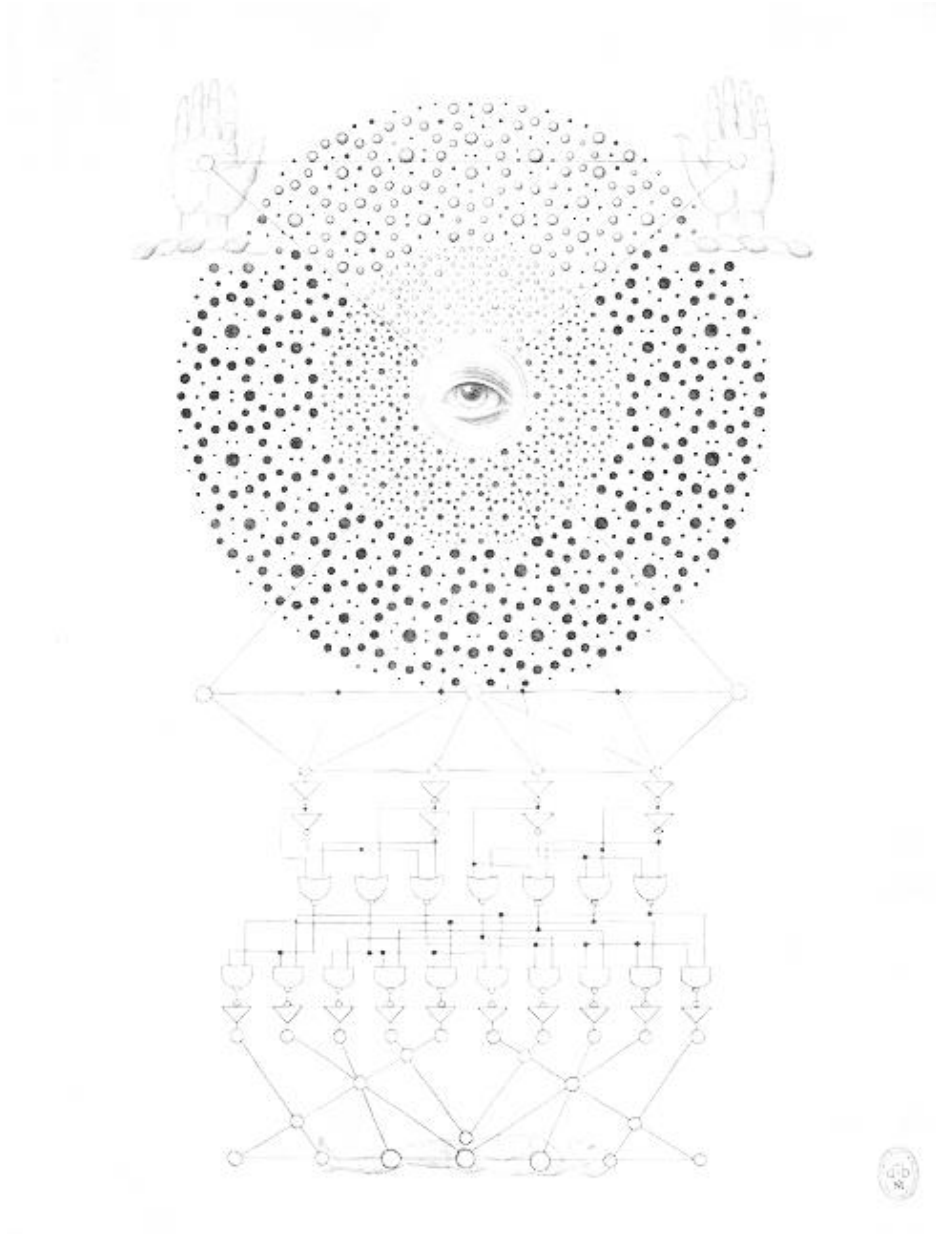


# Machinations numériques & économiques de la lecture



Célian RINGWALD  
M2 Humanité Numérique

*Note : Réalisé dans le cadre du Master Humanité Numérique, ce rapport répond de manière plus large qu'attendue aux demandes pédagogiques des cours de Recherches actuelles en Humanité Numérique et d'Édition. Cartographie critique de la machine de la lecture sur le web. Ce projet d'étude trait d'union, m'aura permis d'étudier une bibliographie utile dans le cadre de mes actuels intérêts de recherche, tout en utilisant des outils numériques de veille.*

*Crédit image : Daniel Martin Dlaz*

Je tiens à remercier mon grand-père Jean-Marie qui m'a fortement aidé dans lors des derniers phases de rédaction de ce rapport, à le rendre intelligible même aux yeux de ma grand-mère.

Je remercie aussi Damien Rose de Data Observer pour ces ultimes corrections.

## SOMMAIRE

Introduction.....	4
1. Un marché comptant plus de 6 millions de références.....	5
2. Filtres numériques du marché du livre .....	6
2.1. Edition du livre à l'ère numérique.....	7
A. <i>Booxby, premier acteur positionné sur la prédiction de Bestsellers</i> .....	7
B. <i>StoryFit et l'évaluation des scénarios</i> .....	8
C. <i>Qualifiction</i> , et son robot qui attribuait des prix littéraires .....	10
2.2. Editorialisation du contenu .....	11
A. <i>Amazon et le filtrage collaboratif</i> .....	11
B. Kadaxis, la clef de l'indexation.....	12
C. JellyBooks, des Ebooks gratuits mais pas que... ..	13
D. Inkitt, du numérique au livre papier.....	14
CONCLUSIONS, OUVERTURES .....	15
ANNEXE A. Phénomène étalés : Recherche actuelle en humanités numériques.....	17
ANNEXE B. Fiche de lecture : théorie économique des singularités et de ses dispositifs de jugements .....	19
ANNEXE C. METHODOLOGIE DE VEILLE.....	20
BIBLIOGRAPHIE .....	23

## Introduction

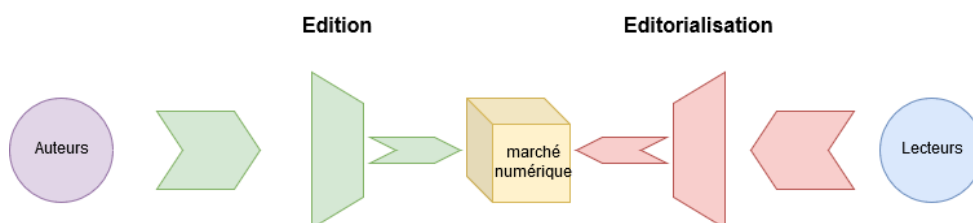
L'économie du livre a été cette dernière décennie bouleversée par la démocratisation de pratiques numériques modifiant nos habitudes d'écriture et de lecture. L'augmentation du nombre de lecteurs potentiels a été permise grâce à la hausse de l'alphabétisation à travers la scolarisation dans les pays développés. L'industrialisation de l'imprimerie a permis à des maisons d'édition d'asseoir une place dominante durant le siècle dernier<sup>1</sup>.

Cependant le « choc informatique » en démocratisant la circulation des connaissances et des savoirs a totalement bouleversé ce marché, les pratiques de lectures, de production d'informations et de diffusion de celles-ci. Dans une société de l'information informatisée nous lisons donc aujourd'hui beaucoup plus dans des contextes très larges (loisirs, travail, communication) et ceci à travers des supports de plus en plus détachés de leurs dimensions physiques<sup>2</sup>. Ces nouveaux paradigmes offrent d'autant plus qu'elles demandent aux lecteurs de s'en approprier les codes : nouvelles perspectives graphiques, dynamiques, lecture délinéarisée....

L'économie du livre est un exemple concret de marché de biens singuliers, comme l'a décrit Lucien Karpik en 2007 (Une fiche de lecture synthétique et jointe en Annexe à ce rapport). Les livres ont une valeur multidimensionnelle et possèdent une forte identité culturelle, ils sont donc pour ses acteurs (consommateurs, comme producteurs et revendeurs) sujets à une grande incertitude. Cependant l'informatisation de celui-ci invite à étendre cette théorie.

Théorisée sous la forme du « phénomène de la longue traine » en 2004 par Chris Anderson, cette tendance d'ouverture et de diffusion d'une offre de contenus toujours plus grande permise par le numérique a contraint les acteurs du marché du livre à construire de nouveaux outils. Ce concept a d'ailleurs donné lieu à une présentation de Jean-Philippe Magué dans le cadre du séminaire de M2 humanité numérique nommé Recherche actuelle. Un compte rendu de celle-ci est disponible en Annexe, cette note est d'ailleurs adaptée au marché de la lecture, complémentaire à ce travail.

Nous n'aborderons pas dans ce rapport les aspects anthropologiques, culturels, sociaux, ou encore historiques de ce grand bouleversement, bien que la somme de tous ces points de vue pourrait expliquer plus en détails ce qui est ici évoqué. Basé sur un corpus d'articles récoltés sur internet, par les outils de DataObserver ainsi que des sources scientifiques, cet état des lieux se penchera dans un premier temps sur l'étude des solutions mises en place en termes d'édition et d'éditorialisation de ces biens culturels par les acteurs de la Booktech, entreprises innovantes proposant de révolutionner le marché du livre.



<sup>1</sup> MOLLIER Jean-Yves, *Une autre histoire de l'édition française*, Paris, 2015

<sup>2</sup> BELISLE Claire, *Lire dans un monde numérique*, Lyon, 2011

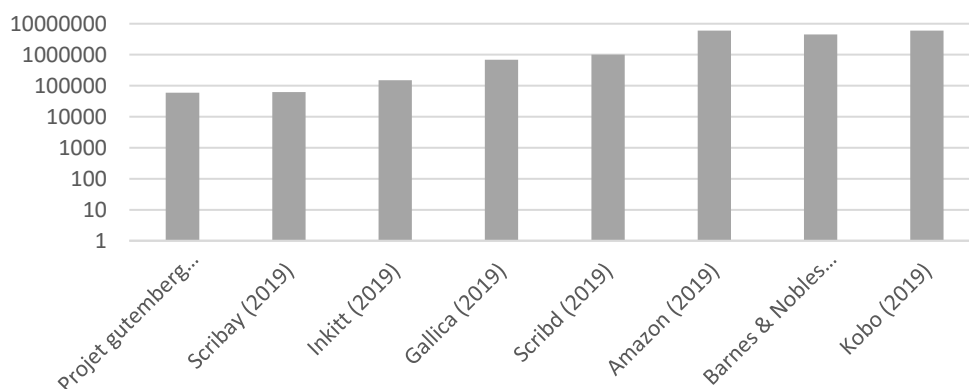
## 1. Un marché comptant plus de 6 millions de références...

Le marché numérique est aujourd'hui pris à parti par deux tendances. Premièrement une hausse du nombre de lecteurs donc de lecteurs potentiels transitant autour de gros acteurs de ventes en ligne. Deuxièmement, une hausse du nombre d'auteurs permise par une hausse des pratiques d'écriture, sociales et de loisirs, promues par les ateliers d'écriture<sup>3</sup> ainsi que la démocratisation du phénomène d'autoédition à travers internet. Cet aspect relance la question de la place de l'auteur : de la mort de celui-ci<sup>4</sup> à la naissance de son double numérique. C'est d'ailleurs tout le sujet du projet Lifranum lancé récemment par le laboratoire Marge<sup>5</sup>.

Mais commençons par aborder notre problématique en analysant le nombre de références disponibles à la lecture sur différentes plateformes. Le nombre de références mises sur le marché de la lecture (qu'elles soient sous forme numérique ou papier) ne cesse d'augmenter. A titre de référence plus de 565 millions de nouvelles ont été écrites par près de 4 millions d'auteurs et sont disponibles sur le réseau de micro-writing : Wattpad. La lecture d'un roman de 160 pages prend près de 2h et demi de lecture<sup>6</sup>, je vous laisse donc imaginer le nombre de vies qu'il faudrait pour lire l'entièreté de ces nouvelles.

On remarque sur les quelques chiffres revendiqués par les différentes sociétés que la distribution du catalogue par « librairie numérique » semble aussi suivre une répartition hypergéométrique. Trois ordres de grandeurs se détachent ici : un premier groupe composé d'Amazon, Kobo (Fnac) ainsi que Barnes & Nobles dont le catalogue approche les 6 millions de références. A ce titre, Gallica et Scribd ont un catalogue comptant près de dix fois moins que le premier groupe. Enfin, nous avons Scribay, Inkit et le projet Gutenberg qui ont eu cent fois moins de références que les places commerciales d'Amazon, de la Fnac ou encore de Barnes et Nobles.

Number of ebooks (log scale)



<sup>3</sup> LAMY Christine, « Les ateliers d'écriture », *Le Journal des psychologues*, Revigny-sur-Ornain, 2009

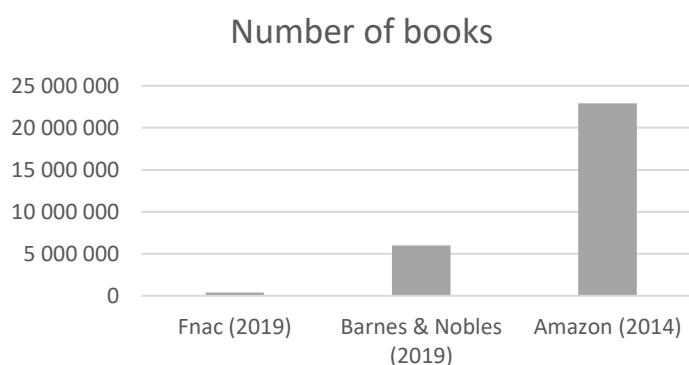
<sup>4</sup> BARTHES Roland, *The death and return of the author*, Edinburgh, 1998

<sup>5</sup> <https://marge.univ-lyon3.fr/lifranum>

<sup>6</sup> <https://howlongtoread.com/>

Cette répartition s'explique par le fait que ces catalogues ont été construits différemment : Wattpad est le résultat d'une communauté très prolifique. Les éditeurs et revendeurs historiques tels que la Fnac et Barnes & Nobles ainsi qu'Amazon, ont un catalogue beaucoup plus large.

L'on remarque que les démarches « ouvertes » telles que les projets Gutenberg et Gallica se défendent bien vis-à-vis des références mises à disposition du grand public. La différence entre ces deux démarches peut s'expliquer par le fait que le projet Gutenberg est l'œuvre de bénévoles, peu équipés, alors que Gallica agrège différents catalogues universitaires. Si le référencement de ebooks ne semble pas placer Amazon en situation de monopole ceci est vrai lorsque l'on parle du catalogue de ventes physiques à distance, puisqu'à travers son statut de revendeur elle agrège la production de plusieurs maisons d'éditions :



Ces catalogues gigantesques de références numériques disponibles ont comme l'explique Olivier Donnat<sup>7</sup> provoqué une guerre de l'attention, celle-ci entraînant une surproduction et impliquant un véritable gaspillage comme l'ont montré récemment les scandales liés à la destruction d'inventaires chez Amazon<sup>8</sup>. C'est une des raisons pour lesquelles le marché a été obligé de se rationaliser. Les entreprises sur lesquelles nous allons nous pencher ont le point commun de toutes se positionner comme des systèmes de filtrage vis-à-vis de l'offre, mais aussi dans la manière dont celle-ci va être proposée.

## 2. Filtres numériques du marché du livre

L'écosystème de la Booktech est plutôt récent puisque né entre 2011 et 2017. Il tire parti de la numérisation de l'écriture et de la lecture afin d'optimiser le cycle de vente et de publication d'un ouvrage, de sa sélection à sa mise en vente. Proposant des systèmes prédictifs il joue un rôle clef vis à vis des éditeurs. En effet ces industries sont encore frileuses quant à l'investissement que peut représenter la mise en place d'outils informatiques perfectionnés.

Nous dresserons ici un état des lieux des technologies décisionnelles, analytiques et prévisionnelles proposées par la communauté scientifique mais aussi par les entreprises que j'ai pu suivre lors de ma veille. Puis nous évoquerons les filtres numériques du marché du livre mis en place au niveau éditorial à travers les différents dispositifs de classement, de mise en avant de l'offre proposée par Amazon, Kadaxis, JellyBooks et Inkitt.

<sup>7</sup> Voir Annexe B

<sup>8</sup> [https://www.lemonde.fr/television-radio/article/2019/01/11/amazon-vendeur-de-destruction-massive\\_5407944\\_1655027.html](https://www.lemonde.fr/television-radio/article/2019/01/11/amazon-vendeur-de-destruction-massive_5407944_1655027.html), consulté le 15/12/2019

## 2.1. Edition du livre à l'ère numérique

L'étude du succès littéraire n'est pas nouvelle, mais la construction de systèmes informatiques permettant à partir du contenu d'un ouvrage de prédire ses ventes potentielles s'est largement développé durant ces 10 dernières années. En 2013, Vikas Ganjigunte Ashok, Song Feng et Yejin Choi<sup>9</sup>, ont étudié l'impact du style de 40 000 ouvrages disponibles en fonction de leurs succès. C'est à l'aide de l'extraction de variables telles que les mots (unigrams), groupes de mots composés de deux mots (bigrams), la distribution des étiquettes morphosyntaxiques à travers plusieurs niveaux de hiérarchies dans les textes, les sentiments évoqués dans ceux-ci que les chercheurs étudient alors la génétique des textes. Ils ont ensuite comparé ces descripteurs vis-à-vis du succès de leurs ouvrages, mesuré selon deux points de vue : l'opinion commune (nombre de ventes Amazon) et l'opinion experte (prix littéraires reçus). Ils remarquent à travers leurs études que certaines catégories grammaticales semblent permettre de discriminer fortement les ouvrages à succès des autres.

Cet aspect pour le moment déserté du champ universitaire et littéraire français, comme le souligne Alexandre Gefen<sup>10</sup>, a eu si l'on peut dire son succès dans le monde anglosaxon et a récemment encore donné lieu à de nombreuses publications scientifiques<sup>11</sup>.

Dans « Success in Books : A Big Data Approach to Bestsellers »<sup>12</sup> les auteurs étudient le cycle de vie du succès des œuvres listées par un important dispositif de cicérones américain : la « Best seller list » du New York Times (cette liste est éditée de manière hebdomadaire depuis 1942). Ils y comparent la réception de ces ouvrages dans le temps vis-à-vis de dispositifs de classements (les prix littéraires), vis-à-vis de leurs réceptions par le marché (le nombre de ventes enregistrées) mais aussi vis-à-vis de la carrière de leurs auteurs. Au Japon dans sa thèse Karl-Johan propose un modèle neuronal permettant de prédire les ventes à court terme d'un bestseller<sup>13</sup>.

En vue de l'incertitude et des frais que peuvent représenter la mise en marché d'un nouvel ouvrage, l'idée à fait « couler son ancre ». Ceci notamment chez les maisons d'édition qui pour faire face à la hausse de la production de récits semblent être en demande d'outils décisionnels leur permettant de faire le tri. Elles se sont alors rapprochées d'universitaires spécialisés dans le traitement automatique du langage. Archer and Jockers les auteurs du « Bestseller Code »<sup>14</sup> en sont l'exemple.

Comme le dit Elisabeth Mol journaliste aux Lettres Numériques : « Chaque année, les éditeurs reçoivent un très grand nombre de manuscrits, dont seulement 2 % seront publiés en moyenne, parmi lesquels 1 % deviendront des bestsellers »<sup>15</sup>, il y avait donc un marché pour ce type de service.

### A. Booxby, premier acteur positionné sur la prédiction de Bestsellers

<sup>9</sup> ASHOK, Vikas Ganjigunte, FENG Song, CHOI Yejin, « Success with Style: Using Writing Style to Predict the Success of Novels », Stony Brook, 2013

<sup>10</sup> GEFEN, Alexandre. « On ne sait jamais rien du sort d'un livre », *Revue critique de fiction française contemporaine*, Gand, 2017

<sup>11</sup> <https://www.theguardian.com/books/2014/jan/10/key-hit-books-discovered-highly-successful-literature>, consulté le 14/10/2019

<sup>12</sup> BURCU Yucesoy, WANG Xindi Wang, HUAND Junming, BARABASI Albert-László, « Success in Books: A Big Data Approach to Bestsellers ». *EPJ Data Science*, 2018

<sup>13</sup> ALM Karl-Johan. « Short Story Popularity Prediction Using Neural Networks with Time Series-Based Circular Dependencies », Saito, 2016

<sup>14</sup> ARCHER Jodie, JOKERS Matthew L., *The Bestseller Code: Anatomy of the Blockbuster Novel*, New York City, 2016

<sup>15</sup> <http://www.lettresnumeriques.be/2019/02/15/predire-les-chiffres-de-vente-grace-a-li-a/>, consulté le 25/10/2019

Site web: <http://www.booxby.com/>



*"Thanks to the increasing ease and accessibility of self-publishing, the number of books published each year in the United States has skyrocketed 400% in just the last 10 years."<sup>16</sup>*

Booxby est la première startup à s'être spécialisée dans l'analyse du potentiel d'un ouvrage à partir de ses caractéristiques linguistiques. Créée par Holly Payne (écrivaine, scénariste et éditrice), elle est basée à San Francisco. Financée par des investisseurs provenant du milieu de la tech : Mitch Lowe ayant travaillé notamment chez Netflix et Mark Bregman, impliqué dans plusieurs sociétés de gestion informatique de flux financiers. Elle travaille avec le National Science Foundation équivalent du CNRS aux Etats-Unis ainsi qu'avec Thamar Solorio de l'université de Houston. C'est d'ailleurs ce qui lui a permis de construire un système d'extraction de nouvelles structures linguistiques de texte basé sur 14 000 nouvelles. C'est sur cela qu'est basé le système décisionnel permettant de mesurer le potentiel d'un nouvel ouvrage que propose l'entreprise. Le système se base sur l'accès de lecture d'un ouvrage (readability), la phonologie de celui-ci, la distribution de ses étiquettes morphosyntaxiques. Il évalue ainsi la similarité d'un nouveau texte avec un catalogue d'ouvrage déjà connu ce qui lui permet d'évaluer le succès de celui-ci. Se présentant comme née du goût de la lecture, l'entreprise propose ses services aux auteurs, éditeurs ainsi qu'aux agents s'attaquant à la base du cycle d'édition d'un ouvrage. Elle compte parmi ses clients certaines antennes locales d'Hachette<sup>17</sup>, mais aussi d'Ingram Content<sup>18</sup> grand groupe d'édition et de distribution américain. Malgré la volonté d'afficher et de donner accès à leurs applications au lecteur, il demeure difficile pour lui d'avoir accès à celles-ci. J'ai pour ma part réalisé la démarche d'inscription mais n'ai jamais été recontacté.

**Résultats de veille :** 13 articles écrits en anglais dont la publication est située entre 2018 et 2019, provenant de sources universitaires, de magazines en ligne liés au monde technologique et relevant du registre de l'interview et de la communication.

#### *B. StoryFit et l'évaluation des scénarios*

Site web: <https://storyfit.com/>

<sup>16</sup> <http://www.independentpublisher.com/article.php?page=2212>, consulté le 29/11/2019

<sup>17</sup> [https://www.huffpost.com/entry/why-book-publishing-seeks-artificial-intelligence\\_b\\_59482841e4b0961faacbe5d6?ncid=engmodushpimg00000004](https://www.huffpost.com/entry/why-book-publishing-seeks-artificial-intelligence_b_59482841e4b0961faacbe5d6?ncid=engmodushpimg00000004), consulté le 09/12/2019

<sup>18</sup> <https://www.forbes.com/sites/johngreathouse/2019/07/27/this-entrepreneurs-journey-started-when-she-was-hit-by-a-drunk-driver-and-then-read-a-book/#1ac3f0bf1a90>, consulté le 15/12/2019





StoryFit est une startup créée en 2015, qui propose elle aussi une lecture assistée par automates afin de mesurer le succès potentiel d'un scénario pour sa publication et sa sortie cinéma. Cette entreprise créée par Monica Landers (ancienne productrice d'une chaîne d'information : ABC News) est financée par des gros fonds d'investissements<sup>19</sup>, impliquée notamment dans le financement de grosses plateformes de crowdfunding et d'entreprises technologiques et médiatiques américaines. Mark Bessen CTO de l'entreprise depuis 2017, était datascientist chez Apple et a notamment participé au projet d'écriture du "Bestseller Code".

Pour proposer ses services, StoryFit<sup>20</sup> a analysé de nombreux textes (scénarios, livres) sous de nombreux angles. La société évalue ainsi le contexte dans lequel se déroule ces histoires : géographique, thématique, référence à d'autres livres, films, la référence technologique, événements ayant lieu (poursuites, danses, rencontres...). Mais elle pousse aussi l'analyse des personnages : leurs noms, leurs activités, leurs milieux sociaux, leurs personnalités, leurs besoins ainsi que l'évolution de leurs émotions. Elle confronte ces informations aux données marketing disponibles afin de mesurer et maximiser la rentabilité des 12 premières semaines de sortie d'un potentiel film basé sur ce scénario.<sup>21</sup>

Bien que la communication de l'entreprise soit fortement ciblée sur la mesure du succès de scénarios, il semblerait qu'elle travaille pour le moment plutôt pour les éditeurs (près de 70 d'après cet article<sup>22</sup>). Elle propose ses solutions sous la forme d'applications web (ciblage de mots clefs utilisés pour la communication, comparaison d'un nouveau texte avec le corpus d'œuvres à succès, analyse des variables extraites par le logiciel), mais propose aussi des études sur mesure.

Résultats de veille : 48 articles, provenant principalement des réseaux sociaux, mais aussi de presse spécialisée pour entreprises, des journaux locaux et nationaux, tous publiés en 2019.

<sup>19</sup> <https://www.bdmifund.com/>, consulté le 17/10/2019

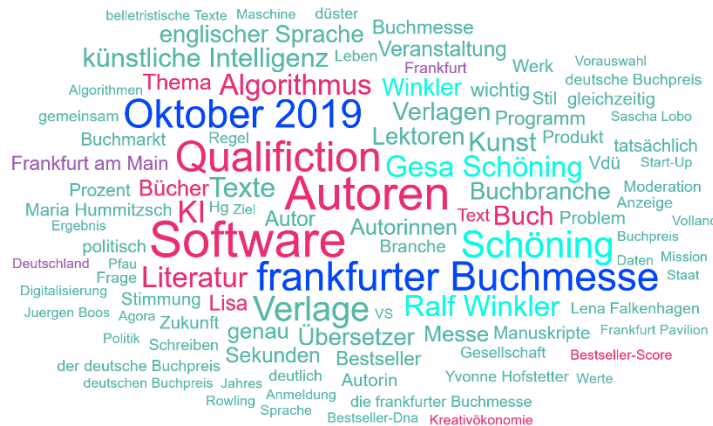
<sup>20</sup> <http://www.lettresnumeriques.be/2019/02/15/predire-les-chiffres-de-vente-grace-a-li-a/>, consulté le 18/10/2019

<sup>21</sup> <https://fr.slideshare.net/booknetcanada/how-can-ai-champion-great-storytelling-mary-brendza-storyfit-monica-landers-storyfit-tech-forum-2019>, consulté le 11/10/2019

22 <https://www.thebookseller.com/futurebook/storyfit-offers-ai-insights-drive-publishers-decisions-about-books-992766>, consulté le 12/11/2019

### C. *Qualifiction*, et son robot qui attribuait des prix littéraires

Site web: <https://www.qualifiction.info/>



Cette entreprise a été créée en 2017 à Hambourg par Gesa Schöning libraire et Ralf Winkler datascientist. Elle a été financée grâce à l'aide de fonds du ministère allemand de l'économie, et de la région. Dans la même voie que les précédentes startups, Qualifiction propose d'offrir un service de prédiction de succès littéraires. L'entreprise a eu droit à une grande couverture durant ma veille. En effet celle-ci a été primée cette année durant la foire du livre de Frankfort (plus grosse foire du livre au monde) d'un prix dédié aux entreprises innovantes. Mais elle a aussi créé l'évènement à travers la sélection d'une liste courte de BestSellers de demain<sup>23</sup>. Cinq livres ont pu ainsi être mis en avant à travers un corpus de 200 livres amateurs, et se sont vu promis d'être publiés par l'éditeur partenaire.

C'est en se basant sur un score fourni par LiSa (für Literaturscreening und Analytik) que ces ouvrages ont été primés. Ce « BestSellers score » se base<sup>24</sup> : sur l'évolution narrative de thématiques (400 pour le moment), des sentiments évoqués, le niveau de suspense offert par l'écriture, de la réflexion attendue du lecteur, de statistiques textuelles (longueurs des phrases, richesse du vocabulaire...), l'analyse des personnages (leurs interactions, leur sexe, leur façon de communiquer), la présence de scènes de la vie quotidienne. Mais elle tente de valoriser aussi l'innovation vis-à-vis du catalogue du même genre. Ces caractéristiques sont ensuite comparées au marché impliquant des variables telles que le prix, le format, le budget marketing envisagé ce qui permet alors à LiSa de fournir des chiffres en termes de prévisions de vente.<sup>25</sup>

Résultat de veille : 238 articles, dont 63 provenant des réseaux sociaux (tweets de communication suite aux événements), 86 provenant de la presse allemande et européenne réalisant un article sur l'IA et l'industrie éditoriale. Principalement en allemand.

<sup>23</sup> <https://www.buchmesse.de/veranstaltungen/doppelte-premiere-ki-verlag-qualifiction-praemieren-bestseller-von-morgen>, consulté le 05/12/2019

<sup>24</sup> <https://www.slideshare.net/idalab/data-science-meets-art-predicting-bestselling-books>, consulté le 15/10/2019

<sup>25</sup> <https://www.livreshebdo.fr/article/qualifiction-analyse-le-potentiel-commercial-des-manuscripts>, consulté le 05/10/2019

## 2.2. Editorialisation du contenu

Les techniques d'éditorialisation de contenus aujourd'hui sont présentes partout sur le net, que ce soit à travers les moteurs de recherche (classement des résultats en fonction d'une requête) mais aussi au sein de plateformes numériques commerciales et sociales. Elles mettent en avant des produits vis-à-vis d'autres en essayant de se baser sur le parcours de leurs utilisateurs. Nous introduirons cette partie à travers l'exemple significatif d'Amazon et d'autres plateformes proposant des contenus culturels. Puis présenterons trois entreprises de la Booktech qui proposent des services liés à ces problématiques.

### *A. Amazon et le filtrage collaboratif*

Comment parler éditorialisation de contenus sans passer par la case Amazon ? Bien que son catalogue soit accessible à travers une typologie de produits ainsi que d'une entrée de recherche, l'entreprise met en avant, de manière dynamique, des produits à l'internaute lors de sa navigation.

C'est en 1998 que l'entreprise a mis en place un algorithme permettant de réaliser ces propositions, décrit en 2001 celui-ci est appelé filtrage collaboratif. Il se base sur les paniers d'achats des consommateurs similaires à un profil donné, dans le but de lui proposer ce que d'autres acheteurs ont regardé ou acheté. C'est principalement sur ce système que se sont construits les modèles de propositions des autres plateformes du web telles que Netflix, Youtube et bien d'autres.

Malgré l'efficacité globale de ce système, notamment au niveau de sa rapidité de réponse face à une nouvelle requête, celui-ci connaît quelques biais. Comme le soulignent Smith et Linden dans un second article réflexif publié en 2017<sup>26</sup>, celui-ci ne prend pas en compte la dimension temporelle des comportements de consommation. C'est notamment pour améliorer ce modèle que Netflix a lancé entre 2007 et 2009 le Netflix Prize<sup>27</sup>, qui a permis de rassembler des entreprises et chercheurs autour de la problématique. Ceux-ci ont proposé d'intégrer des combinaisons de modèles de machine learning faisant intervenir des régularisations mathématiques ainsi que des techniques de réduction de dimensions<sup>28</sup>, ou encore faisant intervenir la dimension temporelle des comportements de visionnage<sup>29</sup>.

Comme nous le voyons ces algorithmes, de plus en plus complexes, essaient à travers des méthodes statistiques d'optimiser la rentabilité en fonction du modèle économique de ces entreprises. Ils peuvent alors être focalisés sur la maximisation du taux de clique, le taux de conversion, le revenu généré ou encore l'engagement des utilisateurs, comme le présente très synthétiquement Dietmar Jannach<sup>30</sup>.

---

<sup>26</sup> SMITH Brent, LINDEN Greg, « Two Decades of Recommender Systems at Amazon.com », IEEE Computer Society, Chypre, 2017

<sup>27</sup> <https://www.netflixprize.com/>, consulté le 20/12/2019

<sup>28</sup> BELL Robert M, KOREN Yehuda, VOLINSKY Chris. « The BellKor Solution to the Netflix Prize », Florham Park, 2007

<sup>29</sup> BELL Robert M, KOREN Yehuda, VOLINSKY Chris. « The BellKor 2008 Solution to the Netflix Prize », Florham Park, 2008

<sup>30</sup> DIETMAR Jannach, JUGOYAC Michael, « Measuring the Business Value of Recommender Systems ». *ACM Transactions on Management Information Systems* 10, 2019

Cependant la mise en avant de certains produits, puisque automatiques, peut paraître redondante pour un utilisateur habitué. C’est pourquoi certaines entreprises comme Critéo<sup>31</sup> a investi la question de l’ennui estimé, notamment à travers des solutions algorithmiques auto-apprenantes permettant de proposer des propositions établies par des modèles concurrents en fonction de la routine estimée de l’utilisateur.

## B. Kadaxis, la clef de l'indexation

Site web : <https://kadaxis.com>



Créé en 2007 par Chris Sim, Kadaxis est une startup proposant des services de conseils permettant d'indexer des contenus textuels à travers le choix de mots clefs. Véritable lien entre le consommateur littéraire et l'offre disponible, les mots clefs sont la porte d'accès aux contenus de nombreux sites de e-commerce. Cependant c'est à l'auteur ou au revendeur de les définir lui-même. Ils sont essentiels car sans eux les consommateurs ne pourraient retrouver un livre donné à travers la barre de recherche.

L'entreprise a d'abord proposé des services gratuits, tels qu'une application destinée aux auteurs, permettant d'extraire des mots clefs sur la base d'un texte<sup>32</sup>, mais aussi une application permettant de faire des découvertes littéraires, destinées aux lecteurs<sup>33</sup>. Ces applications ont permis à l'entreprise d'obtenir du contenu afin d'affûter ces algorithmes mais aussi de gagner en notoriété.

Afin de pouvoir proposer des mots clefs, l'entreprise se base sur une extraction automatique de thématiques<sup>34</sup> qu'elle aligne avec le système de classification BISAC (typologie d'indexation d'ouvrages proposés par le BISG, groupe d'étude de l'industrie du livre américaine). Elle propose aussi des solutions permettant d'intégrer ce système à une plus large échelle dans des systèmes algorithmiques tiers. Matérialisé sous la forme d'API (application programming

<sup>31</sup> WARLOP Romain, LAZARIC Alessandro, MARY Jérémie « Fighting Boredom in Recommender Systems with Linear Reinforcement Learning », *Advances in Neural Information Processing Systems 31*, 2018

<sup>32</sup> <http://authorcheckpoint.com/>

<sup>33</sup> <http://bookdiscovery.com>

<sup>34</sup> BLEI David M., Y. Ng Andrew, JORDAN Micheal. « Latent Dirichlet Allocation », *Journal of Machine Learning Research* 3, 2003, Massachusetts.

interface) payantes, ces points d'accès prennent en entrée un contenu (un ouvrage en occurrence) et offrent en sortie une liste de mots clefs adaptés à l'indexation de celui-ci.

L'entreprise compte aujourd'hui parmi ses clients de nombreuses maisons d'édition anglosaxonnes mais aussi de nombreux groupes de presse, ainsi que des éditeurs scientifiques.

Résultats de veille : 318 articles, 295 provenant de Twitter, le reste étant partagé entre des blogs, des articles de presse ainsi que de retours clients.

### C. JellyBooks, des Ebooks gratuits mais pas que...

Site Web: <https://www.jellybooks.com>



“Data helps *shape* decisions. It does not *make* decisions”

JellyBooks est l'entreprise de notre focus qui est la plus ancienne. Créée en 2011, aujourd'hui elle est cotée en bourse. Son créateur Andrew Rhomberg est un habitué des salons occidentaux traitant de la publication numérique. Sa société a attiré un grand nombre d'utilisateurs en proposant de nombreux livres numériques gratuits, tout en offrant aux lecteurs de nouvelles idées de lecture. Ce système de recommandation a aujourd'hui été abandonné au profit d'un modèle économique basé sur l'analyse des comportements de lecture btenu avec le consentement du lecteur qui a en contrepartie accès à des livres électroniques gratuits. Ce système de tracking permis par la technologie ePub3 qui permet notamment d'intégrer du javascript. Ces livres numériques alors téléchargeables depuis la plateforme de JellyBook sont lisibles à travers diverses applications et liseuses.

Grace à ce système de suivis l'entreprise enregistre alors les dates de lecture, le système de lecture utilisé, les temps de lecture, les surlignages de textes, les clicks effectués ainsi que les changements de polices. Toutes ces statistiques sont disponibles pour le lecteur, mais sont aussi revendues aux éditeurs. Andrew Rhomberg est intervenu notamment aux assises du livre numérique qui a eu lieu en octobre 2019. Il y a abordé les problématiques liées à la guerre de l'attention mise en place par les différents acteurs du divertissement. En effet ces nouveaux modèles de lecture ont donné lieu à de nouveaux modèles de financement : comme le prix à la page tournée.

Résultats de veille : 152 articles, provenant essentiellement des réseaux sociaux, fortement co-cités avec les autres entreprises, souvent en lien avec des évènements. (Conférences, financements...)

#### D. Inkitt, du numérique au livre papier

Site Web: <https://www.inkitt.com/>



Inkitt est la seconde start-up allemande de notre focus. Cette dernière a monté son modèle sur une plateforme sociale de lecture comme Wattpad et reprend les concepts mis en place par JellyBooks réalisant des analyses comportementales de la lecture des textes amateurs disponibles sur la plateforme. Créé en 2014 par Ali Albazaz, cette entreprise a été fondée à travers le financement de divers fonds d'investissement notamment américains. Cet éco-système complet est une véritable compilation des différentes solutions présentées précédemment et permet de lier entre elles des données qui n'étaient jusque-là pas encore connectées : les données liées à la lecture, les données sociales liées aux interactions entre les membres (auteurs et lecteurs) mais aussi des données concernant le processus d'écriture de nouvelles. La startup propose ses services à des auteurs ayant vocation à être édités, alors que Wattpad semble plutôt orienté vers l'écriture récréative et visant un public plus adolescent.

Les auteurs, amenés à rassembler leur communauté sur la plateforme de lecture, peuvent s'ils ont accumulé assez de lectures (afin que l'entreprise ai assez de données comportementales) s'inscrire à des concours leurs permettant d'être publié par l'entreprise. L'entreprise édite alors les nouvelles sélectionnées sous format papier, mais aussi sous la forme d'une lecture immersive (lectures interactives avec animations). Bien entendu toutes ces métriques sont utilisées afin d'ordonner les résultats des recherches des lecteurs mais aussi leur proposer du contenu correspondant à leur profil. A ce titre l'entreprise se situe finalement à la frontière de l'édition et de l'éditorialisation.

Résultats de veille : 1002 articles, dont une majorité finalement liée à l'activité de promotion des auteurs des nouvelles mises en ligne.



## CONCLUSIONS, OUVERTURES

Nous avons pu à travers ce rapport décrire le paysage formé par des entreprises occidentales innovantes se proposant d'intégrer les outils numériques dans les problématiques d'édition et d'éditorialisation. Ces startups jouent le rôle de laboratoire d'expérimentation et sont soutenus financièrement par de grands acteurs déjà établis de la tech et des éditeurs, tout en mobilisant les chercheurs universitaires. Les technologies qu'elles proposent sont porteuses d'enjeux vis à vis de nos pratiques d'écriture et de lecture.

Ces systèmes analysent des caractéristiques de plus en plus complexes afin de découvrir ce qui fait le succès commercial d'un ouvrage : vocabulaire, style d'écriture, personnages, fil narratif mais aussi la phonétique. Elles offrent la possibilité pour les éditeurs de ne plus seulement baser leur choix sur une lecture sensible mais plutôt à travers une lecture distante. Le fait que les sociétés ici évoquées proposent aussi leurs services aux auteurs démontre l'importance de l'auto-édition.

Ces outils peuvent avoir pour effet de normaliser l'offre : en basant ces analyses sur les succès passés mais aussi à travers la mise à disposition de ceux-ci aux auteurs qui pourraient se baser sur ces systèmes pour « améliorer » la génétique de leurs textes avant soumission aux éditeurs. Ces systèmes questionnent aussi la notion de bulles de filtre mise en avant par Eli Pariser, d'autant que dans notre cas les machines façonnent non seulement la disposition des contenus mais aussi leurs compositions même.

Les dispositifs de jugements que Karpik mettait en avant en 2017, ont été dédoublés de leurs versions numériques : la remise de prix à travers une note décernée par un logiciel (exemple de LiSa, Qualifiction), les confluences à travers les systèmes de propositions dynamiques (systèmes d'Amazon), les systèmes de classement d'une réponse à une requête donnée, les réseaux parfois directement intégrés dans le processus d'écriture (exemple d'Inkitt), enfin nous pouvons aussi voir les systèmes d'aide à la décision comme des cicérones informatiques.

Les exemples de JellyBooks et Inkitt nous ont révélé la présence de plus en plus courante de système de tracking (suivis de lecture) sur les applications de lecture, qu'elles soient utilisées à travers une liseuse ou un navigateur web. Ceci remet alors en cause la lecture qui était jusqu'alors envisagée *in silentio*, car celle-ci laisse des traces qui sont par la suite exploitées par des acteurs commerciaux.

Nous avons pu voir que les données permettant d'alimenter ces systèmes sont aussi de nature très différente : données littéraires (le contenu des livres eux-mêmes), données commerciales (d'achats de livres), données sociales (issues des réseaux sociaux) ainsi que des données comportementales (sur l'acte de lecture). Autrefois réparties entre différents acteurs ces informations tendent à être échangées entre ceux-ci comme le proposaient par exemple Boim, Rubi et Al. dans leur présentation « Methods for Boosting Recommender Systems »<sup>35</sup>. Cette volonté d'agréger les informations est d'ailleurs illustrée par Inkitt qui à travers son

---

<sup>35</sup> RUBI Boim, MILA Tova, « Methods for Boosting Recommender Systems ». *IEEE 27th International Conference on Data Engineering Workshops*, Hannover 2011

produit capte des données sociales, littéraires mais aussi de lecture. Les entreprises de booktech pourraient donc devenir le lieu de convergence et d'agrégation de ces informations.

Au niveau de l'édition, les démarches décrites sont très orientées vers la recherche de nouveaux « best-sellers », proposant des solutions diminuant la prise de risques pour les éditeurs en rationalisant le processus de mise sur le marché de nouvelles références. En tentant de maximiser le profit à court terme de ces « succès potentiels », c'est donc sur la partie de gauche de la longue traîne décrite par Anderson que ces entreprises se positionnent. Il est probable que ces entreprises se doivent de fournir des outils informatiques similaires dans les années futures afin de pouvoir résoudre la rationalisation de la partie de droite de cette longue traîne. Nous pouvons imaginer que les techniques utilisées en trading pour monter des portefeuilles d'investissement permettant d'optimiser une rentabilité à long terme pourraient être adaptées à cette problématique.

Ce travail m'aura permis d'introduire le sujet et laisse derrière lui de nombreuses questions : Comment se situe le marché français vis-à-vis de ces nouvelles technologies ? Quels modèles concurrents ont été adoptés par la société de l'information Asiatique ? Comment évaluer l'impact réel de ces technologies sur nos pratiques de lecture ?



## ANNEXE A. Phénomène étalés : Recherche actuelle en humanités numériques

La production de savoir est exponentielle, et ne cesse d'augmenter depuis la naissance de l'écriture, il semblerait que ce constat ne soit pas caractéristique de notre époque, puisque le nombre de livres produits a déjà suivi cette dynamique au moment du développement de l'imprimerie, augmentant déjà à l'époque une explosion de sa diffusion et de sa démocratisation. Il semblerait que le phénomène du « Big data » ne serait que la révélation d'un phénomène sans échelle que nous saisissons qu'à travers la lecture offerte par des outils nous permettant d'en apercevoir les contours. L'informatique est ici le microscope offrant l'opportunité de visualiser une vue sur ces phénomènes étalés, effets d'échelles ne pouvant être naturellement perceptible cognitivement.

M. Magué à travers sa présentation de recherche ouvre celle-ci par la présentation des effets de Longues traines popularisé par Chris Anderson, scientifique et journaliste américain qui fut en charge de la rédaction de Wired jusqu'en 2012. Cette notion a été décrite par un article du magazine publié, en 2004, l'auteur applique ici le phénomène déjà observé à travers les travaux de Pareto vis-à-vis de la distribution des revenus (20% de la population ayant 80% des revenus), de Yule au sujet des taxinomies animales, ou encore Zipf à travers la distribution de fréquence des mots d'un texte. Elle est notamment d'ailleurs très utilisée en SEO (Search Engine Optimisation)<sup>36</sup> afin d'optimiser le référencement de contenus à travers le choix de combinaisons de mots clefs. Cette technique permet de cibler des groupes de mots, ouvrant sur un marché moins concurrentiel.

Anderson rapproche le phénomène des longues traines au commerce en ligne qui permet notamment de proposer un catalogue très large ouvrant alors la rentabilité d'une présence sur des marchés de niches. Notamment parce que le consommateur s'est rendu compte que son goût pouvait l'éloigner des masses médias, qui n'offraient pas d'alternatives. Selon l'auteur c'est à cause du contexte physiquement limité des enseignes classiques qu'il fallait pouvoir assurer une place rentable à un produit dans un catalogue. La qualité semble aussi selon lui ne pas toujours justifier de sa diffusion, ceci pouvant s'expliquer par le manque d'adéquation avec un public local, c'est pourquoi les commerces numériques ouvrent leurs offres à ce qui peut aussi être vu comme un bide. Ceci permet alors parfois de redécouvrir une œuvre bien après sa création dans un contexte différent.

Il propose à travers cet article trois règles « essentielles » pour qui voudrait se lancer dans le marché : Premièrement **tout doit être disponible**, car il est toujours possible de trouver, voir retrouver son public, notamment à travers la réédition. Il propose ensuite de **faire baisser le prix d'achat de manière** drastique, justifiant que les œuvres ne sont actuellement plus consommées de la même manière qu'au paravent et que celle-ci se fait au détail (exemple du CD où l'on écoute quelques musiques seulement sur la totalité) et que les prix actuels ne traduisant plus le coût du cycle de diffusion actuel. D'autant que d'après lui les ventes de la longue traine ont déjà pu être amorties par le passé. Enfin il semblerait que le problème à résoudre serait de pouvoir **fournir des outils de recherches** permettant au consommateur de lui proposer des produits dépendant de ces goûts.

Il évoque aussi dans son article le téléchargement pirate des contenus culturels et ceux-ci ne peuvent avoir selon lui qu'un impact que très mineur sur le marché. En effet la qualité variable de ces contenus est très peu rentable à vue du travail de recherche nécessaire sur un réseau de sites peu fiables pour pouvoir les télécharger.

---

<sup>36</sup> <https://digitiz.fr/blog/longue-traine-seo/>

## Remise en cause et confirmation du phénomène sur le marché du livre

Le constat de Chris Anderson a été remis en cause par de nombreux auteurs dont Lionel Maurel<sup>37</sup>, le marché numérique aurait tendance à renforcer l'effet best-sellers, ne favorisant en contrepartie pas tant la diversité culturelle. Olivier Donnat dans son rapport : « Évolution de la diversité consommée sur le marché du livre, 2007-2016 » propose une analyse détaillée par secteur de la situation. Il évalue dans son rapport cette théorie vis-à-vis de la situation du marché français sur près de 10 ans.

Il remarque dans un premier temps une hausse considérable du nombre d'acheteurs en ligne dans le pays passant de 33% en 2007 à 60% en 2017. Cependant le marché du livre serait d'après lui peu touché par la chute des ventes physiques (mis à part les ventes de cartes et de dictionnaire), car le livre conservant son statut de livre-objet laisserait à la lecture numérique ce qui n'achète pas. D'ailleurs comme le remarque Claire Bélisle<sup>38</sup> les usages de lecture ont beau se démocratiser, les utilisateurs des TICS n'ont pas la sensation d'être face à une « vraie lecture » lorsqu'ils évoquent la lecture sur écran.

Sur le marché français, il semblerait que le volume global des ventes de livres ait légèrement baissé depuis 2007, mais la diversité des ventes (en nombre d'auteurs et en termes de références) semble avoir augmenté ces dernières années, ce qui a fait baisser le nombre moyen de vente par livre et par auteur. Cette variété est d'autant plus marquée pour les livres pour la jeunesse ainsi que pour la bande dessinée.

Cependant la diversité proposée augmente plus vite que celle consommée, l'offre est alors fortement renouvelée (il y a plus de références qui sortent du marché qu'il n'y en a qui entrent). Ce paradoxe semblerait d'après l'auteur être expliqué par la guerre de l'attention provoquée sur les points de vente physiques qui a été mis en place par les éditeurs afin de compenser les pertes impactées par le marché numérique.

Olivier Donnat souligne que le nombre de références vendues à très peu d'exemplaires a augmenté, notamment grâce à l'entrée de petites maisons d'édition ou de maisons d'édition indépendantes, proposant des ouvrages de niche-et dont la vente est permise grâce à internet.

Les grosses maisons d'édition ont réussi à suivre à cette dynamique, et n'ont pas perdu en chiffre d'affaires : elles ont su adapter le portefeuille d'offre en proposant elles aussi une plus grande diversité de références. Elles ont pu déléguer une grande partie du travail de prospection à l'auteur lui-même et celui de la vente à des revendeurs numériques.

Le phénomène de la longue traine serait finalement remarquable, car la diversité des ventes aurait bien augmenté diminuant les ventes des acteurs de la tranche moyenne sans avoir sur les ventes de best-sellers.

Cependant comment les consommateurs accèdent-ils à cette diversité ? Comment peut-il faire un choix à travers une offre grandissante de références ? Et de l'autre côté, comment l'industrie du livre trouve-t-elle des solutions afin de pouvoir sélectionner ce qui pourrait devenir un potentiel best sellers vis-à-vis d'un nombre toujours plus grandissant d'auteurs ?

---

37 Lionel Maurel - Le contrecoup le plus négatif de la guerre au partage et son véritable objectif : <https://scinfolex.com/2014/03/08/le-contrecoup-le-plus-negatif-de-la-guerre-au-partage-et-son-veritable-objectif/>

38 Lire dans un Monde Numérique – p 11

## ANNEXE B. Fiche de lecture : théorie économique des singularités et de ses dispositifs de jugements

Selon Karpik le marché est composé de biens et services que nous pourrions regrouper en trois groupes : les biens homogènes (les biens sont identiques), les biens différenciés (les biens sont identiques mais ne sont pas vendus de la même manière), ainsi que les biens singuliers. Ces derniers biens se font atypiques quant à la diversité du choix qu'ils offrent, et dont l'appréciation peut être sujet à des points de vue **multidimensionnels** et finalement **incommensurables**, les biens culturels entrent alors dans ce cadre.

Le prix dans ce contexte n'est alors pas garant de sa qualité, nous sommes donc pour le consommateur, comme pour le producteur face à une prise à un produit opaque.

L'**incertitude** semble d'autant plus caractériser ces biens, que nous entrevoyons comme culturels à travers ce dossier mais pouvant tout aussi bien être illustrés par le choix d'un médecin parmi un panel de professionnels de la santé pour une opération. Il n'est en effet pas possible de juger à posteriori de sa qualité et demande pour faire son choix de mobiliser beaucoup de connaissances.

L'auteur insiste sur le fait que ces biens ne peuvent exister que dans un marché équipé de dispositifs de jugement, auquel le consommateur délègue l'assemblage de supports de décisions lui permettant in-fine de faire un choix. Il existe selon Karpik cinq catégories de dispositifs :

- les appellations (labels)
- les cicérones (critique d'experts)
- les classements (prix, palmarès)
- les confluences (à travers de la médiation)
- ainsi que le réseaux (avis interpersonnels d'une communauté).

Vis-à-vis de ces dispositifs Karpik indique que le consommateur serait donc placé dans un contexte de coordination lui permettant de réagir une typologie de 6 registres :

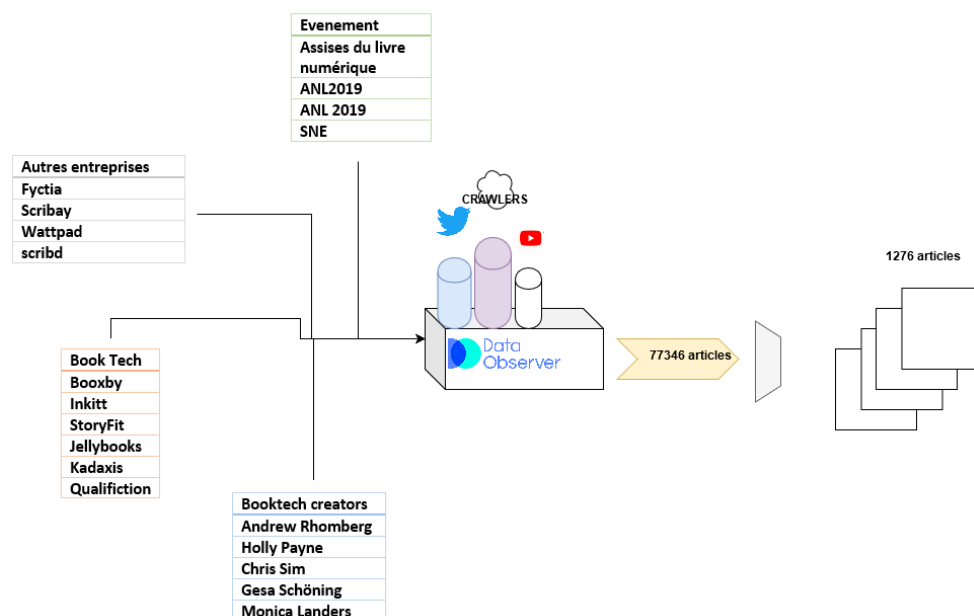
impersonnel	• l'authenticité (le produit est reconnu par label)
	• le « Méga » (la production du bien et telle qu'il n'est pas possible de passer à côté)
	• l'opinion experte (celui-ci est soutenu par un prix littéraire)
	• l'opinion commune (reconnu par l'audimat, le nombre de vente)
personnel	• la conviction (régime de croyance de l'acheteur)
	• le régime de coordination professionnelles (à travers son emplois)

**La valeur économique** de ces biens est lors fixée selon 3 facteurs symboliques : celle du *classement* (évoqué plus haut), qui aura une conséquence sur le prix (même si cela peut l'être de manière artificielle). Elle peut aussi être influencée par un effet de *palmarès* (le volume des ventes) qui en vue d'une politique de vente à l'échelle, le prix en général raisonnable n'y est ici pas forcément gage de qualité. Enfin le dernier facteur est celui de la *gradation*, il met en œuvre plusieurs dispositifs de jugement parfois contradictoire, sa valeur sur le marché est généralement neutralisée (partout pareil).

## ANNEXE C. METHODOLOGIE DE VEILLE

Ce dossier de veille a été réalisé en partie à travers l'étude d'un corpus de presse obtenu à l'aide des outils de Data Observer.

Afin de collecter les articles ayant de l'intérêt j'ai donc choisi de partir de la liste des noms de ce qu'un annuaire de startup appelait la TechBook<sup>39</sup>, mais aussi à travers la liste des entrepreneurs dirigeant ces sociétés afin de recueillir leurs propos ainsi qu'à travers le requêtage de noms d'entreprises pouvant être considérées comme faisant partie du contexte commercial de celles-ci. Enfin actualité oblige j'ai lancé une collecte concernant les assises du livre numérique qui avait invité plusieurs membres de cette communauté entrepreneuriale. Démarrée en octobre 2019 la collecte de données à durée deux mois et a permis de ramener un large panel d'articles, en tout 77 346 articles ont été récoltés, dont une grande part à été attribuée aux mots clefs liés à des plateformes engageant une large communauté notamment Wattpad et Scribd. Après curation j'ai pu obtenir une collection de 1276 articles compilant de nombreuses sources qui ont pu me servir pour construire mon enquête sur l'utilisation de nouveaux procédés de filtrage de l'offre de lecture.

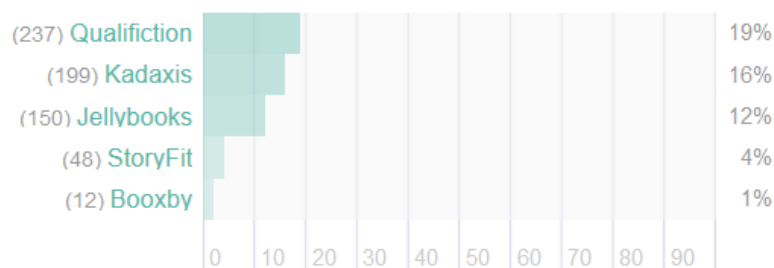


<sup>39</sup> <http://www.ai-startups.org/top/booktech/>, consulté le 01/10/2019

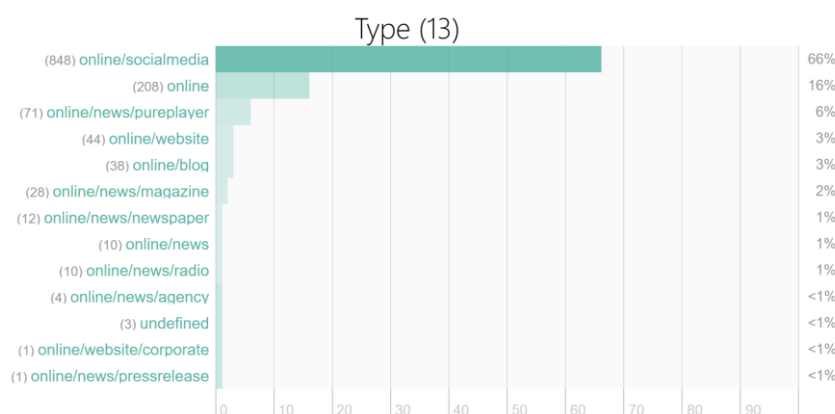
Ces articles ont été publiés entre 2016 et 2019, cependant la plupart de ceux-ci sont concentrés sur la période de collecte :



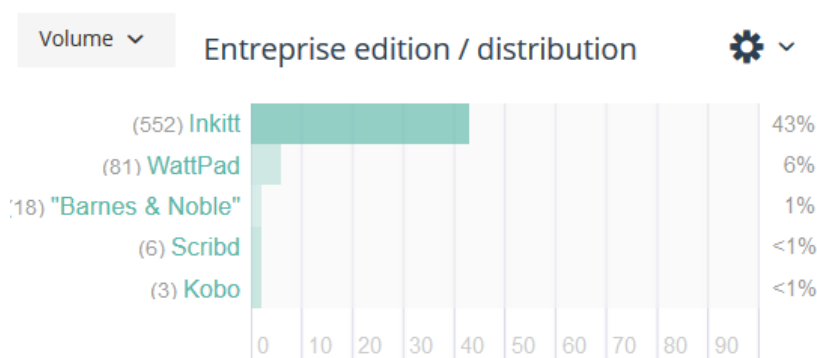
J'ai donc dû ajouter à la main des articles plus anciens me permettant de récupérer les communications antérieures à ma période de collecte. La plupart des sites vitrines de ces entreprises proposent une revue de presse, que j'ai pu intégrer. Cependant comme le montre la répartition du nombre d'articles par entreprise, Qualiction entreprise la plus récente de notre panel a obtenu le plus gros succès médiatique de notre corpus en volume, mais ceci dû à son intervention au salon du livre de Francfort qui s'est déroulé au début de notre collecte.



Un grand volume de notre corpus provient des réseaux sociaux, où la vélocité et la diffusion d'information est très importante, cependant notre collecte nous aura aussi permis d'obtenir 34% d'articles provenant de sources médiatiques mais aussi corporate.



L'on remarque qu'une grande partie du volume des publications s'explique par le fort engagement des communautés d'auteurs et de lecteurs d'Inkitt et Wattpad : 43% des publications de notre corpus.



Ce travail de collecte m'aura permis de construire un corpus de communication très utile mais cependant insuffisant pour mener mon étude, j'ai donc dû construire une bibliographie grâce à Zotéro me permettant de situer ces nouveaux acteurs vis-à-vis des travaux universitaires récents sur le sujet.

J'ai pu notamment grâce à ces sources faire remonter quelques chiffres :

	Amazon	Wattpad	Scribd	Projet Gutenberg	Fnac	Kobo	Gooreads	Inkitt	Gallica	Scribay	Barnes & Nobles
Number of books	22 900 000		500 000	600 000 000	400 000		280 16		500 000		6 000 000
Number of ebooks	1 200 000	565 000 000	60 000 000			6 000 000	2 600 000 000		3 000 000	700	4 500 000
Audience par jours	34 174 010	2 170 585	2 129 886	179 202				250 000	41 625		
Nombre utilisateurs	20 000 000	70 000 000	95 000 000			26 000 000	90 000 000	1 000 000		6 000	
Nombre auteurs		4 000 000		1 000			3000	70 000 000		400	
suivi de lecture	non	oui	oui	non	nsp	nsp	oui	oui	non	oui	non
suivis d'achat	oui	oui	oui	non	oui	oui	non	oui	non	oui	oui
ivis téléchargement	oui	oui	oui	oui	oui	oui	non	oui	oui	oui	oui

## BIBLIOGRAPHIE

## Source entreprise :

<https://www.crunchbase.com/>

<https://datacommons.technation.io/>

Traduction réalisée par la FING de l'article d'Anderson :

<http://www.internetactu.net/2005/04/12/la-longue-traine/>

**THE FATHER OF LONG TAILS, Interview with Benoît Mandelbrot, Hans Ulrich Obrist**

[https://www.edge.org/3rd\\_culture/obrist10/obrist10\\_index.html](https://www.edge.org/3rd_culture/obrist10/obrist10_index.html)

Azoulay, Nicole. « Lucien Karpik, L'économie des singularités, Nrf – Gallimard, 2007. »

*Revue de la régulation. Capitalisme, institutions, pouvoirs*, n° 3/4 (30 septembre 2008).

<http://journals.openedition.org/regulation/4853>.

Caron, Yves. « Du numérique dans l'édition, 2. Assises du livre numérique 2019 ». Billet. *Les Éditions Universitaires d'Avignon* (blog). Consulté le 11 novembre 2019.

<https://eua.hypotheses.org/2449>.

Favereau, Olivier. « L'économie des singularités, L. Karpik. Gallimard, Paris (2007). 373 p. » *Sociologie du travail* 51, n° Vol. 51-n° 2 (19 mai 2009): 320-22.

François-Ronan Dubois, « Jean-Yves Mollier, Une autre histoire de l'édition française », *Lectures* [En ligne], Les comptes rendus, 2015, mis en ligne le 04 novembre 2015, consulté le 17 novembre 2019. URL : <http://journals.openedition.org/lectures/19359>

Gefen, Alexandre. « “On ne sait jamais rien du sort d'un livre”. Théorie et pratique des best-sellers ». *Revue critique de fiction française contemporaine* 0, n° 15 (10 novembre 2017): 87-98.

Kroha, Petr, et Matthias Friedrich. « Comparison of Genetic Algorithms for Trading Strategies », 2014. [https://doi.org/10.1007/978-3-319-04298-5\\_34](https://doi.org/10.1007/978-3-319-04298-5_34).

Lamy, Christine. « Les ateliers d'écriture ». *Le Journal des psychologues* n° 272, n° 9 (2009): 36-39.

Alphabétisation, Rapport dvv - Éducation des Adultes et Développement 71/2008  
<https://fr.unesco.org/themes/alphabetisation-tous>

Filhon Alexandra, « De la culture écrite à l'illettrisme », *Langage et société*, 2014/3 (n° 149), p. 105-125. DOI : 10.3917/ls.149.0105. URL : <https://www.cairn.info/revue-langage-et-societe-2014-3-page-105.htm>

FIGHTING BOREDOM CRITEO : <https://papers.nips.cc/paper/7447-fighting-boredom-in-recommender-systems-with-linear-reinforcement-learning.pdf>

WATT PAD recommendation system: <https://fr.slideshare.net/MohammadIslam20/session-based-recommendations-at-wattpad>

WATTPAD RANKING : <https://support.wattpad.com/hc/en-us/articles/360000769623-Story-Ranking-FAQ>

INTERVIEW WATTPAD RANKING CREATOR : <https://www.wattpad.com/41422004-cracking-the-wattpad-code-insider-secrets-the-pros>

AMAZON : <https://www.linkedin.com/pulse/amazons-recommendation-engine-secret-sauce-mario-gavira/>

SCRIBD “Best sellers”: <https://www.scribd.com/bestsellers/Health-Fitness>

Number of amazon books : <https://www.quora.com/How-many-books-does-Amazon-have-for-sale>

Number of books on amazon : <https://www.quora.com/How-many-books-does-Amazon-have-for-sale>

Article sur scribay : <https://www.actualitte.com/article/lecture-numerique/scribay-reseau-social-et-terre-d-accueil-pour-tous-les-auteurs/55118>

Comment sont publié les livres sur inkitt : <https://www.bookworks.com/2016/05/algorithmic-publishing-by-inkitt-surfaces-best-sellers/>

Comment sont publié les livres sur inkitt : <https://www.quora.com/How-are-books-published-on-Inkitt>

Longue traine et amazon : <https://rogerpacker.com/amazon-boss-jeff-bezos-reveals-over-1000-indie-authors-earned-more-than-100k-through-kdp-in-2017/>

Histoire des technologies numérique de la lecture : <http://recherche.julie-blanc.fr/timeline-publishing/>

Les 5 nuances de gris de l’ édition numérique : <https://web.archive.org/web/20150320045905/https://www.digitalbookworld.com/2013/five-shades-of-book-discovery>

Afflerbach, Peter. « The Influence of Prior Knowledge and Text Genre on Readers’ Prediction Strategies ». *Journal of Reading Behavior* 22, no 2 (juin 1990): 131-48.  
<https://doi.org/10.1080/10862969009547700>.

Ashok, Vikas Ganjigunte, Song Feng, et Yejin Choi. « Success with Style: Using Writing Style to Predict the Success of Novels », s. d., 12.

Bell, Robert M, Yehuda Koren, et Chris Volinsky. « The BellKor 2008 Solution to the Netflix Prize », s. d., 21.

Bell, Robert M, Yehuda Koren, Chris Volinsky, et T Labs. « The BellKor Solution to the Netflix Prize », s. d., 15.

Blei, David M. « Latent Dirichlet Allocation », s. d., 30.



Boim, Rubi, et Tova Milo. « Methods for Boosting Recommender Systems ». In 2011 IEEE 27th

International Conference on Data Engineering Workshops, 288-91. Hannover, Germany: IEEE, 2011. <https://doi.org/10.1109/ICDEW.2011.5767667>.

BookNet Canada. « How Can AI Champion Great Storytelling? - Mary Brendza (Storyfit), M... ». Présentations et discours publics, 15:33:29 UTC.

<https://fr.slideshare.net/booknetcanada/how-can-ai-champion-great-storytelling-mary-brendza-storyfit-monica-landers-storyfit-tech-forum-2019>.

Bourreau, Marc, Sisley Maillard, et François Moreau. « Une analyse économique du phénomène de la longue traîne dans les industries culturelles ». *Revue française d'économie* Volume XXX, no 2 (23 novembre 2015): 179-216.

« Comprendre la valeur perçue de la customisation de masse. Une distinction entre la valeur du produit et la valeur de l'expérience de co-design - Aurélie Merle, Jean-Louis Chandon, Elyette Roux, 2008 ». Consulté le 27 octobre 2019.

<https://journals.sagepub.com/doi/abs/10.1177/076737010802300301>.

Dicerto, Sara. « Multimodal pragmatics : building a new model for source text analysis. », 2015.

Durand, Emmanuel. « Chapitre 6 - La culture est-elle soluble dans la data ? » *Nouveaux Debats*, 1 décembre 2016, 89-108.

Ganjigunte Ashok, Vikas, Song Feng, et Yejin Choi. « Success with Style: Using Writing Style to Predict the Success of Novels ». In *Proceedings of the 2013 Conference on Empirical*

*Methods in Natural Language Processing*, 1753–1764. Seattle, Washington, USA: Association for Computational Linguistics, 2013. <https://www.aclweb.org/anthology/D13-1181>.

Gefen, Alexandre. « “On ne sait jamais rien du sort d'un livre”. Théorie et pratique des best-sellers ». *Revue critique de fixxion française contemporaine* 0, no 15 (10 novembre 2017): 87-98.

Hedström, Peter. « Sociology. Experimental Macro Sociology: Predicting the next Best Seller ». *Science* (New York, N.Y.) 311, no 5762 (10 février 2006): 786-87.

<https://doi.org/10.1126/science.1124707>.

idalab GmbH. « idalab seminar #8 Data Science meets Art! - Predicting bestselling bo... ». *Data & Analytics*, 14:51:39 UTC. <https://www.slideshare.net/idalab/data-science-meets-art-predicting-bestselling-books>.

« Internet Search Result Probabilities: Heaps' Law and Word Associativity\* | Request PDF ». Consulté le 28 octobre 2019.

[https://www.researchgate.net/publication/220469421\\_Internet\\_Search\\_Result\\_Probabilities\\_Heaps'\\_Law\\_and\\_Word\\_Associativity](https://www.researchgate.net/publication/220469421_Internet_Search_Result_Probabilities_Heaps'_Law_and_Word_Associativity).

Jannach, Dietmar, et Michael Jugovac. « Measuring the Business Value of Recommender Systems ». *ACM Transactions on Management Information Systems* 10, no 4 (10 décembre 2019): 1-23. <https://doi.org/10.1145/3370082>.

Digital Trends. « Japanese A.I. Writes Novel, Passes First Round for Literary Prize », 23 mars 2016. <https://www.digitaltrends.com/cool-tech/japanese-ai-writes-novel-passes-first-round-nationnl-literary-prize/>.

Lary, David, Alexey N Nikitkov, et Dan N Stone. « Which Machine-Learning Models Best Predict Online Auction Seller Deception Risk? », s. d., 42.

Lebrun, Tom. « = Contrairement à la plupart des industries culturelles, le milieu du livre récolte ses données depuis longtemps », s. d., 16.

Linden, G., B. Smith, et J. York. « Amazon.Com Recommendations: Item-to-Item Collaborative Filtering ». IEEE Internet Computing 7, no 1 (janvier 2003): 76-80. <https://doi.org/10.1109/MIC.2003.1167344>.

Ménard, Marc. « Systèmes de recommandation de biens culturels ». Les Cahiers du numerique Vol. 10, no 1 (12 juin 2014): 69-94.

———. « Systèmes de recommandation de biens culturels ». Les Cahiers du numerique Vol. 10, no 1 (12 juin 2014): 69-94.

Mwinyi, Ibrahim Hussein, Husnu S Narman, Kuo-Chi Fang, et Wook-Sung Yoo. « Recommendation System Based on Predictive Approach », s. d., 7.

Park, Deuk Hee, et Hyea Kyeong Kim. « A Review and Classification of Recommender Systems Research », s. d., 5.

Payne, Holly Lynn. « A.I. Will Revolutionize Publishing (Not By Writing Books, But By Reading Them) ». Medium, 22 janvier 2018. <https://medium.com/@hollylynnpayne/a-i-will-revolutionize-publishing-not-by-writing-books-but-by-reading-them-f9a16ff888a2>.

Rodrigues, Fátima, et Bruno Ferreira. « Product Recommendation Based on Shared Customer's Behaviour ». Procedia Computer Science, International Conference on ENTERprise Information Systems/International Conference on Project MANagement/International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN / HCist 2016, 100 (1 janvier 2016): 136-46. <https://doi.org/10.1016/j.procs.2016.09.133>.

———. « Product Recommendation Based on Shared Customer's Behaviour ». Procedia Computer Science, International Conference on ENTERprise Information Systems/International Conference on Project MANagement/International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN / HCist 2016, 100 (1 janvier 2016): 136-46. <https://doi.org/10.1016/j.procs.2016.09.133>.

Schnabel, Tobias, Paul N. Bennett, et Thorsten Joachims. « Improving Recommender Systems Beyond the Algorithm ». ArXiv:1802.07578 [Cs], 21 février 2018. <http://arxiv.org/abs/1802.07578>.

Smith, Brent, et Greg Linden. « Two Decades of Recommender Systems at Amazon.Com ». IEEE Internet Computing 21, no 3 (mai 2017): 12-18. <https://doi.org/10.1109/MIC.2017.72>.

———. « Two Decades of Recommender Systems at Amazon.Com ». IEEE Internet Computing 21, no 3 (mai 2017): 12-18. <https://doi.org/10.1109/MIC.2017.72>.

Thurman, Neil, et Steve Schifferes. « THE FUTURE OF PERSONALIZATION AT NEWS WEBSITES: Lessons from a Longitudinal Study ». *Journalism Studies* 13, no 5-6 (octobre 2012): 775-90. <https://doi.org/10.1080/1461670X.2012.664341>.

Warlop, Romain, Alessandro Lazaric, et Jérémie Mary. « Fighting Boredom in Recommender Systems with Linear Reinforcement Learning », s. d., 12.

Wong, Joon Ian. « Amazon's Algorithms Are Misleading Customers and Causing Them to Spend Way More than They Should ». *Quartz*. Consulté le 14 décembre 2019. <https://qz.com/786300/amazon-buy-box-its-algorithms-are-misleading-customers-and-causing-them-to-spend-way-more-than-they-should/>.

Yucesoy, Burcu, Xindi Wang, Junming Huang, et Albert-László Barabási. « Success in Books: A Big Data Approach to Bestsellers ». *EPJ Data Science* 7, n° 1 (6 avril 2018): 7. <https://doi.org/10.1140/epjds/s13688-018-0135-y>.

Zanker, Markus, Laurens Rook, et Dietmar Jannach. « Measuring the Impact of Online Personalisation: Past, Present and Future ». *International Journal of Human-Computer Studies*, 50 years of the International Journal of Human-Computer Studies. Reflections on the past, present and future of human-centred technologies, 131 (1 novembre 2019): 160-68. <https://doi.org/10.1016/j.ijhcs.2019.06.006>.

Payne, Holly Lynn. « A.I. Will Revolutionize Publishing (Not By Writing Books, But By Reading Them) ». *Medium*, 22 janvier 2018. <https://medium.com/@hollylynnpayne/a-i-will-revolutionize-publishing-not-by-writing-books-but-by-reading-them-f9a16ff888a2>.

« Prédire les chiffres de vente grâce à l'I.A ? | Lettres Numériques ». Consulté le 9 décembre 2019. <http://www.lettresnumeriques.be/2019/02/15/predire-les-chiffres-de-vente-grace-a-li-a/>.

Wang, Xindi, Burcu Yucesoy, Onur Varol, Tina Eliassi-Rad, et Albert-László Barabási. « Success in Books: Predicting Book Sales before Publication ». *EPJ Data Science* 8, n° 1 (17 octobre 2019): 31. <https://doi.org/10.1140/epjds/s13688-019-0208-6>.

whitefox. « Q&A with Authors of "The Bestseller Code" Jodie Archer and Matthew Jockers ». *Medium*, 7 septembre 2016. <https://medium.com/@wearewhitefox/q-a-with-authors-of-the-bestseller-code-jodie-archer-and-matthew-jockers-67ed885fd150>.