider ou faire invalider un brevet, estimer sa valeur[8] (VAN ZEEBROECK, 2011 ; « SME tailor-designed patent portfolio analysis » p.d.), protéger et valoriser son portefeuille (etc.) le document brevet offre pléthore d'informations utilisables dans des contextes très variés (JAKOBIAK, 1994).

4.3.1 Usages autarciques

Le cycle de vie de dépôt d'une demande sous-entend des procédures de consultation des bases de brevet. La figure 4.5 reprend les principales phases de la procédure de dépôt (cf. figure 4.1.2.1 et le tableau de construction du document dans le temps : tableau 4.1 p.102) en mentionnant les acteurs clés (en vert) des phases d'instruction de la procédure et en opposant la partie « publique » de la partie « secrète » de la procédure. Ainsi, avant même la phase de demande, la consultation des bases brevet est de mise pour vérifier la nouveauté d'une idée. Le chercheur peut lors de cette phase établir ce point mais aussi, pour identifier des solutions sans faire de développement, positionner une stratégie de recherche, ou encore définir un programme de recherche et développement. En ce sens, poser un état de l'art technologique. La phase de rédaction que l'on suggère de faire accompagner par un juriste, peut demander aussi de procéder à des recherches dans la base : la rédaction des revendications de l'invention peut s'appuyer (et doit se démarquer) des revendications d'autres inventions.
La phase administrative de recherche d'antériorité s'appuie sur des experts qui situent l'invention potentielle dans les publications[9] antérieures. Les experts instruisent le

8. Dans la plupart des travaux que j'ai pu parcourir celle-ci est monétaire. La valeur d'un brevet libre de droit répondant à un problème sociétal n'est évidement pas ciblée par ces travaux.
9. Pas forcément des écrits d'ailleurs. Selon les offices l'antériorité peut se jouer à des échanges oraux...

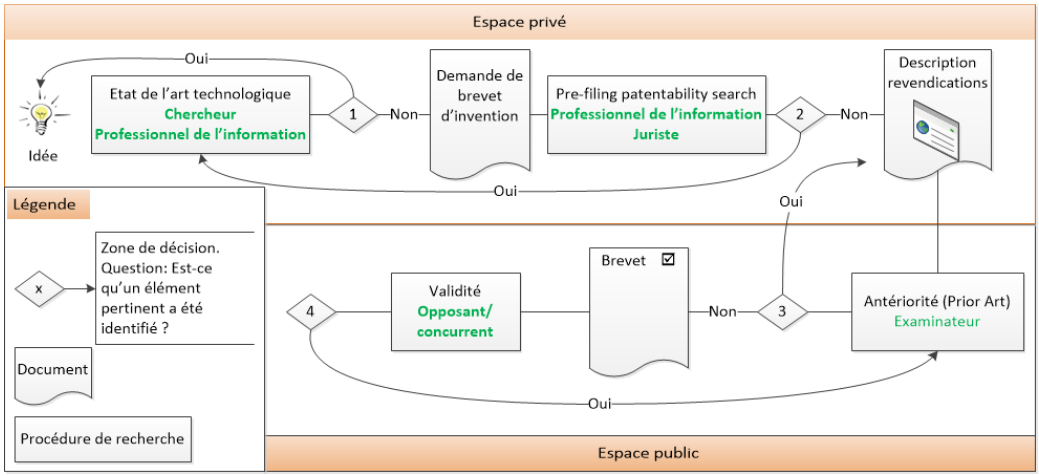


FIGURE 4.5 : Vue de haut niveau du cycle de vie d'une idée à breveter et implication d'acteurs différents, d'après (LUPU et HANBURY, 2013, p. 15).

document en associant des citations complémentaires et révisent éventuellement le classement. La dernière phase de la figure se situe après la validation du brevet par l'office. Cette phase permet à autrui d'attaquer la nouveauté ou l'aspect inventif en démontrant qu'il s'agit d'une contrefaçon ou qu'elle entre dans le champ d'un brevet déjà existant.

Ainsi, au niveau de l'instruction technique du document, chercheurs, examinateurs, juristes, documentalistes sont les acteurs de cette activité de documentation dont les usages se différencient par évidence selon leur objectifs de recherche d'information. Enfin, la procédure d'opposition et d'annulation combine chaque type d'acteurs.

4.3.2 Usages exogènes et étendus

En complément des activités en propriété industrielle précédentes prises au sens professionnel, l'appréhension de la documentation brevet s'articule sur deux plans : l'information directe extraite des documents ou de leur description, et l'information indirecte tirée de l'analyse statistique. L'information comprend (OMPI, 2013, p. 3) :

- des informations techniques provenant de la description et des dessins de l'invention ;
- des informations juridiques provenant des revendications définissant la portée du brevet et de la situation juridique (validité) du brevet dans certains pays ;

– des informations d'ordre commercial provenant des données de référence relatives à l'inventeur, à la date de dépôt, au pays d'origine, etc. ;
– des informations pertinentes pour la politique des pouvoirs publics provenant de l'analyse de l'évolution des dépôts et susceptibles d'être utilisées par les responsables de l'élaboration des politiques, par exemple dans le cadre de la stratégie industrielle d'un pays.

MONFORT-WINDELS (2008) développe les modalités de l'analyse de l'information brevet pour ces usages. L'auteur adresse évidement l'intelligence économique, l'analyse de documents brevet et la plupart des items d'information précédents qui incitent à produire des informations stratégiques propre à alimenter des processus de veille très variés (OUBRICH et BARZI, 2012).

4.3.3 Le brevet pour l'intelligence compétitive

Dans un domaine organisationnel technique donné, la recherche dans la documentation brevet permet d'identifier (MONFORT-WINDELS, 2008) :

– qui est propriétaire des technologies ?
– quels sont les inventeurs clés ?
– quelle est l'étendue des marchés potentiels ?
– qui sont les concurrents ?
– dans quelle gamme technologique sont-ils actifs ?
– comment évoluent-ils ?
– quelle est la dynamique de la technologie (nouveaux arrivants, vitesse d'expansion, fréquence de contribution) ?

Par combinaison, croisement et analyse des données bibliométriques, l'analyse de la documentation brevet peut conduire aussi à :

– l'identification d'experts d'un domaine particulier que l'on peut définir comme les inventeurs les plus prolixes mais aussi ceux dont les inventions couvrent de multiples domaines technologiques ;

- la définition de stratégies de recherche que l'on peut situer et positionner, par exemple, en regard de manques dans un environnement technologique particulier ou en complément de ses propres expertises et technologies ;
- le positionnement de stratégies de développement (DOU, HAUDEVILLE et WOLFF, 2015) pour lesquelles il conviendra de déterminer quels sont les déposants détenteurs de technologies complémentaires ou concurrentielles pouvant devenir des partenaires potentiels ;
- l'analyse des structures relationnelles [10] de réseaux concurrentiels peut accompagner l'élaboration de stratégies d'alliances pour les entreprises : en identifiant les collaborations effectives transcrites par le dépôt conjoint de brevets...

Au-delà de ces questionnements, un critère d'évidence est celui de la **flexibilité** (open) nécessaire à l'élaboration d'une réponse d'intelligence économique dans un contexte organisationnel. Si les questions précédentes constituent, le plus souvent, une approche générique de la veille informationnelle, la **contextualisation du questionnement initial** [11] conduit à des situations variées imposant des appropriations (voire détournement) de scénarios de recherche tels que pré-déterminés par ces questions et laisse supposer des **besoins fonctionnels pas forcément identifiés**.

Questions et réponses

Le brevet peut être fouillé pour obtenir des réponses ou des questions.

Ainsi, au sein d'un processus d'intelligence compétitive (extension *offensive* de l'intelligence économique (S. H. MILLER, 2001)), que ce soit en entreprise, dans le domaine de la recherche académique ou au niveau institutionnel, l'extraction d'information brevet accompagne la stratégie de développement (technologique, commercial (NORDMAN et TOLSTOY, 2016) ou encore de gestion des ressources humaines (PORTER, 2007)).

10. Ce qui revient à transposer sur les réseaux l'item précédent.
11. Par ex. en termes de technologie dont on peut disposer en interne, de compétences individuelles, de l'historique de développement de la structure, de partenaires existants, etc.

4.3.4 Extension et nouveaux usages

Source d'inspiration en substitut d'une activité de R&D interne (TRUMBACH, PAYNE et KONGTHON, 2006), la recherche d'innovation (en regard d'un état organisationnel donné) au sein de la documentation brevet constitue immédiatement un potentiel très important. Cette pratique s'impose (QUONIAM, REYMOND et REY, 2014) en accompagnement de l'innovation frugale [12], mais également en source d'inspiration pour tenter de catalyser l'innovation (CORBEL et MBONGUI-KIALO, 2012 ; MBONGUI-KIALO, 2012 ; DOU et LEVEILLÉ, 2015), de faciliter la créativité (QUONIAM et REYMOND, 2014b ; DOU et LEVEILLÉ, 2015) et alimenter des remue-méninges dans des domaines variés. Le document brevet peut instruire sur les stratégies commerciales des concurrents ou aider à en produire une (B. H. HALL, A. JAFFE et TRAJTENBERG, 2005).

Ce sont là des nouveaux usages, en pleine expansion (RADAUER et WALTER, 2010 ; JOERGES et NOWOTNY, 2012 ; DOU et LEVEILLÉ, 2015), qui étendent la variété des cibles (PME/PMI (TRUMBACH, PAYNE et KONGTHON, 2006), startups, recherche académique, innovation sociale et frugale...) et en conséquence ouvrent une piste de recherche relativement balisée. En effet, des travaux récents de BREITZMAN et MOGEE (2002) et ZHANG, L. LI et T. LI (2015) proposent une comparaison de différentes bases de données et logiciels d'analyse qui satisfont les besoins fonctionnels de l'analyse brevet **professionnelle**. Ceux-ci sont établis pour les professionnels de l'information brevet (voir le panorama en section 4.3.1 p. 114). Les différents scénarios d'utilisation que ces auteurs proposent laissent en marge les praticiens de l'IE des PME, de l'innovation sociale ou frugale.

12. Le lecteur intéressé peut se référer au développement théorique de ce concept (ZESCHKY, WIDENMAYER et GASSMANN, 2011 ; AGARWAL et BREM, 2012 ; BHATTI, KHILJI et BASU, 2013) originaire de l'Inde qui désigne une démarche consistant à répondre à un besoin en utilisant un minimum de ressources. La documentation brevet (demandes non abouties et brevets expirés) constitue dans ce cas de figure une encyclopédie technologique de réponses déjà établies à des problèmes variés. En ce sens, faciliter l'accès à cette documentation, développer et répandre son utilisation constitue une réponse sociétale (QUONIAM et REYMOND, 2014a).

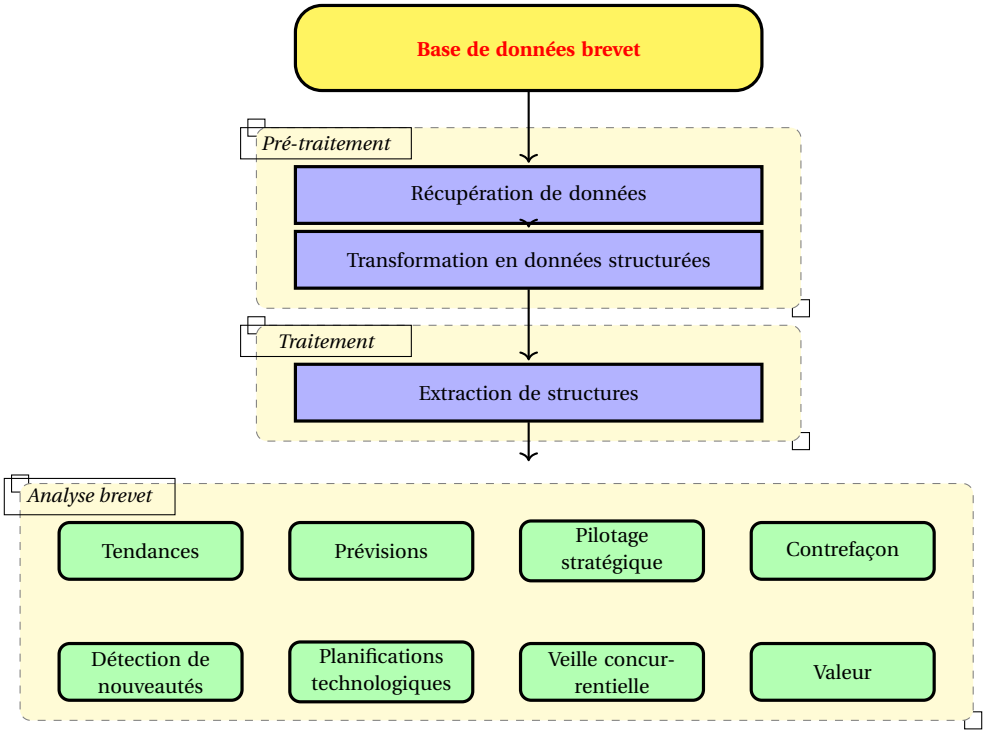


FIGURE 4.6 : Workflow de l'analyse brevet adapté de ABBAS, ZHANG et KHAN (2014).

La médiation au document brevet

Le **développement de scénarios et la définition de critères d'appropriation** constitueraient une application de recherche en SIC dont l'implication sociétale serait non négligeable :

- réutilisation de l'existant (économie de la R&D) ;
- pratique documentaire différente ;
- alternative non académique...

4.4 L'aspect opérationnel

Le schéma 4.6 adapté de ABBAS, ZHANG et KHAN récapitule les divers objectifs de l'analyse brevet. Les auteurs scindent le processus en trois étapes : la première, de prétraitement, inclut la collecte de données et leur transformation en données structurées [13], la seconde procède au traitement pour l'extraction des structures, la dernière, interactive, spécifie les différents scénarios d'analyse.

13. On désigne dans cette phase les processus de tri, de sélection des données. La structuration correspond notamment à l'application de descripteurs de ces dernières. Ce sont les phases « collecte et agrégation de l'artefact médiateur : cf. figureŕeffig :ArtefactMediateur p. 93.

Au plan pratique, il est nécessaire de définir un questionnement comme point de départ : un l'état de l'art sur un domaine technologique ou une ressource naturelle, une analyse de concurrents, une analyse de marché, un problème technique [14] à résoudre. Ce questionnement conduit à une (ou plusieurs) requête(s), dont le résultat est un « univers brevet ».

4.4.1 Construction d'une requête

Point d'entrée incontournable, la recherche documentaire dans le document brevet rejoint le domaine de la recherche d'information en général qui pose le problème de retrouver les documents pertinents à la requête d'un usager [15].

En sus du problème des mots-clés jamais unanimes (BERNERS-LEE, 1989), les brevets sont décrits par des termes génériques (SAAD et NÜRNBERGER, 2012) dans l'objectif d'élargir le cadre des inventions (ou éviter de le réduire). En conséquence, la recherche sur un mot clé générique donnera des ensembles de résultats très brouillés et, réciproquement, la recherche sur un mot-clé précis un ensemble probablement incomplet. L'approche complémentant la réponse des index via les métadescripteurs des classifications permet d'approcher des corpus pertinents. En conséquence, la construction d'une requête doit suivre un mode itératif, introduisant la boucle de rétroaction traditionnelle pour élaborer un cycle performatif de cette élaboration.

ZHANG, L. LI et T. LI (2015) proposent une procédure type de la recherche brevet telle que traduite sur la figure 4.7. Celle-ci se décompose en quatre étapes :

Étape 1 : Construire la requête initiale. Une première action déterminée par le type de recherche brevet délimite le type et l'étendue de la requête ;

Étape 2 : Collecte des résultats et première évaluation.

14. Bien que totalement en phase avec une recherche dans la documentation brevet, la résolution de problèmes qui ne sont quelquefois pas posés n'est pas abordée ici. Nous notons simplement la possibilité d'explorer un univers brevet adéquat pour identifier d'éventuels problèmes.

15. Les premiers travaux remontent à l'origine des ordinateurs (SALTON, 1971 ; FAIRTHORNE, 1956) et ont évolué avec l'apparition de l'hypertexte (BAEZA-YATES, RIBEIRO-NETO et al., 1999 ; BOURDONCLE, 1999 ; LARDY, 2001). L'amélioration des dispositifs d'indexation est une voie de recherche, une autre est celle de l'accompagnement des usagers à déterminer leur besoin et/ou améliorer l'utilisation de l'existant et c'est celle que je suis actuellement.

Étape 3 : Si besoin, selon la pertinence des résultats obtenus (trop ou pas assez de résultats, ensemble bruité par des documents non pertinents, ...) l'on détermine les raisons de ce résultat pour améliorer la requête. Ce peut être en ajoutant des contraintes (hyponymie, dates, classifications) ou au contraire en élargissant la requête (synonymes, hyperonymes, union de requêtes). Puis retour à l'étape 2.

Étape 4 : Traitement du résultat de recherche, synthèse et rédaction du rapport de présentation des résultats.

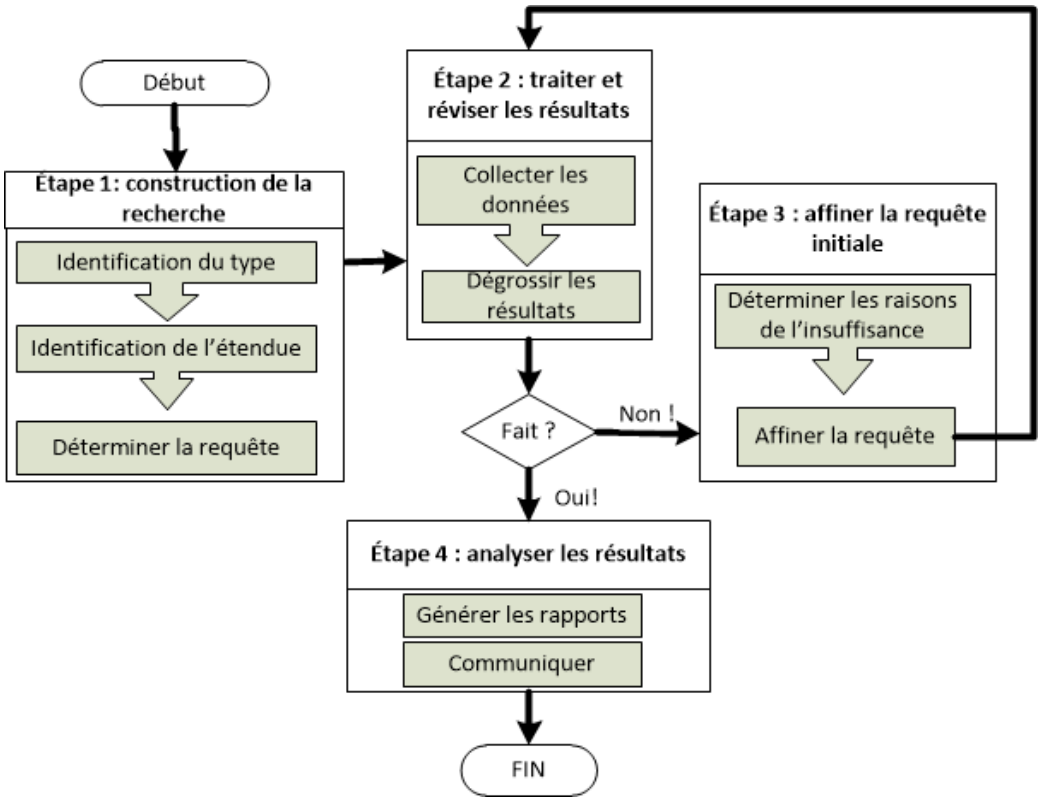


FIGURE 4.7 : Procédure de recherche brevet. La fin est souvent le début...

Á l'étape 2, le besoin fonctionnel d'établir un état descriptif du corpus documentaire obtenu relativement rapidement est d'évidence. Il s'agit d'élaborer une description de l'univers brevet en s'appuyant sur une large gamme d'indicateurs bibliométriques. L'appréciation de mesures simples et globales permet de caractériser l'étendue : nombre de documents, période concernée, technologies adressées ciblent globalement l'étendue des publications.

L'étape 4, d'un point de vue factuel, est une opération qui se réalise sur un nombre important de documents rendant l'opération manuelle longue a priori. Et, en sus de

cet aspect quantitatif, le document brevet est difficile à lire par son mélange de descriptions techniques et juridiques, mais aussi son jargon (PARANJPE, 2012).

4.4.2 Instrumentation

L'identification de signaux forts (tenants de technologies, experts ou déposants), de signaux faibles (opportunités technologiques, inventions ré-exploitables, nouvelles applications), de réseaux de partenaires ou encore la mise en place de simple veilles concurrentielles ou technologiques, tient à la capacité d'utilisation de logiciels d'analyse. D'autres opérations plus complexes peuvent s'avérer nécessaires pour par exemple estimer la valeur (« SME tailor-designed patent portfolio analysis » p.d.) d'un brevet. Ces dernières opérations sont qualifiées d'utilisation avancée de la documentation brevet (ADAMS, 2012) : les données bibliométriques des brevets telles la taille de la famille associée, l'étendue géographique, la variété technologique couverte, les citations, permettent d'alimenter une réflexion stratégique particulière (souvent qualifiée de « valeur »). Ainsi, les logiciels vont exploiter tant les métadonnées que les données textuelles des brevets pour accompagner l'exploration de la documentation. Cette exploration suppose des fonctionnalités variées selon les scénarios associés implicitement aux usagers précédents.

4.4.3 Des besoins fonctionnels complexes

Pour la plupart des contextes d'utilisation et des usagers, la difficulté de recherche dans la documentation brevet est un fait : le document seul est technique, juridique et administratif. Ensuite la volumétrie pour une requête donnée est telle que le traitement manuel systématique en est exclu. Les instruments développent des fonctions d'accompagnement à l'appréhension des corpus documentaires offrant, d'une part, la capacité à traiter de gros volumes de données et, d'autre part, une herméneutique scénarisée pour les besoins autarciques et développée le plus souvent à partir des travaux historiques de bibliométrie (NARIN, 1994). Parcourons l'essentiel.

4.4.3.1 Métriques descriptives

Décrire un corpus documentaire s'appuie sur un ensemble générique d'éléments informationnels calculés à partir du document. La taille du corpus (nombre de documents), la période et les évolutions de publication (historiques), les auteurs, déposants et technologies associées constituent un point de départ d'évidence pour la caractérisation. L'on pourra alors combiner ces indicateurs élémentaires en croisant selon des points de vues spécifiques : selon les auteurs, les dates, les pays pour élaborer des hits parades (plus gros portefeuilles dans un domaine, experts notoires d'une technologie), des historiques mettant en exergue les évolutions des portefeuilles des concurrents ou la dynamique des technologies. Graphiques , cartographies et visualisations interactives facilitent l'élaboration et la sélection des informations en offrant des représentations pertinentes de ces caractéristiques.

4.4.3.2 Appréhension de la relation

PIERRET (2006) s'est intéressé à la découverte de connexions cachées, porteuses de nouveaux savoirs en utilisant une méthode originale d'exploitation des bases de données bibliographiques suivant le modèle transitif de SWANSON dit « découverte de connaissances dans les bases de données bibliographiques (*Knowledge Database Discovery*) et fondé sur le principe que si deux éléments sont en relation avec un même troisième élément, il est probable que les deux éléments initiaux ait un lien de relation... En posant comme hypothèse que, si un brevet mentionne deux auteurs ou plus alors ces derniers sont co-auteurs [16]. Il est ainsi possible de considérer les collaborations extrapolées de la documentation technique. Sur un autre plan comparable, la citation intrinsèque à la construction du document brevet est une information de mise en relation de la construction technologique : un brevet citant des documents brevets antérieurs pose la technologie décrite comme composée par les technologies citées.

En conséquence, le pré-traitement de ces données informationnelles indirectes conduit

16. Et, par extension, sur les déposants (ils sont alors partenaires) ou encore sur les technologies - via la CIB - (pour identifier les croisements technologiques) ?

à un pouvoir descriptif de la relation présupposée :

- Qui fait quoi, où, avec qui, pour qui ?
- Quelles sont les technologies utilisées, y-a-t-il des pistes innovantes possibles (manques éventuels dans les croisements technologiques) ?
- Quelle est l'envergure technologique d'un concurrent, la panoplie des technologies qu'il détient ?
- etc.

La détermination de la réponse à ces questions s'appuie sur l'utilisation de **la relation** existante entre des éléments de natures différentes (par ex. les inventeurs d'un domaine et les technologies associées) afin d'en extraire une information nouvelle (les inventeurs combinant plusieurs technologies dans des brevets différents). Cette utilisation n'est pas immédiate. En effet, COUTENCEAU et BARBARA (2014, p. 55) remarquaient que la présentation de données relationnelles sous forme de listes possède ses limites car ces listes « ne relient pas toujours les brevets qui interagissent entre eux (par ex. les brevets d'un portefeuille d'entreprise et les technologies associées) ». Le tableau listant les déposants et leurs brevets doit être croisé avec le tableau listant lesdits brevets et les technologies associées. ABITEBOUL (2013) pose une démonstration algorithmique de ce phénomène non évident : le calcul relationnel (s'appuie sur des « listes ») ne permet pas d'exprimer l'existence d'une relation par un nombre d'intermédiaires[17] quelconque car il nécessiterait une disjonction infinie alors que le problème est résolu à l'aide de graphes. En effet, cet ensemble de questions est plus facilement traité sous forme de graphes (réseaux ou de matrices (DOU, MOHELLEBI et KISTER, 2012)) qui représentent les entités différentes (déposants, technologie dans notre exemple) faisant abstraction du document brevet qui sert de pivot pour la construction du graphe. Représenté graphiquement, ce dernier favorisera la lecture de ces informations de relations[18].

17. Il s'agit d'obtenir par ex. « le brevet B_1 est relié au brevet B_2 car ils partagent la propriété X_1, puis au brevet B_3 qui partage avec B_2 la propriété X_2 » à partir de la liste des relations Brevets B_i et des propriétés X_i. Il suffit de supposer de la relation avec la propriété 1000 au lieu de la troisième pour s'en convaincre... Si l'on cherche en plus les relations de grandeur non connues, c'est à dire une chaine de brevets de longueur non connue à l'avance construite sur ce principe de partage des propriétés, cela laisse une immensité de possibilités avec des listes.

18. Ce fait a été démontré également dans le contexte de la recherche d'information, dont une branche de formalisation mathématique s'appuie sur l'analyse des concepts formels en support de la construc-

4.5 Recherche académique à partir du document brevet

Les besoins fonctionnels précédents s'alimentent des avancées de la recherche portant sur la documentation brevet. Trois domaines sont globalement identifiables : la technométrie, la technoscience et enfin les développements de techniques d'analyse, de production de nouveaux indicateurs issus de l'avancement des méthodes de traitement de données.

4.5.1 Extension de la bibliométrie

Au plan académique c'est en informétrie, dans le domaine particulier de la technométrie (POLANCO, 1995), que ce document particulier a permis le développement de travaux inspirés de la bibliométrie [19]. Ainsi, peu après sa publication originale dans *Science* par laquelle il fonda un renouveau de la bibliométrie originaire de Lotka, Bradford et Zipf que De Solla Price mettait en application en scientométrie, GARFIELD publia un article moins connu dans *the Journal of Patent Office Society* (1957) intitulé « Breaking the subject index barrier – a citation index for chemical patents ». Fondé sur l'index juridique des citations de Shepards (ADAIR, 1955), GARFIELD généralisa sa notion d'index de citations en développant un outil distinct des index d'organisation hiérarchiques traditionnels (i.e les classements alphabétiques ou thématiques). L'index de citation devient une transcription de réseaux d'associations (d'individus - les inventeurs, de sociétés - les déposants) qui se prête à l'identification d'informations à partir des métadonnées des documents, informations de nature différente des contenus.

La riche structuration de la documentation, le processus de description, d'enrichissement et de vérification en font une source documentaire de premier choix pour les études, qu'elles soient théoriques ou appliquées, appuyées sur les métriques de l'in-

tion des requêtes possibles comme des combinaisons de catégories (descripteurs) et des connecteurs logiques. Chaque association « document-Champ de description » constitue un contexte formel potentiel.

19. Voir à ce propos le travail de ROSTAING (1996)

formation. Les documents brevets sont utilisés aussi dans des applications différentes : ils sont la source de mesure des systèmes de production de la connaissance et aussi le point d'entrée de mesure des processus économiques de l'innovation. Les scientomètres l'utilisent pour établir des informations pertinentes pour la politique des pouvoirs publics provenant de l'analyse de l'évolution des dépôts et susceptibles d'être utilisées par les responsables de l'élaboration des politiques, par exemple dans le cadre de la stratégie industrielle d'un pays (BARRÉ, LAVILLE et al., 1995), pour la production d'indicateurs de mesure d'activité (BARRÉ et LAVILLE, 1994 ; BASSECOULARD et ZITT, 2005 ; LEYDESDORFF, ALKEMADE et al., 2015), il témoigne de la production scientifique.

4.5.2 Technoscience et émergence

Le brevet est perçu, aujourd'hui, comme un des principaux vecteurs de l'innovation (LEYDESDORFF, ALKEMADE et al., 2015) dans la mesure où il permet la « circulation des connaissances » dont on peut tracer les évolutions par le suivi indirect (MOED et HALEVI, 2014 ; CHERRABI et al., 2015). LEYDESDORFF (2015) suggère de développer les méthodes et modèles d'analyse plutôt que les aspects théoriques afin d'étudier la dynamique de l'innovation technologique.

Construire des Oligopticons

LEYDESDORFF suggère de « préciser l'information », que la structuration de cette information captée selon des algorithmes de traitement permet de développer des angles de vue. Ceci rejoint le concept des *oligopticons* de VENTURINI et LATOUR (2010). Nous comprendrons par là « la construction d'artefacts de médiation » tel que modélisé en p. 93.

Ainsi, le document brevet se prête à nombre d'analyses, par déduction, induction ou abduction, pour la production d'informations de niveaux très différents : mesurer les dynamiques de l'innovation (GRILICHES, 1998 ; DE KERMADEC, 2001), étudier les rapports sciences-société (NARIN, HAMILTON et OLIVASTRO, 1997 ; CALLON, J.-P. COURTIAL et PENAN, 1993 ; GIBBONS et al., 1994) ou les liens sciences et technologie (SCHMOCH, 1997 ; MEYER, 2000 ; MEYER, 2006 ; TIJSSEN, 2001). BROCK et al. (2012) développent en ce sens une pragmatique de l'émergence technoscientifique en application de la théo-

rie de l'acteur réseau CALLON et LATOUR (2013).

4.5.3 Développement de techniques d'analyse

Discutant de l'intérêt de la mise en place de services d'information brevet ouverts aux PME/PMI, RADAUER et WALTER (2010) justifient, d'une part, une évolution récente des besoins informationnels notamment des PME/PMI et, d'autre part, du nécessaire traitement sémantique pour affiner les résultats. Les besoins fonctionnels de description de corpus sont couverts, l'automatisation d'autres fonctionnalités d'accompagnement à l'exploration est attendue. Les techniques de fouille de données textuelles apportent une facilitation de l'approche informationnelle brevet.

ABBAS, ZHANG et KHAN (2014) décrivent une taxinomie des méthodes de traitement des documents brevet en situant des travaux de recherche récents appliqués au domaine. Le développement de la puissance de calcul permet aujourd'hui d'introduire la fouille de données textuelle (traitement naturel du langage, approches sémantiques) pour des opérations de classement, de tri et de sélection automatisées. Les auteurs mentionnent également la visualisation de données facilitant la lecture de données complexes.

Un panorama rapide de la littérature du domaine montre les voies de recherche actuellement très actives. En s'appuyant sur le réseau des citations, ÉRDI et al. (2013) identifient des branches technologiques à des ensembles de brevets et, prenant la dimension temporelle, ils prédisent la dynamique des évolutions de ces branches. Les auteurs montrent que leur technique de segmentation permet de caractériser le développement technologique et montrer comment un brevet cité par d'autres croise différents champs industriels afin de détecter les technologies émergentes.

MOMENI et ROST (2016), par un procédé de classification comparable, suivent les chemins d'évolution des technologies pour développer une méthode d'identification de technologie disruptive afin de limiter les risques des choix de développement des organisations.

THOMA (2014) produit un index des brevets à l'aide d'un indicateur composite combinant vingt indicateurs selon sept dimensions pour appliquer une méthode d'analyse

factorielle afin d'améliorer l'estimation de la valeur des brevets.

CHOI et HWANG (2014) proposent d'unifier les deux voies d'analyse brevets traditionnellement opposées : l'approche par les réseaux (citations) et l'approche par les contenus (résumés, thématique, terminologie et mots-clés). Les auteurs combinent des réseaux de brevets issus des citations avec les réseaux de mot-clés associés issus des contenus (ibid.) qui leur permet d'identifier les termes centraux et stables d'une technologie et identifier les relations entre les éléments technologiques essentiels.

En fournissant une mesure de la nouveauté technologique les auteurs séparent les combinaisons d'anciennes technologies des nouvelles connaissances « pures ». Ainsi, VERHOEVEN, BAKKER et VEUGELERS (2016) revisitent la notion de mesure de la nouveauté technologique et montrent que les deux dimensions sont statistiquement corrélées tout en portant chacune des informations de nature différente.

TANNEBAUM et RAUBER (2014) accompagnent les examinateurs de brevets à étendre leur requête (cf. section 4.4.1p. 120) à l'aide des requêtes d'autres examinateurs sur la base de données brevets UPSTO.

Au plan des contenus, le traitement en langage naturel est appliqué afin d'accompagner l'analyse des contenus des documents brevet. BRÜGMANN et al. (2015) développement des modules de reconnaissance d'entités nommées, de segmentation, de comparaison de revendication [...] pour améliorer la description automatisée des résumés. De leur côté, VENUGOPALAN et RAI (2015) développent la classification thématique des brevets dans les domaines des marchés et produits. Cette classification permet l'identification de brevets (avec un taux annoncé de 98%) qui ne correspondent pas à des produits. Les auteurs montrent alors une application de cette technique à l'élaboration d'une méthode d'estimation sociétale des retombées de la connaissance.

4.6 Résumé et conclusions

Si l'on peut trouver des exceptions parfois déroutantes (cf. figure 4.8), la documentation brevet est ainsi d'intérêt pour toute structure de recherche et développement, y

compris académiques, afin de déterminer une vision stratégique. Pour le développement de marchés, la prospective pour les décideurs de tous niveaux le recours à cette documentation peut satisfaire des besoins informationnels variés (REYMOND, 2016a). Le document brevet est par co-construction, un document révélant des qualités rares en documentation. La granularité de la description et la fiabilité des contenus impliquent que l'on peut construire un regard sur un ensemble de documents brevet selon des perspectives différentes : l'espace et le temps qui caractériseront le territoire et la temporalité de(s) l'invention(s), les descripteurs du document qui précisent l'auctorialité, la propriété, les procédures de dépôt suivies et, une perspective méta via le système international de classement qui ouvre un regard indépendant de la langue de description. Enfin, alors que les références internes à la documentation brevet révèlent la composition technologique de l'invention (information indirecte induite des citations et des classifications), les références externes (académiques) révèlent, même partiellement, les liens sciences/technologie.

L'ouverture des bases de données sur internet conduit à une expansion des usagers, le développement de techniques d'analyse accompagne de nouveaux usages qui entraînent des besoins fonctionnels nouveaux. Au-delà de la technométrie macroscopique (ETZKOWITZ et LEYDESDORFF, 1997), les documents brevet regorgent d'informations (ROSTAING, 1996, p. 95–115) qui peuvent être fouillées (PORTER, 2007) pour extraire les connaissances des données descriptives des inventions (ou demandes de), à des fins de marketing (B. H. HALL, A. JAFFE et TRAJTENBERG, 2005), d'intelligence économique (JAKOBIAK, 1994 ; DOU GOARIN, 2013) en général. La notion de fouille entend ici un ensemble de procédés d'enrichissement et d'extraction à partir des données initiales d'information complémentaires, extrapolées : une augmentation des données qui précède et accompagne l'exploration de la foison informationnelle (cf. figure 3.2 p. 93). Que ce soient, par exemple, les noms et adresses des inventeurs qui géolocalisent les inventions, les classifications et les revendications qui soulignent les développements technologiques, les citations qui permettent d'estimer l'impact et la valeur des brevets (B. H. HALL, A. JAFFE et TRAJTENBERG, 2005) ou, sur un autre plan, déterminent la composition technologique d'un brevet. Aucune de ces données ne sont immédiates : que vaut une liste d'adresses (issues du monde entier) en regard d'un

planisphère positionnant celles-ci ? Idem une liste de codes de classification comparée à leur description en langage normalisé ? etc. L'intérêt d'une instrumentation qui accompagne ce procédé de lecture de données « cachées » se pose ainsi en évidence. De surcroît, la contextualisation de ces données en référence à une situation particulière (par ex. un lieu géographique, on peut tenter d'identifier des experts d'un domaine mais localement...) nécessite de pouvoir adapter l'instrument.

Enfin, la fouille de données permet de découvrir des tendances dans les données à l'aide des statistiques. L'application de ces travaux de fouille dépasse le domaine des brevets car elle se retrouve dans les systèmes de recommandations, les procédures de segmentation et classification, la recherche d'information, le filtrage sélectif, ou encore la prédiction. Technologies sémantiques (ontologies, thésaurus), technologies de classification automatiques, règles d'inférences et formalisation de prédicats promettent nombre d'améliorations de l'augmentation de données qui en est à ses balbutiements.

Médiation au document brevet et SIC

Ce sont ainsi des sources de mécanismes que l'on peut qualifier de renouveau herméneutique en perpétuel progrès. L'accompagnement des usagers à l'accès à cette ressource documentaire est de mise et justifie un axe de recherche en SIC : la médiation aux usages de la documentation qui alimente à son tour l'élaboration de nouvelles fonctionnalités.

De plus, pour satisfaire la flexibilité nécessaire à alimenter un processus dynamique d'intelligence économique, l'utilisation d'une instrumentation variée d'exploration de cette vaste documentation peut quelquefois nécessiter une combinaison unique. L'instrumentation doit être suffisamment ouverte pour se décliner et s'adapter à une contextualisation fine de la collecte et du traitement de l'information, mais également apte à couvrir, évidemment en second temps, les nécessaires phases de communication et de partage d'information pour les strates décisionnelles. Ainsi, pour répondre aux multiples questions de l'IE ou, plus largement, de questionnement des SHS (JOERGES et NOWOTNY, 2012), la flexibilité nécessaire soulignée précédemment peut se traduire par la création de chaînes de traitement spécifiques allant de la création corpus, leurs mesures simples jusqu'à l'analyse fine et profonde de ces derniers.

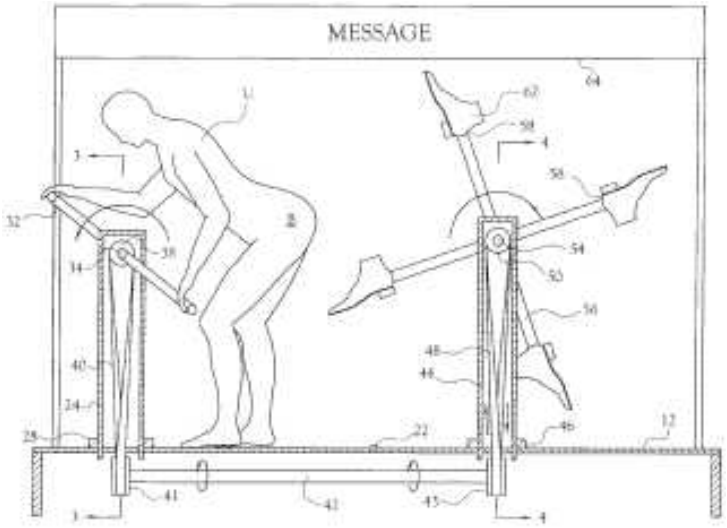


FIGURE 4.8 : La machine à se botter les fesses d'après ARMSTRONG, J. W. (2001). Le concept sera étendu par la suite comme « facteur de travaux de groupe, d'auto-thérapie, et d'inspiration de la créativité »par REESE L. (2006).

Nouvelle herméneutique SIC

Créer des chaînes de traitement (*oligopticon*) pour construire des corpus, les fouiller et les analyser selon des perspectives sociétales, c'est aussi accompagner ces usages pour alimenter les débats de l'open knowledge et open science.

CHAPITRE 5

P2N : artefacts médiateurs au document brevet

Patents and patent applications usually contain the first disclosure of new technologies and processes and serve to link theory with practice, providing 'real world' examples of the application of scientific research.

Increasingly, scientific discoveries are reported first in the patent literature, rather than in academic journals. To ensure that science students have the skills that match the information resources they will use as professionals, patent searching must become part of their information literacy instruction.

— D. MAC MILLAN, Patently Obvious : the place for patents in information literacy in the sciences (2006)

ב

AVEC le Pr. Luc QUONIAM, nous avons initié en 2014 la mise en œuvre, les bribes d'un extracteur de documents brevets associés à une requête. Le premier collecteur a été mis au point avec les étudiants de Master en intelligence économique et territoriale de l'université de Toulon. L'objectif était d'une part d'initier les étudiants à des compétences uniques en intelligence économique : la capacité à développer une automatisation de procédés de collecte et de traitement de données internet, leur permettre d'appréhender la documentation brevet et enfin de les former à l'analyse de données complexes (REYMOND, 2016a). Patent2net - P2N (2016b) est depuis un outil développé en Python sous licence libre (CECILL-B) par une communauté internationale de chercheurs. L'outil extrait les données associées aux documents brevets (contenus ou description), opère une série de traitements afin de présenter ces données à une gamme d'outils facilitant la lecture : depuis les tableaux de données, les représentations graphiques ou cartographiques aux instruments plus élaborés de classification automatique des contenus, d'analyse accompagnée des contenus ou encore de représentation en cartes heuristiques. Les exports vers les instruments de gestion bibliographique standards sont aussi élaborés. Libre et open source, P2N se prédispose ainsi à la réalisation d'études de cas à différentes échelles : depuis l'étude de portefeuilles d'états (CHERRABI et al., 2015) ou de régions, à l'analyse particulière de portefeuilles d'entreprises, d'un domaine (par ex. l'enseignement à distance (EDUARDO STOROPOLI, 2016) ou d'un cas particulier cf. « la dengue » (RIBEIRO FERRAZ et al., 2016).

P2N s'inscrit dans la classe des artefacts aux données dépeints par BATTLES (2013), faisant lumière sur les données ouvertes que sont les documents brevets par la réalisation d'un ensemble de macroscopes (BÖRNER, 2011) et d'oligopticons particuliers qui sont dérivables et personnalisables à souhait. Le caractère structuré et hétérogène de la base de données de l'Office Européen des Brevets a rendu plus aisée la tâche de réalisation d'artefacts médiateurs à cette documentation. L'utilisation pratique de ces artéfacts va de l'intelligence compétitive (industrie, marketing, économie, inno-

vation, créativité, cf. section 4.3.3 p. 116) à la recherche académique (politique de la recherche, scientométrie, bibliométrie, etc. cf. section 4.5 p. 125).

5.1 Description

Le chapitre précédent a souligné le potentiel informationnel du document, ses utilisations et la gamme informationnelle inscrite dans le document brevet. L'instrumentation décrite ci-après a pour objectif de rendre visible et d'explorer pour exploiter ce potentiel.

5.1.1 Choix techniques et technologiques

Pour la facilité de lecture et d'appropriation du code, le langage Python a été utilisé. Le choix de ce langage repose sur plusieurs critères de lisibilité (un niveau très proche de la langue anglaise), de portabilité, et d'une grande communauté qui a développé foison de bibliothèques de bas niveau open-source (ZIADÉ, 2009, chapitres 1 et 2). Ces éléments en font un langage de prédilection tant pour les applications académiques (DOWNEY et al., 2002; SWINNEN, 2012) que professionnelles à toute échelle de grandeur (depuis les scripts qui contrôlent l'installation des systèmes d'exploitation ou logiciels sur certains téléphones portables ou ordinateurs, aux scripts qui formulent la réponse que renvoient les index du web).

Le format des exports de données s'appuie sur les standards ouverts, notamment les fichiers à séparateur de valeurs fixes (csv) et les documents XML ou json. Enfin, les données descriptives des documents brevets sont au format compatible avec les outils de gestion de référence les plus usités par appui sur le format BIBTEX.

5.1.2 Architecture générale

P2N prend sa source sur les bases de données de l'Office Européen des brevets (EPO) (PATENTDATA@EPO.ORG, 2014) via l'interface de programmation (API) (KALLAS, 2006).

Les scripts de P2N se chargent de collecter contenus et descripteurs de documents brevets associés à une requête : cet ensemble est appelé Univers Brevet (UB) dans ce qui suit. Des outils de traitement formatent les données d'un UB pour les explorer par toute une gamme d'instruments apportant des herméneutiques spécifiques aux documents : que ce soit au niveau des données descriptives, leurs associations (plan bibliographique) ou encore au plan des contenus mais aussi au plan méta (classifications cf. section 4.1.3.3 107).

5.1.2.1 Fonctionnement

Le fonctionnement de l'outil est structuré en trois phases décrites par le principe du processus d'analyse des brevets de la figure 4.6 en p. 119 de ABBAS, ZHANG et KHAN.

- Pré-traitement : la construction de la requête (la requête qui sera transmise à l'API d'EspaceNet).
- Traitement : requêtes récursives, cette phase permet de collecter les données bibliographiques pour chaque brevet, les données textuelles et les données familles des brevets.
- Post-traitement : c'est la phase de préparation des données des brevets pour leur utilisation (formatage CSV, JSON, HTML, IRaMuTeQ, Graphes dynamiques ...).

L'utilisateur doit avant tout modifier le fichier de paramétrage pour y inscrire a minima sa requête et le répertoire associé dans lequel seront téléchargées les données (cf. l'annexe B en p. 13).

5.1.2.2 Collecte de données

Les scripts de collecte (*gather*.py) servent la captation des données de l'OEB et s'appuient sur la bibliothèque python epo-ops (SONG, 2014). Le collecteur d'univers brevets de P2N est compatible avec le langage CQL *(Contextual Query Language)*. L'Univers brevet collecté est le même que celui affiché en réponse par le moteur interne

du site web de l'OEB, le moteur *Smart Search* (2015) ou encore l'interface de recherche avancée.

Le langage CQL permet la construction de requêtes élaborées à partir d'opérateurs booléens de paramétrage de critères sur les différents champs (titre, résumés, auteur, déposant, classification, date). En options (dans le fichier de paramétrage), le collecteur peut étendre la collecte aux contenus (résumés, descriptions et revendications) ainsi qu'aux familles de brevets associés à chacun des éléments de l'Univers brevet résultant. Voir la section 4.4.1 en p. 120.

Ainsi, la requête CQL (ta=passiflor* or ta='passion fruit*') AND pd<2015 désigne l'univers de documents brevets publié avant 2015 (commutateur pd) et contenant dans le titre ou le résumé (commutateur ta) les chaînes de caractère commençant par « passiflor » ou par « passion fruit ».

La collecte

Une fois la requête établie, le script CollecteEtTraite [a] lance la collecte des données puis le traitement. L'attente peut être longue puisque ce script respecte les règles des robots de collecte (environ une requête toutes les 2 secondes. En fonction de l'univers brevet cela peut prendre plusieurs heures pour télécharger jusqu'à 2000 documents [b] et descriptions associées.

a. Le nom est variable selon les systèmes d'exploitation.

b. La limite de taille de corpus peut être franchie en décomposant les requêtes selon des critères temporels et de classification pour élaborer des ensembles couvrant complet de l'Univers brevet qui soient de taille inférieure à cette limite.

5.2 Présentation des résultats

Les résultats d'une collecte s'explorent avec un navigateur (Firefox de préférence). Un nuage dynamique de répertoires de données permet de visualiser l'ensemble des requêtes réalisées. Le lecteur est invité à se rendre à l'adresse dédiée à la publication de résultats de collecte et d'analyse de P2N : `http://patent2netv2.vlab4u.info/`.
La sélection d'un des répertoires ouvre la page générale de la collecte.

5.2.1 Page générale

La figure C.1 de l'annexe C p. 17 décrit la page d'accueil présentant les résultats d'une requête brevet. On y trouve :

- Une zone de caractérisation de l'univers brevet (date de collecte, nombre de documents, abrégés téléchargés, etc.).
- Une zone listant des outils d'analyse qui vont ouvrir d'autres pages de navigateur.
- Une zone dédiée aux exports pour des outils complémentaires (libres et téléchargeables aux liens inscrits à la date de rédaction) associant pour chacun les fichiers à télécharger.

5.2.2 Le formatage des données descriptives

Pour l'exploration en tableaux, listes, cartes utilisables via un navigateur Web, différents outils de médiation préparent les données pour la navigation interactive. La sous-suite formateExport* produit des données compatibles avec les API DataTable (*DataTable API* 2007) de tableaux de données structurées et de croisés dynamiques interactifs type pivottable(KRUCHTEN, 2013). Ces deux instruments sont en soi une suite d'instruments.

5.2.2.1 Tableau dynamique de données de champs bibliographiques

Ce premier outil s'appuie sur l'API DataTable [1] qui permet de donner la capacité au navigateur de présenter les données en tableaux dynamiques (choix des colonnes) pour que l'usager puisse réaliser les opérations de sélection, extraction, ordonnancement, export, combinaisons par opérateurs entre colonnes et choix interactifs de ces dernières. Les opérations de tri par inventeur, mandataire, titre, date, etc. permettent de réaliser les opérations courantes de découverte des redondances et d'exploration basique des corpus résultants.

La figure C.2 en p. 19 décrit l'interface de l'outil réalisé.

1. Cf. http://datatables.net/.

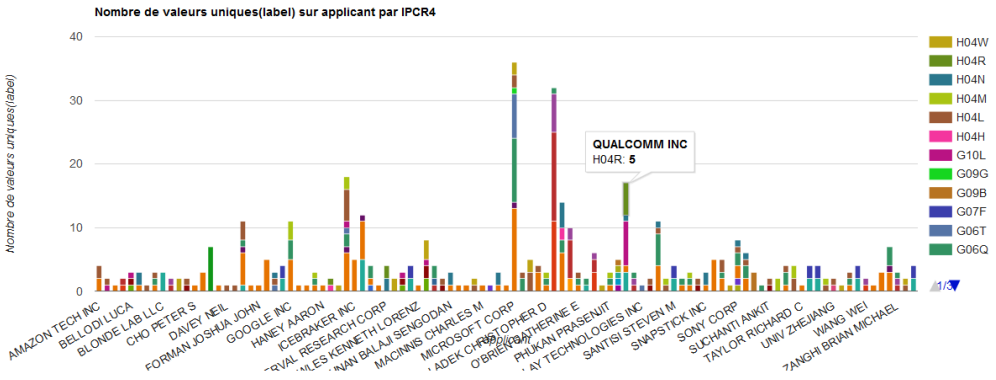


FIGURE 5.1 : Cartographie par tableau pivot, le croisement de deux champs descriptifs. Figurent sur ce diagramme le déposant (applicant) et le code de classification (niveau4) de l'univers brevet ta='social interaction' and PD<2016.

5.2.3 L'outil de construction de tableaux croisés dynamiques

Le second formatage transforme les données descriptives en fichiers HTML et Json, compatibles avec l'outil open-source écrit en langage javascript PivotTable.js [2]. Pivotable.js peut générer des histogrammes, des camemberts, des cartes de chaleur et fournit une interface drag and drop pour interagir avec les données. L'outil permet de produire des graphiques synthétiques comme par ex. la figure 5.1 qui met en exergue les éléments les plus signifiants (l'inventeur le plus prolixe, le déposant détenteur du domaine, les dates de changement de technologie, le brevet le plus cité, la plus grande famille, etc.) au sein de l'univers brevet en cours. La figure C.3 en p. 20 décrit l'interface de l'outil réalisé.

Notez également la dimension interactive de chaque diagramme et tableau produits qui accroissent l'expérience utilisateur en facilitant l'accès à des données informatives riches : la souris positionnée sur un élément provoque une bulle informative précisant les données (ainsi, en figure 5.1, apparait que la société QUALCOMM INC a déposé cinq brevets demandes classées en H04R dans cet univers).
Si nous restons génériques pour qualifier cet univers brevet l'on comprend aisément la nécessaire transposition et contextualisation à une entreprise ou une question d'intelligence économique qui imposerait un filtrage particulier de ce diagramme.

2. Cf. http://nicolas.kruchten.com/pivottable/examples/.

Données ouvertes et publiables

La dimension ouverte des données collectées conduit à des traitements aisés approfondis et paramétrables à souhait. Les formats HTML, CSS, et Json permettent d'appuyer une construction de documents dynamiques et interactifs. L'aboutissement est d'offrir une interface d'expérience utilisateur efficace sur le rapport $\frac{\text{synthèse de présentation}}{\text{quantité de données accessibles}}$, sur le principe des *Data Driven Documents* (D3) (BOSTOCK, OGIEVETSKY et HEER, 2011).

Ces outils permettent ensemble de saisir nombre de questions d'intérêt en intelligence compétitive : que ce soit défensif (par ex. quels sont les principaux concurrents dans mon domaine technologique pour cet univers brevet ?) ou offensif (par ex. avec qui vais-je pouvoir tenter un rapprochement pour cette technologie ? Quelles sont les opportunités de marché ?), l'utilisation de ces données nécessitera peut-être de filtrer sur des critères précis, de les explorer pour les exploiter et les présenter.

5.2.4 Les cartographies géographiques

Une dimension complémentaire est apportée par les formatages de cartographies géographiques appuyés sur la librairie D3plus d'extension de la précédente (SIMOES, 2013) qui met en lumière les marchés, les zones de développement et les origines des dépôts (attaques) de demande de protection pour un univers donné (cf. figure 5.2).

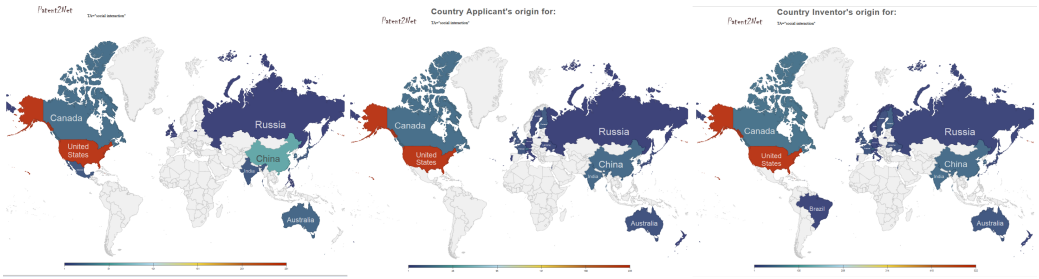


FIGURE 5.2 : Les dépôts de demande selon les procédures de dépôt (hors procédures régionales et mondiales) (à gauche). Le planisphère du milieu met en valeur les pays d'origine des inventeurs, celui de gauche les déposants pour l'univers brevet « social interaction ».

Enfin, un traitement permet un export vers Zotero (RRCHNM, 2015) pour la création de rapports.

En plus de ces outils génériques traitant les données descriptives, P2N ouvre une interopérabilité avec des outils traitant les contenus. P2N ouvre quatre connecteurs principaux pour l'analyse plus complexe. La première, à l'origine du nom de l'outil, est la sous-suite de création des réseaux.

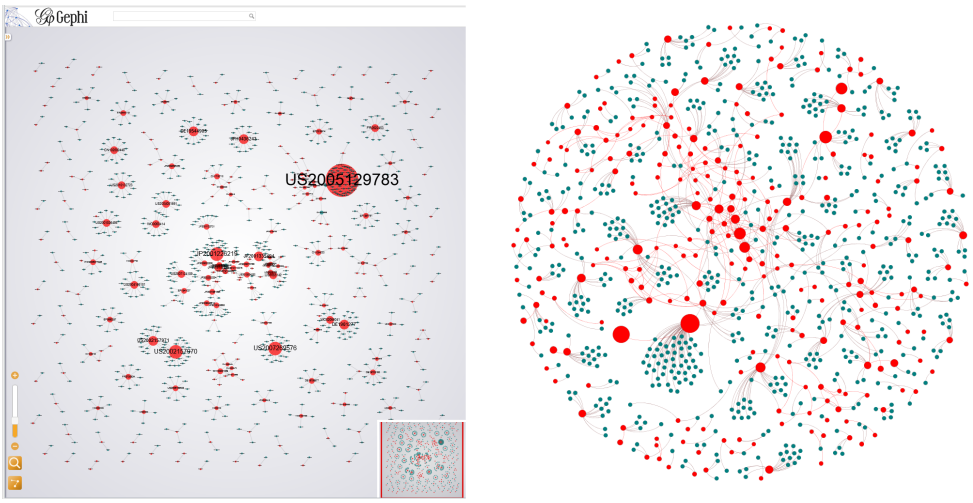
5.3 Réseaux

Nous avons vu l'intérêt de porter un regard à la mise en relation entre les éléments descriptifs (cf. section 4.4.3.2 en p. 123) pour, par exemple, identifier les collaborateurs, les associés. En permettant la visualisation et l'exploration des connections entre les différents champs, les réseaux ouvrent une herméneutique privilégiée à la lecture de ces données complexes, une heureuse alternative aux données en liste : qui travaille pour qui, qui collabore avec qui, quelles technologies sont croisées, quels sont les domaines d'expertise, les références ou encore les citations sont autant de questions dont la réponse figure dans l'affichage des réseaux. Il faudra toutefois représenter, distinguer des nœuds sur ces critères relationnels pour faciliter l'exploration.

5.3.1 Interfaces de médiation

Le *Graph Exchange XML Format* (GEXF) a été choisi pour son interopérabilité avec les logiciels d'affichage de réseau (HEYMANN et al., 2007), son ouverture (par l'utilisation du langage XML) et notamment le logiciel Gephi (BASTIAN, 2009) en complément d'un lecteur/afficheur de données gexf développé en javascript et inclus dans le traitement par navigateur web (qui fournit des fonctionnalités moindres). L'utilisateur a ainsi deux options :

– l'utilisation du navigateur et de l'utilitaire d'affichage des réseaux Gexf-js (VELT, 2011). Comme par exemple celui présenté à gauche sur la figure 5.3 qui permet de « lire » les données de citations (dans cet exemple) et facilement les explorer. L'interface dispose en outre d'un outil de recherche (sur le nom du nœud), présente les liens entrants et sortants ainsi que des informations complémentaires



(a) Interface par navigateur montrant les réseaux de citations. La taille des nœuds est proportionnelle au nombre de ses voisins (degré)

(b) Positionnement du même réseau avec Gephi. La taille des nœuds est proportionnelle au nombre de liens reçus (récursivement par l'algorithme du Page Rank)

FIGURE 5.3 : Comparaison des deux interfaces de visualisation des réseaux

caractéristiques (degré, url (cf. infra)) ;

– l'utilisation d'un logiciel de traitement de graphe (application de mode de représentation, opérations de calcul, détermination des nœuds de transition, etc.). Une machinerie plus lourde mais qui s'applique à de très gros réseaux et permet des opérations plus fines dont la présentation dépasse le cadre de ce document. Les graphes exportés permettent, à l'aide du logiciel, la lecture, l'exploration et la manipulation des graphes (de grande taille). L'outil permet aussi de cibler l'analyse à partir de certains seuils (par ex. sur le nombre d'interconnexions/collaborations), ou dans une certaine période temporelle. Ainsi le même réseau (source, nombre de nœuds, liens) que celui de la figure 5.3 (à gauche) est repositionné à l'aide de Gephi pour mettre en valeur les nœuds qui reçoivent le plus de liens selon l'algorithme du Page Rank (BRIN et PAGE, 1998).

5.3.2 Augmentation des données

Dans les deux cas, les nœuds de chaque réseau sont augmentés, dès que possible, par des URL permettant de faciliter la compréhension du réseau en enrichissant la dimen-

sion informationnelle de l'interface[3] :

- l'URL du site de l'OEB paramétré pour que s'affiche la liste des brevets (nœuds correspondant à un inventeur ou un déposant),
- la description d'un brevet (pour un nœud brevet/label),
- le lien vers la page de description de la classification sur le site de l'Office Mondial de la Propriété Intellectuelle dédié à la définition des descripteurs de classement des brevets.

La figure 5.3 (à gauche) montre le réseau de citations de l'univers précédent (en rouge les brevets collectés, en bleu les brevets citant les précédents). Le positionnement des nœuds est effectué selon l'algorithme *Neato* du logiciel *Graphviz* (BILGIN et al., 2015), la taille des nœuds est proportionnelle à leur degré. Pour faciliter la lecture, une barre d'exploration s'ouvre sur la gauche présentant des données sur les nœuds (nom, degré, liens entrants et sortants). La figure 5.3 (à droite) montre le même réseau pour lequel les nœuds sont proportionnels à leur Page-Rank (BRIN et PAGE, 1998) et colorés par type (citants en bleu/cités en rouge) montrant des brevets du même univers selon un autre point de vue. Les labels des nœuds ont été supprimés (pas faisable sur l'interface javascript) pour clarifier la présentation du réseau. Les différents outils de segmentation temporelle ou sur critères spécifiques (nombre de liens, valeurs d'un paramètre, etc.) ne sont pas affichés ici et permettent un filtrage complémentaire du réseau affiché.

5.3.3 Graphes réalisés : explorations et hypothèses sous-jacentes

In fine, ce sont dix réseaux qui sont créés par P2N : collaborations, références et croisements inter champs descriptifs. La convention de nommage suivante a été respectée pour les distinguer :

$$UBChamp1Champ2.gexf$$

3. Au plan pratique, cela nécessitera de disposer de deux écrans a minima. Certains réseaux, par leur taille, mériteraient également un affichage sur des dispositifs à très haute résolution et de grande taille.

5.3.3.1 Réseaux de collaborations

Le tableau 5.1 présente les trois réseaux qui représentent les collaborations au sein d'un domaine brevet. Pour chaque champ descripteur, chaque co-entrée est associée avec la précédente, l'hypothèse est faite que si chacun figure dans une demande de brevet, c'est qu'ils ont collaboré. Le résultat est le recensement de ces collaborations effectives dans le domaine de l'UB en question. Le croisement technologique s'appuie sur le même principe. En extrapolant les multiples classements [4] de chaque brevet comme des croisements de technologie [5].

TABLE 5.1 : Les réseaux de collaboration

Réseau	**Code**	**Description**
Inventeurs	Inventor	Présente quels sont les experts ayant collaboré et à quel degré (nombre de brevets co-demandés)
Déposant	Applicant	Présente quelles sont les organisations ayant collaboré et à quel degré (nombre de brevets co-demandés)
Technologies	CrossTech	Présente quels sont les croisements technologiques effectifs, à quel degré ils apparaissent (co-présence des classements dans la description des brevets).

5.3.3.2 Réseaux de croisements

Le tableau 5.2 présente les différents réseaux de croisement. L'hypothèse est faite que la co-présence des entrées de chaque champ sous-entend une information spécifique

4. Ceux-ci sont possibles et non obligatoires.

5. Il s'agit là par contre d'une supposition. Le classement étant une opération manuelle est sujet non seulement à quelques variantes d'appréciation (WONGEL, 2005) et n'est pas à l'abri non plus des aléas humains des offices. Le réseau peut ainsi être bruité par des classements secondaires qui sont présents plus pour l'application potentielle qu'une description de la construction de la technologie.

issue de la relation mise en évidence par les liens entre les nœuds.

TABLE 5.2 : Les réseaux de croisements

Réseau	**Code**	**Description**
Inventeurs et mandataires	Applicant Inventor	Présente qui travaille pour qui (au moment précédant le dépôt)
Déposants et technologies	Applicant CrossTech	Présente les domaines de compétences technologiques des entreprises
Inventeurs et technologies	Inventor CrossTech	Présente les expertises technologiques des inventeurs
Pays et technologies	Country CrossTech	Présente les dominantes technologiques des pays

La dimension temporelle prend tout son sens pour ces types de réseaux qui décrivent des relations effectives entre des entités différentes. Au plan de la classification, l'hypothèse est faite que le dépôt d'un brevet dans un domaine technologique sous-entend une expertise de ce domaine.

5.3.3.3 Réseaux de citations

La dernière catégorie de réseaux est la plus fréquente des analyses bibliométriques en référant aux citations des brevets. Le document brevet s'inscrit dans l'ensemble des documents brevets en portant référence aux autres documents brevets. Bien que difficiles à expliquer dans un cadre général, ces citations de brevet traduisent des compositions de technologies (brevets citant d'autres brevets), un reflet de la qualité/valeur/intérêt (HARHOFF, SCHERER et al., 2003 ; THOMA, 2014) d'un brevet par le nombre de citations reçues.

Les citations peuvent aussi être utilisées en tant que témoins de la circulation des flux de savoir (HU et A. B. JAFFE, 2003 ; ALCACER et GITTELMAN, 2006 ; BACCHIOCCHI et MONTOBBIO, 2010), la valeur d'un marché (B. H. HALL, A. JAFFE et TRAJTENBERG, 2005).

L'importance d'un brevet (ou celle attachée par son déposant tout au moins) peut se mesurer par le nombre de ses équivalents. C'est ainsi un ensemble de trois réseaux dirigés (le sens des relations porte l'origine et la destination de la citation) qui est créé (cf. tableau 5.3) pour représenter les indicateurs précédents d'un univers brevet donné.

TABLE 5.3 : Les réseaux de références

Réseau	**Code**	**Description**
Références	References	Ensemble des références des brevets de l'UB internes à l'UB incluant les références bibliographiques externes (académiques, journaux, etc.)
Citations	Citations	Ensemble des documents de l'UB cités par d'autres documents brevets (pas forcément de l'UB)
Équivalents	Equivalents	Réseaux de brevets équivalents

Chaine herméneutique ouverte

Et si d'autres réseaux se révélaient être d'intérêt pour des interrogations en intelligence compétitive? Quels sont les réseaux de partenaires par pays? Ou encore par domaine technologique? La modification des chaînes de traitement préétablies est relativement simple et permettra la création d'une instrumentation dérivée propre à cette question.

5.4 Textométrie

Développée dans les années soixante comme composante de l'analyse des discours, la lexicométrie regroupe différentes perspectives d'herméneutique textuelle. Le développement de ces techniques d'analyse permettent de dépasser la notion de transformation de texte en « sacs de mots » qu'on lui a longtemps reproché. Si d'emblée

la représentation d'un corpus par les termes (non vides) statistiquement les plus fréquents permet de se faire une idée des contenus abordés [6], les techniques développées en analyse textuelle (LEBART et SALEM, 1994) dans la lignée de l'analyse des données de BENZÉCRI (BEAUDOUIN, 2016) complètent par quelques solutions matures applicables directement aux contenus des brevets pour apprécier rapidement les masses documentaires.

En sus des données descriptives des documents brevets, l'EPO fournit aussi, si disponibles, des données textuelles correspondant aux résumés, aux descriptions et aux revendications.

P2N traite ces contenus pour les ouvrir à des traitements de textométrie et d'analyse de données textuelles. Il s'agit d'opérer des calculs au niveau des termes utilisés dans ces parties textuelles des documents brevets. L'analyse se situe d'abord au niveau lexico-sémantique par la production de l'inventaire lexical et terminologique de l'univers brevet traité afin d'en extraire l'essentiel.

Puis il s'agit de construire les réseaux de notions identifiées par les associations terminologiques au sein des documents. La flexibilité du point d'entrée (les notions) est nécessaire pour que l'utilisateur opère à l'utilisation contextuelle de cette exploration. Pour faciliter la lecture, il s'agira alors de visualiser les termes les plus fréquents, les associations lexicales par des techniques issues des travaux de la méthode Leximappe (CALLON, J.-P. COURTIAL et W. TURNER, 1991). En complément, la technique de classification hiérarchique ascendante associée à la méthode Alceste (A. REINERT, 1983; M. REINERT, 1986; M. REINERT, 1990) facilitera le regroupement en classes des différentes associations terminologiques pour tenter de regrouper autour d'éléments sémantiques forts identifiés dans le corpus. Pour compléter l'instrumentation, chaque type de texte (résumé, description, revendication) sera marqué par les descripteurs issus des données bibliographiques (date, classification, auteur). Ce marquage permettra l'introduction de « points de vue » spécifiques aux visualisations : par date, par déposant, par classification pour faciliter la lecture et le traitement.

Les données sont formatées pour l'outil IRaMuTeQ (Interface de R pour les Analyses

6. En particulier dans le document brevet : les contenus sont relus et validés, décrivant l'invention les revendications et devant résumer au mieux cette dernière.

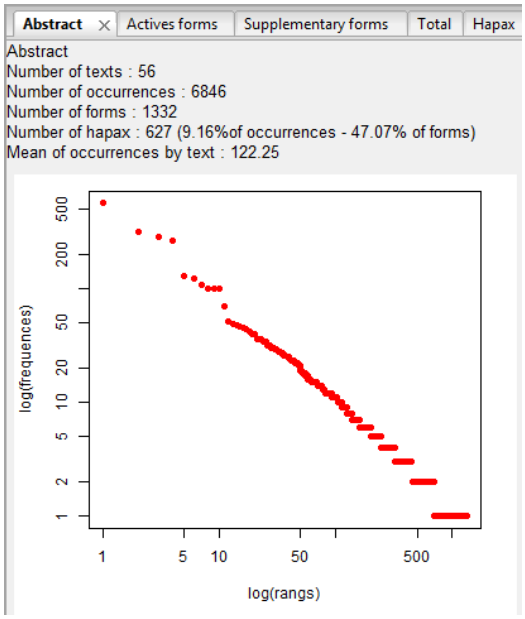


FIGURE 5.4 : Statistiques élémentaires sur un univers brevet

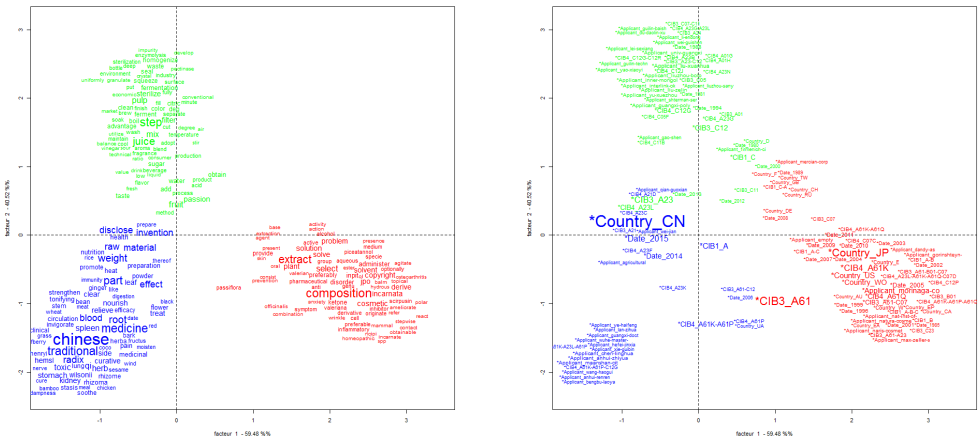
Multidimensionnelles de Textes et de Questionnaires) de Pierre RATINAUD (RATINAUD, 2008) qui délivre une interface simple à la réalisation de ces opérations. Je fournis à titre d'exemples quelques visualisations issues des traitements par cet outil de cas particuliers. Ces travaux en sont à leur début, de nombreuses pistes sont en vue pour performer l'ensemble du dispositif (termes vides, construction de dictionnaires, etc.).

5.4.1 Statistiques élémentaires

La figure 5.4 délivre quelques statistiques élémentaires : nombre de textes, nombre de termes, formes identifiées, hapax et moyenne des occurrences par texte. Le nombre d'hapax est relativement élevé mais traduit d'une part la spécificité de certains brevets mais aussi les erreurs fréquentes des OCR. La caractérisation est toutefois utile pour apprécier l'ensemble documentaire en vigueur.

5.4.2 Les projections multidimensionnelles

Le marquage des textes par leurs métadonnées bibliométriques permet de construire la mise en relation des associations lexicales (les mots qui se retrouvent à côté d'autres au sein d'une fenêtre prédéterminée et paramétrable dans le logiciel) et de segmenter



(a) Lexique (b) Métadonnées

FIGURE 5.5 : Représentation du lexique et des métadonnées associées selon la même projection et positionnement sur le plan

selon des vues propres aux métadonnées (dates, mandataires, inventeurs, etc.).

Ainsi, la figure 5.5 fournit une projection multidimensionnelle du corpus textuel associé à l'univers brevet « Passiflore » distinguant les trois classes identifiées par couleur (à gauche). À droite, se trouve la représentation des métadonnées associées selon la même projection et positionnement sur le plan.

Les figures 5.5 montrent un des résultats de traitement d'analyse factorielle des correspondances par IRaMuteQ appliqué à l'univers brevet de la passiflore. Trois catégories sont identifiables par couleurs et permettent de lire l'utilisation de cette dernière pour la médecine (essentiellement chinoise) en bleu, pour les jus de fruits (en vert), et pour la composition chimique (pharmaceutique). La figure de droite permet de conforter cette lecture (à l'aide de la classification (champs CIB)), de positionner les termes les plus proches de leur métadonnées associées (pays, date) en prenant garde aux erreurs de projection. Le lecteur intéressé par cet outil en réfèrera à (MARTY, MARCHAND et RATINAUD, 2013). IRaMuTeQ réalise nombres de cartographies après un traitement lexical classique (lemmatisation, fenêtres) paramétrable et appuyé sur des dictionnaires : nuage de mots, classification ascendante hiérarchique de REINERT (1990), et analyses multidimensionnelles selon les cooccurrences terminologiques (LAFON, 1981). Les descripteurs permettent l'accès à la construction de points de vue spécifiques (une date, un déposant ou les deux) et les distinctions de nature textuelle (revendications, résumés ou descriptions) permettent d'élaborer des stratégies d'analyse

différentes offrant une herméneutique numérique du texte relativement aboutie pour reprendre l'expression de MAYAFFRE (2002).

Textométrie augmentée

L'essentiel est de considérer la possibilité d'intervenir sur ces traitements lexicaux en amont, tant sur la préparation des données que sur les dictionnaires terminologiques utilisés. L'ensemble peut être mis en phase avec des bases de connaissances pour répondre à une logique de traitement contextualisée encore plus fine.

5.5 Classification automatique des documents

Développé à l'origine pour la classification des résultats de recherche (en particulier l'algorithme Lingo, (OSINSKI, STEFANOWSKI et WEISS, 2004)), l'outil Carrot2 (WEISS et OSINSKI, 2004), est agencé en complément de P2N pour réaliser séparément une classification des résumés, description et revendications des univers brevets. Carrot2 permet, de façon paramétrable en utilisant différents algorithmes, de réaliser une classification automatique des textes (OSISKI et WEISS, 2005) puis de les présenter de façon graphique et interactive. Nous renvoyons le lecteur intéressé à (GONZALES-AGUILAR et RAMÍREZ-POSADA, 2012) pour une présentation complète.

L'interface de Carrot2 affiche ici le traitement par l'algorithme kmeans appliqué sur les résumés de l'UB passiflore, avec 15 en seuil de fréquence d'un terme pour son utilisation. La visualisation dans la partie droite présente chaque classe en taille proportionnelle à sa dominance du corpus.

La figure 5.6 montre le traitement par Carrot 2 des résumés issus du corpus passiflore. La classification correspond à la découverte de similarités fondées sur la présence terminologique dans les textes (pris ici dans leur ensemble). L'outil est paramétré ici pour découvrir 11 classes de textes sur lesquelles il positionne les étiquettes issues des termes les plus fréquents de la classe. C'est une autre lecture que la précédente dont la complémentarité se pose en évidence : ici l'instrument sert à rapprocher les textes entre eux, et les étiqueter. On peut lire à travers les classes à droite de la figure une

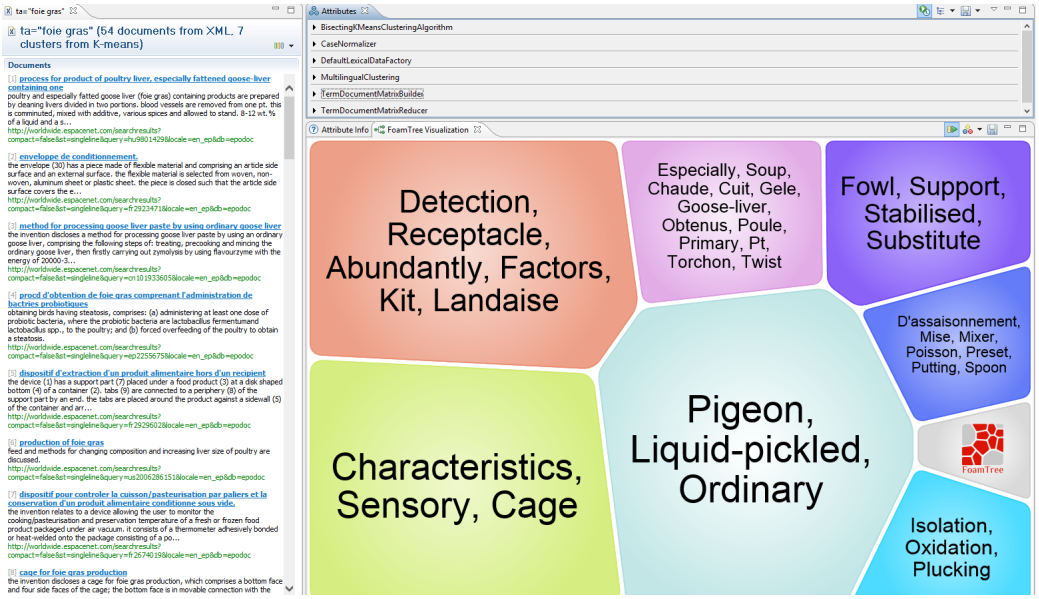


FIGURE 5.6 : L'interface de Carrot2 : à gauche la liste des documents traités, en haut les attributs de classification et la représentation type « diagramme de Voronoï » des classes identifiées proportionnellement à sa dominance du corpus.

précision sur les trois classes identifiées par le traitement textométrique. Le traitement précédent visait à souligner les termes les plus fréquemment associés entre eux au sein des phrases, le recours à la classification automatique porte sur l'utilisation de chaque texte dans sa globalité. L'articulation des deux approches est complémentaire pour faciliter la lecture globale du corpus. Sans détailler les fonctionnalités fines, il paraît d'évidence l'utilisation dérivée de cette instrumentation : soit une approche des signaux forts (les classes identifiées), soit des signaux faibles (ces documents inclassables). Une fois encore, dériver cette chaîne permet de construire aisément des chaînes spécifiques pour répondre par exemple à une mise en perspective historique (réaliser des corpus par la date de publication), ou une perspective singulière (corpus par pays, par inventeur, par déposant) pour découvrir et cerner les spécialisations, voire de positionner des textes internes par rapport à ceux issus du monde brevet.

5.6 L'approche méta

Une des propriétés remarquables du document brevet est sa description par la classification internationale (cf. section 4.1.3.3 p. 107). Réalisée par les auteurs, puis vérifiée

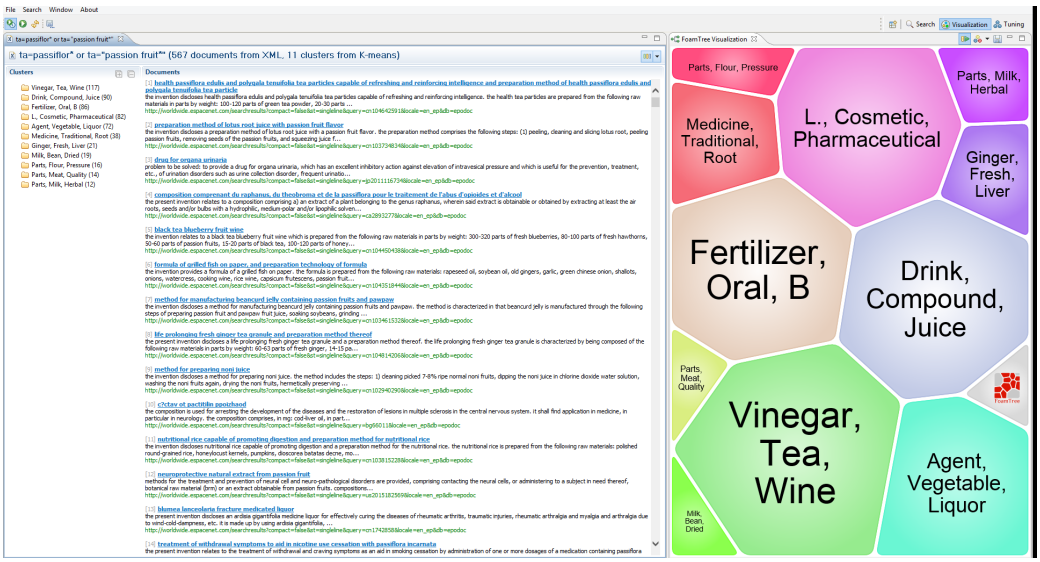


FIGURE 5.7 : Traitement par Carrot 2 des résumés issus du corpus passiflore par l'algorithme kmeans avec 15 en seuil de fréquence d'un terme pour son affichage.

et amendée éventuellement par les experts, elle constitue une vue indépendante de la langue sur une invention. CHUNG-HUEI et CHAN-YI (2015) prônent l'utilisation du document brevet comme ressource documentaire dans les bibliothèques et notent la nécessité de la prise en compte de la classification et son organisation hiérarchique pour analyser les documents.

P2N ouvre cette voie en réalisant, pour un univers brevet, une projection des classifications de chacun des documents sous forme de carte heuristique au format ouvert Freeplane (POLIVAEV, 2000). La figure 5.8 montre la projection des données de classification en une carte heuristique (éditable) au format Freeplane. Les codes de classification sont étendus par leur description en anglais pour faciliter la lecture, la taille de la police est fonction de la fréquence du code de classification, sa couleur fonction de la classe (huit au total dans la hiérarchie qui comporte dans la version actuelle plus de 100000 entrées). L'instrument permet alors de manipuler, commenter, organiser, dérouler la hiérarchie de la classification, de composer aisément une cartographie du domaine, identifier des opportunités et les saturations technologiques éventuelles. Dans une étude récente, (DIRNBERGER, 2016) détaille les apports de ces représentations pour l'analyse brevet professionnelle. Ces représentations sont reconnues (ibid.) pour faciliter la compréhension, la créativité, et facilitent la communication et la médiation de ces documents (présentation, interactions, adjonction de figures, commen-

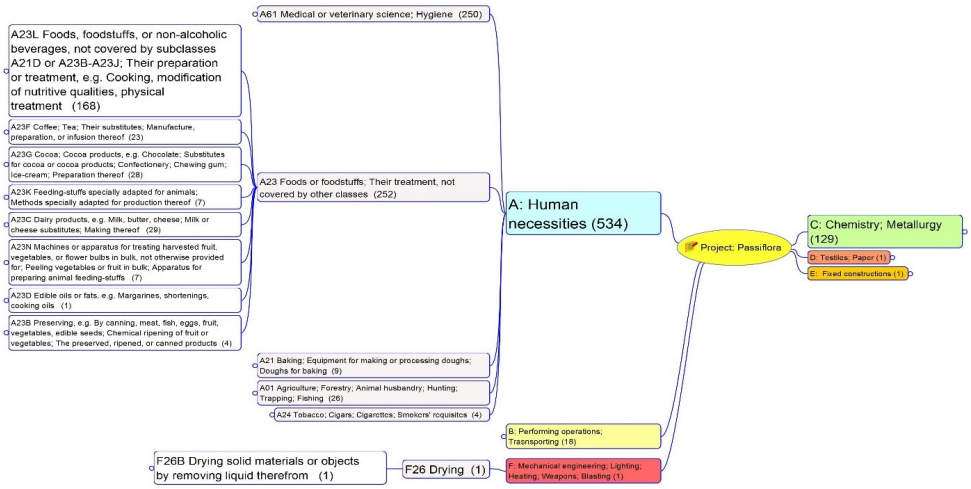


FIGURE 5.8 : Carte heuristique Freeplane de la plassiflore. Les codes de la CIB sont distingués par couleur (niveau 1), et explicités en langue anglaise

taires, etc.). La figure 5.8 montre l'univers de la passiflore sous l'angle des descripteurs de la classification.

5.7 Conclusion du chapitre

La documentation brevet est une source majeure de l'intelligence compétitive et la base européenne des brevets alimente le phénomène croissant open-data. Plus de 150000 publications mensuelles nouvelles sont recensées à ce jour. P2N offre une gamme d'outils modulaires et open-source, facilement dérivables et exploitables pour l'analyse des documents brevets mais aussi la mise en œuvre de chaînes de traitement spécifiques à un contexte particulier (organisation, entreprise, collectivité territoriale). La mise en uvre ouverte (modifiable et adaptable) de ces outils de médiation de cette documentation se positionne ainsi comme une instrumentation de construction de moyens de pilotage sémiotiques rendant concrètes les relations qu'entretiennent open data, datamining et algorithmique (CARMES et NOYER, 2014). L'enjeu en est la production d'indicateurs spécifiques, permettant une lecture selon un point de vue particulier utile à une situation de veille informationnelle, à une organisation, à un instant de cette documentation et éventuellement en combinaison avec d'autres sources. Les différents exemples présentés sous-tendent l'utilisation de ressources complémen-

taires pour élaborer des analyses fines des données informationnelles. Le code open-source de P2N et les chaînes de traitements mise en œuvres sont ré-exploitables pour être complétées, affinées et augmentées pour améliorer la valeur informationnelle des connaissances extraites. In fine, cette approche resitue l'open data et ses outils d'intelligence économique comme l'instrument qui sert à en fabriquer d'autres (C. DESCHAMPS et MOINET, 2011), pour opérer une herméneutique renouvelée, une approche objectivée et pragmatique de la volumétrie, de la volatilité et de la variété des données : méthode directe pour « passer de l'information à la connaissance », du « savoir pour agir » au « connaître est agir » (MOINET, 2011). À vocation pédagogique et académique, P2N souffre de quelques écueils et évolue au gré des promotions tout en offrant matière aux formations engagées dans les humanités digitales en ouvrant la pratique à nombre de compétences clés du domaine : tant au plan de la composition avec les chaînes existantes, qu'au niveau de l'apprentissage de logiciels clés de textométrie, classification ou de traitement des réseaux. La dimension ouverte des données collectées conduit à des traitements aisés approfondis et paramétrables à souhait pouvant être utilisés dans de nombreux contextes (juridique, historique, sociétal, économique...).

CHAPITRE 6

L'apport infométrique

La médiation entre l'homme et le monde devient un monde, la structure du monde (Simondon, 1969).
C'est donc toute une coalescence théorique qui doit, selon moi, être mise en chantier, de la transmission, vers la médiation et, plus fondamentalement, vers notre rapport au monde...
de sorte qu'on n'est peut-être pas loin de pouvoir penser que le terme « information » devrait dorénavant être compris comme la substance même de ce rapport.

— C. BALTZ, Tous Shanoniens ? (2007)

LES INSTRUMENTS ET PROCÉDÉS des deux chapitres précédents explicitent des briques médiatrices. Ces briques sont fondées séparément sur les différentes strates informationnelles issues de la structure élémentaire légale de la documentation brevet : texte, métadonnées et classification descriptives.

Et, comme nous l'avons vu, les diverses instrumentations ouvrent à un éclairage nouveau des univers brevets en aidant leur distinction, leur classement et associations que la quantité informationnelle rend difficile à appréhender. La focalisation sur l'essentiel est un enjeu.

Parmi les outils, IRAMuTeQ s'appuie sur les propriétés lexicométriques des textes pour identifier les relations et thèmes en réalisant les traitements de la méthode des mots associés. Carrot2 opère une classification automatique des documents. Cette classification est fondée sur la représentation vectorielle des documents, représentations auxquelles sont appliqués différents algorithmes. Freeplane permet de projeter les classements des différents brevets et de réaliser une cartographie technologique fondée sur les classes de la Classification Internationale des Brevets (CIB). Cependant chacun de ces outils bride l'explorateur dans une partie seulement de l'information disponible : d'un côté les associations terminologiques, de l'autre les contenus des résumés vus comme des « sacs de mots », et enfin le classement des brevets dans une des branches (ou plusieurs) de la CIB en une information séparée des précédentes.

Nous tentons de dépasser ici ces limites d'une part, en combinant différentes sources informationnelles complémentaires et d'autre part, en réalisant une hybridation des méthodes d'aide à la lecture [1].

Chaque combinaison s'élaborera de métriques de pertinence des réalisations afin d'apprécier les résultats, d'explorer les voies potentielles et de justifier les choix opérés. Nous rappelons ainsi quelques notions élémentaires utiles par la suite : les techniques discriminatoires de texte fondées sur les lois de puissance en focalisant en particulier

1. Il s'agit en quelque sorte de dépasser les apports des techniques de traitement du logiciel IRAMuTeQ, de Carrot et de Freeplane par une méthode hybride.

sur le traitement des données textuelles. En effet, un univers brevet est constitué d'un ensemble de données textuelles (les documents brevets) qui sont eux-même constitués de données textuelles complémentaires (des segments de texte ci-après) : résumé, description, revendications et métadonnées. Parmi les métadonnées, la classification internationale constitue une référence informationnelle complémentaire pouvant être convoquée pour accompagner l'interprétation de l'univers brevet.

Dans ce qui suit, nous construisons les briques élémentaires constitutives de l'outil pour apprécier son apport en :

- déterminant l'intérêt de la suppression (ou pas) des mots vides,
- identifiant la contribution des termes des différentes parties textuelles au gain informationnel selon leur origine (titre, résumé, classe),
- mesurant l'impact des mots associés (séries de mots consécutifs).

Nous aurons besoin de développer un certain formalisme qui sera introduit progressivement pour suivre les différentes constructions opérées.

6.1 La recherche des termes importants

L'annexe A présente les notions élémentaires sous-jacentes aux questionnements présentés ici. Nous complétons par ce qui constitue l'essentiel du développement de ce chapitre.

6.1.1 L'approche statistique

Au sein d'un niveau même de texte, il est admis que les termes de forte fréquence sont peu informationnels car en général ce sont les mots vides (*stopwords*) tels les articles, les pronoms, les prépositions et les conjonctions et, réciproquement, que les hapax[2] sont des mots rares faiblement représentatifs des textes (LUHN, 1957). Ces derniers détiennent cependant potentiellement la singularité, la rareté, les signaux faibles au

2. Termes apparaissant une fois (occurrence 1) dans un corpus.

sein d'un texte. L'école française de bibliométrie (LHEN et al., 1995) segmente de fait en trois parties (Trivial, Information, Bruit) la projection rang / fréquence des termes d'un corpus (cf. figure 6.1 p. 159). Les auteurs tentent d'approcher (cf. section A.1 p. 3 de l'annexe A) la zone « information » (représentée en violet sur la figure 6.1) d'un corpus textuel par une construction expérimentale à l'aide de l'entropie de RENYI[3] et de la diversité de HILL.

D'une façon un peu similaire, avec le souci d'améliorer les systèmes d'indexation, GOFFMAN dans une communication personnelle avec PAO (1978)[4] remarque le point d'inflexion des courbes zipfiennes. L'auteur remarque que ce point délimite le passage de la haute fréquence à la basse fréquence pour les termes du corpus, ce qui laisse supposer que les termes d'intérêt seraient autour de ce point. Le développement mathématique qui va s'en suivre et la vérification dans le domaine médical de ce fait (BOYCE et LOCKARD, 1975) ont validé cette supposition et permettent d'identifier précisément le point d'inflexion de GOFFMAN sur la courbe. Les auteurs suggèrent ainsi d'utiliser les quelques voisins, les mots de rang autour de ce point comme ensemble terminologique hautement informatif du corpus. En guise d'illustration, la figure 6.1 p. 159 montre schématiquement une distribution théorique[5] Rang/Fréquence suivant la loi de ZIPF, les trois parties des catégories de texte de LHEN et al. (1995) ainsi que la courbe théorique d'apport sémantique de GOFFMAN.

Le point d'inflexion de GOFFMAN se calcule en utilisant le dénombrement de termes à occurrences calculées et rangées de façon croissante :
Soit I_1 le nombre de termes d'occurrence 1 (les hapax) ; I_n le nombre de termes d'occurrences n, k une constante, et T le nombre total de mots du texte. La seconde loi de ZIPF établit que

$$I_n = \frac{k \times T}{n^2 - \frac{1}{4}} \tag{6.1}$$

Alors, BOOTH (1967) montre alors que le rapport I_n et I_1 est indépendant de la taille du texte et de la constante k. BOYCE et LOCKARD (1975) considèrent que la zone de

3. Cf. section A.1.1 en p. 3.
4. Vu dans (CASTILLO SEQUERA, 2010, p. 121).
5. On retrouvera l'allure de cette distribution dans les données empiriques utilisées par la suite.

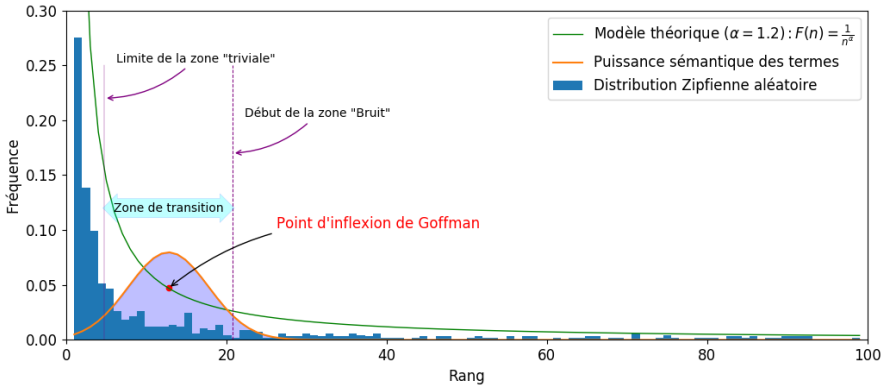


FIGURE 6.1 : Point d'inflexion de GOFFMAN, zones triviales et informationnelles d'une distribution rang/fréquence théorique

transition démarre lorsque $I_n = 1$. Ainsi, le rang du terme marquant l'inflexion est solution de l'équation $n^2 + n - 2 \times I_1 = 0$ et s'obtient par la formule :

$$n = \frac{-1 \pm \sqrt{1 + 8 * I_1}}{2} \tag{6.2}$$

Autour de ce point (une seule solution positive), les termes sont significativement plus informatifs que les autres, situés dans la zone de transition de GOFFMAN entre les termes de hautes fréquences et ceux de basse fréquence (URBIZAGÁSTEGUI ALVARADO et RESTREPO ARANGO, 2011).

6.1.2 La mise en relation

L'avènement de l'approche relationnelle de la théorie de l'acteur réseau (LATOUR, 1993 ; DEVISME, 2007) et de la traduction (J.-P. COURTIAL, 2010) trouvent manifestation par le traitement des mots associés [6] (RUIZ-BAÑOS, BAILÓN-MORENO, JIMENEZ-CONTRERAS et al., 1999 ; RUIZ-BAÑOS, BAILÓN-MORENO, JIMÉNEZ-CONTRERAS et al., 1999). Ces traitements mettent en exergue au sein des textes les thématiques : une prépondérance de certains termes qui se situent de façon centrale au niveau du réseau construit ou encore à l'interface entre différentes zones du réseau. De son côté RUIZ-BAÑOS remarque que le comportement statistique des descripteurs d'articles scientifiques issus

6. Notons que ces éléments fondent un développement d'unification des lois infométriques (BAILÓN-MORENO, JURADO-ALAMEDA et al., 2005b ; BAILÓN-MORENO, 2003).

de la base Francis entre 1980 et 1993 ne correspond pas à ce qu'il devrait être selon ces lois (RUIZ-BAÑOS, BAILÓN-MORENO, JIMENEZ-CONTRERAS et al., 1999 ; RUIZ-BAÑOS, BAILÓN-MORENO, JIMÉNEZ-CONTRERAS et al., 1999). L'auteur note des déviations très importantes entre le théorique (statistique textuelle) et ce qui est observé.

LEYDESDORFF s'oppose à l'utilisation des termes pour la considération des réseaux et l'interprétation qui peut en être faite sur les plans relationnels. Son argument principal tient à ce que les termes utilisés peuvent prendre des sens différents selon la position relative des mots (LEYDESDORFF et HELLSTEN, 2006) et les objectifs d'interprétation (LEYDESDORFF, 1992 ; LEYDESDORFF, 1997). En considérant qu'un même mot peut s'utiliser dans une phrase quelconque, dans un titre ou encore dans un texte décrivant un concept présenté dans les titre et texte précédents, son degré sémantique prend des valeurs différentes. En suivant LEYDESDORFF, l'hypothèse que l'on peut construire est que l'utilisation de ces termes sans précautions induit un bruit biaisant l'interprétation des traductions ou, d'une autre façon, que la vertu informationnelle que l'on peut leur attribuer doit être modérée.
Dans le cas présent, nous considérons un univers sémantique proche comportant un attracteur (les mots-clés utilisés dans la requête) commun. De fait, le biais qui serait induit par des mots associés pour les raisons précédentes au plan sémantique est de ce point de vue réduit. Si l'hypothèse de LEYDESDORFF se justifie dans un cadre général, le cas particulier de corpus tels que construits par requête sur les bases brevet diminue la probabilité de co-utilisation fortuites de termes et en conséquent de sémantique différente pour des mots-associés. En second lieu, la considération de mots-associés de taille supérieure à 2 diminue drastiquement la probabilité d'utilisation sémantique différente.

6.1.2.1 Combinaison des approches

Les espaces vectoriels sont un modèle fréquent de représentation des documents considérés comme structurés ou pas (BARONI et LENCI, 2008). Réciproquement ces derniers peuvent être aussi utilisés pour réaliser une représentation sémantique (LANDAUER et DUMAIS, 1997) des mots. Pour ce faire, on considére le contexte pris soit comme

le document en entier (DEERWESTER et al., 1990), soit encore comme une zone autour du mot considéré ce qui donne lieu à des niveau sémantiques différents. À nouveau, l'unité pour la zone considérée varie selon les approches (MARTIN, LIERMANN et NEY, 1998) et s'entend comme une phrase, un paragraphe, ou une fenêtre [7]. La similarité des mots co-occurrents est alors représentée par leur proximité. Ainsi, l'utilisation des bigrammes a optimisé l'efficacité de classement des algorithmes de classification contrairement à ceux utilisant seulement les unigrammes.

Un petit retour sur l'information de SHANNON montre que l'unité élémentaire « information » n'est pas définie, l'entropie mesure seulement le degré de nouveauté (Cf. A.1.1 p. 3), ou la variété-information ce qui laisse la possibilité d'identifier la notion de gain informationnel sur des plans différents. Ainsi, BALPE, LELU et PAPY (1996) soulignent que la définition de l'unité d'information dépend de l'usage qui est attendu et celle-ci est dépendante (BISKRI et DELISLE, 2001) de l'objectif de lecture et de compréhension que nous nous donnons. En conséquence, nous pouvons considérer la notion d'apport informationnel au niveau des mots (approche statistique textuelle simple), au niveau des combinaisons de mots (N-Grammes, ou mots associés).

Ainsi, d'un point de vue terminologique, n-gramme fait référence aux séquences de mots consécutifs d'au moins deux mots ($n = 2$). Mots et n-grammes sont désignés par la suite par l'appellation utilisée en linguistique (BESTGEN, 2014) « termes ». Par les termes, l'identification de réseaux de cooccurrences lexicales, permettent d'approcher une dimension thématique et sémantique des discours (MAYAFFRE, 2014).

6.1.2.2 Le cas des documents structurés

Un autre couplage est nécessaire pour compléter ce qui est introduit ici. L'hypothèse complémentaire aux faits précédemment soulignés est que les niveaux textuels induisent un contexte sémantique spécifique pour les termes utilisés. Ainsi, le même terme T_i a-t-il un impact différent s'il est employé dans un titre, un résumé ou autre en regard d'un corpus textuel?

7. Un ensemble de N mots prenant le mot en question en son centre.

Résumé et questionnement

Dans un ensemble de textes, certains termes (les mots associés) sont plus importants que d'autres au sens statistique. L'introduction de termes révèlent les thèmes et relations entre ceux-ci au sein du texte.
Toutefois, on peut supposer que le niveau de texte (résumé, titre...) impliquerait un « poids » informationnel différent.
Nous allons élaborer une série d'opérations visant à vérifier ces éléments et introduire une métrique apte à identifier cette éventuelle différenciation.

6.2 Méthodologie

Dans ce qui suit, il s'agit de construire un ensemble d'instruments développant la lisibilité des textes en s'appuyant sur la qualité intrinsèque des bases de données brevet, qui décrivent les inventions sur de multiples niveaux : description, revendications, résumés et titres sont des sous-textes d'un document et en décomposent le premier niveau. Chaque invention étant elle-même rangée par au moins un méta-descripteur positionné dans le schéma de classement international, nous développons à partir de ce descripteur, la reconstruction textuelle de ce classement associé à chaque invention pour (re)construire le second niveau textuel. Le dernier niveau est reconstruit par les mots associés dont la longueur (séries de mots construisant des termes) est variable (de 1 à 5 dans nos expérimentations) et va permettre de se distancier de la simple considération « sac de mots » pour les différents textes qu'introduisent implicitement la représentation vectorielle des documents.

Afin d'opérer les calculs afférents à la mesure des résultats, nous introduisons le formalisme nécessaire à ces différentes opérations. On supposera une relation implicite qualitative entre les différents niveaux de textes : d'une part il y a exclusion mutuelle sur leurs objectifs, et d'autre part, ils ne sont pas totalement séparés au niveau lexical. In fine, les niveaux de texte se complètent pour maximiser le gain informationnel.

6.2.1 Prétraitement

Nous faisons abstraction des prétraitements d'évidence réalisés par P2N. Nous partons d'un univers brevet (U) qui désigne un ensemble de N textes.

6.2.1.1 Augmentation textuelle par la description issue des classements

DIBIAGGIO et NESTA (2005) soulignent que la classification pourrait être l'unité d'analyse la plus appropriée pour l'exploitation de l'information brevet contenue dans les bases brevet. Ainsi, la classification précise quelques éléments à chaque demande de brevet de façon implicite. Proposée par le déposant, elle est revue et corrigée par les examinateurs. La classe primaire est obligatoire et s'appuie sur le schéma international (WIPO). Les éventuelles classes secondaires ont deux objectifs possibles : soit il s'agit d'une classification exploratoire (LEYDESDORFF, 2008) soit elles viennent renforcer la classification primaire. Dans tous les cas, le titre de la sous-classe doit être aussi précis que possible sur le contenu de la sous-classe. Avec ces contraintes, le travail des examinateurs est alors une activité intellectuelle, **de rajout de connaissance** aux documents brevet (LARKEY, 1999).

L'exploitation de cette connaissance se fait soit a priori pour identifier et construire une requête, soit a posteriori en lecture complémentaire de la description d'un brevet. Nous proposons ici de combiner cette description à la description textuelle des brevets afin de la rendre explicite pour son exploitation par les outils de traitement automatique.

L'effort documentaire est réalisé sur deux plans : le classement dans une des branches de la classification et (antérieurement) la rédaction du texte descriptif de chaque classement possible. Ces deux éléments sont en soit une source d'information documentaire extérieure et complémentaire aux constituants du document brevet mais ne sont disponibles qu'à l'extérieur de ceux-ci.

ə

Un texte descriptif normalisé

La description experte unanime et internationale des inventions constitue un élément informationnel complémentaire aux résumés réalisés par les inventeurs. Une alternative normalisée à la description de l'invention.

Un effort de rangement

L'effort de classement d'une invention dans un champ de la classification constitue une action documentaire forte, un complément d'information potentiel : le fait qu'une invention soit classée dans une catégorie (hiérarchie de) cadre cette invention de façon univoque par rapport à sa description et son résumé que l'on peut supposer plus précis.

Biais potentiel

Au risque d'introduire un biais, ne pouvant distinguer automatiquement ce qui relève de l'exploratoire de ce qui est complémentaire lors du classement des inventions dans les textes des classes (IPC), nous utiliserons l'ensemble des classes disponibles pour chaque document brevet.

6.2.1.2 Distinction de l'origine du vocabulaire

Soit $U = \{Res_i, Des_i, Tit_i, Cla_i, IPC_i\}$ l'univers brevet un ensemble de documents composés des contenus textuels respectivement les résumés, les descriptions, les titres, les revendications (claims) et le texte de classes (IPC) IPC_i. Ce dernier se construit récursivement comme l'ensemble des textes de description du code de classement le plus précis disponible (IPCR11) [8] associé au texte décrivant les nœuds parents de la hiérarchie jusqu'à la branche primaire (section).

Les revendications et descriptions n'étant disponibles dans la base de l'EPO que sous certaines conditions, nous restreignons la construction des corpus textuels U dans ce qui suit à

$$U = \{Res_i, Tit_i, IPC_i\}_{i \in \{1,\cdots,N\}}. \tag{6.3}$$

8. Si le niveau 11 n'est pas disponible, alors le premier niveau supérieur disponible et parents sont utilisés.

De même nous appliquons un procédé de filtrage pour rendre consistant l'ensemble : sont exclus les entrées vides ou dans un autre encodage (titres en coréen, japonais ou chinois par ex. comme cela peut se produire parfois dans la base).

Ainsi pour un document brevet, le nouveau texte le représentant est construit par l'addition de son titre, son résumé et la description textuelle des classes dont il relève. Soit au plan formel : $\forall D \in U, D = Res + Tit + IPC$

Les termes pouvant être utilisés en tant que titres, dans le résumé ou encore dans le texte du classement, nous distinguons pour la suite leur origine que le schéma ensembliste (cf. diagramme de VENN 6.2 p. 165) résume avec les notations précédentes : Res, Tit et IPC désignant les textes provenant respectivement des résumés, des titres et des descriptions textuelles de classement. Notons l'apparition d'une zone supplémentaire qui recouvre les effets de bord de la procédure d'augmentation précédente : lors de l'apposition d'un texte supplémentaire à un autre, les co-occurrences peuvent générer des termes qui n'étaient ni dans le premier ni dans le second texte. Cette zone est désignée par 'Bord' dans le schéma. Les zones d'intersection désignent les parties du vocabulaire appartenant à la fois à un niveau d'origine et à un autre. Les zones triviales sont dénommées par la provenance du vocabulaire.

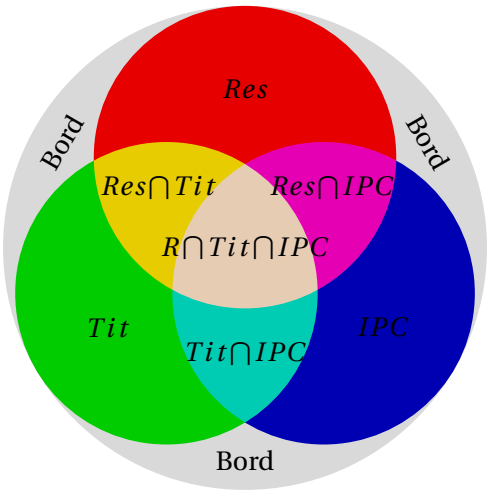


FIGURE 6.2 : Ensembles terminologiques et leur niveau textuel d'origine ou construits. Le symbole ∩ désigne l'intersection.

6.2.1.3 Utilisation des mots associés

ə

Soit $\mathscr{V}$ le vocabulaire associé à un texte privé de ses mots vides[9]. $\mathscr{V}$ est de cardinal N_0. Pour l'ensemble du texte (en espérant qu'il soit de taille raisonnable), construisons l'ensemble des co-occurrences terminologiques, la co-occurrence se définissant ici comme l'apparition séquentielle d'un mot à un autre. Ainsi, le texte « les enfants vont à la plage » donnera les co-occurrences de degré 2 « enfants vont » et « vont plage ». Chacune de ces co-occurrences est alors ajoutée au vocabulaire qui n'est plus constitué d'un ensemble de simple mots mais de *termes*.

Construisons alors le modèle récursif de cette construction au degré n. Soit $\mathscr{V}$ le « vocabulaire » constitué par les co-occurrences de degré n (n-grammes) pour $n > 1$ fixé. Le cardinal de $\mathscr{V}$ est alors au mieux égal à N_1 (pas de mots associés) augmenté des bi-grammes ($n = 2$), puis des tri-grammes ($n = 3$) et ainsi de suite. Avec l'exemple précédent, l'ensemble des termes du texte au final sera le suivant : ['enfants', 'vont', 'plage', 'enfants vont', 'vont plage', 'enfants vont plage'], soit six termes au total.

Critères

L'étude ci-après fera intervenir quatre critères d'analyse pour décider de :

- l'utilisation des mots vides ;
- l'inclusion des mots associés (N-Grammes ou pas) et jusqu'à quel degré (valeur de N) ;
- l'utilisation de l'origine textuelle de contenus distinguant les différents apports de chaque constituant.
- les arguments d'infométrie classique permettront de déterminer la fréquence relative des termes et de statuer sur leur utilisation comme critère de segmentation.

6.2.2 Les corpus de test

En appui à des corpus déjà étudiés en collaboration (REYMOND et QUONIAM, 2016), les cas de la *peau de banane* et de la *passiflore* serviront de corpus d'étude. Ces deux

9. Nous verrons un peu plus loin la raison de cette exclusion, standard dans les traitements linguistiques.

corpus sont construits autour d'un univers sémantique particulier en tentant de déterminer ce que l'on peut réaliser autour de ressources naturelles.

Reproductibilité

Notons dès à présent que la limite à ces deux corpus n'a d'intérêt que pour la construction du mode opératoire. Chacun des traitements réalisés par la suite pourront être dispensés sur d'autres corpus.

6.2.2.1 La peau de banane

L'univers brevet de la peau de banane [10], désigné par la suite par U_{Banana} est construit par la requête « ta="banana* peel*" OR ta="banana* skin*" » sur l'EPO. Le premier juin 2017, 711 demandes sont rapportées par le collecteur. La famille est étendue à 756 demandes. 696 résumés en anglais seront identifiés.
Le procédé de filtrage réduira cet ensemble à 513 documents.

L'ensemble compte 7378 mots uniques, pour 113614 occurrences réparties selon le tableau 6.1 p. 167 :

Origine	Occurrences
Titres	4448
Résumés	79673
IPC	29493

TABLE 6.1 : Répartition des termes de U_{Banana} selon les origines des textes

6.2.2.2 La passiflore

L'univers brevet de la passiflore [11], désigné par la suite par $U_{Passiflora}$ est construit par la requête « ta=passiflor* or ta="passion fruit*" » sur l'EPO. Le 5 juillet 2017, 944 demandes sont rapportées par le collecteur. La famille est étendue à 1365 demandes. 890 résumés en anglais sont identifiés.
Le procédé de filtrage réduira cet ensemble à 837 documents.

10. L'ensemble des données est accessible http ://patent2netv2.vlab4u.info/DATA/banana_peel.html.
11. L'ensemble des données est accessible http ://patent2netv2.vlab4u.info/DATA/Passiflora.html.

L'ensemble compte 10349 mots uniques, pour 183134 occurrences réparties selon le tableau 6.2 p. 168.

Titres	7496
Résumés	120188
IPC	55450

TABLE 6.2 : Répartition des termes de $U_{Passiflora}$ selon les origines des textes

6.3 Analyses préliminaires

Les analyses des deux corpus précédents entendent statuer sur l'apport informationnel :

- des mots vides ;
- des mots fréquents du corpus ;
- des mots associés.

6.3.1 Mots vides ou pas ?

La figure 6.3 montre la distribution des occurrences des termes de l'$U_{Passiflore}$. Les termes sont rangés selon leur fréquence d'apparition (les hautes fréquences sont à gauche, les hapax à droite) dans une présentation logarithmique selon les deux axes. On constate que la tendance générale des deux courbes est très proche de la courbe modèle de Mandelbrot-Lévy (cf. figure A.1 en p. 6). Les deux courbes permettent de saisir l'influence des mots vides [12] (*stopwords* - sw) qui apparaissent clairement dans les hautes fréquences par un écart prononcé entre les courbes rouges et noires. L'histogramme incrusté dans la figure 6.3 représente la ventilation de l'incidence selon les origines du vocabulaire. Les nombres affichés pour chacun des sous ensembles représentent les occurrences terminologiques dans l'ensemble résultant avant et après de la suppression des mots vides [13]. L'incidence des mots vides (sw) est clairement plus

12. Définis par les linguistes comme le vocabulaire n'apportant pas d'information (adjectifs, pronoms, etc.).

13. Les documents de l'univers brevet étant en anglais, la liste des mots-vides construite pour les outils de linguistique de la bibliothèque *Natural Language ToolKit* de python 2.7 a été utilisée :

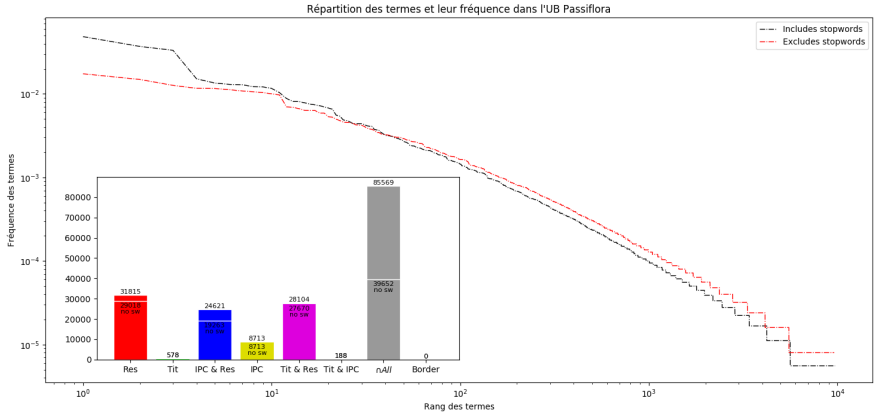


FIGURE 6.3 : L'influence des mots vides sur les distributions terminologique de l'$U_{Passiflora}$

forte sur les sous-ensembles des titres résumés et IPC (IPC & Res) et au croisement de l'ensemble des catégories de texte comme l'on peut s'y attendre.

Ces éléments sont confirmés par la ressemblance des courbes et des ventilations et répartitions des histogrammes par la figure 6.4 en p. 169 représentant la même distribution sur la peau de banane.

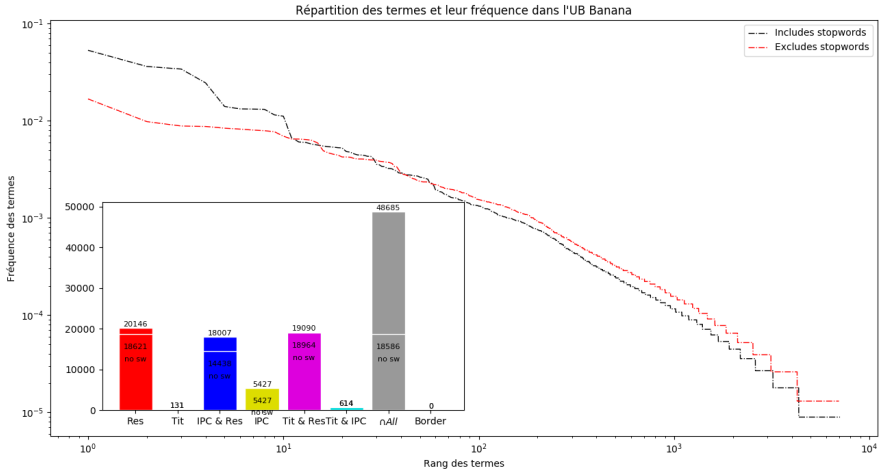


FIGURE 6.4 : L'influence des mots vides sur les distributions terminologique (Banana)

Les $n-grammes$ n'ayant pas encore été introduits, l'ensemble bordure est vide.

`nltk.corpus.stopwords.words('english')`. Cf. (BIRD, KLEIN et LOPER, 2009, p. 60) pour une description détaillée.

6.3.2 La zone de transition de Goffman

ə

Afin de compléter l'analyse, regardons l'incidence des mots vides sur les termes de la zone de transition de GOFFMAN. Arbitrairement nous prenons un ensemble de dix termes autour du point d'inflexion. Ces termes sont ceux que l'on retrouve dans les tableaux 6.3 et 6.4 des univers respectifs Passiflora et Banana. Leur position relative sur les représentations rang fréquence sont données respectivement par les graphiques 6.3 p. 170 et 6.6

avec stopwords (rangs 59-68)	**sans**
discloses	herbal
powder	specific
nutritive	subclasses
weight	pulp
following	cooking
preservation	beverage
medicine	sugar
that	vinegar
qualities	modification
water	chemical

TABLE 6.3 : Les termes autour du point d'inflexion de Goffman (+-5) de $U_{Passiflora}$

avec stopwords (rangs 47-56)	**sans**
veterinary	non
hygiene	specific
tea	blood
part	purposes
discloses	milk
feed	dental
compositions	comprises
chinese	medicine
organic	macromolecular
fertilisers	one

TABLE 6.4 : Les termes autour du point d'inflexion de Goffman (+-5) de U_{Banane}

Élimination des mots-vides

Les corpus seront dorénavant expurgés des mots vides.

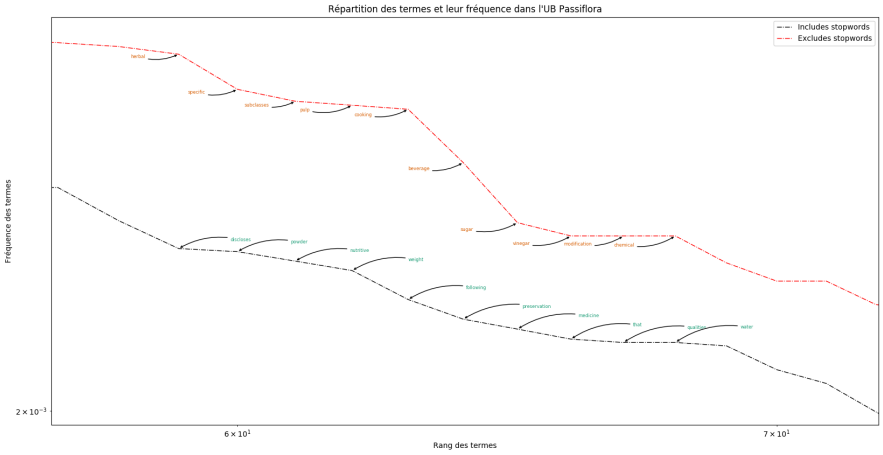


FIGURE 6.5 : Les 10 termes de la zone de transition (Passiflora)

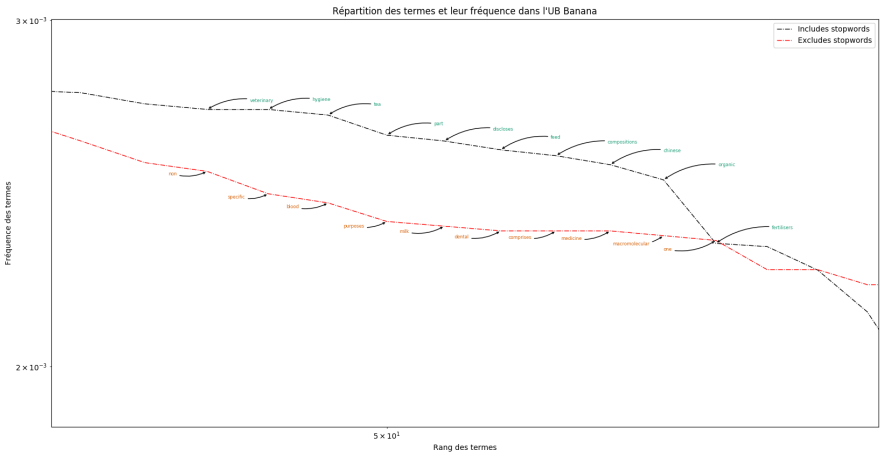


FIGURE 6.6 : Les 10 termes de la zone de transition de (Banana)

6.3.3 Inclusion des N-Grammes

L'inclusion des N-grammes [14] va clairement modifier la répartition statistique terminologique : certains bi-grammes (degré 2 i.e. $N = 2$) ont une fréquence statistique plus importante que les termes autour du point d'inflexion de GOFFMAN. Le tableau 6.5 p. 172 montre ce décalage dans U_{Banane}.

L'incidence majeure des N-Grammes est visible sur la taille du vocabulaire, l'ensemble des termes qui croît exponentiellement avec le degré des N-grammes. Le tableau 6.6 montre l'évolution sur l'ensemble des termes du résumé selon le degré des N-grammes.

14. Dans ce qui suit, le degré des N-grammes désigne la valeur de N.

sans stopwords (rangs 47-56)	bi-grammes (rangs 49-58)
non	blood
specific	purposes
blood	milk
purposes	comprises
milk	medicine
dental	dental
comprises	macromolecular
medicine	one
macromolecular	products
one	acid

TABLE 6.5 : Les termes autour du point d'inflexion de GOFFMAN (+-5) (Banana)

En comparaison, les autres segments terminologiques sont relativement plus petits. Les figures 6.8 p. 173 et 6.7 p. 173 montrent ces autres sous-ensembles de termes qui, entre eux, sont plus comparables en taille mais qui suivent globalement la même évolution exponentielle par la taille. La considération de ces éléments montre l'incidence de l'effet de bord de l'opération d'accolement des textes issus des classements de la CIB des différents documents brevets avec l'ensemble *Border* qui prend une taille non négligeable pour des valeurs de N>3.

Il est intéressant de noter que la séparation est conservée et que des termes qui, pris séparément, appartiennent à la zone d'intersection de l'ensemble des segments terminologiques ($\cap All$), se retrouvent « ventilés » par des mots associés dans leur zone d'origine. En prenant pour hypothèse que la zone d'intersection relève du « banal » en regard des zones de titres ou IPC à valeur descriptive plus forte ou contrôlée que la précédente, ces termes constituent une piste potentielle [15] de termes auxquels l'on peut attribuer une forte valeur sémantique.

Degré	Banana	Passiflora
1	5099	7843
2	34219	47374
3	76022	101913
4	123124	162749
5	172300	226473

TABLE 6.6 : Évolution de la taille du nombre de termes issus des résumés dans les deux corpus selon le degré des N-Grammes

Il est intéressant de constater aussi l'incidence des N-grammes sur la position de la zone de transition de GOFFMAN. La figure 6.9 en p. 174 montre dans l'U_{Banane} la posi-

15. Si le temps ne nous a pas permis de la suivre par la suite, le formalisme et les travaux de structuration des analyses effectuées préparent grandement le terrain pour ce faire.

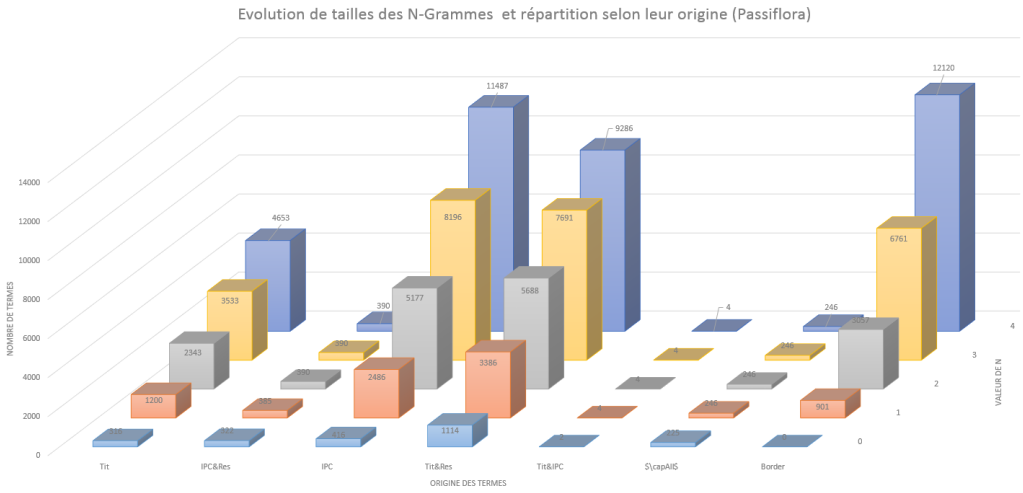


FIGURE 6.7 : Évolution des tailles des ensembles de vocabulaire selon leur origine (Passiflora)

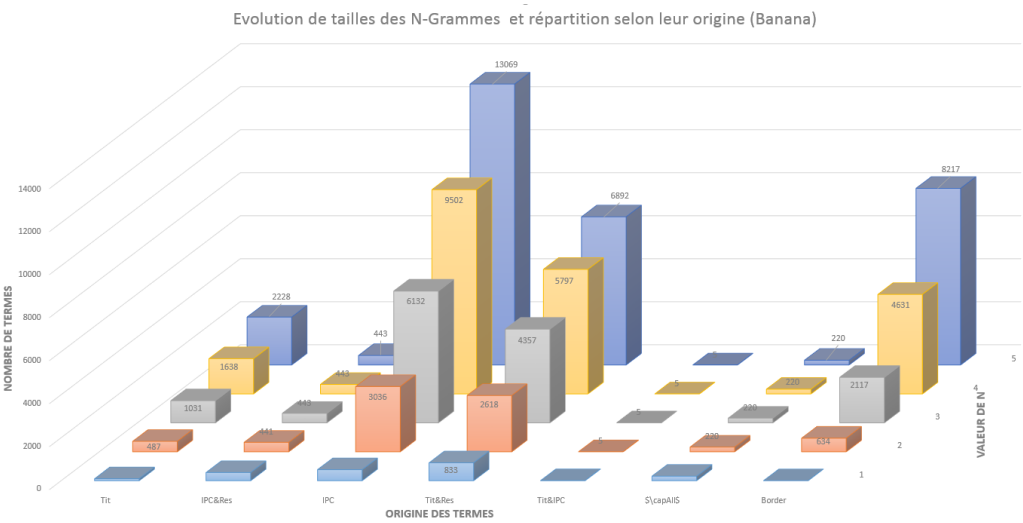


FIGURE 6.8 : Évolution des tailles des ensembles de vocabulaire selon leur origine (Banana)

tion relative des courbes rangs/fréquence selon le degré des N-grammes et la zone de GOFFMAN (10 termes) marqués par des symboles en rouge. Le second graphique encadré montre l'évolution des rangs sur les deux corpus, les rangs au sein des corpus de la passiflore en fonction du degré des N-Grammes augmentant de façon plus marquée comme attendu de par la taille relative des corpus.

Ainsi, le degré des N-grammes influe sur la zone de transition. Il en est de même des zones de coupure de l'information triviale (C_t) et bruitée (C_b) de ces distributions (cf. section A.1.1 p. 3) comme le montre le tableau 6.7 p. 174 décrivant l'évolution des rang de coupure selon le degré des N-Grammes.

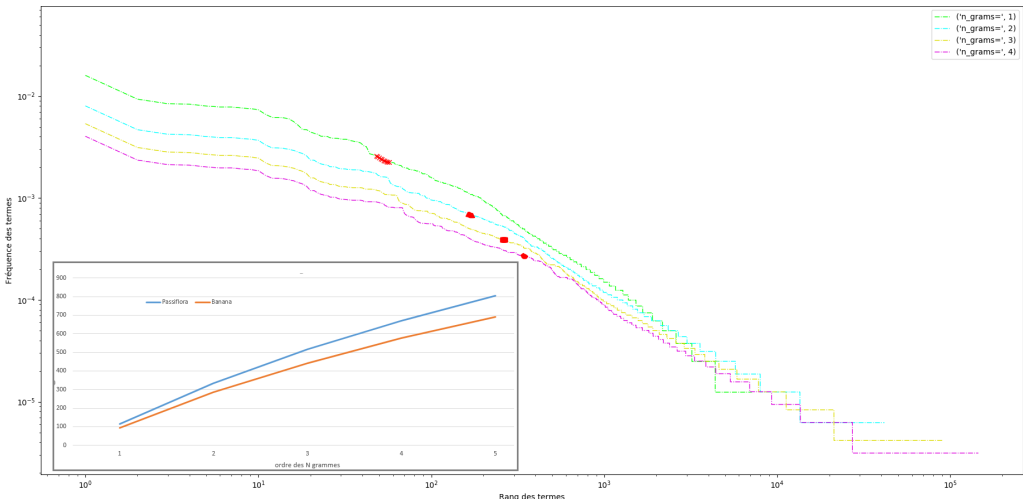


FIGURE 6.9 : Incidence N-gramms (Banana)

	Banana		Passiflora	
Degré	C_t	C_b	C_t	C_b
5	4763	30661	4267	48449
4	3161	26061	2936	39841
3	1680	20417	1885	30018
2	787	13085	931	18593
1	222	3189	299	4211

TABLE 6.7 : Rangs de coupure des zones information et bruit selon le degré des N-Grammes dans les deux corpus

6.3.4 De l'origine des termes

Le graphique 6.10 p. 175 montre la distribution des termes en montrant leur répartition selon leur segment d'origine représentée par leur couleur. L'impact de l'origine sur les zones des termes statistiquement « importants » est variable : l'implication croissante des termes issus de la classification que leur couleur dénote (bleu foncé (IPC), Violet (IPC et Résumés) et bleu ciel (Titres et IPC)) est clairement visible.

La figure 6.11 en p. 175 corrobore cette influence marquée des termes issus de la classification sur la zone de transition à l'aide des N-grammes. Marquant une tendance très comparable à celle du second corpus (figure 6.10) accentué [16] encore par le fait que le corpus est de taille supérieure (environ 50%).

Ces éléments nous permettent de procéder aux choix suivants pour la suite :

16. Sous-tendant par là un effet de masse dont il faudrait établir le seuil critique.

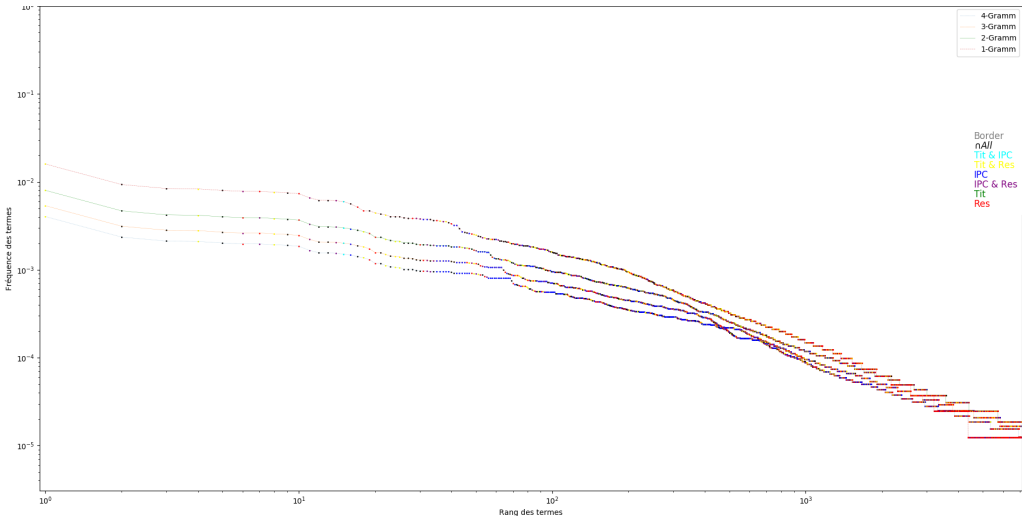


FIGURE 6.10 : Ventilation des termes selon leur origine au sein des distributions (Banana)

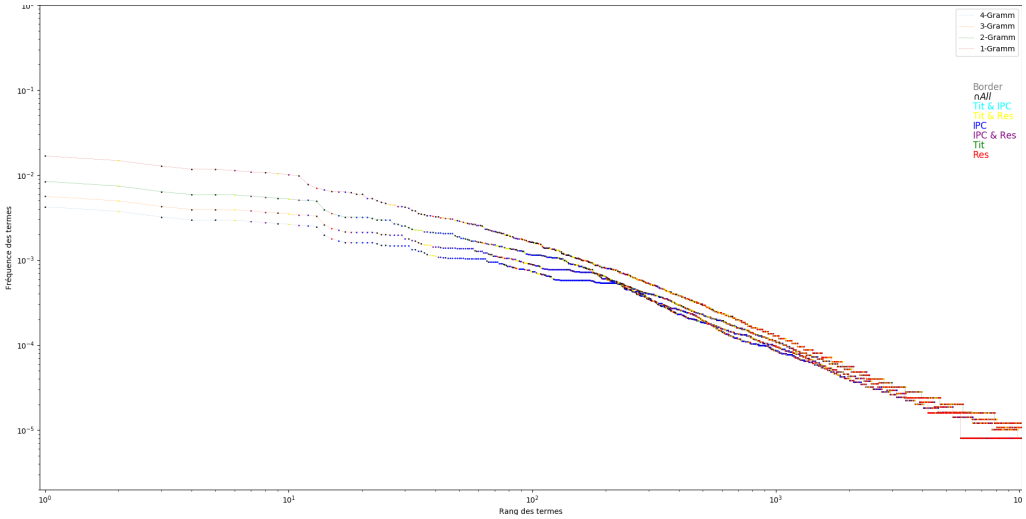


FIGURE 6.11 : Ventilation des termes selon leur origine au sein des distributions (Passiflora)

- les corpus seront nettoyés des mots vides,
- les mots associés seront utilisés jusqu'au niveau 3 pour limiter les effets de bord tout en permettant de favoriser l'incidence des termes issus de la classification (IPC) ;
- les termes de la zone de transition de GOFFMAN seront « favorisés » ;
- les termes relevant de la catégorie IPC également.

6.4 Médiation intellective documentaire

ə

Les analyses précédentes permettent de statuer sur les apports informationnels effectifs des différents procédés d'élimination des mots vides, de l'utilisation des mots associés qui permettent de dépasser la limite de la représentation sac de mots (par une représentation sac de mots complétée par un sac de suites de mots) et l'implication forte des termes saillants mais secondaires au plan statistique.

Le traitement va être le suivant :

1. reconstruction des documents par concaténation des différents segments :
   - Classement IPC
   - Titre
   - résumé
2. suppression des mots-vides ;
3. application d'un algorithme de classification automatique. Ce dernier est initialisé pour identifier un nombre de classes (arbitrairement choisi ici à 11 [17]) exclusives [18]. L'algorithme appliqué est celui des $k-moyennes$ ou $k-means$ (ARTHUR et VASSILVITSKII, 2007).
4. le corpus documentaire est alors représenté et associé à une interface. La représentation colore et identifie les différents documents par un symbole de couleur correspondant à la classe du document. Le survol du symbole par la souris complète cette information par le titre du document. Le symbole est aussi cliquable pour relier directement vers le document original (dans sa version reconstruite).

L'interface (cf. fig 6.12 p. 177) dispose ainsi de deux menus :

- le premier (en haut à droite) permet l'identification des couleurs des classes et en présente les mots les plus fréquents. Chaque classe est sélectionnable de sorte à mettre en valeur les documents correspondants.

17. Il pourrait être intéressant de fixer ce nombre non pas de façon arbitraire mais sur la base des corpus traités en utilisant, par exemple, le nombre de branches du classement IPC atteintes par le corpus.
18. Un document appartient à une classe et une seule.

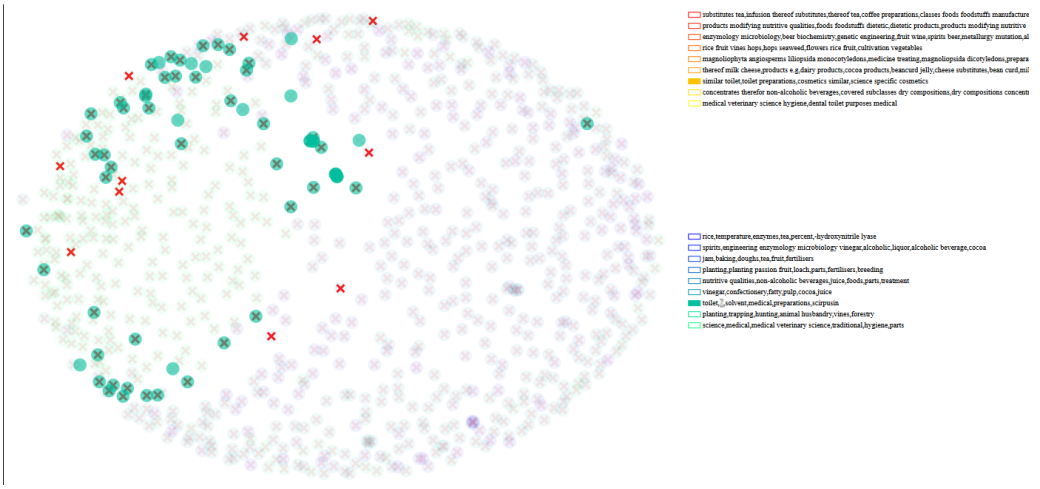


FIGURE 6.12 : Interface d'exploration par double clustérisation. Cas de la Passiflora

– le second présente les termes la zone étendue de GOFFMAN de l'ensemble des documents, termes choisis parmi les termes de l'intersection des segments IPC et des résumés pour maximiser leur apport sémantique. Tout comme le précédent, les entrées de ce menu sont sélectionnables pour mettre en valeur les documents présentant ces termes. Cette sélection affiche les documents contenant l'un des termes du menu de façon transversale aux classes construites par l'algorithme de k-« moyennes ».

Dans la figure 6.12 p. 177, une classe de chaque menu est sélectionnée. Les documents correspondant à la sélection du premier menu sont marqués d'une croix rouge et ceux correspondant au second menu sont colorés selon la couleur de la classe.

L'interface permet ainsi la double exploration du corpus. La classification classique – ici k-means++ (ibid.) – rassemble par degré de ressemblance les vecteurs représentant les documents dans l'espace des termes qui constituent le corpus en neuf[19] classes disjointes. Chacune des classes est alors représentée par les termes les plus fréquents dans la classe (menu en haut à droite de la figure 6.12).
La seconde classification introduit les termes associés issus de la zone de GOFFMAN pour offrir une autre facette du corpus. Le même principe de segmentation est alors appliqué en combinaison avec la première interface, ce qui permet d'en croiser les ef-

19. à terme, ce nombre sera proportionnel au nombre d'entrées distinctes de la CIB de niveau 3 pour tenter de coller dynamiquement au corpus.

fets.

ə

Par construction, le corpus constitué des descriptifs de la classification internationale des brevets, des titres et des résumés des documents brevets s'identifie comme « augmenté » des qualités de chacun des niveaux de texte. Chacune des deux interfaces associées au résultat produit par les classifications automatiques expose deux facettes différentes du corpus, ouvrant un éclairage complémentaire aux autres fonctions mises en œuvre dans P2N. Les étapes successives de traitement des données informationnelles laissent libre court à l'introduction de procédés spécifiques à une herméneutique particulière à développer que ce soit au plan des données construites ou reconstruites (introduction de dictionnaires de synonymes, de définitions complémentaires, autres exclusions terminologiques...) ou au plan des procédés de sélection (choix des termes de l'interface, critère de choix des termes de la zone de GOFFMAN, etc.). Cette mise en œuvre illustre la conception théorique développée dans la première partie en laissant ouvert une extension des recherches propres à des utilisations inédites de la documentation brevet, usages laissés implicites dans ce qui vient d'être développé et qui dépassent les usages habituels de cette documentation.

CHAPITRE 7

Conclusion

Si maintenant on prend quelque recul à l'égard de l'imaginaire techno-scientifique, on peut reconnaître que l'outil n'impose aucune fatalité, sinon celle des préjugés qui ont présidé à sa fabrication (et encore...) : rien n'empêche alors de changer de préjugés, et d'adopter un point de vue stratégiquement opportuniste [...]. Les critères sont alors ceux de l'efficacité et de la sagesse pratique (régulée par l'éthique)

— François RASTIER, Aparté Pédauque 2 (27 janvier 2005)

٢

L'INTERNET ÉTEND NOTRE MONDE, l'autonomie observable du réseau le rend comparable à un organisme vivant qui associe humains et non humains, associations que l'hypertexte retranscrit, et que le web mémorise en données. L'hypertexte se pose en métaphore de la Communication, mais aussi des processus sociotechniques, et de « toutes les sphères de la réalité où les significations sont en jeu » (LÉVY). Les flux de données de types et de formats variés, de dynamique propre et d'origine différentes sont aujourd'hui la source directe ou indirecte de potentiels captas à capacité informatives.

Aucun instrument ne peut rendre compte de tous les points de vue à élaborer ou des faits à discerner. LATOUR et LÉVY invitent à produire une herméneutique propre de cet écosystème, un mode de lecture qui devra se situer à un niveau conceptuellement adaptable par tracé des observables, par des (re)constructions synthétiques faisant appel pour leur entendement à tous nos sens (dont la vision). C'est ainsi que, par essence, cette herméneutique s'appuiera sur une instrumentation apte à extraire des flux de l'écosystème pour rendre possible la lecture de faits. LATOUR et VENTURINI (2006, p. 255) définissent ainsi des *oligoptiques*, les instruments capables de saisir une réalité (précise) du monde au service d'un infra-langage dont les termes « ne désignent pas ce qui est cartographié mais la façon de cartographier quelque chose à partir d'une nouvelle définition du territoire ».

La construction d'oligoptiques ou de macroscopes (BÖRNER, 2011) en SIC s'inscrit dans la lignée d'une instrumentation comparable à celles élaborées pour les sciences du vivant. Les oligoptiques font partie de l'instrumentation constitutive de la documentation d'un fait ou d'un ensemble de faits pour décrire un phénomène, réaliser une scène qui devient *in fine* observable au travers de ces captas. En ce sens, les artefacts médiateurs sont des opérateurs intellectifs des événements (numériques) qui se placent au service de la production d'informations factuelles.

Le lien avec les SIC se pose en évidence, participe d'une extension, en appui sur des données numériques, devenue nécessaire pour les Sciences de l'Information et de la

Communication et qui les amène directement aux humanités digitales. Le couplage incontournable (information, fonction de médiation) réunit en premier lieu information/communication. Cette écriture instruit aussi les prémisses d'opérateurs d'intellection des Humanités. Que ce soient le filtrage, l'agrégation, la transposition... tout élément de construction de l'émergence informationnelle se retrouve dans ce couple (information, fonction de médiation) qui transporte les éléments de son herméneutique. Pour les SIC, les artefacts se posent en médiateurs offrant la documentation de faits et libèrent un niveau informationnel nouveau. Artefacts extensibles, ouverts, partageables et adaptables à souhait sur des phénomènes informationnels et communicationnels pas forcément compliqués mais d'envergure. Ces dispositifs permettent le passage à l'échelle des mégadonnées (*big data*) et ouvrent enfin la perspective d'études scientifiques inédites en adressant la complexité humaine non plus au moyen d'une limitation à une perspective circonstanciée d'un cadre spécifique mais par le « statistiquement vrai » ou « fortement probable » des sciences naturelles.

Les notions centrales des SIC, telles par exemple l'identité (marque, organisationnelle ou individuelle) ou la mobilité (diaspora, culturelle, professionnelle...) sont représentatives d'un ensemble de questions vives de caractérisation des formes sociales, culturelles, organisationnelles et info-communicationnelles actuelles dont l'écosphère est le témoin et quelquefois la manifestation. Les mettre en regard, les assembler permet de réfléchir sur des continuités ou des discontinuités conceptuelles mais aussi empiriques qui sont au centre des réflexions des chercheurs en communication. En conséquence, l'enjeu est décisif quant à l'élaboration d'instruments et artefacts de médiation alliant la collecte d'information pour la reconstruction de faits établis de ces questions. L'exploration des résultats sur des faits observables quantifie l'épaisseur symbolique et médiatique. L'exploitation des données informationnelles et de leur fonction de médiation permet de relativiser la portée et l'incidence, laissant espérer la généralisation (sous réserves) que l'on pourra alors cerner... en degré d'intellectivité.

٢

La médiation de l'anthropocène aux Humanités : les SIC

L'anthropocène est une source potentielle de captas propres à alimenter une nouvelle forme de documentation d'événements informationnels et communicationnels. L'enjeu scientifique en est l'inscription des pratiques des humanités digitales au sein des SIC, pratiques qui dépassent la simple donnée informative pour s'attacher aux modalités de production : les médiations intellectives adressent les mégadonnées (ouvertes ou pas). Réciproquement, la captation d'événements ouvre à la reconstruction d'unités de référence libérant la perspective d'analyse comparée sur des élaborations de références normalisées.

En application directe, la documentation brevet est une source majeure de l'intelligence compétitive. La base européenne des brevets alimente le phénomène croissant de l'open-data : plus de 150000 publications mensuelles nouvelles sont recensées à la date de rédaction de ce document. Cette immense source informationnelle est négligée bien que d'intérêt en intelligence économique, marketing, innovation, et l'ensemble des sciences de l'ingénieur et de la santé en général. La base regorge d'innovations et de sources potentielles à l'inspiration, mais bien évidement d'informations stratégiques et concurrentielles. Par son volume et sa dynamique, cette documentation constitue un flux de données structurées, faciles à appréhender. Pourtant, fouiller cette documentation, extraire de l'information pertinente en regard de problématiques très variées, demeure encore d'une difficulté majeure.

P2N offre une gamme d'outils modulaires et open-source, facilement dérivables et exploitables : des artefacts de médiation pour l'analyse des documents brevets. La modularité permet la mise en œuvre de chaînes de traitements spécifiques (d'autres fonctions de médiation) à un contexte particulier (organisation, entreprise, collectivité territoriale...).

Ainsi, au-delà de la documentation d'un domaine technologique particulier [1], de l'analyse compétitive (qui travaille sur quoi, avec qui et pour qui avec quelle expertise), P2N peut être utilisé pour la production de corpus documentaires, ainsi que les chaînes de traitement spécifiques pour leur analyse. Ce peut être, par exemple, pour la construction de perspectives historiques de l'étude de l'évolution d'un domaine, ou bien l'éla-

1. Par exemple les cas de la passiflore (837 documents) et de la peau de banane (513 documents) qui sont utilisés à titre d'exemple dans le chapitre précédent.

boration d'une biographie, etc. Les différents traitements sont ouverts depuis la captation de la source brevet, leur exploration jusqu'à la présentation des résultats obtenus. Chacune de ces chaînes est ré-exploitable afin de pouvoir procéder à une contextualisation propre à des questionnements d'intelligence économique, en phase avec la dynamique informationnelle générale et la dynamique organisationnelle interne s'adaptant aux multiples usages de la veille.

Documentation brevet et médiations intellectives

Apprendre à utiliser la documentation brevet est un enjeu de taille, et la complexité impose une médiation nécessaire qui se pose au sein de recherches en SIC.
Modifiables et adaptables à souhait, les outils de médiation de P2N se positionnent enfin comme une instrumentation de construction de moyens de pilotage sémiotiques rendant concrètes les relations qu'entretiennent open data, datamining et algorithmique.

Sur un autre plan, la gamme instrumentale déployée adresse globalement les formations des humanités en général en permettant d'aborder et de pratiquer l'appréhension de données numériques au travers de corpus choisis pour :

- la pratique et l'exploration de graphes de grande taille ;
- l'utilisation des techniques des mots associés, de la classification ascendante hiérarchique et des nuages de mots clés pour explorer des corpus documentaires conséquents ;
- l'utilisation d'outils de classification automatique des documents ;
- la pratique du traitement automatique du langage ;
- l'assemblage et la combinaison de ces opérateurs.

Bibliographie générale

Livres (et chapitres de)

ABITEBOUL, S. (2013). *Sciences des données : de la logique du premier ordre à la toile*. Collège de France. T. 226. Paris : OpenEdition books (cf. p. 48, 124).

ADAMS, S. (2012). *Information sources in patents*. De Gruyter Saur. ISBN : 978-3-11-023512-8 (cf. p. 103, 113, 122).

ANDRÉ, J. (2003). *Petites leçons de typographie*. Irisa (cf. p. vii).

BACHIMONT, B. (2007). *Ingénierie des connaissances et des contenus : le numérique entre ontologies et documents*. Hermes science publ. 278 p. (cf. p. 78).

BAEZA-YATES, R., B. RIBEIRO-NETO et al. (1999). *Modern information retrieval*. T. 463. {ACM} press New York (cf. p. 120).

BALTZ, C. (2009). *Information, Shannon en questions : retour sur un concept majeur*. Paris : Europia. 108 p. ISBN : 978-2-909285-47-4 (cf. p. 48).

BATTAIL, G. (1997). *Théorie de l'Information : Application Aux Techniques de Communication*. Masson (cf. p. 44).

BENGHOZI, P.-J. et al. (2012). *L'internet des objets : quels enjeux pour l'Europe? = The internet of things : what challenges for Europe?* Paris : Éditions de la Maison des sciences de l'homme. ISBN : 978-2-7351-1258-6 (cf. p. 34).

BENOIT, D. (1995). *Introduction aux sciences de l'information et de la communication*. Paris : PUF (cf. p. 47).

BERRY, D. (2012). *Understanding digital humanities*. Palgrave Macmillan. 318 p. ISBN : 978-0-230-37193-4 (cf. p. 77).

BERRY, G. (2008). *Pourquoi et comment le monde devient numérique - (chaire d'innovation technologique Liliane Bettencourt)*. Collège de France. Paris : OpenEdition books (cf. p. 36).

BIRD, S., E. KLEIN et E. LOPER (2009). *Natural Language Processing with Python : Analyzing Text with the Natural Language Toolkit*. " O'Reilly Media, Inc." 504 p. (cf. p. 169).

BOUGNOUX, D. (2001). *Introduction aux sciences de la communication*. La Découverte. T. 2 (cf. p. 67).

BOUHAÏ, N., H. HACHOUR et I. SALEH (2014). *Les frontières du numérique*. Editions L'Harmattan. ISBN : 978-2-336-35600-6 (cf. p. 32, 33).

BOWKER, G. et al. (2014). *Social science, technical systems, and cooperative work : beyond the great divide*. Taylor \& Francis. ISBN : 978-1-317-77876-9 (cf. p. 70).

BRILLOUIN, L. (1959). *La Science et La Théorie de l'information*. Masson (cf. p. 44).

BROWN SPENCER, G. (1969). *Laws of Form*. Allen & Unwin London. 165 p. (cf. p. 70).

BUCKLAND, M. (1991). *Information and information systems*. Praeger. 225 p. ISBN : 978-0-275-93851-2 (cf. p. 69).

BURGIN, M. (2010). *Theory of information : fundamentality, diversity and unification*. T. 1. World Scientific. 672 p. (cf. p. 44, 70, 71, 73).

BUYYA, R. et A. DASTJERDI (2016). *Internet of things : principles and paradigms*. Elsevier Science. ISBN : 978-0-12-809347-4 (cf. p. 66).

CALLON, M., J.-P. COURTIAL et H. PENAN (1993). *La scientométrie*. Paris : Presses Universitaires de France (cf. p. 126, 2).

CALLON, M. et B. LATOUR (2013). *La science telle qu'elle se fait : anthologie de la sociologie des sciences de langue anglaise*. LA DECOUVERTE. ISBN : 978-2-7071-7769-8 (cf. p. 127).

CASE, D. O. et L. M. GIVEN (2012). *Looking for Information : A Survey of Research on Information Seeking, Needs, and Behavior*. 4e éd. Emerald Group Publishing. 491 p. (cf. p. 44).

CHAZELLE, B. (2013). *L'algorithmique et les sciences*. T. 229. Paris : OpenEdition books (cf. p. 72).

CLAVERIE, B. (2010). *L'homme augmenté : néotechnologies pour un dépassement du corps et de la pensée*. Paris : l'Harmattan (cf. p. 35).

COLIN, S. (2015). *Humanizing big data : marketing at the meeting of data, social science and consumer insight*. 1re éd. Kogan, Page. 209 p. ISBN : 978-0-7494-7212-2 (cf. p. 34).

COSINSCHI, E. et M. COSINSCHI (2009). *Essai de logique ternaire sémiotique et philosophique*. Lang. ISBN : 978-3-0343-0048-3 (cf. p. 46).

COUTENCEAU, C. et F. BARBARA (2014). *L'intelligence économique au service de l'innovation*. Eyrolles. ISBN : 978-2-212-55773-2 (cf. p. 110, 124).

DE KERMADEC, Y. (2001). *Innover grâce au brevet : une révolution déclenchée par internet*. Insep Editions. 152 p. (cf. p. 100, 106, 126).

DE ROSNAY, J. (1975). *Le macroscope : vers une vision globale*. Éditions du Seuil. ISBN : 978-2-02-004567-4 (cf. p. 46).

— (2007). *2020, les scénarios du futur : comprendre le monde qui vient*. Des idées \& des hommes. ISBN : 978-2-35369-013-8 (cf. p. 27, 36).

De SOLLA PRICE, D. J. (1986). *Little Science, Big Science... and Beyond*. Columbia University Press New York (cf. p. 2).

DEBRAY, R. (2000). *Introduction à la médiologie*. Presses universitaires de France. 223 p. ISBN : 978-2-13-050105-3 (cf. p. 27).

DESCHAMPS, J. (2010). *Science de l'information : de la discipline à l'enseignement*. Editions des Archives contemporaines. ISBN : 978-2-8130-0028-6 (cf. p. 46).

DESROSIÈRES, A. (2000). *La politique des grands nombres : histoire de la raison statistique*. La Découverte. ISBN : 978-2-7071-3353-3 (cf. p. 66).

— (2008). *L'argument statistique : gouverner par les nombres*. Presses de l'école des mines. ISBN : 978-2-35671-005-5 (cf. p. 65).

DOWNEY, A. et al. (2002). *How to Think like a Computer Scientist : Learning with Python*. Wellsley, Mass. : Green Tea Press. ISBN : 978-0-9716775-0-0. (Visité le 29/04/2017) (cf. p. 135).

EGGHE, L. et R. ROUSSEAU (1990). *Introduction to infometrics : quantitative methods in library, documentation and information science*. Amsterdam, New-York, Oxford, Tokyo : Elsevier Science Publisher (cf. p. 2).

ESCARPIT, R. (1991). *L'information et la communication : théorie générale*. Paris : Hachette (cf. p. 47).

ETZKOWITZ, H. et L. LEYDESDORFF (1997). *Universities and the global knowledge economy : a triple helix of university-industry-government relations*. Pinter. ISBN : 978-1-85567-421-9 (cf. p. 129).

FERRY, L. (2016). *La révolution transhumaniste*. Plon (cf. p. 35).

FLORIDI, L. (2011). *The philosophy of information*. Oxford. Oxford University Press. 405 p. ISBN : 978-0-19-923238-3 (cf. p. 44).

GIBBONS, M. et al. (1994). *The new production of knowledge : the dynamics of science and research in contemporary societies*. Sage (cf. p. 100, 126).

GRANSTRAND, O. (1999). *The economics and management of intellectual property*. Edward Elgar Publishing (cf. p. 100).

GROTHENDIECK, A. et J. L. VERDIER (1972). *Théorie Des Topos et Cohomologie Etale Des Schémas*. T. 269. Berlin, Heidelberg : Springer Berlin Heidelberg. ISBN : 978-3-540-05896-0. DOI : 10.1007/BFb0081551. (Visité le 27/03/2017) (cf. p. 39).

HAFNER, K. (1996). *Where wizard stay up late : the origins of the internet*. New York : Touchstone. ISBN : 0-684-81201-0 (cf. p. 27).

JEANNERET, Y. et B. OLLIVIER (2004). *Les sciences de l'information et de la communication : savoirs et pouvoirs*. CNRS. ISBN : 978-2-271-06244-4 (cf. p. 47).

HESS, C. et E. OSTROM (2007). *Understanding knowledge as a commons : from theory to practice*. Cambridge, MA [etc.] : The MIT Press. ISBN : 0-262-08357-4 (cf. p. 113).

HUITEMA, C. (1995). *Et dieu créa l'internet...* Paris : Eyrolles. ISBN : 2-212-08855-8 (cf. p. 27).

IBEKWE-SANJUAN, F. (2012). *La science de l'information : origines, théories et paradigmes*. Lavoisier. ISBN : 978-2-7462-8912-3 (cf. p. 45, 46, 48).

« Introduction générale » (2014). In : RENUCCI, F., B. LE BLANC et S. LEPASTIER. *L'autre n'est pas une donnée : altérités, corps et artefacts*. Hermès 68. Paris : CNRS, p. 11–14. ISBN : 978-2-271-08074-5 (cf. p. 84).

JAFFE, A. B. et M. TRAJTENBERG (2002). *Patents, citations, and innovations : a window on the knowledge economy*. MIT press (cf. p. 109).

JAKOBIAK, F. (1994). *Le brevet source d'information*. Dunod. 188 p. (cf. p. 114, 129).

JOERGES, B. et H. NOWOTNY (2012). *Social studies of science and technology : looking back, ahead*. Springer Netherlands. ISBN : 978-94-010-0185-4 (cf. p. 118, 130).

JUANALS, B. (2003). *La culture de l'information : du livre au numérique*. Lavoisier. ISBN : 978-2-7462-0691-5 (cf. p. 45).

KELLY, M. et J. BIELBY (2016). *Information cultures in the digital age : a festschrift in honor of rafael capurro*. Springer Fachmedien Wiesbaden. 479 p. ISBN : 978-3-658-14681-8 (cf. p. 71).

KOURILSKY, F. (2002). *Ingénierie de l'interdisciplinarité. Un nouvel esprit scientifique*. Paris : L'Harmattan (cf. p. 68).

LAMIZET, B. (1995). *Les lieux de la communication*. Liège : Mardaga (cf. p. 89, 91).

LATOUR, B. (2006). *Changer La Société-Refaire de La Sociologie*. Trad. par N. GUILHOT. Paris, France : La Découverte Poche. 406 p. (cf. p. 40, 41, 180).

LE COADIC, Y.-F. (2004). *La science de l'information*. Paris : PUF (cf. p. 46).

LE DEUFF, O. (2014). *Le temps des humanités digitales : la mutation des sciences humaines et sociales*. Limoges : Fyp éditions. ISBN : 978-2-36405-122-5 (cf. p. 76).

Le formulaire brevet. Comment remplir votre dossier de dépôt de brevet ? (2015). Institut National de la Propriété Industrielle. INPI. 32 p. (cf. p. 103, 107).

LE MOIGNE, J.-L. (2003). *La modélisation des systèmes complexes*. Dunod. ISBN : 2-10-004382-X (cf. p. 40).

LEBART, L. et A. SALEM (1994). *Statistique Textuelle*. Dunod. Paris (cf. p. 147).

LÉVY, P. (1997). *L'intelligence collective : pour une anthropologie du cyberspace*. La Découverte Paris (cf. p. 36).

MAIGRET, É. (2015). *Sociologie de la communication et des médias. 3e édition*. Armand Colin. ISBN : 978-2-200-28361-2 (cf. p. 50).

MANOVICH, L. (2013). *Software takes command*. Bloomsbury Academic. ISBN : 978-1-62356-745-3 (cf. p. 74).

MATTELART, A. (2003). *Histoire de la société de l'information*. La découverte. ISBN : 2-7071-4159-3 (cf. p. 87).

MAYER-SCHÖNBERGER, V. et K. CUKIER (2013). *Big data : a revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt. ISBN : 978-0-544-00269-2 (cf. p. 34).

MCCARTY, W. (2014). *Humanities computing*. Palgrave Macmillan. ISBN : 1-137-44042-2 (cf. p. 76).

MEUNIER, J. et D. PERAYA (2010). *Introduction aux théories de la communication : analyse sémio-pragmatique de la communication médiatique*. De Boeck Supérieur. 464 p. ISBN : 978-2-8041-6044-9 (cf. p. 79).

MILLERAND, F., J. RUEFF et S. PROULX (2010). *Web social : mutation de la communication*. Presses de l'Universite du Quebec. 374 p. ISBN : 978-2-7605-2498-9 (cf. p. 80).

MOINET, N. (2011). *Intelligence économique : mythes et réalités*. Paris : CNRS (cf. p. 154).

MOLES, A. (1972). *Théorie de l'information et Perception Esthétique*. Denoël, Gonthier. 327 p. (cf. p. 48, 49).

MORIN, E. (1977). *La méthode. Tome I. La nature de la nature*. Édition du Seuil (cf. p. 46, 85).

MORIN, E. et J.-L. LE MOIGNE (1999). *L'intelligence de la complexité*. L'Harmattan. ISBN : 2-7384-8085-3 (cf. p. 35, 40, 91).

MUCCHIELLI, A. (2006). *Les sciences de l'information et de la communication*. Hachette Éducation. ISBN : 978-2-01-181389-3 (cf. p. 47).

MUCCHIELLI, A. (2000). *La nouvelle communication : épistémologie des sciences de l'information-communication*. Paris : Armand Colin. ISBN : 2-200-01696-4 (cf. p. 51).

NICOLESCU, B. (1996). *La transdisciplinarité*. Ed. du Rocher (cf. p. 68).

OLIVESI, S. (2014). *Sciences de l'information et de la communication : objets, savoirs, discipline*. PUG. ISBN : 978-2-7061-2180-7 (cf. p. 47).

OLLIVIER, B. (2000). *Observer la communication : naissance d'une interdiscipline*. CNRS éd. (cf. p. 67).

OMPI (2013). *Les brevets comme moyen d'accès à la technologie. Introduction*. Genève, Suisse : OMPI (Organisation Mondiale de la Propriété Intellectuelle). 12 p. ISBN : 978-92-805-1922-8 (cf. p. 106, 115).

— (2015). *Guide de l'ompi sur l'utilisation de l'information en matière de brevets*. Genève, Suisse : Organisation Mondiale de la Propriété Intellectuelle. 44 p. ISBN : 978-92-805-2652-3 (cf. p. 102–104).

OTLET, P. (1934). *Traité de la documentation. le livre sur le livre : théorie et pratique*. Brussels : Éditiones Mundaneum (cf. p. 28, 36).

PISANI, F. et D. PIOTET (2011). *Comment le web change le monde : des internautes aux webacteurs*. Pearson. ISBN : 978-2-7440-6448-7 (cf. p. 33).

POINCARÉ, H. (1911). *Les sciences et les humanités*. Fayard (cf. p. 76).

QUETELET, A. (1842). *Etudes Sur l'homme*. Wouters, Raspoet (cf. p. 86).

QUÉTELET, A. (1829). *Recherches Statistiques Sur Le Royaume Des Pays-Bas*. Tarlier (cf. p. 87).

RASTIER, F. (2011). *La mesure et le grain : sémantique de corpus*. Paris. Honoré Champion. 272 p. ISBN : 978-2-7453-2230-2 (cf. p. 66).

ROMELE, A. et M. SEVERO (2015). *Traces numériques et territoires*. Presses des Mines. 270 p. ISBN : 978-2-35671-436-7. DOI : 10.4000/books.pressesmines.1984. (Visité le 29/04/2017) (cf. p. 78).

ROSTAING, H. (1996). *La bibliométrie et ses techniques*. Sciences de la société (cf. p. 125, 129).

SALTON, G. et M. J. MCGILL (1983). *Introduction to modern information retrieval*. New York : McGraw-Hill (cf. p. 8).

SALTON, G. (1971). *The smart retrieval system : experiments in automatic document processing*. Upper Saddle River, NJ, USA : Prentice-Hall (cf. p. 120).

SEGAL, J. (2003). *Le zéro et le un - histoire de la notion scientifique d'information au 20e siècle*. Éditions Syllepse (cf. p. 44).

SERRES, M. (2004). *Les limites de l'humain : textes des conférences et des débats*. Editions L'Age d'homme. 234 p. ISBN : 978-2-8251-1897-9 (cf. p. 36).

STIEGLER, B. (2004). *Mécréance et discrédit : la décadence des démocraties industrielles*. Galilée. ISBN : 978-2-7186-0660-6 (cf. p. 66).

— (2006). *Mécréance et discrédit : les sociétés incontrôlables d'individus désaffectés*. Galilée. ISBN : 978-2-7186-0706-1 (cf. p. 66).

SWINNEN, G. (2012). *Apprendre à programmer avec Python 3 : avec 60 pages d'exercices corrigés ! : objet, multithreading, bases de données, événements, programmation Web, programmation réseau, unicode, impression PDF, Python 2.7 & 3.2, tkinter, cherrypy*. Paris : Eyrolles. ISBN : 978-2-212-13434-6 (cf. p. 135).

THOMAS, B. (2009). *Gouverner sans gouverner. Une archéologie politique de la statistique*. Paris : PUF (cf. p. 65).

VARET, G. et M.-M. VARET (1995). *Maîtriser l'information à travers sa terminologie*. T. 9. Besançon : Presses Universitaires de la Faculté de Lettres (cf. p. 53).

VERHAEGEN, P. (2010). *Signe et communication*. De Boeck Supérieur. 272 p. ISBN : 978-2-8041-1743-6 (cf. p. 56, 79).

VIAL, S. (2013). *L'être et l'écran : Comment Le Numérique Change La Perception*. Presses Universitaires de France. ISBN : 978-2-13-062786-9 (cf. p. 85).

WATZLAWICK, P. et al. (1972). *Une logique de la communication*. Editions du Seuil. Paris, France (cf. p. 50).

WEINBERGER, D. (2007). *Everything is miscellaneous : the power of the new digital disorder*. Henry Holt et Company. ISBN : 978-1-4299-2795-6 (cf. p. 33).

WELGER-BARBOZA, C. (2012). *Les digital humanities aujourd'hui : centres, réseaux, pratiques et enjeux*. Sous la dir. de P. MOUNIER. OpenEdition Press. (Visité le 01/04/2015) (cf. p. 76, 77).

WIPO (2009). *Guide to technology databases*. World Intellectual Property Organization. Geneva, Switzerland : WIPO Publication. 84 p. ISBN : 978-92-805-2225-9. (Visité le 20/10/2015) (cf. p. 111).

WOLTON, D. (2012). *Indiscipliné : 35 ans de recherches*. Editions Odile Jacob. ISBN : 978-2-7381-8100-8 (cf. p. 50, 95).

ZIADÉ, T. (2009). *Programmation Python : Conception et Optimisation*. Eyrolles. ISBN : 978-2-212-85242-4 (cf. p. 135).

ZIPF, G. K. (1949). *Human behavior and the principle of least effort : an introduction to human ecology*. Cambridge MA : Addison-Wesley (cf. p. 2).

Collections (et extraits)

BALPE, J.-P., A. LELU et F. PAPY, éds. (1996). *Techniques avancées pour l'hypertexte*. Paris : Hermes Sciences Publications (cf. p. 53, 161).

BALTZ, C. (2007a). « Tous shannoniens ». In : *Racines oubliées des sciences de la communication*. Sous la dir. d'A.-M. LAULAN. T. 2. Hermès 48. Paris : CNRS Éd, p. 87–93. ISBN : 978-2-271-06530-8 (cf. p. 47).

BARATS, C., éd. (2013). *Manuel d'analyse du web en sciences humaines et sociales*. Armand Colin. ISBN : 978-2-200-28684-2 (cf. p. 31, 80).

BASSECOULARD, E. et M. ZITT (2005). « Patents and publications ». In : *Handbook of Quantitative Science And Technology Research*. Sous la dir. de H. F. MOED, W. GLANZEL et U. SCHMOCH. Springer, p. 665–694 (cf. p. 126).

BEAUDE, B. (2008). « Internet, un lieu du monde ». In : *L'Invention Du Monde*. Sous la dir. de J. LÉVY. Les Presses de Sciences Po. Paris, France : Lévy, Jacques, p. 111–131 (cf. p. 36).

BERNS, T. (2013). « Quand le réel nous gouverne ». In : *Gouverner par les standards et les indicateurs : de Hume aux ranking*. Sous la dir. de B. FRYDMAN et A. VAN WAEYEN-

BERGE. Bruxelles : Bruylant, p. 383–390. URL : http://hdl.handle.net/2013/ULB-DIPOT:oai:dipot.ulb.ac.be:2013/195719 (visité le 01/04/2015) (cf. p. 65, 66).

BESNIER, J.-M. et J. PERRIAULT, éds. (2013). *Interdisciplinaritè : entre disciplines et indiscipline*. Hermès. Paris : CNRS. 278 p. ISBN : 978-2-271-07966-4 (cf. p. 68).

BHATTI, Y., S. KHILJI et R. BASU (2013). « Frugal innovation ». In : *Globalization, change and learning in south asia*. Elsevier Ltd., p. 123–145. ISBN : 9780857094643 (ISBN). URL : http://www.scopus.com/inward/record.url?eid=2-s2.0-84902421591&partnerID=40&md5=3bd871df8550b26af0d01a305418e7ee (cf. p. 118).

CALLON, M., J.-P. COURTIAL et W. TURNER (1991). « La méthode leximappe : un outil pour l'analyse stratégique du développement scientifique et technique ». In : *Gestion de La Recherche. Nouveaux Problèmes, Nouveaux Outils*. Sous la dir. de D. VINCK. Bruxelles : De Boeck, p. 207–277 (cf. p. 147, 2).

CASILLI Antonio, A. (2014a). « Fracture numérique ». In : *Dictionnaire Des Inégalités*. Sous la dir. d'ALAIN BIHR, ROLAND PFEFFERKORN. Armand Colin, p. 172–173. URL : https://halshs.archives-ouvertes.fr/halshs-01055876 (cf. p. 77).

COURTIAL, J.-P. (2010). « Traduction et résonance morphique ». In : *Débordements*. Sous la dir. de M. AKRICH et al. Presses des Mines, p. 107–127. ISBN : 978-2-911256-38-7. DOI : 10.4000/books.pressesmines.732. URL : http://books.openedition.org/pressesmines/732 (visité le 15/05/2017) (cf. p. 159).

DAGENAIS, B. (2007). « Edgar Morin et la pensée complexe ». In : *Racines oubliées des sciences de la communication*. Sous la dir. d'A.-M. LAULAN. Hermès 48. Paris : CNRS Éd, p. 179–184. ISBN : 978-2-271-06530-8 (cf. p. 85).

DEVÈZE, J. (2004). « Abraham Moles, un exceptionnel passeur transdisciplinaire ». In : *Critique de la raison numérique*. Sous la dir. de V. PAUL et J. PERRIAULT. Hermès 39. Paris : CNRS Éd, p. 188–200. ISBN : 978-2-271-06245-1 (cf. p. 83).

DROMARD, D. et D. SERET (2006). « Internet ». In : France : Encyclopædia Universalis (cf. p. 27, 36).

DUFOULON, S., éd. (2012). *Internet ou la boite à usages*. L'Harmattan (cf. p. 32).

GALINON-MÉLÉNEC, B., éd. (2011). *L'homme trace. Perspectives anthropologiques des traces contemporaines*. Paris : CNRS éditions. 412 p. (cf. p. 58).

GARDIÈS, C., éd. (2012). *Approche de l'information-documentation. Concepts fondateurs*. Toulouse : Cépaduès (cf. p. 91).

GOLD, M. ., éd. (2012). *Debates in the digital humanities*. University of Minnesota Press. ISBN : 978-0-8166-7794-8 (cf. p. 77).

GRILICHES, Z. (1998). « Patent statistics as economic indicators : a survey ». In : *R&D And Productivity : the Econometric Evidence*. University of Chicago Press, p. 287–343 (cf. p. 100, 126).

GURRY, F. et C. FINK, éds. (2016). *Le système international des brevets*. Publication de l'OMPI. Genève, Suisse. 93 p. ISBN : 978-92-805-2761-2 (cf. p. 104, 105).

JACQUET, G., F. VENANT et V. BERNARD (2005). « Polysémie lexicale ». In : Hermès science publications, p. 99–132 (cf. p. 47).

JEANNERET, Y. (2011). « Complexité de la notion de trace. De la traque au tracé ». In : *L'Homme-Trace : Perspective Anthropologiques Des Traces Contemporaines*. Sous la dir. de B. GALINON-MÉLÉNEC. CNRS Editions. Paris, France : CNRS éditions, p. 59–86 (cf. p. 55, 56).

JUANALS, B. et J.-M. NOYER (2007). « Dell H. Hymes : vers une pragmatique et une anthropologie communicationnelle. » In : *Racines oubliées des sciences de la communication*. Sous la dir. d'A.-M. LAULAN. T. 2. Hermès 48. Paris : CNRS Éd, p. 117–123. ISBN : 978-2-271-06530-8 (cf. p. 85, 86, 88).

KRASTEVA, A., éd. (2013). *E-citoyennetés*. L'Harmattan (cf. p. 32).

LATOUR, B. (1993). « "les 'vues' de l'esprit". Une introduction à l'anthropologie des sciences et des techniques ». In : *Sciences de l'information et de la communication*. Sous la dir. de D. BOUGNOUX. Larousse. Larousse, p. 572–596 (cf. p. 159).

LELEU-MERVIEL, S. (2003). « Les désarrois des maîtres du sens à l'ère du numérique ». In : *Hypertextes, Hypermédias : Créer Du Sens À L'Ère Numérique*. Sous la dir. de J. BALPE et I. SALEH. Londres/Paris : Hermès/Lavoisier, p. 17–34 (cf. p. 53).

LELEU-MERVIEL, S. et P. USEILLE (2008). « Quelques révisions du concept d'information ». In : *Problématiques Émergentes Dans Les Sciences de L'Information*. Sous la dir. d'HERMÈS. Traité des sciences et techniques de l'information. Lavoisier, p. 25–56. URL : https://hal.archives-ouvertes.fr/hal-00695777 (cf. p. 44, 59, 71, 73).

MIEGE, B. (2004). « Les tics : un champ marqúe par la complexite et un entrelacs d'enjeux ». In : *L'Information Communication, Objet de Connaissance*. Bruxelles : Édition de Boeck Université, p. 113–123 (cf. p. 51).

NELSON, T. (1967). « Getting it out of our system ». In : *Information Retrieval : A Critical View*. Sous la dir. de G. SCHECHTER. Washington, D.C. : Thompson Book Company, p. 191–210 (cf. p. 30).

NOYER, J.-M., éd. (1995). *Les sciences de l'information -bibliométrie, scientométrie, infométrie* -. Rennes : Presses Universitaires de Rennes. ISBN : 2-86847-150-1 (cf. p. 80).

OSISKI, S. et D. WEISS (2005). « Carrot2 : design of a flexible and efficient web information retrieval framework ». In : *Advances In Web Intelligence*. Springer, p. 439–444 (cf. p. 150).

PERRIAULT, J. (2007). « Le rôle de l'informatique dans la pensée en information et en communication ». In : *Racines oubliées des sciences de la communication*. Sous la dir. d'A.-M. LAULAN. T. 2. Hermès 48. Paris : CNRS Éd, p. 127–129. ISBN : 978-2-271-06530-8 (cf. p. 83).

PROULX, S. (2007). « Naissance des sciences de la communication dans le contexte militaire des années 1940 aux États-Unis ». In : *Racines oubliées des sciences de la communication*. Sous la dir. d'A.-M. LAULAN. T. 2. Hermès 48. Paris : CNRS Éd, p. 61–67. ISBN : 978-2-271-06530-8 (cf. p. 47).

QUONIAM, L. (2013). « Le brevet : objet de recherche en sciences de l'information et de la communication ». In : Traité des sciences et techniques de l'information. Hermès science publications, p. 95–114. ISBN : 978-2-7462-4535-8. URL : https://books.google.fr/books?id=m-vYngEACAAJ (cf. p. 101).

RÉGIMBEAU, G. (2012). « Médiations ». In : *Approche de L'Information - Documentation. concepts fondateurs*. Sous la dir. de C. GARDIÈS. Toulouse : Cépaduès (cf. p. 92).

ROUET, G. et F. SOULAGES, éds. (2013). *Frontières géoculturelles et géopolitiques*. L'Harmattan (cf. p. 32).

ROUQUETTE, S. (2010). « Internet : un espace médiatique fragmenté ». In : *Entre communautés et mobilité : une approche interdisciplinaire des médias*. Sous la dir. de S. AGOSTINELLI, D. AUGEY et F. LAURIE. Presses des Mines, p. 135–151 (cf. p. 28).

SALEH, I. et al., éds. (2013). *H2ptm13*. Paris : Hermès/Lavoisier (cf. p. 33).

SOULAGES, F., éd. (2013). *Géoartistique et géopolitique*. Paris : l'Harmattan (cf. p. 32).

SZONIECKY, S. (2013). « Les frontières des écosystèmes d'informations numériques ». In : *Géoartistique & Géopolitique : Frontières*. Sous la dir. de F. SOULAGES. Arts, esthétique, vie culturelle. Paris : l'Harmattan, p. 65–76. URL : https://hal-univ-paris8.archives-ouvertes.fr/hal-01098423 (cf. p. 32).

TURNER, W. A. et al. (1988). « Packaging Information for Peer Review : New Co-Word Analysis Techniques ». In : *Handbook of Quantitative Studies of Science and Technology*. Sous la dir. d'A. F. J. van RAAN. Elsevier (cf. p. 2).

VITALIS, A. (2007). « Actualité de Jacques Ellul : la communication dans le contexte d'une société technicienne ». In : *Racines oubliées des sciences de la communication*. Sous la dir. d'A.-M. LAULAN. T. 2. Hermès 48. Paris : CNRS Éd, p. 163–170. ISBN : 978-2-271-06530-8 (cf. p. 47, 83).

WOLTON, D. (2007). « Conclusion. De l'information aux sciences de la communication ». In : *Racines oubliées des sciences de la communication*. Sous la dir. d'A.-M. LAULAN. T. 2. Hermès 48. Paris : CNRS Éd, p. 189–202. ISBN : 978-2-271-06530-8 (cf. p. 84).

— (2014). « Incommunication et altérité. Entretien ». In : *L'autre n'est pas une donnée : altérités, corps et artefacts*. Sous la dir. de F. RENUCCI, B. LE BLANC et S. LEPASTIER. Hermès 68. Paris : CNRS, p. 212–217. ISBN : 978-2-271-08074-5 (cf. p. 88).

Articles de revues

ABBAS, A., L. ZHANG et S. U. KHAN (2014). « A literature review on the state-of-the-art in patent analysis ». In : *World Patent Information* 37, p. 3–13. ISSN : 0172-2190. DOI : http://dx.doi.org/10.1016/j.wpi.2013.12.006. URL : http://www.sciencedirect.com/science/article/pii/S0172219013001634 (cf. p. 119, 127, 136).

ADAIR, W. C. (1955). « Citation indexes for scientific literature? » In : *American Documentation (Pre-1986)* 6.1, p. 31 (cf. p. 125, 2).

ADAMIC, L. A. et B. A. HUBERMAN (1999). « Internet : growth dynamics of the world wide web ». In : *Nature* 401. URL : http://www.hpl.hp.com/research/idl/papers/webgrowth/nature9sept99.pdf (cf. p. 27).

ADNOT, P. (2006). « La valorisation de la recherche dans les universités ». In : Les rapports d'information et de contrôle du gouvernement (341 (2005-2006)). URL : https://www.senat.fr/rap/r05-341/r05-34115.html#toc142 (cf. p. 110).

ALCACER, J. et M. GITTELMAN (2006). « Patent citations as a measure of knowledge flows : the influence of examiner citations ». In : *The Review of Economics And Statistics* 88.4, p. 774–779 (cf. p. 109, 145).

AVRAMESCU, A. (1979). « Actuality and Obsolescence of Scientific Literature ». In : *Journal of the Association for Information Science and Technology* 30.5, p. 296–303 (cf. p. 2).

BACCHIOCCHI, E. et F. MONTOBBIO (2010). « International knowledge diffusion and home-bias effect : do USPTO and EPO patent citations tell the same story? » In : *Scandinavian journal of economics*, p. 441–470. ISSN : 03470520, 14679442. DOI : 10.1111/j.1467-9442.2010.01614.x. URL : http://doi.wiley.com/10.1111/j.1467-9442.2010.01614.x (visité le 02/03/2017) (cf. p. 145).

BADILLO, P.-Y. et N. PÉLISSIER (2015). « Usages et usagers de l'information numérique ». In : *Revue Française Des Sciences de L'Information Et de La Communication* 6. URL : http://rfsic.revues.org/1448 (visité le 01/04/2015) (cf. p. 27).

BADIR, S. (2005). « La notion de texte chez hjelmslev ». In : Texto ! Inédits. URL : http://www.revue-texto.net/Dialogues/Debat_Hjelmslev/Badir_Notion.html (cf. p. 56).

BAHL, L. R., F. JELINEK et R. L. MERCER (1983). « A Maximum Likelihood Approach to Continuous Speech Recognition ». In : *IEEE transactions on pattern analysis and machine intelligence* 2, p. 179–190 (cf. p. 3).

BAILÓN-MORENO, R., E. JURADO-ALAMEDA et al. (2005a). « Bibliometric Laws : Empirical Flaws of Fit ». In : *Scientometrics* 63.2, p. 209–229. ISSN : 0138-9130, 1588-2861. DOI : 10.1007/s11192-005-0211-5. URL : http://link.springer.com/10.1007/s11192-005-0211-5 (visité le 02/07/2017) (cf. p. 6).

BAILÓN-MORENO, R., E. JURADO-ALAMEDA et al. (2005b). « The Unified Scientometric Model. Fractality and Transfractality ». In : *Scientometrics* 63.2, p. 231–257 (cf. p. 159, 6).

BALTZ, C. (2007b). « Tous shannoniens ? » In : *Hermès, La Revue* 48.2, p. 87–93. ISSN : 9782271065308. URL : http://www.cairn.info/revue-hermes-la-revue-2007-2-page-87.htm (cf. p. 48).

BARBUT, M. (1988). « Des bons et des moins bons usages des distributions parétiennes en analyse des données ». In : *Histoire & Mesure* 3.1, p. 111–128. ISSN : 0982-1783. DOI : 10.3406/hism.1988.1296. URL : http://www.persee.fr/web/revues/home/prescript/article/hism_0982-1783_1988_num_3_1_1296 (visité le 23/05/2017) (cf. p. 2, 7).

— (1989). « Note sur l'ajustement des distributions de Zipf-Mandelbrot en statistique textuelle ». In : *Histoire & Mesure* 4.1, p. 107–119. ISSN : 0982-1783. DOI : 10.3406/hism.1989.879. URL : http://www.persee.fr/web/revues/home/prescript/article/hism_0982-1783_1989_num_4_1_879 (visité le 20/05/2017) (cf. p. 2, 4, 5).

BARONI, M. et A. LENCI (2008). « Concepts and Properties in Word Spaces ». In : *Italian Journal of Linguistics* 20.1, p. 55–88. URL : http://www.wordspace.collocations.de/lib/exe/fetch.php/course:baroni_lenci_2008.pdf (cf. p. 160, 8).

BARRÉ, R., F. LAVILLE et al. (1995). « L'observatoire des sciences et des techniques : activités - définition - méthodologie ». In : *Solaris* 2. URL : http://biblio-fr.info.unicaen.fr/bnum/jelec/Solaris/d02/2barre.html (cf. p. 126).

BARRÉ, R. et F. LAVILLE (1994). « La bibliométrie des brevets : une mesure de l'activité technologique ». In : *Economie Et Statistique* 275.1, p. 71–81 (cf. p. 100, 126).

BASKERVILLE, R. L. et M. D. MYERS (2002). « Information systems as a reference discipline ». In : *Mis Quarterly* 26.1, p. 1. ISSN : 02767783. DOI : 10.2307/4132338. JSTOR : 10.2307/4132338?origin=crossref (cf. p. 44, 45, 47).

BASTIN, G. et J. FRANCONY (2016). « L'inscription, Le Masque et La Donnée. Datafication Du Web et Conflits d'interprétation Autour Des Données Dans Un Laboratoire Invisible Des Sciences Sociales ». In : *Revue d'Anthropologie des connaissances* 10.4 (cf. p. 34).

BATES, M. J. (1999). « The invisible substrate of information science ». In : *Journal of the American Society For Information Science* 50.12, p. 1043–1050 (cf. p. 71).

BELLEFLAMME, P. (2006). « Patents and incentives to innovate ». In : *Ethical Perspectives* 13.2, p. 267–288. ISSN : 1370-0049. DOI : 10.2143/EP.13.2.2016634. URL : http://poj.peeters-leuven.be/content.php?url=article&id=2016634 (visité le 18/02/2017) (cf. p. 111).

BERNERS-LEE, T., J. HENDLER et O. LASSILA (2001). « The semantic web ». In : *Scientific American* 5. URL : http://www.sciam.com/article.cfm%20?%20articleID=00048144-10D2-1C70-84A9809EC588EF21 (cf. p. 30).

BERNERS-LEE, T. (1989). « Information management : a proposal ». In : *W3C - World Wide Web Consortium*. URL : http://www.w3.org/History/1989/proposal.html (visité le 30/01/2015) (cf. p. 28, 29, 120).

— (1995). « Hypertext and our collective destiny ». In : URL : http://www.w3.org/Talks/9510%5C%5C_Bush/Talk.html (visité le 01/04/2015) (cf. p. 28).

BESTGEN, Y. (2014). « Construction Automatique d'un Lexique de n-Grammes Pour La Fouille d'opinion ». In : *Document numérique* 17.1, p. 103–123. ISSN : 12795127. DOI : 10.3166/dn.17.1.103-123. URL : http://dn.revuesonline.com/article.jsp?articleId=19421 (visité le 04/06/2017) (cf. p. 161).

BISKRI, I. et S. DELISLE (2001). « Les N-Grams de Caractères Pour l'aide à l'extraction de Connaissances Dans Des Bases de Données Textuelles Multilingues ». In : *Proceedings of TALN-2001*, p. 93–102 (cf. p. 161).

BJÖRNEBORN, L. et P. INGWERSEN (2001). « Perspectives of webometrics ». In : *Scientometrics* 50.1, p. 65–82 (cf. p. 80).

BONDÍ, A. (2008). « Hjelmslev et la « fonction sémiotique » : du modèle structural au modèle cognitif ». In : *Histoire Épistémologie Langage* 30.2, p. 199–212. ISSN : 0750-8069. DOI : 10.3406/hel.2008.3173. URL : http://www.persee.fr/doc/hel_0750-8069_2008_num_30_2_3173 (cf. p. 56, 57, 69).

BONINO, D., A. CIARAMELLA et F. CORNO (2010). « Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics ». In : *World Patent Information* 32.1, p. 30–38. URL : https://ideas.repec.org/a/eee/worpat/v32y2010i1p30-38.html (cf. p. 102).

BOOTH, A. D. (1967). « A "Law" of Occurrences for Words of Low Frequency ». In : *Information and control* 10.4, p. 386–393 (cf. p. 158, 4).

BÖRNER, K. (2011). « Plug-and-play macroscopes ». In : *Communications of the acm* 54.3, p. 60. ISSN : 00010782. DOI : 10.1145/1897852.1897871. URL : http://portal.acm.org/citation.cfm?doid=1897852.1897871 (visité le 29/04/2017) (cf. p. 134, 180).

BOSTOCK, M., V. OGIEVETSKY et J. HEER (2011). « D3 data-driven documents ». In : *IEEE Transactions on Visualization And Computer Graphics* 17.12, p. 2301–2309. ISSN : 1077-2626. DOI : 10.1109/TVCG.2011.185. URL : http://dx.doi.org/10.1109/TVCG.2011.185 (cf. p. 140).

BOUGRINE, H. (2001). « Le dépôt et la délivrance du brevet ». In : *Cahiers de La Documentation* 1, p. 5–10. URL : http://www.abd-bvd.be/wp-content/uploads/2001-1_Bougrine.pdf (cf. p. 103).

BOURDONCLE, F. (1999). « Panorama et perspectives des outils de recherche d'information textuelle sur internet ». In : *Idt. Information, Documentation, Transfert Des Connaissances* 16, p. 219–226 (cf. p. 120).

BOYCE, B. et M. LOCKARD (1975). « Automatic and Manual Indexing Performance in a Small File of Medical Literature. » In : *Bulletin of the Medical Library Association* 63.4, p. 378 (cf. p. 158).

BRADFORD, S. C. (1934). « Sources of information on specific subjects ». In : *British Journal of Engineering* 137, p. 85–86 (cf. p. 2).

BREITZMAN, A. F. et M. E. MOGEE (2002). « The many applications of patent analysis ». In : *Journal of Information Science* 28.3, p. 187–205. DOI : 10.1177/016555150202800302. URL : http://jis.sagepub.com/content/28/3/187.abstract (cf. p. 118).

BRENNER, J. (2014). « Information : a personal synthesis ». In : *Information* 5.1, p. 134–170. ISSN : 2078-2489. DOI : 10.3390/info5010134. URL : http://www.mdpi.com/2078-2489/5/1/134/ (visité le 01/02/2017) (cf. p. 71).

BRENNER, J. E. (2012). « Mark burgins theory of information ». In : *Information* 3.4, p. 224–228. ISSN : 2078-2489. DOI : 10.3390/info3020224. URL : http://www.mdpi.com/2078-2489/3/2/224/ (visité le 01/02/2017) (cf. p. 71).

BRIN, S. et L. PAGE (1998). « The anatomy of a large-scale hypertextual web search engine ». In : *Www7 / Computer Networks* 30 (1-7), p. 107–117. URL : http://wx.bsjh.tcc.edu.tw/~t2003013/wiki/images/8/8f/Anatomy.pdf (cf. p. 142, 143).

BROOKES, B. C. (1968). « The derivation and application of the bradford-zipf distribution ». In : *Journal of Documentation* 7, p. 299–335 (cf. p. 2).

BROOKES, B. C. (1984). « Ranking Techniques and the Empirical Log Law ». In : *Information processing & management* 20 (1-2), p. 37–46 (cf. p. 4, 5).

BRÜGMANN, S. et al. (2015). « Towards content-oriented patent document processing : intelligent patent analysis and summarization ». In : *World patent information* 40, p. 30–42. ISSN : 01722190. DOI : 10.1016/j.wpi.2014.10.003. URL : http://linkinghub.elsevier.com/retrieve/pii/S0172219014001410 (visité le 21/02/2017) (cf. p. 128).

BUSH, V. (1945). « As we may think ». In : *The Atlantic Monthly* 1.176, p. 101–108. URL : http://www.theatlantic.com/doc/194507/bush (cf. p. 28).

CALLON, M., J. P. COURTIAL et al. (1983). « From translation to problematic networks : an introduction to co-word analysis ». In : *Social Science Information* 22, p. 191–235 (cf. p. 2).

CAPART, G. (2006). « Should universities file patent applications ? » In : *Ethical Perspectives* 13.2, p. 221–230. ISSN : 1370-0049. DOI : 10.2143/EP.13.2.2016631. URL : http://poj.peeters-leuven.be/content.php?url=article&id=2016631 (visité le 18/02/2017) (cf. p. 111).

CAPURRO, R. et B. HJØRLAND (2003). « The concept of information ». In : *Annual Review of Information Science And Technology* 37.1, p. 343–411 (cf. p. 44, 56).

CARMES, M. et J.-M. NOYER (2014). « L'irrésistible montée de l'algorithmique ». In : *Les Cahiers Du Numérique* 10.4. Sous la dir. d'E. SOULIER, p. 63–109. ISSN : 1622-1494, 2111-434X (cf. p. 67, 153).

CASILLI Antonio, A. (2014b). « Les données numériques : un enjeu d'éducation et de citoyenneté ». In : URL : https://hal.archives-ouvertes.fr/hal-01068525 (cf. p. 76).

CHEN, Y.-L. et Y.-T. CHIU (2013). « Cross-language patent matching via an international patent classification-based concept bridge ». In : *Journal of information science* 39.6, p. 737–753. ISSN : 0165-5515, 1741-6485. DOI : 10.1177/0165551513494641. URL : http://jis.sagepub.com/cgi/doi/10.1177/0165551513494641 (visité le 21/10/2015) (cf. p. 108).

CHOI, J. et Y.-S. HWANG (2014). « Patent keyword network analysis for improving technology development efficiency ». In : *Technological forecasting and social change* 83, p. 170–182. ISSN : 00401625. DOI : 10.1016/j.techfore.2013.07.004. URL : http://linkinghub.elsevier.com/retrieve/pii/S004016251300156X (visité le 21/02/2017) (cf. p. 128).

CITTON, Y. (2015). « Humanités numériques : une médiapolitique des savoirs encore à inventer ». In : *Multitudes* 59.2, p. 169. ISSN : 0292-0107, 1777-5841. DOI : 10.3917/mult.059.0169. URL : https://hal.archives-ouvertes.fr/hal-01373171/file/M59-CITTON-HumanitesNumMediapolitique-Oct2015.pdf (visité le 24/11/2016) (cf. p. 76).

CORBEL, P. et S. MBONGUI-KIALO (2012). « L'utilisation de l'information brevet au sein des bureaux d'études : du potentiel au réel ». In : *Revue Internationale D'Intelligence Économique* 4.2, p. 139–152. ISSN : 2101647X. DOI : 10.3166/r2ie.4.139-152. URL : http://r2ie.revuesonline.com/article.jsp?articleId=18610 (visité le 04/02/2017) (cf. p. 118).

CUKIER, K. et V. MAYER-SCHÖNBERGER (2013). « Mise en données du monde, le déluge numérique ». In : URL : http://www.monde-diplomatique.fr/2013/07/CUKIER/49318 (visité le 01/04/2015) (cf. p. 89).

DAVALLON, J., N. NOEL-CADET et D. BROCHU (2003). « L'usage dans le texte : les traces d'usage du site gallica ». In : *Lire, Écrire, Réécrire, Objets, Signes Et Pratiques Des Médias Informatisé*, p. 47–89 (cf. p. 55).

DAVIS, L. J. (1997). « Constructing Normalcy ». In : *The disability studies reader* 3 (cf. p. 86).

DE AGUILERA, M. (2006). « Les modèles de la communication et leur objet d'étude ». In : *Communication Et Organisation* 30, p. 110–125. ISSN : 1168-5549, 1775-3546. DOI : 10.4000/communicationorganisation.3457. URL : http://communicationorganisation.revues.org/3457 (visité le 25/11/2016) (cf. p. 48).

DEANE, P. D. (1988). « Polysemy and cognition ». In : *Lingua* 75.4, p. 325–361. ISSN : 00243841. DOI : 10.1016/0024-3841(88)90009-5. URL : http://linkinghub.elsevier.com/retrieve/pii/0024384188900095 (visité le 09/04/2017) (cf. p. 47).

DEERWESTER, S. et al. (1990). « Indexing by Latent Semantic Analysis ». In : *Journal of the American society for information science* 41.6, p. 391 (cf. p. 161).

DESCHAMPS, C. et N. MOINET (2011). « L'émergence d'internet dans les outils d'intelligence économique ». In : *Le Temps Des Médias* 1.16, p. 288. DOI : 10.3917/tdm.016.0147 (cf. p. 154).

DESROSIÈRES, A. (1988). « Masses, individus, moyennes : la statistique sociale au xixe siècle ». In : *Hermès* 2, p. 41–66 (cf. p. 65).

DEVISME, L. (2007). « Latour : Mettre Les Sciences Au Travail. » In : *EspacesTemps.net*. Livres. URL : http://www.espacestemps.net/articles/latour-mettre-les-sciences-au-travail/ (cf. p. 159).

DIBIAGGIO, L. et L. NESTA (2005). « Patents Statistics, Knowledge Specialisation and the Organisation of Competencies ». In : *Revue d'économie industrielle* 110.1, p. 103–126 (cf. p. 163).

DIRNBERGER, D. (2016). « The use of mindmapping software for patent search and management ». In : *World patent information* 47, p. 12–20. ISSN : 01722190. DOI : 10.1016/j.wpi.2016.08.004. URL : http://linkinghub.elsevier.com/retrieve/pii/S0172219016300941 (visité le 23/07/2017) (cf. p. 152).

DOU, H., D. MOHELLEBI et J. KISTER (2012). « L'importance du traitement bibliométrique des brevets pour développer l'activité industrielle. Exemple des bitumes en algérie ». In : *Rist (Revue Scientifique Et Technique)* 19.1 (cf. p. 124).

DOU, H., B. HAUDEVILLE et D. WOLFF (2015). « L'analyse de l'information brevet comme catalyseur de l'innovation et du développement des entreprises ». In : *Vie & Sciences de L'Entreprise* 199.1, p. 72. ISSN : 2262-5321, 2262-5372. DOI : 10.3917/vse.199.0072. URL : http://www.cairn.info/revue-vie-et-sciences-de-l-entreprise-2015-1-page-72.htm (visité le 04/02/2017) (cf. p. 100, 117).

DOU, H. et V. LEVEILLÉ (2015). « Utilisation de l'information brevet pour faciliter la créativité et le développement technologique. Application au développement durable ». In : *Revue Internationale D'Intelligence Économique* 7.1, p. 25–45. ISSN : 2101647X. DOI : 10.3166/r2ie.7.25-45. URL : http://r2ie.revuesonline.com/article.jsp?articleId=21045 (visité le 04/02/2017) (cf. p. 118).

DOUCETTE, D. et al. (2007). « Toward a new science of information ». In : *Data Science Journal* 6, p. 198–205 (cf. p. 44).

DOUKAS, C. et al. (2011). « Digital cities of the future : extending home assistive technologies for the elderly and the disabled ». In : *Telematics And Informatics* 28.3, p. 176–190 (cf. p. 27).

DRUCKER, J. (2010). « Graphesis ». In : *paj : The Journal of the Initiative for Digital Humanities, Media, and Culture* 2.1 (cf. p. 73).

— (2011). « Humanities Approaches to Graphical Display ». In : *Digital Humanities Quarterly* 5.1, p. 1–21 (cf. p. 73).

DUBOIS, J. (1973). « Code, texte, métatexte ». In : *Littérature* 12.4, p. 3–11. ISSN : 0047-4800. DOI : 10.3406/litt.1973.1985. URL : http://www.persee.fr/doc/litt_0047-4800_1973_num_12_4_1985 (cf. p. 53).

DURAND-BARTHEZ, M. (2013). « Former à l'information brevets dans l'enseignement supérieur ». In : *Revue Internationale D'Intelligence Économique* 5.1, p. 25–38 (cf. p. 101, 113).

ECMA, E. (1999). « 262 : Ecmascript Language Specification ». In : *ECMA (European Association for Standardizing Information and Communication Systems), pub-ECMA : adr,* (cf. p. 30).

EKNOYAN, G. (2007). « Adolphe Quetelet (1796 1874) the average man and indices of obesity ». In : *Nephrology dialysis transplantation* 23.1, p. 47–51. ISSN : 0931-0509, 1460-2385. DOI : 10.1093/ndt/gfm517. URL : https://academic.oup.com/ndt/article-lookup/doi/10.1093/ndt/gfm517 (visité le 28/04/2017) (cf. p. 86).

ÉRDI, P. et al. (2013). « Prediction of emerging technologies based on analysis of the us patent citation network ». In : *Scientometrics* 95.1, p. 225–242. ISSN : 0138-9130, 1588-2861. DOI : 10.1007/s11192-012-0796-4. URL : http://link.springer.com/10.1007/s11192-012-0796-4 (visité le 21/02/2017) (cf. p. 127).

FAIRTHORNE, R. (1956). « The patterns of retrieval ». In : *American Documentation* 7.2, p. 65–70 (cf. p. 120).

FEDOROWICZ, J. (1982). « The Theoretical Foundation of Zipf's Law and Its Application to the Bibliographic Database Environment ». In : *Journal of the Association for Information Science and Technology* 33.5, p. 285–293 (cf. p. 4).

FURNER, J. (2015). « Information Science Is Neither ». In : *Library trends* 63.3, p. 362–377. ISSN : 1559-0682. DOI : 10.1353/lib.2015.0009. URL : https://muse.jhu.edu/content/crossref/journals/library_trends/v063/63.3.furner.html (visité le 13/05/2017) (cf. p. 44).

GELLNER, E. (1984). « Le statut scientifique des sciences sociales ». In : *Revue Internationale Des Sciences Sociales* XXXVI.4, p. 599–619. URL : http://unesdoc.unesco.org/images/0006/000636/063623fo.pdf (visité le 01/04/2015) (cf. p. 63).

GIRARD, J.-Y. (1971). « Une extension de l'interpretation de gödel a l'analyse, et son application a l'elimination des coupures dans l'analyse et la theorie des types ». In : *Studies In Logic And the Foundations of Mathematics* 63, p. 63–92 (cf. p. 52).

GONZALES-AGUILAR, A. et M. RAMÍREZ-POSADA (2012). « Carrot2 : búsqueda y visualización de la información ». In : *El Profesional de La Información* 21.1, p. 105–112.

URL : http://project.carrot2.org/publications/gonzales-ramirez-2012.pdf (cf. p. 150).

GUICE, J. (1998). « Looking backward and forward at the internet ». In : *The Information Society* 14.3 (cf. p. 27).

HALL, B. H., A. JAFFE et M. TRAJTENBERG (2005). « Market value and patent citations ». In : *Rand Journal of Economics*, p. 16–38 (cf. p. 118, 129, 145).

HARHOFF, D., F. M. SCHERER et al. (2003). « Citations, family size, opposition and the value of patent rights ». In : *Research Policy* 32.1596, p. 1343–1363. URL : http://www.globelicsacademy.org/pdf/BronwynHall_5.pdf (cf. p. 145).

HARHOFF, D. et S. WAGNER (2009). « The duration of patent examination at the European Patent Office ». In : *Management Science* 55.12, p. 1969–1984 (cf. p. 109).

HIDALGO-NUCHERA, A., S. IGLESIAS-PRADAS et Á. HERNÁNDEZ-GARCÍA (2009). « Utilización de las bases de datos de patentes como instrumento de vigilancia tecnológica ». In : *El Profesional de La Informaciñón* 18.5, p. 511–519 (cf. p. 100).

HIRATA, D. et al. (2015). « O uso de informações patentárias para a valozização de residuos industriais : o caso do lodo de tratemento de esgoto doméstico ». In : *Revista de Ciências Da Administração* 1.1, p. 55–71 (cf. p. 112).

HOFKIRCHNER, W. (2009). « How to achieve a unified theory of information ». In : *Triplec : Communication, Capitalism & Critique. Open Access Journal For A Global Sustainable Information Society* 7.2, p. 357–368 (cf. p. 69).

— (2013). « Emergent information. when a difference makes a difference... » In : *Triplec : Communication, Capitalism & Critique. Open Access Journal For A Global Sustainable Information Society* 11.1, p. 6–12 (cf. p. 91).

HU, A. G. et A. B. JAFFE (2003). « Patent citations and international knowledge flow : the cases of Korea and Taiwan ». In : *International journal of industrial organization* 21.6, p. 849–880. ISSN : 01677187. DOI : 10.1016/S0167-7187(03)00035-3. URL : http://linkinghub.elsevier.com/retrieve/pii/S0167718703000353 (visité le 03/05/2017) (cf. p. 145).

HUBERT, J. J. (1981). « General Bibliometric Models ». In : (cf. p. 2).

JEANNERET, Y. (2001). « Informatic literacy : manifestations, captations et déceptions dans le texte informatisé ». In : *Spirales* 28, p. 11–32 (cf. p. 52).

JEANNERET, Y. (2002). « Communication, transmission, un couple orageux ». In : *Sciences Humaines* 36, p. 24–27 (cf. p. 49).

— (2007). « Usages de l'usage, figures de la médiatisation ». In : *Communication & Langages* 151.1, p. 3–19 (cf. p. 92).

JELINEK, F. (1976). « Continuous Speech Recognition by Statistical Methods ». In : *Proceedings of the IEEE* 64.4, p. 532–556 (cf. p. 3).

JÉRÔME, S. (2006). « Cash for knowledge ? Ethicals implications of patenting academic research. Quatrième forum éthique de la fondation universitaire ». In : *Cahiers de La Documentation* 1, p. 35–41 (cf. p. 110, 111).

KALLAS, P. (2006). « Open patent services ». In : *World Patent Information* 28.4, p. 296–304. URL : http://EconPapers.repec.org/RePEc:eee:worpat:v:28:y:2006:i:4:p:296-304 (cf. p. 112, 135).

KESSLER, M. M. (1963). « Bibliographic couplic beetween scientific papers ». In : *American Documentation* 14.1 (cf. p. 2).

KYHENG, R. (2005). « De la sémantique des textes au web sémantique ». In : *Texto ! Textes Et Cultures* X.2. URL : http://www.revue-TEXTO.net/Redaction/Dossier_EE/Kyheng/Kyheng_Semantique.html (cf. p. 56).

LAFON, P. (1981). « Analyse lexicométrique et recherche des cooccurrences ». In : *Mots* 3.1, p. 95–148. ISSN : 0243-6450. DOI : 10.3406/mots.1981.1041. URL : http://www.persee.fr/web/revues/home/prescript/article/mots_0243-6450_1981_num_3_1_1041 (visité le 11/05/2017) (cf. p. 149).

LANDAUER, T. K. et S. T. DUMAIS (1997). « A Solution to Plato's Problem : The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. » In : *Psychological review* 104.2, p. 211 (cf. p. 160).

LARDY, J.-P. (2001). « La recherche d'information sur le web ». In : *Résonnances* 4 (6–7). URL : http://archivesic.ccsd.cnrs.fr/sic%20_00000052 (cf. p. 120).

LEINER, B. M. et al. (1997). « The past and future history of the internet ». In : *Commun. Acm* 40.2, p. 102–108. ISSN : 0001-0782. DOI : http://doi.acm.org/10.1145/253671.253741 (cf. p. 27).

LELEU-MERVIEL, S. (2010a). « De l'infra-conceptuel à des données à horizon de pertinence focalisée ». In : *Questions de Communication* 18, p. 171–184 (cf. p. 51).

— (2013). « Traces, information et construits de sens. Déploiement de la trace visuelle de la rétention indicielle à l'écriture ». In : *Intellectica* 59, p. 65–88 (cf. p. 59).

LERTNATTEE, V. et T. THEERAMUNKONG (2006). « Class normalization in centroid-based text categorization ». In : *Inf. Sci.* 176.12, p. 1712–1738. ISSN : 0020-0255. DOI : 10.1016/j.ins.2005.05.010. URL : http://dx.doi.org/10.1016/j.ins.2005.05.010 (cf. p. 8).

LEVY, P. (2014). « The philosophical concept of algorithmic intelligence ». In : *Spanda* 2 (Collective Intelligence) (cf. p. 61).

LÉVY, P. (1991). « L'hypertexte, instrument et métaphore de la communication ». In : *Réseaux. La communication : une interrogation philosophique* 9.46, p. 59–68. ISSN : 0751-7971. DOI : 10.3406/reso.1991.1831. URL : http://www.persee.fr/doc/reso_0751-7971_1991_num_9_46_1831 (cf. p. 37).

— (2015). « Le medium algorithmique ». In : *Sociétés* 129.3, p. 79. ISSN : 0765-3697, 1782-155X. DOI : 10.3917/soc.129.0079. URL : http://www.cairn.info/revue-societes-2015-3-page-79.htm (visité le 26/11/2016) (cf. p. 59).

LEYDESDORFF, L. (1992). « A Validation Study of "LEXIMAPPE" ». In : *Scientometrics* 25.2, p. 295–312. ISSN : 1588-2861. DOI : 10.1007/BF02028087. URL : http://dx.doi.org/10.1007/BF02028087 (cf. p. 160).

LEYDESDORFF, L. et S. BENSMAN (2005). « Classification, powerlaws, and the logarithmic transformation ». In : URL : http://www.leydesdorff.net/log05/log05.pdf (cf. p. 2).

LEYDESDORFF, L. (1997). « Why Words and Co-Words Cannot Map the Development of the Sciences ». In : *JASIS* 48, p. 418–427 (cf. p. 160).

— (2008). « Patent Classifications as Indicators of Intellectual Organization ». In : *Journal of the Association for Information Science and Technology* 59.10, p. 1582–1597 (cf. p. 163).

— (2015). « Can technology life-cycles be indicated by diversity in patent classifications? The crucial role of variety ». In : *Scientometrics* 105.3, p. 1441–1451. ISSN : 0138-9130, 1588-2861. DOI : 10.1007/s11192-015-1639-x. URL : http://link.springer.com/10.1007/s11192-015-1639-x (visité le 02/12/2015) (cf. p. 126).

LEYDESDORFF, L., F. ALKEMADE et al. (2015). « Patents as instruments for exploring innovation dynamics : geographic and technological perspectives on photovoltaic cells ». In : *Scientometrics* 102.1, p. 629–651. ISSN : 0138-9130, 1588-2861. DOI : 10.1007/s11192-014-1447-8. URL : http://link.springer.com/10.1007/s11192-014-1447-8 (visité le 04/02/2017) (cf. p. 126).

LEYDESDORFF, L. et L. HELLSTEN (2006). « Measuring the Meaning of Words in Contexts : An Automated Analysis of Controversies about 'Monarch Butterflies,' Frankenfoods, 'and' stem Cells' ». In : *Scientometrics* 67.2, p. 231–258 (cf. p. 160).

LIQUÈTE, V., I. FABRE et C. GARDIÈS (2010). « Faut-il reconsidérer la médiation documentaire? » In : *Les Enjeux de L'Information Et de La Communication*. Dossier 2010. URL : http://w3.u-grenoble3.fr/les_enjeux/pageshtml/art2010-dossier.php (cf. p. 92).

LOTKA, A. (1926). « The Frequency Distribution of Scientific Productivity ». In : *Journal of the Washington Academy of Sciences* 16, p. 317–323 (cf. p. 2).

LUHN, H. P. (1957). « A Statistical Approach to Mechanized Encoding and Searching of Literary Information ». In : *IBM Journal of research and development* 1.4, p. 309–317 (cf. p. 157).

LUPU, M. et A. HANBURY (2013). « Patent retrieval ». In : *Foundations And Trends In Information Retrieval* 7.1, p. 1–97 (cf. p. 115).

LYCETT, M. (2013). « Datafication : making sense of (big) data in a complex world ». In : *European journal of information systems* 22.4, p. 381–386. ISSN : 0960-085X, 1476-9344. DOI : 10.1057/ejis.2013.10. URL : http://link.springer.com/10.1057/ejis.2013.10 (visité le 03/05/2017) (cf. p. 34).

MACMILLAN, D. (2006). « Patently obvious : the place for patents in information literacy in the sciences ». In : *Research Strategies* 20.3, p. 149–161. ISSN : 0734-3310. DOI : http://dx.doi.org/10.1016/j.resstr.2006.06.004. URL : http://www.sciencedirect.com/science/article/pii/S0734331006000073 (cf. p. 110).

MANDELBROT, B. (1954). « Structure formelle des textes et communication : deux études paramétriques ». In : *Word* 10.1, p. 1–27 (cf. p. 4).

MANSFIELD, E. (1991). « Academic research and industrial innovation ». In : *Research Policy* 20.1, p. 1–12 (cf. p. 111).

MARTIN, S., J. LIERMANN et H. NEY (1998). « Algorithms for bigram and trigram word clustering ». In : *Speech Commun* 24.1, p. 19–37. ISSN : 0167-6393. DOI : 10.1016/S0167-6393(97)00062-9. URL : http://dx.doi.org/10.1016/S0167-6393(97)00062-9 (cf. p. 161).

MARTY, E., P. MARCHAND et P. RATINAUD (2013). « Les médias et l'opinion-eléments théoriques et méthodologiques pour une analyse du débat sur l'identité nationale ». In : *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 117.1, p. 46–60 (cf. p. 149).

MATHIEN, M. (2003). « Abraham Moles ou l'information et la communication : Au carrefour des sciences, de la vie quotidienne et de l'esthétique ». In : *Communication* (Vol. 22/2), p. 167–181. ISSN : 1189-3788, 1920-7344. DOI : 10.4000/communication.4684. URL : http://communication.revues.org/4684 (visité le 21/04/2017) (cf. p. 57, 83).

MAY, C. (2006). « Patents, universities and the provision of social goods in the information society ». In : *Ethical Perspectives* 13.2, p. 289–304. ISSN : 1370-0049. DOI : 10.2143/EP.13.2.2016635. URL : http://poj.peeters-leuven.be/content.php?url=article&id=2016635 (visité le 18/02/2017) (cf. p. 111).

MAYAFFRE, D. (2002). « L'Herméneutique Numérique ». In : *L'Astrolabe. Recherche littéraire et Informatique* (numéro spécial), p. 1–11. URL : https://hal.archives-ouvertes.fr/hal-00586512 (cf. p. 150).

MBONGUI-KIALO, S. (2012). « Le brevet : un outil de communication au service de l'innovation ». In : *Revue Internationale D'Intelligence Économique* 4.2, p. 175–185. ISSN : 2101647X. DOI : 10.3166/r2ie.4.175-185. URL : http://r2ie.revuesonline.com/article.jsp?articleId=18612 (visité le 04/02/2017) (cf. p. 118).

MERZEAU, L. (2009). « Du signe à la trace : l'information sur mesure ». In : *Hermes* 53, p. 23–29. URL : https://halshs.archives-ouvertes.fr/halshs-00483292 (cf. p. 77).

— (2013). « L'intelligence des traces ». In : *Intellectica - La Revue de L'Association Pour La Recherche Sur Les Sciences de La Cognition (Arco)* 1.59, p.115–135. URL : https://halshs.archives-ouvertes.fr/halshs-01071211 (cf. p. 54).

MEYER, M. (2000). « Does science push technology? patents citing scientific literature ». In : *Research Policy* 29.3, p. 409–434 (cf. p. 109, 126).

— (2006). « Are patenting scientists the better scholars?: an exploratory comparison of inventor-authors with their non-inventing peers in nano-science and technology ». In : *Research Policy* 35.10, p. 1646–1662 (cf. p. 126).

MICHEL, J. et B. BETTELS (2001). « Patent citation analysis. A closer look at the basic input data from patent search reports ». In : *Scientometrics* 51.1, p. 185–201. ISSN : 1588-2861. DOI : 10.1023/A:1010577030871. URL : http://dx.doi.org/10.1023/A:1010577030871 (cf. p. 109).

MILLE, A. (2013). « Des traces à l'ère du web ». In : *Intellectica - La Revue de L'Association Pour La Recherche Sur Les Sciences de La Cognition (Arco)* 1.59, p. 7–28. URL : https://halshs.archives-ouvertes.fr/halshs-01071211 (cf. p. 54).

MILLER, M. (2016). « La culture de l'abstraction doit être renforcée par la culture générale ». In : *Le Monde. Campus 021*. URL : http://www.lemonde.fr/o21/article/2016/12/06/la-culture-de-l-abstraction-doit-etre-renforcee-par-la-culture-generale_5044330_5014018.html#IJopEWbTSF2mXcsl.99 (cf. p. 76).

MILLER, S. H. (2001). « Competitive intelligence–an overview ». In : *Competitive Intelligence Magazine* 1.11, p. 1–14 (cf. p. 117).

MOED, H. F. et G. HALEVI (2014). « A bibliometric approach to tracking international scientific migration ». In : *Scientometrics* 101.3, p. 1987–2001. ISSN : 0138-9130, 1588-2861. DOI : 10.1007/s11192-014-1307-6. URL : http://link.springer.com/10.1007/s11192-014-1307-6 (visité le 06/02/2017) (cf. p. 100, 126).

MOMENI, A. et K. ROST (2016). « Identification and monitoring of possible disruptive technologies by patent-development paths and topic modeling ». In : *Technological forecasting and social change* 104, p. 16–29. ISSN : 00401625. DOI : 10.1016/j.techfore.2015.12.003. URL : http://linkinghub.elsevier.com/retrieve/pii/S0040162515004072 (visité le 21/02/2017) (cf. p. 127).

MONFORT-WINDELS, F. (2008). « Donner du sens aux brevets. Typologie des utilisations de l'information brevets ». In : *Cahiers de La Documentation* 1, p. 18–22. URL : www.abd-bvd.be/wp-content/uploads/2008-1_Monfort-Windels_3e_partie1.pdf (cf. p. 116).

MORGAN, E. L. (1994). « The world wide and mosaic : an overview for librarians ». In : *The Public-Access Computer Systems Reviews* 5.6, p. 5–26. URL : http://epress.lib.uh.edu/pr/v5/n6/morgan.5n6 (visité le 01/04/2015) (cf. p. 36).

NANNIPIERI, O., I. MURATORE et al. (2016). « Les Statistiques : Un Pharmakon Pour La Communication ? » In : *ESSACHESS-Journal for Communication Studies* 9 (01), p. 79–94 (cf. p. 87).

NARIN, F. (1994). « Patent bibliometrics ». In : *Scientometrics* 30.1, p. 147–155. ISSN : 0138-9130. DOI : 10.1007/BF02017219. URL : http://dx.doi.org/10.1007/BF02017219 (cf. p. 122).

NARIN, F., K. S. HAMILTON et D. OLIVASTRO (1997). « The increasing linkage between u.s. technology and public science ». In : *Research policy* 26.3, p. 317–330. ISSN : 00487333. DOI : 10.1016/S0048-7333(97)00013-9. URL : http://linkinghub.elsevier.com/retrieve/pii/S0048733397000139 (visité le 18/10/2015) (cf. p. 111, 126).

NORDMAN, E. R. et D. TOLSTOY (2016). « The impact of opportunity connectedness on innovation in smes' foreign-market relationships ». In : *Technovation*. ISSN : 0166-4972. DOI : http://dx.doi.org/10.1016/j.technovation.2016.04.001. URL : http://www.sciencedirect.com/science/article/pii/S016649721630027X (cf. p. 117).

OMPI (2016). « Guide d'utilisation de la classification internationale des brevets (version 2016) ». In : URL : http://www.wipo.int/export/sites/www/classifications/ipc/fr/guide/guide_ipc.pdf (visité le 08/08/2015) (cf. p. 107).

ORDUNA-MALEA, E., M. THELWALL et K. KOUSHA (2017). « Web citations in patents : Evidence of technological impact? » In : *Journal of the association for information science and technology* 68.8, p. 1967–1974. ISSN : 23301635. DOI : 10.1002/asi.23821. URL : http://doi.wiley.com/10.1002/asi.23821 (visité le 15/08/2017) (cf. p. 110).

OUAKRAT, A. et J. MÉSANGEAU (2016). « Re-socialiser les traces d'activités numériques : une proposition qualitative pour les sciences de l'information et de la communication ». In : *Revue Française Des Sciences de L'Information Et de La Communication*. Humanités Numériques et Sciences de l'Information et de la Communication 8. URL : https://halshs.archives-ouvertes.fr/halshs-01294749 (cf. p. 68).

OUBRICH, M. et R. BARZI (2012). « Le brevet comme source d'information stratégique : cas de l'activité inventive au maroc ». In : *Revue Internationale D'Intelligence Économique* 4.2, p. 205–222. ISSN : 2101647X. DOI : 10.3166/r2ie.4.205-222. URL : http://r2ie.revuesonline.com/article.jsp?articleId=18614 (visité le 04/02/2017) (cf. p. 100, 116).

PAO, M. L. (1978). « Automatic Text Analysis Based on Transition Phenomena of Word Occurrences ». In : *Journal of the Association for Information Science and Technology* 29.3, p. 121–124 (cf. p. 158).

PARANJPE, P. P. (2012). « Patent information and search ». In : *Desidoc Journal of Library & Information Technology* 32.3 (cf. p. 101, 122).

PÉLISSIER, D. (2015). « « Trace », vous avez vraiment dit « trace » ? » In : *Présence Numérique Des Organisations* 76. URL : https://presnumorg.hypotheses.org/76 (cf. p. 54).

PERVEZ, A. (2009). « Information as form ». In : *Triplec : Cognition, Communication, Co-Opération* 7.1, p. 1–11. ISSN : 1726-670X (cf. p. 70).

PICARD, D. (1992). « De la communication à l'interaction : l'évolution des modèles ». In : *Communication Et Langages* 93.1, p. 69–83. ISSN : 0336-1500. DOI : 10.3406/colan.1992.2380. URL : http://www.persee.fr/doc/colan_0336-1500_1992_num_93_1_2380 (cf. p. 50).

PINÈDE, N. (2014). « Le réseau, un « entre-deux » liens ? De quelques facettes et dynamiques du lien hypertexte ». In : *Sciences de La Société* 91. DOI : 10.4000/sds.1299. URL : http://sds.revues.org/1299 (cf. p. 31).

PLANTIN, J.-C. et F. RUSSO (2016). « D'abord Les Données, Ensuite La Méthode ?: Big Data et Déterminisme En Sciences Sociales* ». In : *Socio* 6, p. 97–115. ISSN : 2266-3134, 2425-2158. DOI : 10.4000/socio.2328. URL : http://socio.revues.org/2328 (visité le 28/04/2017) (cf. p. 87).

POLANCO, X. (1995). « Aux sources de la scientométrie ». In : *Solaris* 2. URL : http://biblio-fr.info.unicaen.fr/bnum/jelec/Solaris/d02/2polanco1.html (cf. p. 125).

PORTER, A. L. (2007). « How tech mining can enhance R&D management ». In : *Research-Technology Management* 50.2, p. 15–20 (cf. p. 117, 129).

QUÉRÉ, L. (1989). « Les boîtes noires de Bruno Latour ou le lien social dans la machine ». In : *Réseaux* 7.36, p. 95–117. ISSN : 0751-7971. DOI : 10.3406/reso.1989.1354. URL : http://www.persee.fr/doc/reso_0751-7971_1989_num_7_36_1354 (cf. p. 99).

— (2000). « Au juste, qu'est-ce que l'information ? » In : *Réseaux*. Communiquer à l'ère des réseaux 18.100, p. 331–357. ISSN : 0751-7971. DOI : 10.3406/reso.2000.2227. URL : http://www.persee.fr/doc/reso_0751-7971_2000_num_18_100_2227 (cf. p. 50).

QUONIAM, L., C. T. KNIES et M. R. MAZIERI (2014). « Patente como objeto de pesquisa em ciências da informação e comunicação ». In : *Encontros Bibli : Revista Eletrônica de Biblioteconomia E Ciência Da Informação*, 19, p. 243–268. ISSN : 1518 - 2924. DOI : 0.5007/1518-2924.2014v19n39p243 (cf. p. 101).

RADAUER, A. et L. WALTER (2010). « Elements of good practice for providers of publicly funded patent information services for SMEs selected and amended results of a benchmarking exercise ». In : *World Patent Information* 32.3, p. 237–245. ISSN : 0172-2190. DOI : http://dx.doi.org/10.1016/j.wpi.2009.09.003. URL : http://www.sciencedirect.com/science/article/pii/S0172219009000982 (cf. p. 118, 127).

« Rapport d'activité 2015 » (2016). In : Commissariat Général à l'Investissement. Réd. par A. JUPPÉ et M. ROCARD, p. 60 (cf. p. 111).

RASTIER, F. (2005). « Sémiotique du cognitivisme et sémantique cognitive : questions d'histoire et d'épistémologie ». In : *Texto ! Inédits*. URL : http://www.revue-texto.net/Inedits/Rastier/Rastier_Semantique-cognitive.html (cf. p. 56).

REINERT, A. (1983). « Une Méthode de Classification Descendante Hiérarchique : Application à l'analyse Lexicale Par Contexte ». In : *Les cahiers de l'analyse des données* 8.2, p. 187–198 (cf. p. 147).

REINERT, M. (1986). « Un Logiciel d'analyse Lexicale ». In : *Les Cahiers de l'analyse des données* 11.4, p. 471–481 (cf. p. 147).

— (1990). « Alceste une méthodologie d'analyse des données textuelles et une application : aurelia de gerard de nerval ». In : *Bulletin de Méthodologie Sociologique* 26.1, p. 24–54 (cf. p. 147, 149).

REYMOND, D. (2016a). « Chaînes de traitement de la documentation brevet : nouvelles herméneutiques pour lintelligence économique ». In : *Revue Internationale d'Intelligence Économique (RIEE)* 8.1 (cf. p. 129, 134).

REYMOND, D. et L. QUONIAM (2016). « A new patent processing suite for academic and research purposes ». In : *World patent information* 47, p. 40–50. ISSN : 01722190. DOI : 10.1016/j.wpi.2016.10.001 (cf. p. 166).

RIBEIRO FERRAZ, R. N. et al. (2016). « Example of Open-Source OPS (Open Patent Services) for Patent Education and Information Using the Computational Tool Patent2Net ». In : *World Patent Information* 46, p. 21–31. ISSN : 0172-2190. DOI : 10.1016/j.wpi.2016.05.002. URL : http://www.sciencedirect.com/science/article/pii/S0172219016300333 (visité le 01/06/2016) (cf. p. 134).

RICHMOND, S. (2016). « Why humans matter (the 4th revolution : how the infosphere is reshaping human reality by Luciano Floridi) ». In : *Kyiv-Mohyla Humanities Journal* 0.3, p. 231. ISSN : 2313-4895. DOI : 10.18523/kmhj73965.2016-3.231-235. URL : http://kmhj.ukma.edu.ua/article/view/73965 (visité le 10/12/2016) (cf. p. 36).

ROUVROY, A. et T. BERNS (2013). « Gouvernementalité algorithmique et perspectives d'émancipation : le disparate comme condition d'individuation par la relation ? » In : *Reseaux* 31.177. Sous la dir. de D. CARDON, p. 163–196. URL : http://works.bepress.com/antoinette_rouvroy/47 (visité le 01/04/2015) (cf. p. 65, 66).

ROUVROY, A. et B. STIEGLER (2015). « Le régime de vérité numérique. De la gouvernementalité algorithmique à un nouvel état de droit ». In : *Socio - Le Tournant Numérique,... Et Après ?* Sous la dir. de D. DIMINESCU et M. WIEVIORKA, p. 113–140 (cf. p. 65).

RUIZ-BAÑOS, R., R. BAILÓN-MORENO, E. JIMENEZ-CONTRERAS et al. (1999). « Structure and Dynamics of Scientific Networks. Part I : Fundamentals of the Quantitative Model of Translation ». In : *Scientometrics* 44.2, p. 217–234 (cf. p. 159, 160, 6).

RUIZ-BAÑOS, R., R. BAILÓN-MORENO, E. JIMÉNEZ-CONTRERAS et al. (1999). « Structure and Dynamics of Scientific Networks. Part II : The New Zipf's Law, the Clusters of Co-Citations and the Model of the Descriptor Presence ». In : *Scientometrics* 44.2, p. 235–265 (cf. p. 159, 160, 6).

SAAD, F. et A. NÜRNBERGER (2012). « Overview of prior-art cross-lingual information retrieval approaches ». In : *World Patent Information* 34.4, p. 304–314. ISSN : 0172-2190. DOI : http://dx.doi.org/10.1016/j.wpi.2012.08.013. URL : http://www.sciencedirect.com/science/article/pii/S0172219012001469 (cf. p. 120).

SALTON, G., A. WONG et C. S. YANG (1975). « A vector space model for automatic indexing ». In : *Commun. Acm* 18.11, p. 613–620. ISSN : 0001-0782. DOI : http://doi.acm.org/10.1145/361219.361220 (cf. p. 8).

SAVOLAINEN, R. (2016). « Elaborating the conceptual space of information-seeking phenomena ». In : *Information Research* 21.3, paper 720. URL : Information%20Research (cf. p. 44, 69).

SCHAFER, V. et B. THIERRY (2013a). « Des algorithmes et du naturel... entretien avec bernard chazelle ». In : *Revue Des Sciences Et Technologies de L'Information* 5, p. 641–652 (cf. p. 31, 60).

— (2013b). « Qui a inventé internet ? une vraie « fausse question »... » In : *Le Temps Des Médias* 1.20, p. 223–235 (cf. p. 27, 28).

SCHMOCH, U. (1997). « Indicators and the relations between science and technology ». In : *Scientometrics* 38.1, p. 103–116. ISSN : 1588-2861. DOI : 10.1007/BF02461126. URL : http://dx.doi.org/10.1007/BF02461126 (cf. p. 126).

SCHRADER, A. M. (1986). « The domain of information science : problems in conceptualization and in consensus-building ». In : *Information Services & Use* 6 (5-6), p. 169–205 (cf. p. 44).

SERRES, A. (2002). « Quelle(s) problématique(s) de la trace ? » In : URL : https://archivesic.ccsd.cnrs.fr/sic_00001397 (cf. p. 54).

SHADBOLT, N., W. HALL et T. BERNERS-LEE (2006). « The semantic web revisited ». In : *IEEE Intelligent Systems*. URL : http://eprints.ecs.soton.ac.uk/12614/01/Semantic_Web_Revisted.pdf (visité le 01/04/2015) (cf. p. 30).

SIDOROVA, A. et al. (2008). « Uncovering the Intellectual Core of the Information Systems Discipline ». In : *Mis Quarterly*, p. 467–482 (cf. p. 44).

SMALL, H. (1973). « Co-citations in the scientific literature : a new measure of the relationship beetween two documents ». In : *Journal of the American Society For Information Science* 24.4, p. 265–269 (cf. p. 2).

« SME tailor-designed patent portfolio analysis » (p.d.). In : 31. ISSN : 0172-2190. DOI : http://dx.doi.org/10.1016/j.wpi.2008.12.003. URL : http://www.sciencedirect.com/science/article/pii/S0172219008001762 (cf. p. 114, 122).

SMITH, A. G. (2004). « Web links as analogues of citations ». In : *Information Research* 9.4. URL : http://informationr.net/ir/9-4/paper188.html (cf. p. 80).

TANNEBAUM, W. et A. RAUBER (2014). « Using query logs of uspto patent examiners for automatic query expansion in patent searching ». In : *Information retrieval* 17 (5-6), p. 452–470. ISSN : 1386-4564, 1573-7659. DOI : 10.1007/s10791-014-9238-7. URL : http://link.springer.com/10.1007/s10791-014-9238-7 (visité le 21/02/2017) (cf. p. 128).

TÉTU, J.-F. et F. RENZETTI (1995). « Internet : Évolution d'un Projet d'espace Public de La Recherche ». In : *Technologies de l'information et société* 7.2, 11 pages. URL : https://halshs.archives-ouvertes.fr/halshs-00396161 (cf. p. 27).

THELWALL, M. (2002). « Evidence for the existence of geographic trends in university website interlinking ». In : *Journal of Documentation* 58.5 (cf. p. 80).

— (2008). « Bibliometrics to webometrics ». In : *Journal of Information Science* 34.4, p. 605–621 (cf. p. 80).

THELWALL, M., R. BINNS et al. (2002). « European union associated university websites ». In : *Scientometrics* 53.1, p. 95–111 (cf. p. 80).

THELWALL, M. et D. WILKINSON (2004). « Finding similar academic web sites with links, bibliometric couplings and colinks ». In : *Inf. Process. Manage.* 40.3, p. 515–526. ISSN : 0306-4573. DOI : http://dx.doi.org/10.1016/S0306-4573(03)00042-6. URL : http://www.scit.wlv.ac.uk/~cm1993/papers/2004%20_IPM%20_Links%20_Bibliometric%20_Couplings%20_Colinks.pdf (cf. p. 80).

THOMA, G. (2014). « Composite value index of patent indicators : factor analysis combining bibliographic and survey datasets ». In : *World patent information* 38, p. 19–26. ISSN : 01722190. DOI : 10.1016/j.wpi.2014.05.005. URL : http://linkinghub.elsevier.com/retrieve/pii/S0172219014000921 (visité le 21/02/2017) (cf. p. 127, 145).

THOMS, L. et M. THELWALL (2005). « Academic home pages : reconstruction of the self ». In : *First Monday* 10.12. URL : http://www.firstmonday.org/issues/issue10_12/thoms/ (cf. p. 80).

TIJSSEN, R. J. (2001). « Global and domestic utilization of industrial relevant science : patent citation analysis of science–technology interactions and knowledge flows ». In : *Research Policy* 30.1, p. 35–54 (cf. p. 126).

TRUMBACH, C. C., D. PAYNE et A. KONGTHON (2006). « Technology mining for small firms : knowledge prospecting for competitive advantage ». In : *Technological Forecasting And Social Change* 73.8, p. 937–949 (cf. p. 118).

URBIZAGÁSTEGUI ALVARADO, R. et C. RESTREPO ARANGO (2011). « La Ley de Zipf y El Punto de Transición de Goffman En La Indización Automática ». In : *Investigación bibliotecológica* 25.54, p. 71–92 (cf. p. 159).

VAN DIJCK, J. (2014). « Datafication, Dataism and Dataveillance : Big Data between Scientific Paradigm and Ideology ». In : *Surveillance & Society* 12.2, p. 197 (cf. p. 34).

VAN OVERWALLE, G. (2006). « Reconciling patent policies with the university mission ». In : *Ethical Perspectives* 13.2, p. 231–247. ISSN : 1370-0049. DOI : 10.2143/EP.13.

2.2016632. URL : http://poj.peeters-leuven.be/content.php?url=article&id=2016632 (visité le 18/02/2017) (cf. p. 111).

VAN ZEEBROECK, N. (2011). « The puzzle of patent value indicators ». In : *Economics of Innovation And New Technology* 20.1, p. 33–62 (cf. p. 114).

VAYRE, J.-S. (2014). « Manipuler les données. Documenter le marché ». In : *Les Cahiers Du Numérique* 10.1, p. 95–125. URL : https://hal.archives-ouvertes.fr/hal-01003135 (cf. p. 79).

VENTURINI, T. et B. LATOUR (2010). « The social fabric : digital traces and quali-quantitative methods ». In : *Proceedings of Future En Seine*. URL : http://www.medialab.sciences-po.fr/publications/Venturini_Latour-The_Social_Fabric.pdf (visité le 01/04/2015) (cf. p. 81, 90, 126).

VENUGOPALAN, S. et V. RAI (2015). « Topic based classification and pattern identification in patents ». In : *Technological forecasting and social change* 94, p. 236–250. ISSN : 00401625. DOI : 10.1016/j.techfore.2014.10.006. URL : http://linkinghub.elsevier.com/retrieve/pii/S0040162514002923 (visité le 21/02/2017) (cf. p. 128).

VERHOEVEN, D., J. BAKKER et R. VEUGELERS (2016). « Measuring technological novelty with patent-based indicators ». In : *Research policy* 45.3, p. 707–723. ISSN : 00487333. DOI : 10.1016/j.respol.2015.11.010. URL : http://linkinghub.elsevier.com/retrieve/pii/S0048733315001857 (visité le 21/02/2017) (cf. p. 128).

WADE, M., M. BIEHL et H. KIM (2006). « Information Systems Is Not a Reference Discipline (and What We Can Do about It) ». In : *Journal of the Association for Information Systems* 7.1, p. 14 (cf. p. 44).

WHITE, H. et B. GRIFFITH (1982). « Authors as markers of intellectual space : co-citation in studies of science, technology, and society ». In : *Journal of Documentation* 38.4, p. 255–272 (cf. p. 2).

WHITE, H. et K. MCCAIN (1989). « Bibliometrics ». In : *Annual Review of Information Science And Technology* 24, p. 119–186 (cf. p. 2).

WONGEL, H. (2005). « The reform of the ipcconsequences for the users ». In : *World Patent Information* 27.3, p. 227–231. ISSN : 0172-2190. DOI : http://dx.doi.org/10.1016/j.wpi.2005.02.002. URL : http://www.sciencedirect.com/science/article/pii/S0172219005000402 (cf. p. 144).

WOOD, L. (1999). « Programming the Web : The W3C DOM Specification ». In : *IEEE Internet Computing* 3.1, p. 48–54. ISSN : 10897801. DOI : 10.1109/4236.747321. URL : http://ieeexplore.ieee.org/document/747321/ (visité le 08/03/2017) (cf. p. 30).

ZESCHKY, M., B. WIDENMAYER et O. GASSMANN (2011). « Frugal innovation in emerging markets ». In : *Research-technology management* 54.4, p. 38–45. ISSN : 08956308, 19300166. DOI : 10.5437/08956308X5404007. URL : https://www.alexandria.unisg.ch/publications/211637/L-en (visité le 26/03/2014) (cf. p. 118).

ZHANG, L., L. LI et T. LI (2015). « Patent mining : a survey ». In : *Sigkdd Explor. Newsl.* 16.2, p. 1–19. ISSN : 1931-0145. DOI : 10.1145/2783702.2783704. URL : http://doi.acm.org/10.1145/2783702.2783704 (cf. p. 118, 120).

Actes de conférence

AGARWAL, N. et A. BREM (2012). « Frugal and reverse innovation - literature overview and case study insights from a german mnc in india and china ». In : *2012 18th international conference on engineering, technology and innovation, ice 2012 - conference proceedings*. 2012 18th international conference on engineering, technology and innovation, ice 2012. ISBN : 9781467322751 (ISBN). DOI : 10.1109/ICE.2012.6297683. URL : http://www.scopus.com/inward/record.url?eid=2-s2.0-84867908700&partnerID=40&md5=e68dae5f8e6bd23d727e2cbbe0d6b5a6 (visité le 18/06/2012) (cf. p. 118).

AGUILLO, I. (2010). « Web, webometrics and the ranking of universities ». In : *Proceedings of the 3Rd European Network of Indicators Designers Conference on Sti Indicators For Policymaking And Strategic Decision*. CNAM, Paris (cf. p. 80).

ANTONIE, M.-L. et O. ZAIANE (2002). « Text Document Categorization by Term Association ». In : IEEE Comput. Soc, p. 19–26. ISBN : 978-0-7695-1754-4. DOI : 10.1109/ICDM.2002.1183881. URL : http://ieeexplore.ieee.org/document/1183881/ (visité le 24/05/2017) (cf. p. 8).

ARTHUR, D. et S. VASSILVITSKII (2007). « K-Means++ : The Advantages of Careful Seeding ». In : *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial et Applied Mathematics, p. 1027–1035 (cf. p. 176, 177, 23, 24).

BALTZ, C. (2013). « Numérique : meurtre de l'information ? » In : sous la dir. d'I. SALEH et al. Paris : Hermès/Lavoisier, p. 345–358 (cf. p. 46).

BATTLES, M. (2013). « Data artefacts : tracking knowledge-ordering conflicts through visualization ». In : *Classification And Visualization : Interfaces To Knowledge : Proceedings of the International Udc Seminar*. Sous la dir. d'A. SLAVIC, A. A. SALAH et S. DAVIES. The Hague, the Netherlands : Würzburg : Ergon Verlag (cf. p. 134).

BEAUDOUIN, V. (2016). « Retour Aux Origines de La Statistique Textuelle : Benzécri et l'école Française d'analyse Des Données ». In : *JADT 2016*. Statistical Analysis of Textual data - JADT 2016. Nice, France : Mayaffre, D. Poudat, C., Vanni, L. et al., p. 17–27. URL : https://hal.archives-ouvertes.fr/hal-01376938 (cf. p. 147).

BEN AMOR, S., F. RENUCCI et H. ZÉNOUDA (2013). « Aux frontières de l'homme-interfacé ». In : sous la dir. d'I. SALEH et al. Paris : Hermès/Lavoisier, p. 345–358 (cf. p. 33).

BERNERS-LEE, T., R. CAILLIAU et al. (1992). « World-wide web : an information infrastructure for high-energy physics ». In : *Proceedings of the Workshop on Software Engineering, Artificial Intelligence And Expert Systems For High Energy And Nuclear Physics*. Sous la dir. de D. PERRET-GALLIX. La Londe-les-Maures, France : World Scientific, Singapore. URL : http://citeseer.ist.psu.edu/12816.html (visité le 07/04/2007) (cf. p. 28).

BOULLIER, D. (2015). « Les sciences sociales face aux traces du big data ? » In : *Société, Opinion Et Répliques*. FMSHWP- 2015-88. URL : https://halshs.archives-ouvertes.fr/halshs-01141120 (cf. p. 77).

BROCK, D. C. et al. (2012). « Applied actant-network theory : toward the automated detection of technoscientific emergence from full-text publications and patents ». In : *2012 Aaai Fall Symposium Series* (cf. p. 126).

CAPURRO, R. (2010). « Beyond humanisms ». In : Information ethics : future of humanities. URL : http://www.capurro.de/humanism.html (visité le 27/01/2017) (cf. p. 71, 73).

CHERRABI, N. et al. (2015). « Étude sur les demandes de dépôt de brevet en Algérie, au Maroc et en Tunisie ». In : *Proceedings of the 5th. International Symposium ISKO-Maghreb. Knowledge Organization in the perspective of Digital Humanities : Researches and Applications*. Knowledge Organization in the perspective of Digital Humanities : Researches and Applications (cf. p. 126, 134).

CHUNG-HUEI, K. et L. CHAN-YI (2015). « An empirical study on utilizing pre-grant publications in patent classification analysis ». In : *15Th International Conference on Scientometrics And Informetrics*. Sous la dir. d'A. SALAH et al. Istanbul, Turkey : Bogaziçi University Printhouse. URL : http://www.issi2015.org/en/Proceedings-of-ISSI-2015.html (cf. p. 152).

COURBET, D. (2004). « Comparaison épistémologique des recherches en sic et sciences de gestion dans le domaine de la communication externe, divergences et terrain commun ». In : *Colloque La Communication d'entreprise : regards croisés sciences de gestion et sciences de l'information et de la communication*. Nice, 6-7 décembre 2001. (cf. p. 91).

DOU GOARIN, C. (2013). « Patent analysis, detection of new markets for employment ». In : *Innovation, Brevets Et Normes : Complémentarités Et Conflits*. Tours, France. URL : https://hal.archives-ouvertes.fr/hal-00913373 (cf. p. 129).

FLON, É. et al. (2009). « Traces d'écriture, traces de pratiques, traces d'identités ». In : *Actes de La Conférence H2Ptm* (cf. p. 55).

LARKEY, L. S. (1999). « A Patent Search and Classification System ». In : *Proceedings of the Fourth ACM Conference on Digital Libraries*. ACM, p. 179–187 (cf. p. 163).

LELEU-MERVIEL, S. (2010b). « Le sens aux interstices, émergence de reliances complexes ». In : *Colloque International Francophone "Complexité 2010"*. Lille, France, p. 22. URL : https://hal.archives-ouvertes.fr/hal-00526508 (cf. p. 59, 60).

LHEN, J. et al. (1995). « La Statistique Des Lois de Zipf, Actes Du Colloque, Les Systèmes d'informations Élaborés ». In : *Les Systèmes d'information Élaborés*. Journée d'étude de la SFBA. Ile Rousse - Corse : Société Française de Bibliométrie Appliquée (cf. p. 158, 3).

MAYAFFRE, D. (2014). « Plaidoyer en faveur de l'analyse de données co(n) Textuelles. Parcours cooccurrentiels dans le discours présidentiel français (1958-2014) ». In : *JADT 2014*. Sous la dir. d'I.-S. NOUVELLE. JADT 2014, Proceedings of the 12th International Conference on Textual Data Statistical Analysis. Paris, France : Emilie Née et al., p. 15–32. URL : https://hal.archives-ouvertes.fr/hal-01181337 (cf. p. 161).

OSINSKI, S., J. STEFANOWSKI et D. WEISS (2004). « Lingo : search results clustering algorithm based on singular value decomposition ». In : *Intelligent Information Processing And Web Mining, Proceedings of the International Iis : Iipwm'04 Conference Held In Zakopane, Poland, May 17-20, 2004*. Sous la dir. de M. A. KLOPOTEK, S. T. WIERZCHON et K. TROJANOWSKI. Advances in Soft Computing. Springer, p. 359–368. ISBN : 3-540-21331-7 (cf. p. 150).

QUONIAM, L. et D. REYMOND (2013). « Patents as a free innovation tool ». In : *International Conference on Information Systems and Technology Management (CONTECSI)* (cf. p. 103).

— (2014a). « Frugal innovation : social responsability impact ». In : *III Simposio Internacional de Gestão de Projetos (III SINGEP)* (cf. p. 118).

— (2014b). « Patent as a tool for frugal innovation, innovative conception, KDD ». In : *L'information et la connaissance au cur de l'innovation et de La croissance* (cf. p. 118).

QUONIAM, L., D. REYMOND et C. REY (2014). « Frugal innovation, innovative conception, KDD ». In : *Journée des utilisateurs Intellixir* (cf. p. 118).

REYMOND, D. et N. PINÈDE (2010). « Website and communication strategy alignment : a librarian science approach to webometrics tools ». In : *Proceedings of the Sixth International Conference on Webometrics, Informetrics and Scientometrics (WIS) & Eleventh COLLNET Meeting* (cf. p. 80).

SCHÖCH, C. (2012). « Nouvelles configurations : textes, outils, méthodes, et infrastructures de recherche dans les études de lettres ». In : *Colloque International : Configuration(S)*. Paris, France. URL : https://hal.archives-ouvertes.fr/hal-00951518 (cf. p. 81).

SZONIECKY, S. et M. LOUÂPRE (2015). « Outillages numériques pour les humanités : cartographier des réseaux d'influences ». In : *Isko - Magreb 2015 : Organisation de La Connaissance Dans La Perspective Des Humanités Numériques : Recherches Et Applications*. Hammamet, Tunisia. URL : https://hal.archives-ouvertes.fr/hal-01220142 (cf. p. 78).

TÉTU, J.-F. (1992). « Le Territoire, Entre Frontières et Réseaux ». In : *VIIIe Congrès de La Société Française Des Sciences de l'information et de La Communication*. Sous la dir. de CREDO/S.F.S.I.C. Lille, France : CREDO/S.F.S.I.C, pp.115–119. URL : https://halshs.archives-ouvertes.fr/halshs-00395687 (cf. p. 32).

Rapports

MOUNIER, M. et P. DACOS (2015). *Humanités numériques. État des lieux et positionnement de la recherche française dans le contexte international*. Institut français, p. 89 (cf. p. 76, 77).

Autres (thèses, HDR, présentations, etc.)

BAILÓN-MORENO, R. (2003). « Ingeniería del conocimiento y vigilancia tecnológica aplicada a la investigación en el campo de los tensioactivos. Desarrollo de un modelo ciencimétrico unificado ». Thèse de doct. España : Departamento de Ingeniería Química, Universidad de Granada. 675 p. (cf. p. 159, 5).

BJÖRNEBORN, L. (2004). « Small-world link structures across an academic web space : a library and information science approach ». Thèse de doct. Copenhagen, Denmark : Department of Information et Library Science (cf. p. 80).

BOOTZ, P. (2001). « Formalisation of a functional model of communication by digital technologies applied to the poetic creation ». Theses. Université Paris VIII Vincennes-

Saint Denis. 572 p. URL : https://tel.archives-ouvertes.fr/tel-00012165 (cf. p. 53).

CASTILLO SEQUERA, J. L. (2010). « Nueva Propuesta Evolutiva Para El Agrupamiento de Documentos En Sistemas de Recuperación de Información ». Alcalá (España) : Universidad de Alcalá. Departamento de Ciencias de la Computación. 270 p. (cf. p. 158).

EDUARDO STOROPOLI, J. (2016). « O uso do knowledge discovery in database (kdd) de informações patentarias sobre ensino a distância : contribuções de ensino superior ». Thèse de doct. Universidade Novo de Julho, São Paulo, Brazil (cf. p. 134).

LELEU-MERVIEL, S. (1996). « La scénistique : méthodologie pour la conception de documents en media multiples suivant une approche qualité ». Habilitation à diriger des recherches. Université Paris VIII Vincennes-Saint Denis. URL : https://tel.archives-ouvertes.fr/tel-00660099 (cf. p. 58).

LEVY, P. (2015). « The emergence of reflexive collective intelligence ». URL : http://pierrelevyblog.com/tag/ieml/ (visité le 01/04/2015) (cf. p. 64).

« Manually self-operated butt-kicking machine » (2006). (US). URL : https://worldwide.espacenet.com/publicationDetails/biblio?FT=D&date=20060504&DB=EPODOC&locale=en_EP&CC=US&NR=2006094518A1&KC=A1&ND=5 (visité le 03/01/2017) (cf. p. 131).

MERZEAU, L. (2011). « Pour une médiologie de la mémoire ». Habilitation à diriger des recherches. Université de Nanterre - Paris X. URL : https://tel.archives-ouvertes.fr/tel-00904667 (cf. p. 54, 55, 76, 90).

MOUNIER, M. et P. DACOS (2015). *Humanités numériques. État des lieux et positionnement de la recherche française dans le contexte international*. Institut français, p. 89 (cf. p. 76, 77).

PIERRET, J.-D. (2006). « Methodologie et structuration d'un outil de decouverte de connaissances base sur la litterture biomedicale : une application basee sur le mesh ». Theses. Université du Sud Toulon Var. URL : https://tel.archives-ouvertes.fr/tel-00011704 (cf. p. 123).

SERRES, A. (2000). « Aux sources d'internet : l'émergence d'arpanet ». Thèse de doct. Villeneuve d'Ascq : Presses Universitaires du Septentrion (2003) (cf. p. 27).

« User-operated amusement apparatus for kicking the user's buttocks » (2001). (US). J. W. ARMSTRONG. URL : https://worldwide.espacenet.com/publicationDetails/biblio?FT=D&date=20010925&DB=EPODOC&locale=en_EP&CC=US&NR=6293874B1&KC=B1&ND=6 (visité le 03/01/2017) (cf. p. 131).

Logiciels et librairies

BASTIAN, M. (2009). *Gephi*. Version 0.9.1. URL : https://gephi.org/ (cf. p. 141).

BILGIN, A. et al. (2015). *Graphviz*. URL : http://graphviz.org/ (cf. p. 143).

DataTable API (2007). URL : http://datatables.net (cf. p. 138, 19).

HEYMANN, S. et al. (2007). *GEXF File Format*. Version 1.2. URL : http://gexf.net/format/ (cf. p. 141).

KRUCHTEN, M. (2013). *PivotTable*. URL : http://nicolas.kruchten.com/pivottable (cf. p. 138, 20).

PATENTDATA@EPO.ORG (2014). *API Console EPO Developer Portal (OPS)*. Version 1.2. URL : https://developers.epo.org/ (cf. p. 135).

POLIVAEV, D. (2000). *Freeplane - Free Mind Mapping and Knowledge Management Software*. Version 1.5.20. URL : http://freeplane.sourceforge.net/wiki/index.php/Main_Page (cf. p. 152).

RATINAUD, P. (2008). *Interface de R Pour Les Analyses Multidimensionnelles de Textes et de Questionnaires*. Version 0.7 alpha 2. URL : http://www.iramuteq.org (cf. p. 148).

REYMOND, D. (2016b). *Patent2Net (P2N)*. Version 2. URL : https://github.com/Patent2net/P2N (cf. p. 134).

RRCHNM (2015). *Zotero*. Version 4. URL : https://www.zotero.org/ (cf. p. 140).

SIMOES, A. (2013). *D3plus*. URL : http://d3plus.org/ (cf. p. 140).

SONG, G. (2014). *Python-Epo-Ops-Client*. Version 3.1. URL : https://github.com/55minutes/python-epo-ops-client (cf. p. 136).

VELT, R. (2011). *GEXF-JS*. URL : https://github.com/raphv/gexf-js/ (cf. p. 141).

WEISS, D. et S. OSINSKI (2004). *Carrot2 - Open Source Search Results Clustering Engine*. URL : http://project.carrot2.org/ (cf. p. 150).

III

Annexes

ANNEXE A

Éléments d'infométrie convoqués

Dans cet environnement turbulent, dont l'évolution dépend des initiatives d'un grand nombre, la capacité de collecter des informations et de les traiter, avant de prendre les décisions, devient cruciale.

— CALLON, COURTIAL, TURNER, La méthode leximappe : un outil pour l'analyse stratégique du développement scientifique et technique (p. 208) (1991)

LES PROCESSUS DE PRODUCTION D'INFORMATION généralisés par EGGHE et ROUSSEAU (EGGHE et ROUSSEAU, 1990) font partie dorénavant des fondamentaux des sciences sociales au sein desquelles les lois de puissance permettent de modéliser de nombre de phénomènes (Économie avec PARETO, linguistique avec ZIPF, productivité des revues avec BRADFORD, biologie avec YULE). Toutes issues ou variantes des lois de Pareto (BARBUT, 1988), les lois empiriques de la scientométrie (CALLON, J.-P. COURTIAL et PENAN, 1993) recouvrent les distributions (ZIPF (1949), LOTKA (1926) et BRADFORD (1934)) en description de structures des phénomènes informationnels (WHITE et MCCAIN, 1989; LEYDESDORFF et BENSMAN, 2005; KESSLER, 1963), mais aussi en appréciation de leur dynamique par des lois de croissance, déclin et d'obsolescence (PRICE (1986), BROOKES (1968) [1]). Globalement, trois traitements de statistique descriptives appliqués à des ensembles de textes (et de leurs producteurs) sont opérés : la production d'indicateurs non relationnels d'activités; les indicateurs de première génération (analyse des co-citations, (ADAIR, 1955; SMALL, 1973; WHITE et GRIFFITH, 1982)) et les indicateurs de seconde génération (analyse des mots associés (CALLON, J. P. COURTIAL et al., 1983; CALLON, J.-P. COURTIAL et W. TURNER, 1991; W. A. TURNER et al., 1988)).

Ces modèles s'appuient sur la théorie de l'information de SHANNON, qui permet d'en démontrer la validité mathématique.

Le modèle fractal de MANDELBROT constitue un pallier intermédiaire. À ces développements théoriques s'opère l'ajustement des lois à des données réelles (BARBUT, 1989).

1. Cf. (HUBERT, 1981), AVRAMESCU (1979)

A.1 Description des ensembles textuels

A.1.1 De l'entropie aux zones pertinentes

Parmi les multiples définitions de l'entropie, l'entropie de RENYI ou entropie d'ordre α se définit par

$$\begin{cases} \text{pour } \alpha \neq 1, H_\alpha = \frac{1}{1-\alpha} \log \sum_{i=1}^{n} (p_i)^\alpha \\ \text{pour } \alpha = 1, H_1 = \lim_{\alpha \longrightarrow 1} H_\alpha = -\sum_{i=1}^{n} p_i \log p_i \end{cases} \tag{A.1}$$

Notons que H_1 est l'entropie de SHANNON (cf. équation A.1.2 p. 5).

La diversité de HILL est l'exponentielle de l'entropie de RENYI :

$$D_\alpha = e^{H_\alpha} \tag{A.2}$$

et est aussi dénommée dans d'autres travaux comme une **mesure de la perplexité** (JELINEK, 1976; BAHL, JELINEK et MERCER, 1983).

Sur une distribution de n formes pour un total de N occurrences, en posant $H_0 = log n$ et $D_0 = \exp H_0 = n$, H_1 l'entropie de SHANNON, $D_1 = \exp H_1$. LHEN et al. (1995) utilisent la concentration de HILL et DAGET définie par

$$C = \sum_{i=1}^{n} \left(\frac{n_i}{N}\right)^2 = \sum_{i=1}^{n} (p_i)^2 \tag{A.3}$$

pour définir la diversité d'ordre 2 : $D_2 = \frac{1}{C}$

Les auteurs statuent alors que la fréquence de séparation (coupure) entre les zones triviales et informationnelles (notée C_b) puis les zones informationnelles et bruitées (notée C_t) d'une distribution de ZIPF (cf. figure 6.1 p. 159) s'obtiennent par les formules suivantes :

$$C_b = D_0 \times \left(\frac{H_0}{H_1} \times \frac{H_0 - H_1}{H_0 - H_2 + \frac{H_1 + H_2}{H_1}} \right) \tag{A.4}$$

$$C_t = D_2 = \frac{1}{C} \tag{A.5}$$

A.1.2 Distribution de Zipf-Mandelbrot

La loi primitive de J.B ESTOUB et G.K. ZIPF nommée loi harmonique, établit que les données empiriques relatives aux fréquences des mots dans un texte peuvent être représentées selon la loi construite comme ci-après.

Les mots issus d'un texte de longueur N sont rangés dans l'ordre de leur fréquence d'apparition (BARBUT, 1989). Alors, la fréquence $f(r)$ du mot de rang r suit approximativement la forme de l'équation :

$$f(r) = \frac{K}{r^B} \text{ avec } B > 0 \tag{A.6}$$

Alors que l'on observe un bon ajustement sur des données empiriques pour les rangs suffisamment élevés, cette loi s'ajuste non seulement très mal sur les premiers rangs (les mots de fréquence d'apparition très faible) mais se révèle seulement « théorique », c'est à dire applicable grossièrement à des grands ensembles de texte. La variété des textes non seulement par leur nombre de mots différents mais également sur l'emploi de ces mots impliquent que cette loi ne donne que l'allure générale de la distribution statistique terminologique et ne peut servir comme outil de discrimination fin de textes. De nombreux travaux se sont attachés à affiner cette loi depuis introduisant des paramètres permettant d'ajuster notamment pour les hautes et basses fréquences (BOOTH, 1967 ; FEDOROWICZ, 1982 ; BROOKES, 1984).

Benoit MANDELBROT (1954) peu satisfait du principe du moindre effort comme explication de la loi de ZIPF, a repris en introduisant le concept d'information de SHANNON et WEAVER, définit comme une mesure de « l'élément de surprise, fructueuse ou non, qu'apporte en moyenne la réception de chaque nouveau mot » et non pas ce qu'apprend chaque mot en moyenne. C'est ainsi que l'auteur signale que le sens du terme « information » de SHANNON est celui de « variété-information » qui se pose par la forme exacte : si les unités T_r ont respectivement les probabilités p_r alors l'informa-

tion H est définie par

$$H = -\sum p_r \log p_r \tag{A.7}$$

La loi « canonique » de MANDELBROT va proposer d'introduire un paramètre supplémentaire ρ, une translation sur le rang qui permet d'améliorer l'ajustement de la loi A.1.2 aux données empiriques pour adopter la forme :

$$f(r) = \frac{K}{(r+\rho)^B} \text{ avec } B > 0 \tag{A.8}$$

Notons que si $\rho = 0$, on retrouve l'équation A.1.2

MANDELBROT définit B comme l'inverse de la « température du texte ». Notons que la preuve fournie par MANDELBROT s'appuie sur l'entropie de SHANNON, s'applique encore à « des textes suffisamment longs » pour lesquels l'unité élémentaire n'est pas définie : ce peut être un caractère ou un « message », suite de symboles discrets sur plusieurs niveaux hiérarchisés : phonèmes, mots, ou combinaisons de mots. La constante K est fonction du corpus d'étude et différente de celle de ZIPF et ESTOUB. Il convient de remarquer que lorsque $B = -1$, il s'agit de l'équation de BROOKES (BROOKES, 1984) qui ajuste la loi de ZIPF aux valeurs des hautes fréquences à l'aide du nombre de termes distincts de la distribution, loi de ZIPF que l'on retrouve (BAILÓN-MORENO, 2003, p. 111) en prenant $B = -1$ et $m = 0$.

En pratique, la pente de la distribution est donnée par une régression linaire et le paramètre de MANDELBROT par :

$$m = \left(\frac{K}{f(1)}\right)^{\frac{1}{|B|}} - 1 \tag{A.9}$$

L'ajustement de ces lois à des données empiriques n'est pas toujours direct (BARBUT, 1989), et il a été remarqué que celles-ci ne sont pas appropriées pour des niveaux de texte différents : RUIZ-BAÑOS remarque que le comportement statistique des descrip-

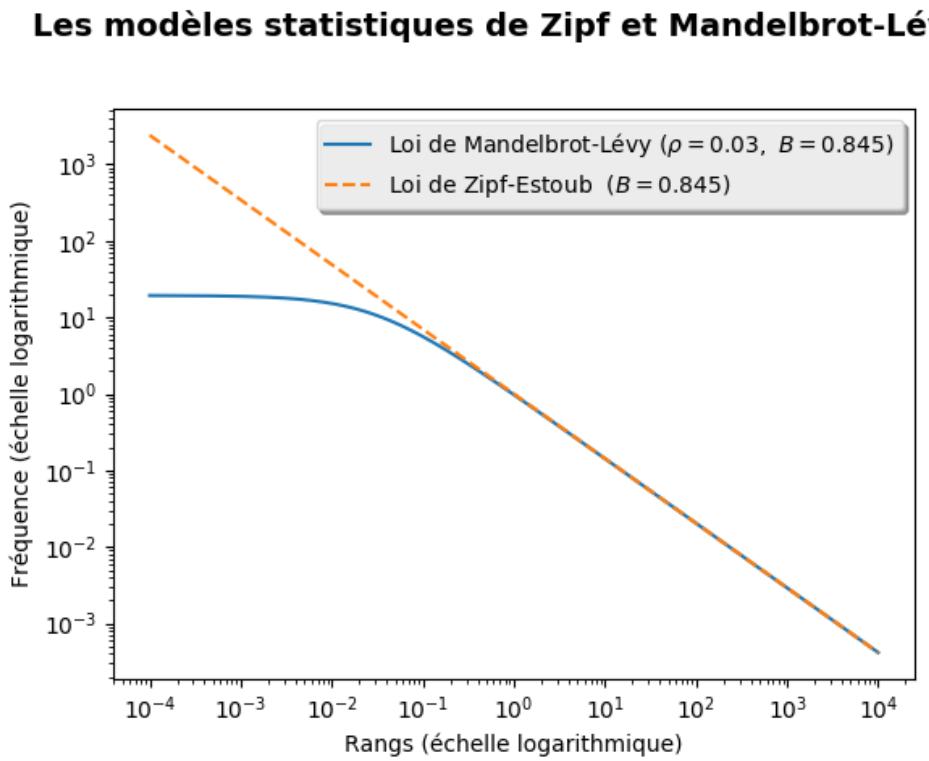


FIGURE A.1 : Représentation des modèles Zipf-Estoub et Mandelbrot-Lévy

teurs d'articles scientifiques issus de la base Francis entre 1980 et 1993 ne correspond pas à ces lois (RUIZ-BAÑOS, BAILÓN-MORENO, JIMENEZ-CONTRERAS et al., 1999 ; RUIZ-BAÑOS, BAILÓN-MORENO, JIMÉNEZ-CONTRERAS et al., 1999) notant des déviations très importantes entre le théorique et ce qui est observé. Les auteurs suggèrent alors un modèle fractal (BAILÓN-MORENO, JURADO-ALAMEDA et al., 2005b) en substitut de la loi de ZIPF [2] montrant que la distribution se divise en trois zones, chacune reproductible par un coefficient exponentiel négatif. La première zone représente les descripteurs principaux qui se situent majoritairement au centre du réseau de thèmes, réseau construit par la méthode des mots associés (BAILÓN-MORENO, JURADO-ALAMEDA et al., 2005a). La seconde zone est celle des descripteurs dits thématiques formant le réseau en relation avec le terme principal. Enfin, le reste des termes sont les descripteurs extra-thématiques qui font parti de l'ensemble du réseau hors thème concret.

2. Et plus généralement de scientométrie.

A.1.3 Un cas particulier des distributions Parétiennes

A.1.3.1 Lois de PARETO

Première loi se définit pour tout $x \leq a > 0$, la proportion $P(X)$ des mots de rang supérieur à x est :

$$P(x) = \left(\frac{a}{x}\right)^{\alpha} \tag{A.10}$$

où l'exposant α est positif et supérieur à 1 (en général). Le paramètre a s'interprète comme un seuil au-delà duquel la loi s'applique.

La seconde loi se définit pour tout $x \leq a > 0$, la proportion $P(X)$ des mots de rang supérieur à x est :

$$P(x) = \left(\frac{a+c}{x+c}\right)^{\alpha} \tag{A.11}$$

où l'exposant α est positif et supérieur à 1 (en général). Le paramètre c est une constante supérieure à $-a$.

A.1.3.2 Caractérisation des distribution parétiennes

Il est préférable de ne pas passer aux logarithmes pour effectuer un ajustement parétien (BARBUT, 1988) La distribution d'une variable est parétienne si et seulement si : pour chaque seuil x de fréquence des mots la fréquence moyenne des termes de rang supérieur à celui de x est égale à β fois $x + \mu$. Les valeurs de β et μ sont données par :

$$\begin{cases} \beta = \frac{\alpha}{\alpha - 1} \\ \mu = c(\beta - 1) \end{cases} \tag{A.12}$$

A.2 Représentation et classifications textuelles

La représentation classique des textes en vecteur (SALTON, WONG et YANG, 1975 ; SALTON et MCGILL, 1983) dans l'espace des mots qu'ils contiennent est désormais classique. La figure A.2 schématise la représentation la plus basique des Q documents $(D_1,\ldots,D_Q)$ dans l'espace vectoriel des mots $(M_1,\ldots,M_N)$ qu'ils contiennent en utilisant leurs occurrences (fonction Occ) :

$$
\begin{array}{cc}
 & \overbrace{}^{Mots} \\
 & \begin{array}{ccccc} M_1 & \cdots & M_i & \cdots & M_N \end{array} \\
\left.\begin{array}{c} \\ \\ \\ \\ \\ \end{array}\right\} Documents & \begin{array}{c} D_1 \\ \vdots \\ D_k \\ \vdots \\ D_Q \end{array} \left(\begin{array}{ccccc} Occ(M_1) & & \cdots & & Occ(M_N) \\ \vdots & & \vdots & & \vdots \\ & & Occ(M_i) & & \\ \vdots & & \vdots & & \vdots \\ Occ(M_1) & & \cdots & & Occ(M_N) \end{array}\right)
\end{array}
\tag{A.13}
$$

Les espaces vectoriels sont un modèle fréquent de représentation des documents considérés comme structurés ou pas (BARONI et LENCI, 2008). Sur cette représentation s'appuient nombre de techniques de reconnaissance, de résumé, de regroupements, et de classifications (LERTNATTEE et THEERAMUNKONG, 2006 ; ANTONIE et ZAIANE, 2002) supervisées ou pas qui dépassent le cadre de ce document.

Références de l'annexe A

ADAIR, W. C. (1955). « Citation indexes for scientific literature? » In : *American Documentation (Pre-1986)* 6.1, p. 31 (cf. p. 125, 2).

ANTONIE, M.-L. et O. ZAIANE (2002). « Text Document Categorization by Term Association ». In : IEEE Comput. Soc, p. 19–26. ISBN : 978-0-7695-1754-4. DOI : 10.1109/ICDM.2002.1183881. URL : http://ieeexplore.ieee.org/document/1183881/ (visité le 24/05/2017) (cf. p. 8).

AVRAMESCU, A. (1979). « Actuality and Obsolescence of Scientific Literature ». In : *Journal of the Association for Information Science and Technology* 30.5, p. 296–303 (cf. p. 2).

BAHL, L. R., F. JELINEK et R. L. MERCER (1983). « A Maximum Likelihood Approach to Continuous Speech Recognition ». In : *IEEE transactions on pattern analysis and machine intelligence* 2, p. 179–190 (cf. p. 3).

BAILÓN-MORENO, R., E. JURADO-ALAMEDA et al. (2005a). « Bibliometric Laws : Empirical Flaws of Fit ». In : *Scientometrics* 63.2, p. 209–229. ISSN : 0138-9130, 1588-2861. DOI : 10.1007/s11192-005-0211-5. URL : http://link.springer.com/10.1007/s11192-005-0211-5 (visité le 02/07/2017) (cf. p. 6).

BAILÓN-MORENO, R. (2003). « Ingeniería del conocimiento y vigilancia tecnológica aplicada a la investigación en el campo de los tensioactivos. Desarrollo de un modelo ciencimétrico unificado ». Thèse de doct. España : Departamento de Ingeniería Química, Universidad de Granada. 675 p. (cf. p. 159, 5).

BAILÓN-MORENO, R., E. JURADO-ALAMEDA et al. (2005b). « The Unified Scientometric Model. Fractality and Transfractality ». In : *Scientometrics* 63.2, p. 231–257 (cf. p. 159, 6).

BARBUT, M. (1988). « Des bons et des moins bons usages des distributions parétiennes en analyse des données ». In : *Histoire & Mesure* 3.1, p. 111–128. ISSN : 0982-1783. DOI : 10.3406/hism.1988.1296. URL : http://www.persee.fr/web/revues/home/prescript/article/hism_0982-1783_1988_num_3_1_1296 (visité le 23/05/2017) (cf. p. 2, 7).

BARBUT, M. (1989). « Note sur l'ajustement des distributions de Zipf-Mandelbrot en statistique textuelle ». In : *Histoire & Mesure* 4.1, p. 107–119. ISSN : 0982-1783. DOI : 10.3406/hism.1989.879. URL : http://www.persee.fr/web/revues/home/prescript/article/hism_0982-1783_1989_num_4_1_879 (visité le 20/05/2017) (cf. p. 2, 4, 5).

BARONI, M. et A. LENCI (2008). « Concepts and Properties in Word Spaces ». In : *Italian Journal of Linguistics* 20.1, p. 55–88. URL : http://www.wordspace.collocations.de/lib/exe/fetch.php/course:baroni_lenci_2008.pdf (cf. p. 160, 8).

BOOTH, A. D. (1967). « A "Law" of Occurrences for Words of Low Frequency ». In : *Information and control* 10.4, p. 386–393 (cf. p. 158, 4).

BRADFORD, S. C. (1934). « Sources of information on specific subjects ». In : *British Journal of Engineering* 137, p. 85–86 (cf. p. 2).

BROOKES, B. C. (1968). « The derivation and application of the bradford-zipf distribution ». In : *Journal of Documentation* 7, p. 299–335 (cf. p. 2).

BROOKES, B. C. (1984). « Ranking Techniques and the Empirical Log Law ». In : *Information processing & management* 20 (1-2), p. 37–46 (cf. p. 4, 5).

CALLON, M., J. P. COURTIAL et al. (1983). « From translation to problematic networks : an introduction to co-word analysis ». In : *Social Science Information* 22, p. 191–235 (cf. p. 2).

CALLON, M., J.-P. COURTIAL et H. PENAN (1993). *La scientométrie*. Paris : Presses Universitaires de France (cf. p. 126, 2).

CALLON, M., J.-P. COURTIAL et W. TURNER (1991). « La méthode leximappe : un outil pour l'analyse stratégique du développement scientifique et technique ». In : *Gestion de La Recherche. Nouveaux Problèmes, Nouveaux Outils*. Sous la dir. de D. VINCK. Bruxelles : De Boeck, p. 207–277 (cf. p. 147, 2).

De SOLLA PRICE, D. J. (1986). *Little Science, Big Science... and Beyond*. Columbia University Press New York (cf. p. 2).

EGGHE, L. et R. ROUSSEAU (1990). *Introduction to infometrics : quantitative methods in library, documentation and information science*. Amsterdam, New-York, Oxford, Tokyo : Elsevier Science Publisher (cf. p. 2).

FEDOROWICZ, J. (1982). « The Theoretical Foundation of Zipf's Law and Its Application to the Bibliographic Database Environment ». In : *Journal of the Association for Information Science and Technology* 33.5, p. 285–293 (cf. p. 4).

HUBERT, J. J. (1981). « General Bibliometric Models ». In : (cf. p. 2).

JELINEK, F. (1976). « Continuous Speech Recognition by Statistical Methods ». In : *Proceedings of the IEEE* 64.4, p. 532–556 (cf. p. 3).

KESSLER, M. M. (1963). « Bibliographic couplic beetween scientific papers ». In : *American Documentation* 14.1 (cf. p. 2).

LERTNATTEE, V. et T. THEERAMUNKONG (2006). « Class normalization in centroid-based text categorization ». In : *Inf. Sci.* 176.12, p. 1712–1738. ISSN : 0020-0255. DOI : 10.1016/j.ins.2005.05.010. URL : http://dx.doi.org/10.1016/j.ins.2005.05.010 (cf. p. 8).

LEYDESDORFF, L. et S. BENSMAN (2005). « Classification, powerlaws, and the logarithmic transformation ». In : URL : http://www.leydesdorff.net/log05/log05.pdf (cf. p. 2).

LHEN, J. et al. (1995). « La Statistique Des Lois de Zipf, Actes Du Colloque, Les Systèmes d'informations Élaborés ». In : *Les Systèmes d'information Élaborés*. Journée d'étude de la SFBA. Ile Rousse - Corse : Société Française de Bibliométrie Appliquée (cf. p. 158, 3).

LOTKA, A. (1926). « The Frequency Distribution of Scientific Productivity ». In : *Journal of the Washington Academy of Sciences* 16, p. 317–323 (cf. p. 2).

MANDELBROT, B. (1954). « Structure formelle des textes et communication : deux études paramétriques ». In : *Word* 10.1, p. 1–27 (cf. p. 4).

RUIZ-BAÑOS, R., R. BAILÓN-MORENO, E. JIMENEZ-CONTRERAS et al. (1999). « Structure and Dynamics of Scientific Networks. Part I : Fundamentals of the Quantitative Model of Translation ». In : *Scientometrics* 44.2, p. 217–234 (cf. p. 159, 160, 6).

RUIZ-BAÑOS, R., R. BAILÓN-MORENO, E. JIMÉNEZ-CONTRERAS et al. (1999). « Structure and Dynamics of Scientific Networks. Part II : The New Zipf's Law, the Clusters of Co-Citations and the Model of the Descriptor Presence ». In : *Scientometrics* 44.2, p. 235–265 (cf. p. 159, 160, 6).

SALTON, G., A. WONG et C. S. YANG (1975). « A vector space model for automatic indexing ». In : *Commun. Acm* 18.11, p. 613–620. ISSN : 0001-0782. DOI : http://doi.acm.org/10.1145/361219.361220 (cf. p. 8).

SALTON, G. et M. J. MCGILL (1983). *Introduction to modern information retrieval*. New York : McGraw-Hill (cf. p. 8).

SMALL, H. (1973). « Co-citations in the scientific literature : a new measure of the relationship beetween two documents ». In : *Journal of the American Society For Information Science* 24.4, p. 265–269 (cf. p. 2).

TURNER, W. A. et al. (1988). « Packaging Information for Peer Review : New Co-Word Analysis Techniques ». In : *Handbook of Quantitative Studies of Science and Technology*. Sous la dir. d'A. F. J. van RAAN. Elsevier (cf. p. 2).

WHITE, H. et B. GRIFFITH (1982). « Authors as markers of intellectual space : co-citation in studies of science, technology, and society ». In : *Journal of Documentation* 38.4, p. 255–272 (cf. p. 2).

WHITE, H. et K. MCCAIN (1989). « Bibliometrics ». In : *Annual Review of Information Science And Technology* 24, p. 119–186 (cf. p. 2).

ZIPF, G. K. (1949). *Human behavior and the principle of least effort : an introduction to human ecology*. Cambridge MA : Addison-Wesley (cf. p. 2).

B

ANNEXE

Paramétrage de collectes et de traitements

Le fichier de paramétrage ci-dessous est coloré syntaxiquement selon les conventions suivantes :

- En noir, les commentaires pour aider l'usager ;
- En bleu les paramètres à renseigner (requête, répertoire des données puis ensemble des commutateurs d'activation des scripts de traitement) ;
- Les lignes en vert séparent des parties dédiées des configurations spécifiques (collecteur, réseaux, etc.) ;
- En pourpre la description de l'action à réaliser ou description du commutateur ;
- En rouge les valeurs obligatoires de certains commutateurs (True, False).

La requête (ligne 5) et le répertoire des données pour les traitements (collecte et résultats) sont des éléments critiques. La plupart des autres commutateurs sont réservés à une utilisation avancée, la configuration initiale étant suffisante pour le bon fonctionnement de l'outil.

```
1 # Patent2Net configuration file
```

```
################################
# request
"insert below your request in cql format as done in the example"
request : ta=passiflor* or ta=""passion fruit*""
################################
# Directory to use
"insert a compliant name (no space or special characters) for
   the ""Patent Universe""(PU)"
DataDirectory : Passiflora
################################
#Patent2Net options  : Set to *True* or False
################################
# Collecting
"patents list corresponding to the request. e.g the "Patent
   Universe"(PU)"
GatherPatent : True
"patents bibliographic data corresponding to patent list"
GatherBiblio : True
"patent content (description, abstract, claims) completing
   patents bibliographic data"
GatherContent : True
OPSGatherContentsv2-Iramuteq : True
"patents families extending the PU to families"
GatherFamilly : False
################################
# Networks. Produce both online and for download and local use
   of Gephi
"Inventors' network "
InventorNetwork : True
"Applicants' network"
ApplicantNetwork : True
"IPCs' network"
CrossTechNetwork : True
"Applicants and Inventors network"
ApplicantInventorNet : True
"Applicants and IPCs' network"
ApplicantCrossTechNetwork : True
"Inventors and IPCs' network "
InventorCrossTechNetwork : True
"Countries and IPCs' network"
CountryCrossTechNetwork : True
"Network of references for the PU (each reference cited by each
   patent within the PU, includes external bibliographic refs) "
ReferencesNetwork : True
"Network of citation for the PU (each patent cited within the PU
    cited in the rest of the patents world) "
CitationsTechNetwork : True
"Network of equivalent patents for each patent within the PU"
EquivalentsTechNetwork : True
"The total network"
CompleteNetwork : True
"Complete families network"
FamiliesNetwork : True
FamiliesHierarchicNetwork : True
################################
" Mindmaps of IPCs both online and for download and local use of
    freeplane"
```

```
P2N-FreePlane : True
################################
"export patents as a bibliographic file in bibTex format"
FormateExportBiblio : True
################################
"Produce the world maps of inventors, applicants, deposit
   country"
FormateExportAttractivityCartography : True
FormateExportCountryCartography : True
################################
"Patents as a table"
FormateExportDataTable : True
FormateExportDataTableFamilies : True
################################
"Prepare interactive pivottable"
FormateExportPivotTable : True
################################
"prepare Abstracts, descriptions, claims for text analysis"
"for Carrot2 and Iramuteq"
FusionCarrot2 : True
FusionIramuteq2 : True
```

C

ANNEXE

Éléments d'interface de P2N

Cf. section 5.2.1 en p. 138 pour la mise en contexte de la figure C.1 de présentation de la page d'accueil d'une collecte P2N.

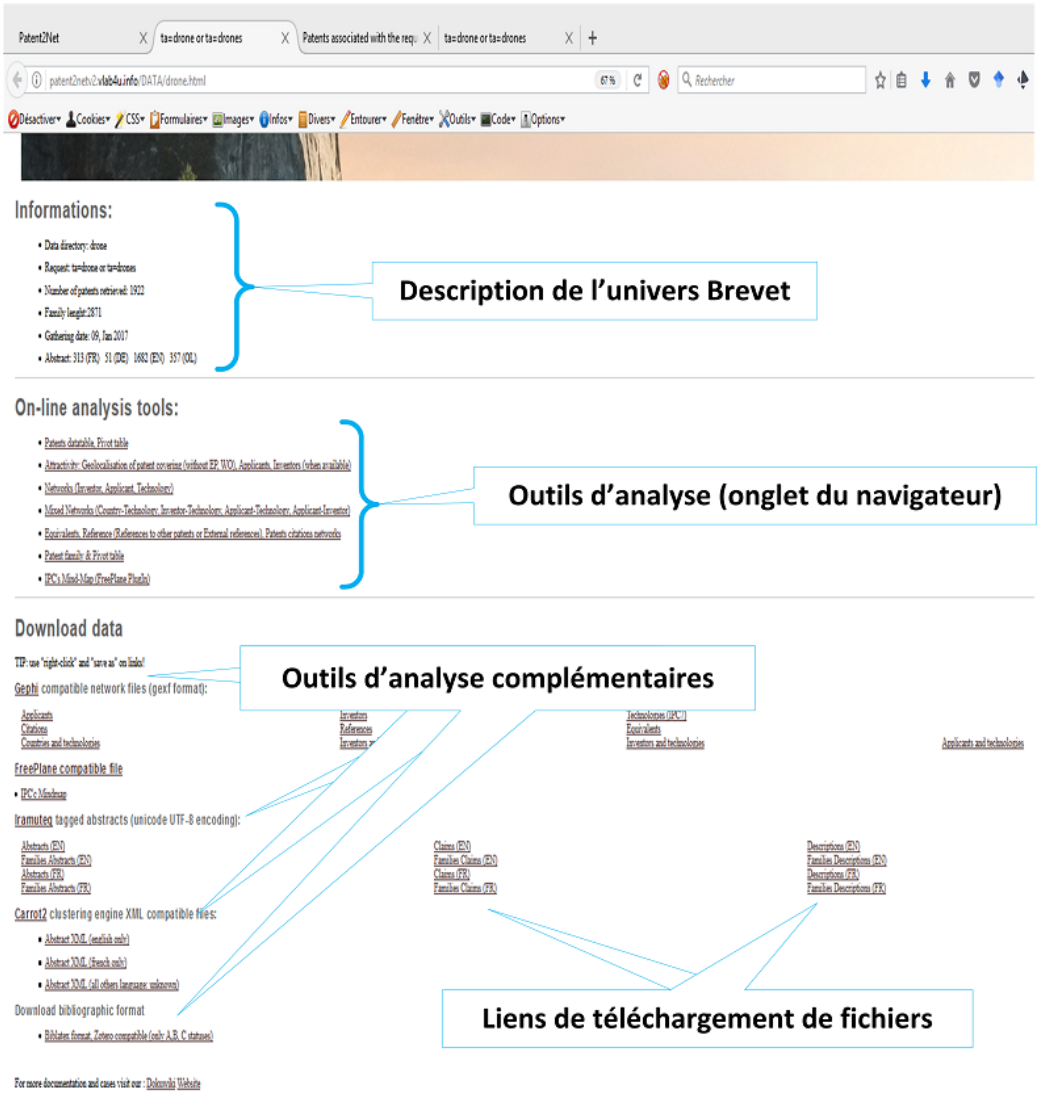


FIGURE C.1 : Description de la page d'accueil d'une collecte

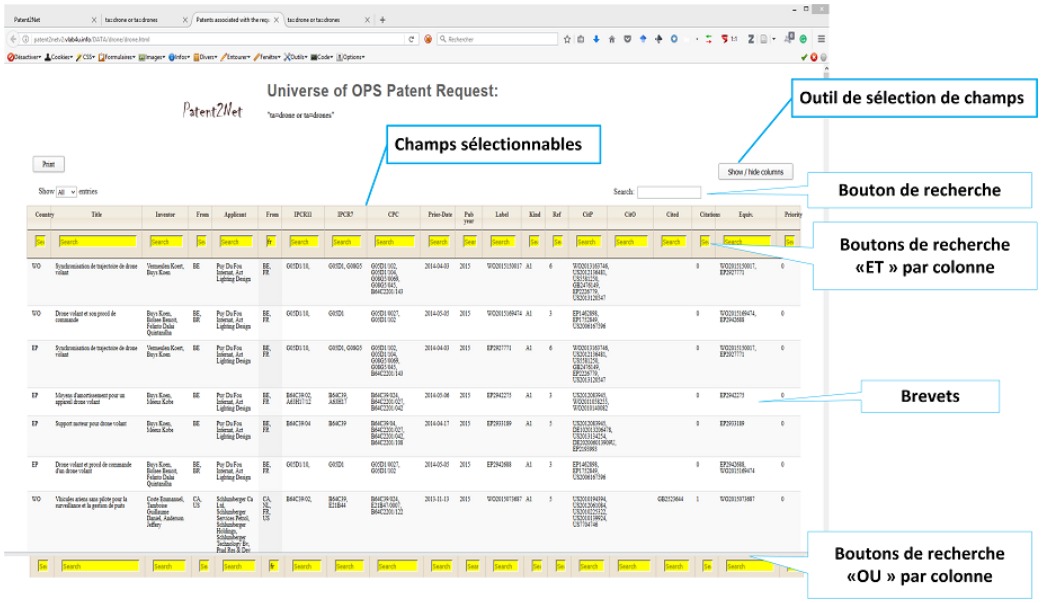


FIGURE C.2 : Description de l'outil de présentation des données en tableau

Cf. section 5.2.2.1 en p. 138 pour la mise en contexte de la figure C.2 de présentation de l'outil de présentation des données en tableau de P2N construit à partir de l'API datatable (*DataTable API* 2007).

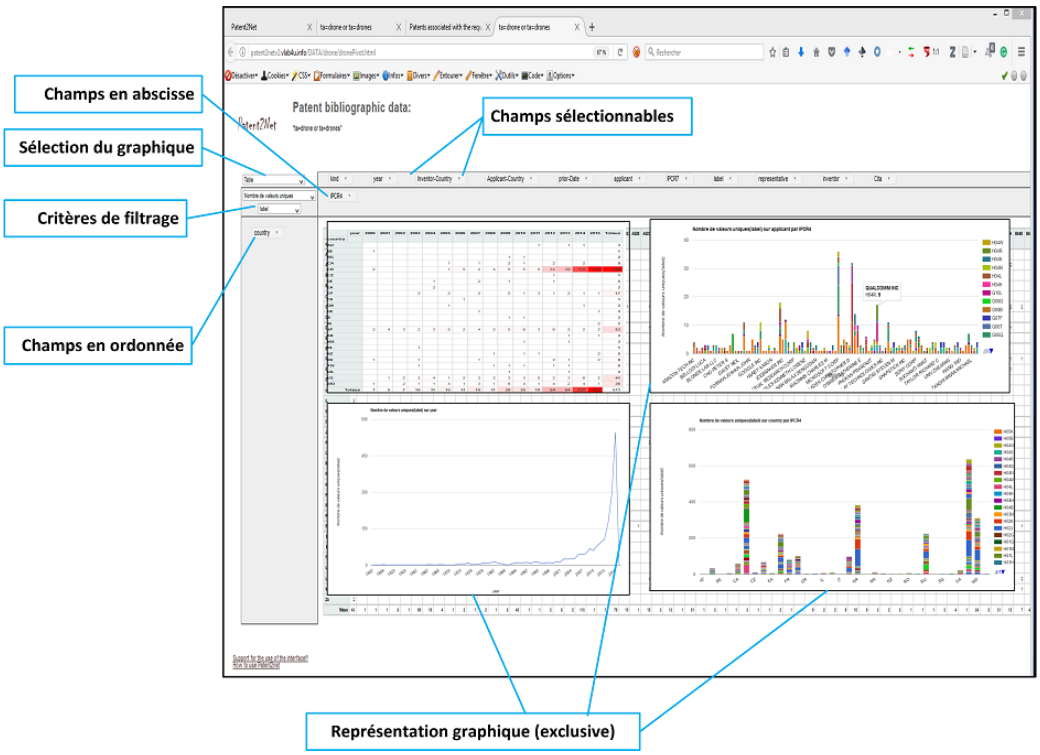


FIGURE C.3 : Description de l'outil de création de graphiques à partir des données

À partir des données descriptives des documents brevets, l'interface propose de construire des graphiques variés (tableaux, courbes, carte de chaleur, diagramme de Voronoï, etc.). La figure C.3 décrit l'interface de P2N construite sur l'API pivotable (KRUCHTEN, 2013).

D

ANNEXE

Mots importants et classification

Toute science est nécessairement schématique et approximative. Toute science est un compromis entre le souci de simplicité et le souci de ressemblance. Une grande simplicité est commode, mais elle risque de ne pas donner une image suffisamment ressemblante des faits ; une ressemblance trop poussée rend le modèle trop complexe et pratiquement inutilisable. Ce que l'on peut dire, c'est qu'à un niveau d'approximation donné, le modèle scientifiquement le meilleur est celui qui est le plus commode. En ce sens il y a autant de théories vraies que de degrés d'approximation donnés. Toute théorie n'est nécessairement qu'un certain compromis entre le souci d'être aussi voisin que possible de l'expérience et celui de ne pas devenir inutilisable à force de complexité au niveau donné de la synthèse où elle se place [...] Le propre de toute théorie scientifique, c'est de faire des hypothèses simplificatrices en éliminant toutes les circonstances accessoires et en ne gardant de la réalité que ses caractères essentiels. En fait une théorie est plus ou moins valable suivant que ce choix, en tout état de cause nécessaire, est plus ou moins bon.

— Alain ALLAIS, Puissance et dangers de l'utilisation de l'outil mathématique en économique. *Econometrica* (1954)

En phase avec l'épigraphe, montrant la relative simplicité des agencements techniques construits pour réaliser les expérimentations du chapitre 6 p. 155–178 des extraits de code de l'artefact élaboré sont présentés ici. Le code est en général commenté, la coloration syntaxique aidant à sa lecture (distinction des fonctions, des variables et de leur valeur).

Chargement des différentes bibliothèques. Le caractère # délimite les commentaires insérés dans le code.

```
from __future__ import print_function

import os
import sys
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.manifold import MDS
from sklearn.cluster import KMeans
from sklearn.metrics.pairwise import cosine_similarity
from nltk.corpus import stopwords as nltkSW
from TAL_P2N_Lib import tokenize_only, tokenize_and_stem
from P2N_Lib import LoadBiblioFile
from P2N_Config import LoadConfig

from collections import OrderedDict

import matplotlib
import matplotlib.pyplot as plt

import mpld3
import numpy as np
import logging
from sklearn.externals import joblib
from optparse import OptionParser
#from sklearn.feature_extraction.text import HashingVectorizer
#from sklearn.decomposition import TruncatedSVD
#from nltk.stem.snowball import SnowballStemmer
#from sklearn.pipeline import make_pipeline
```

Initialisation des variables et des paramètres de traitement.

Calcul de la zone de GOFFMAN (cf. 6.1.1 p. 157) et identification des termes afférents avec les notations des concepts définis en pages 3.

```

D0=len(set(word_freq_df['term'])) #unic forms of corpus
H0=np.log(D0) #
H1=-sum([pi* np.log(pi) for pi in FreqTrie.get_values().transpose()[2]])
    #pi means p_i in latex writing style ;-). Shannon entropy
D1 = np.exp(H1)
R = np.abs(H1/H0) #regularité
```

```
C=sum([ np.emath.power(pi,2) for pi in FreqTrie.get_values().transpose()
    [2]])
H2 = -np.log(C)
D2 =  1/C                          #trivial shortcut
Lt = H1-H2
Li = (H1+H2)/H1
Lb = H0-H1
Cb= D0 -D0*(1/R)*(Lb/(Lb+Lt+Li)) # noise shortcut
CuttingLeft=int(Cb) #Lhen, J, T Lafouge, Y Elskens, L Quoniam, et H Dou.
    « La statistique des lois de Zipf, actes du colloque, Les systèmes
    dinformations élaborés ». In Les systèmes dinformation élaborés. Ile
    Rousse - Corse: Société Française de Bibliométrie Appliquée, 1995.
CuttingRight=int(D2)
minDf = FreqTrie.get_values().transpose()[2][CuttingLeft]
maxDf = FreqTrie.get_values().transpose()[2][CuttingRight]
I1 = len( FreqTrie[FreqTrie['occurrences']==1]) # words of Occurence =1
rankGoffman  = int(round(np.sqrt(1+8+I1)-1)/2)
```

L'algorithme de classification (k-means++ (ARTHUR et VASSILVITSKII, 2007)) est globalement appliqué à l'ensemble du corpus pour opérer un traitement classique (tenant compte de l'intégralité des termes du corpus) sur la matrice de fréquence des termes (Tf) agencée sur l'inverse de la fréquence des termes par document (IdF), les mots vides sont extraits, et les termes (mots associés) sont développés au niveau 4. Le nombre de cluster a été préalablement fixé (arbitrairement jusqu'à présent) au nombre de classes IPC (8) +1 (pour éviter les nombres pairs).

```

vectorizer = TfidfVectorizer(stop_words=stopwords, use_idf=True,
    tokenizer=tokenize_only,
                             ngram_range=(1,4))
X = vectorizer.fit_transform(Contents)
km = KMeans(n_clusters=num_clusters, init="k-means++", max_iter=100,
    n_init=1,
                verbose=True)
km.fit(X)
```

Le résultat *km* sera utilisé pour implanter l'interface permettant de sélectionner les nœuds de chacune des classes élaborées.

```

vectorizer = TfidfVectorizer(stop_words=stopwords, use_idf=True,
    tokenizer=tokenize_only,
                             ngram_range=(1,4), vocabulary = DicoMaxiGood
    .keys())
PreX = vectorizer.fit_transform(Contents)
km2 = KMeans(n_clusters=num_clusters , init='k-means++', max_iter=100,
    n_init=1,
                verbose=True)
```

L'étape suivante extrait les termes de la zone de GOFFMAN calculée en donnant préférence aux bi, tri et quadrigrammes s'il en est pour initialiser la classification (k-means++

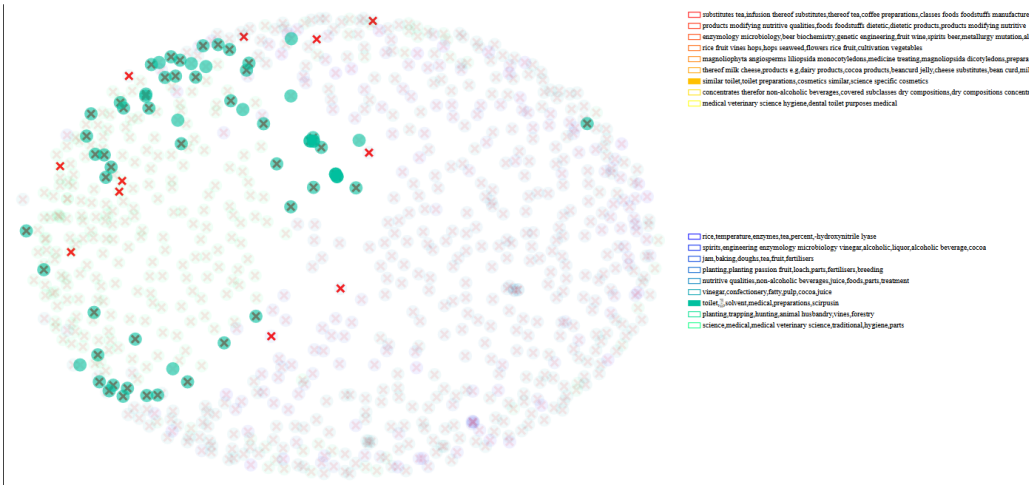


FIGURE D.1 : Interface d'exploration par double clustérisation. Cas de la Passiflora

(ARTHUR et VASSILVITSKII, 2007)) autour de ces termes. Le même algorithme que précédemment est appliqué mais en associant pour cette seconde étape les termes extraits de la zone de GOFFMAN.

Ce second niveau de classification sera utilisé par la seconde zone de contrôle de l'interface (en bas à droite) pour permettre à l'usager de combiner les deux voies d'exploration.

Index

Table des figures

Liste des tableaux

Table des matières