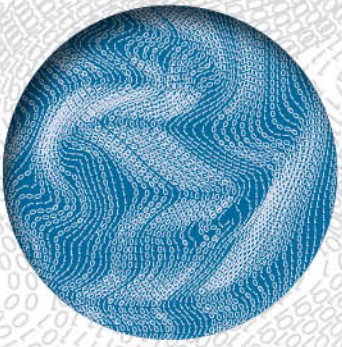
Les robots sont-ils des lecteurs comme les autres ?

Pierre-Carl Langlais

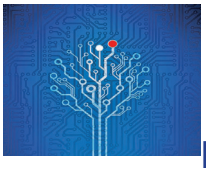
La diffusion numérique des données en SHS

Guide des bonnes pratiques éthiques et juridiques

sous la direction de
Véronique Ginouvès & Isabelle Gras



DIGITALES



DIGITALES

La diffusion numérique des données en SHS

Guide des bonnes pratiques éthiques et juridiques

sous la direction de

Véronique Ginouvès & Isabelle Gras

2018

Presses Universitaires de Provence

© Presses Universitaires de Provence
Aix-Marseille Université

29, avenue Robert-Schuman – F – 13621 Aix-en-Provence CEDEX 1
Tél. 33 (0)4 13 55 31 91

pup@univ-amu.fr – Catalogue complet sur presses-universitaires.univ-amu.fr

DIFFUSION LIBRAIRIES : AFPU DIFFUSION – DISTRIBUTION SODIS

Les robots sont-ils des lecteurs comme les autres ?

Émergence et codification d'une exception au droit d'auteur pour le *text & data mining*

Pierre-Carl Langlais
Université Paris IV- Sorbonne

Abstract: *The large-scale digitization of academic publications and patrimonial resources has recently stirred significant interests in text and data mining across scientific communities. Theses computing and statistical techniques allow to extract structured information or, more relevantly for the humanities, to map the textual characteristics (such as genres or intertextual relations) of very large corpora. Yet, the promises of text and data mining are constrained by legal uncertainties, mostly regarding intellectual property rights. Without copyright holders' agreement, the automated recopy of lawful sources and their treatment by the members of a scientific project are very likely illegal.*

This study aims to address the main legal stakes of mining projects in social science and the humanities and to recount the gradual transformation of informal claims into structured mobilization to implement exceptions. In several countries such as the United States, Japan or the Canada, the right to mine has become a de facto extension of the right to read thanks to pre-existing exceptions. In the European Union, this process requires explicit legal reforms, as the legal frame of the 2001 author rights directive proves too restrictive. Since 2014, three major European countries have passed a text mining exception, the United Kingdom, Germany and France, with the French version remaining for the moment a partly failed attempt. In parallel with theses national initiatives, an European-wide exception will likely be part of the currently debated European Authors Rights Reform. Theses legal evolutions not only helped to secure text and data mining activities in research but seems to have encourage the structuration of emerging scientific practices, as the enforcement of the exception requires to codification common norms and infrastructures.

« Le droit de lire est le droit d'extraire » (« *The right to read is the right to mine* »)

Depuis 2012, ce slogan a défini l'engagement d'institutions, de bibliothèques et de communautés en faveur d'une exception au droit d'auteur pour l'extraction de texte et de données (ou « *text & data mining* »). Au travers de cette mobilisation, un enjeu technique relativement méconnu en dehors des milieux scientifiques s'est

imposé dans l'agenda des politiques publiques. L'exception est entrée dans la loi d'abord au Royaume-Uni en 2014, en France en 2016 (dans le cadre de la loi pour une République numérique) et tout dernièrement en Allemagne. Elle constitue l'une des mesures les plus probables de la future réforme européenne du droit d'auteur.

Au regard de ce contexte, la portée de notre étude est double. Elle tente de clarifier les principaux enjeux de la pratique du *text & data mining* en sciences humaines et sociales ainsi que les principaux obstacles et insécurités juridiques auxquels font face les projets de recherche lorsqu'ils tentent d'explorer de grandes collections de textes ou de données. Dans un deuxième temps, nous retraçons le processus, toujours en cours, de légalisation du *text & data mining* : les exceptions introduites depuis quelques années en France et dans d'autres pays européens restent à ce jour des cadres expérimentaux imparfaits, destinés à évoluer ultérieurement.

Du droit de lire au droit d'extraire

L'assimilation du droit d'extraire au droit de lire a le mérite de souligner la continuité entre de nouvelles méthodes de recherche « automatisée » et des pratiques bien plus anciennes. L'extraction consiste ici usuellement à déléguer à des modèles statistiques et/ou à des séries d'instructions l'exécution de tâches jusqu'à présent effectuées manuellement par des chercheurs et/ou par du personnel associé (employé, étudiant, bibliothécaire, etc.).

Le travail d'investigation scientifique repose anciennement sur la constitution de « fiches » et de bases de données, soit sur le transfert systématique d'un savoir dispersé (Gardey 2008 : 151). Toutes ces pratiques sont tacitement conditionnées par plusieurs « droit de lire » et « droit de reprendre » : elles héritent d'un compromis ancien entre le public et les acteurs de la production éditoriale. Il est légal de reprendre des *faits* d'un texte préexistant à partir d'une source acquise licitement. Les éléments non originaux ne relèvent en effet généralement pas du droit d'auteur : pour reprendre l'adage de la jurisprudence française, les « idées sont de libre parcours ». L'exception de courte citation permet également de reprendre de courts extraits de texte.

Qu'elles prennent la forme d'une série de fiches sur papier ou qu'elles soient informatisées, les bases de données et de textes scientifiques sont déjà normées par toute une série de règles sous-jacentes. Les débats autour du *text & data mining* soulèvent la question du maintien ou de l'évolution de ces régulations dès lors que les conditions du travail scientifique différent radicalement : concrètement, au lieu d'être effectué par des *personnes*, le processus de collecte est délégué à des *outils automatisés*.

Les techniques mobilisées aujourd'hui pour faire du *text & data mining* ne sont pas inédites. Les premières expériences d'analyse statistique des textes remontent à la fin du XIX^e^ siècle [1] et ont débouché très tôt sur la formulation d'outils mathéma-

1 La notion de *stylométrie* est théorisée dès 1898 par Witold Lutoslawski pour décrire une méthodologie de datation des dialogues de Platon à partir du décompte d'occurrences de mots rares (Lutoslawski 1898).

tiques importants, tels que les « chaînes de Markov[2] ». La plupart des modèles de classification des textes ont été introduits au cours des années 1960 et 1970[3] : les outils de référence employés aujourd'hui héritent encore largement de la terminologie élaborée à cette époque (comme le tableau *terme-document*). C'est également au cours de cette période que l'on assiste à la formation de communautés spécialisées en sciences humaines et sociales. En France, la « lexicométrie » née dans le sillage des travaux du laboratoire de Saint-Cloud dispose alors de tous les attributs d'un champ de recherche autonome et notamment d'une revue : *Mots*.

La véritable rupture découle de la numérisation massive d'articles de recherche, de bases de données et de sources primaires. En vingt ans, l'édition scientifique s'est dans sa quasi-totalité convertie à une diffusion électronique. Une grande partie des œuvres passées dans le domaine public sont aujourd'hui disponibles en ligne. Si ces collections massives ne peuvent être consultées et analysées en un temps raisonnable par un chercheur ou par un projet de recherche, la numérisation les rend disponibles pour les « robots ». Tout texte informatisé est aussi une série de données indexées et, à ce titre, il peut aussi faire l'objet de n'importe quelle opération formelle.

Le projet de génomique Text2Genome illustre bien ce « besoin » de documentation automatisée né de la numérisation : les gènes mentionnés dans trois millions d'articles scientifiques ont pu être identifiés et partiellement annotés[4]. L'initiative Content2Mine tente aujourd'hui de généraliser cette première approche en extrayant plusieurs centaines de millions de « faits » de la littérature scientifique[5]. Dans les deux cas, le travail réalisé n'est pas sensiblement différent de la compilation traditionnelle de fiches : à terme, l'extraction produit une « carte d'identité » d'un gène à partir des informations contenues dans la littérature de référence. Seulement, l'utilisation du travail manuel aurait étalé ce travail sur des années, voire des décennies.

Des programmes émergents en humanités numériques tentent de généraliser ces nouvelles approches à de grands corpus patrimoniaux, à l'image de la Venice Time Machine qui compile depuis quelques années des millions de « faits » extraits des archives historiques publiques et privées de Venise (Abbott 2017[6]). Les projets de « lectures distantes » menés par plusieurs chercheurs en littérature ou en histoire illustrent une autre forme de documentation automatisée : non plus l'identification d'entités et de faits mais la classification des textes[7]. Mes recherches s'inscrivent dans ce cadre et visent à retracer l'émergence des genres journalistiques dans les

---

2 Ce modèle a été initialement proposé par le mathématicien Andreï Markov en 1913 pour rendre compte de l'enchaînement des voyelles et des consonnes dans un classique de la littérature russe, *Eugène Onéguine*.

3 Les principales innovations introduites par la suite ont eu pour l'instant un impact plutôt limité. Le « *topics modeling* », très à la mode au début des années 2000, fait aujourd'hui l'objet d'une certaine désaffection.

4 http://bergmanlab.genetics.uga.edu/text2genome/.

5 http://contentmine.org/.

6 https://www.nature.com/articles/n-12446262.

7 Pour une présentation générale de ce champ de recherche en formation voir Underwood (2017). http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html.

grands corpus de presse constitués par Europeana (qui comprend, par exemple, l'ensemble des exemplaires parus de grands quotidiens français de 1815 à 1940)[8].

Les termes employés pour désigner aujourd'hui la modélisation statistique de textes ou de données témoignent de l'importance prise par les « infrastructures informationnelles » constituées en préalable à toute analyse : *text & data mining*, *knowledge discovery* dans les sciences techniques et médicales, *lecture distante* (*distant reading*) dans les humanités… Par contraste avec les travaux pionniers de la lexicométrie dans les années 1970, la préparation du corpus devient tout aussi déterminante que la modélisation statistique. Il existe aujourd'hui tout un savoir en formation autour de la gestion des données et métadonnées (« *data management* »), qui fait partie intégrante du *text & data mining* et fonde sa spécificité au regard de ses antécédents historiques.

Pourquoi le *text & data mining* est-il (généralement) illégal ?

L'élaboration des infrastructures informationnelles se heurte cependant à plusieurs difficultés pratiques qui justifient l'introduction d'une nouvelle exception au droit d'auteur. Dans la plupart des cas, il est en effet impératif de disposer en mémoire locale de l'intégralité des ressources étudiées. Chaque projet de *text mining* fait face à des besoins spécifiques et doit mettre en œuvre un ensemble de techniques inédit : les scripts modulables en python ou en R sont en pratique privilégiés au détriment d'applications intégrées. Typiquement, tous les éléments du texte ne sont pas également intéressants : les éléments syntaxiques sont déterminants pour reconnaître les entités et leurs attributs (tel un gêne et ses propriétés) ; ils sont subsidiaires, voire parasitaires, lorsqu'il s'agit de classer automatiquement les textes (on procède alors généralement au retrait de ces « mots vides »). Pour les textes anciens, le recours à une retranscription automatisée de qualité variable (ou OCR, pour *optical character recognition*) justifie l'emploi de corrections ciblées de certaines erreurs structurelles.

En principe, cette copie intégrale est autorisée. En France tout accès licite ouvre un droit de « copie privée » : l'utilisateur peut créer une ou plusieurs copies d'un ouvrage, d'un film ou d'une base de données tant qu'il ne les destine pas à un « usage collectif » et qu'il respecte certaines restrictions[9]. Un chercheur ayant accès à un article scientifique d'Elsevier (par exemple, via les abonnements souscrits par son institution) est ainsi totalement habilité à en faire une copie pour son usage personnel. Le *text & data mining* consiste seulement, sous cet angle, à généraliser cet usage préexistant : les immenses corpus de Text2Genome constituent autant de « copies privées », employées pour les besoins spécifiques d'un petit groupe de chercheur et non destinées à être republiées.

---

8 C'est notamment l'objet du projet Numapresse, http://www.numapresse.org/.

9 Notamment, le déverrouillage des systèmes de gestion de droits numériques (DRM) peut être tenu pour un usage illégitime depuis l'arrêt Mulholland Drive de la Cour de cassation (2006). http://www.journaldunet.com/juridique/juridique060303.shtml.

En pratique, la récupération des textes est très compliquée : « selon les conditions imposées par la grande majorité des contrats d'abonnements aux revues scientifiques, les abonnés peuvent lire les articles sous *paywall*, mais ne peuvent extraire automatiquement leur contenu » (Murray-Rust, Molloy & Cabell 2014 : 16). Un projet phare comme Text2Genome n'a été rendu possible qu'au terme de 3 ans de tractations avec les principaux éditeurs scientifiques (Van Noorden 2013 [10]). La combinaison de plusieurs incertitudes juridiques a pour effet de rendre la plupart des projets de *text & data mining* quasiment « illégaux » et/ou d'être contraints par des négociations complexes [11].

Les cinq sections suivantes font le point sur ces blocages, qui concernent aussi bien les sources utilisées, les modalités de consultations, les formes de réutilisation, les acteurs mobilisés, et le statut des « résultats ».

Qu'est qu'une source licite ?

Par définition tant que l'extraction, qu'elle soit « manuelle » ou « automatisée », respecte les différents droits protégeant une ressource elle est légale : le droit de lire ne saurait différer du droit d'extraire. Les conditions d'accès aux contenus numérisés sont néanmoins beaucoup plus floues et incertaines.

De nombreuses organisations revendiquent des droits indus sur leurs collections ou leurs systèmes d'informations : on parle alors de « *copyfraud* [12] ». Ces réclamations sont très fréquentes sur les contenus passés dans le domaine public qui constituent normalement une ressource privilégiée pour l'analyse automatisée en sciences humaines et sociales : alors que les droits patrimoniaux ont expiré, des bibliothèques numériques et des acteurs commerciaux continuent de revendiquer une propriété sur les textes. Plusieurs législations récentes ont eu pour effet de légitimer *a posteriori* des pratiques de « *copyfraud* ». En France, la loi Valter a ainsi entériné la requalification des collections patrimoniales en « informations publiques » susceptibles de restrictions de communication (Dulong de Rosnay et Langlais 2017 : 63-72).

Ces différents paramètres d'accès contribuent à créer des bases « trouées ». La recherche de sources historiques ou de corpus littéraire est fréquemment confrontée au « trou noir du web », soit la période 1940-1990 où les documents sont toujours protégés par le droit d'auteur mais ne présentent pas un intérêt commercial suffisant pour que les détenteurs des droits entreprennent une numérisation. De la même manière, l'exploration des corpus scientifiques dépend des abonnements souscrits auprès de tel ou tel éditeur (alors même que la totalité d'un champ de recherche s'étend généralement bien au-delà d'une seule collection).

---

10 https://www.nature.com/news/text-mining-spat-heats-up-1.12636.

11 Dans l'étude synthétisant les positions de Savoirscom1 que j'ai rédigée avec Lionel Maurel, nous parlions également d'« illégalité collatérale » (Langlais et Maurel 2014 : 12) : http://www.savoirscom1.info/wp-content/uploads/2014/01/Synthe%CC%80se-sur-le-statut-le%CC%81gal-du-content-mining.pdf.

12 Le terme a été introduit par le juriste américain Jason Mazzone en 2006. Voir en particulier Mazzone (2011).

Qu'est qu'une consultation licite ?

Sur le web, les droits de consultations sont souvent « virtuels » : le lecteur ne dispose pas *effectivement* de l'ensemble des données ou des documents mais peut *potentiellement* les consulter et/ou les télécharger à distance. Cette distinction n'est généralement pas problématique s'agissant d'une lecture « humaine ». Le rythme de consultation des ressources est suffisamment lent pour ne pas être entravé par la nécessité de les récupérer manuellement. Par contraste, les projets de *text & data mining* requièrent, comme nous l'avons vu, la consultation de très grandes collections de textes ou de données en un temps raisonnable. Il est possible de programmer des robots pour importer automatiquement les ressources souhaitées mais le rythme de consultation est alors bien supérieur à celui d'un humain. Cette récupération en série (ou *scraping*) est fréquemment assimilée à un abus, dans la mesure où elle sollicite davantage le serveur qui doit gérer potentiellement un grand nombre de requêtes pour un seul internaute.

Certains de nos projets de *text & data mining* ont ainsi été « bloqués » par des hébergeurs variés (Google, Elsevier, Gallica, etc.) sur la base de cet usage « abusif ». Tout en n'étant pas toujours injustifiées, ces mesures techniques de protection constituent également un outil commode pour contrôler, voire empêcher la récupération des textes et des données. Ainsi, à l'issue d'un projet de récupération de métadonnées sur Google Scholar, nous concluions : « Les choses sont simple : c'est *catch me if you can*. Sauf à truquer la surveillance de Google, Google Scholar est complètement inaccessible. » (Langlais 2014[13])

Au fil des années, les chercheurs en *text mining* ont développées des parades plus ou moins efficaces pour anticiper ces blocages : utilisations de navigateurs différents d'une consultation sur l'autre (via le « *user agent* »), définition d'un temps de latence aléatoire entre chaque consultation (pour mimer le comportement d'un lecteur humain), recours à des IP différentes de celles de son poste fixe (via un VPN ou le réseau TOR)[14]. Toutes ces tactiques sont vraisemblablement illégales : elles contreviennent aux conditions d'usages (« *terms of service* ») définies par chaque site.

Qu'est qu'une réutilisation licite ?

Les bases de données et les compilations peuvent faire l'objet d'une protection au titre de la propriété intellectuelle dès lors qu'elles justifient d'une originalité intrinsèque[15]. Ce principe, anciennement appliqué en droit français n'institue qu'une restriction limitée : comme l'a montré, à la fin des années 1980, la jurisprudence Le Monde vs. Microfor, la plupart des bases de données se situent vraisemblablement sous le seuil de l'originalité (Frochot 1988[16]).

L'introduction d'un « droit *sui generis* des bases de données » en droit européen à la fin des années 1990 a remis en question ce compromis. Ce droit est davantage

13 https://scoms.hypotheses.org/216.

14 Nous donnons un aperçu de ces parades ici https://scoms.hypotheses.org/216 (Langlais 2014).

15 En France, cette protection est envisagée par la jurisprudence depuis le début du XIXe siècle (Rideau 2008) : http://www.copyrighthistory.org/cam/tools/request/showRecord?id=commentary_f_1869.

16 Référence republiée à cette adresse : http://www.les-infostrateges.com/article/880432/affaire-microfor-le-monde.

inspiré de la logique anglo-saxonne du « *copyright* » que de la tradition européenne du droit d'auteur : il permet au « producteur » de la base d'empêcher les réutilisations « substantielles ». Cette « substantialité » reste à ce jour mal définie et a suscité des décisions variables en jurisprudence française et européenne [17]. Néanmoins, elle introduit une insécurité juridique supplémentaire, d'autant que le droit des bases de données est théoriquement éternel (il peut durer tant que le producteur de la base investit pour son entretien).

Toutes les bases de données peuvent invoquer cette protection sous réserve de justifier de cet investissement substantiel. Par définition, sur le web, toute ressource (qu'il s'agisse d'un texte, d'une image, d'un enregistrement, etc.) constitue *aussi* une base de données. Le droit des bases de données est ainsi devenu une justification majeure de l'appropriation de contenus dans le domaine public (ou « *copyfraud* »).

La propriété intellectuelle n'est pas le seul cadre légal encadrant la réutilisation et les traitements statistiques effectués sur les corpus. Les données personnelles font l'objet de régulations de mieux en mieux délimitées [18] que nous n'évoquons ici que très partiellement et qui affectent tout particulièrement les projets de recherche en sciences humaines et sociales. Ces régulations ne visent pas seulement l'enregistrement d'informations privées associées à une personne physique, mais également les recoupements potentiels autorisés par l'emploi de techniques de classification automatisées. Le nouveau Règlement général sur la protection des données de l'Union européenne établit que les données personnelles « traitées de manière licite, loyale et transparente au regard de la personne concernée » (art. 5). La recherche fondamentale peut bénéficier de certaines dérogations dans la mesure où elle est effectuée dans « l'intérêt public » (Rumbold et Pierscionek 2017). Néanmoins, pour être en conformité avec ces encadrements les projets de *text & data mining* peuvent ainsi être amenés à anonymiser les données (soit, à remplacer les noms des personnes par des variables arbitraires), voire à masquer certaines informations facilitant la réidentification. L'implémentation française du RGPD dans la loi française tel que prévue par les décrets du 1er août 2018 ne mentionne pour l'instant aucune exception pour la recherche, même si le droit européen pourrait primer à cet égard [19].

Qui peut licitement accéder à la ressource ?

Le caractère composite des compétences et savoirs mobilisés (expertise dans un certain champ scientifique [biologie, chimie, physique, littérature, etc.], analyse statistique, traitement informatique, constitution et gestion de bases de données) nécessite souvent des collaborations interdisciplinaires. Tous les acteurs impliqués n'ont pas nécessairement les mêmes droits d'accès. Notamment, la circulation des textes scientifiques est très « compartimentée » : les grands acteurs de l'édition universitaire (Elsevier, Springer) concluent usuellement des accords (*big deals*) avec

17 Sur ces incertitudes, voir Caspers et Guibault (2016) qui citent plusieurs controverses non résolues autour de la définition du droit des bases de données.

18 L'Union européenne a ainsi introduit en 2016 un Règlement général sur la protection des données, qui sera appliqué en 2018 : http://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A32016R0679.

19 https://scinfolex.com/2018/07/18/donnees-personnelles-et-recherche-scientifique-quelle-articulation-dans-le-rgpd.

des bibliothèques universitaires sur certaines collections délimités. Concrètement, des chercheurs en informatique ont peu de chance d'accéder à des revues en biologie ou en littérature (et vice versa). La source licite de certains collaborateurs du projet risque ainsi d'être illicite pour d'autres.

Par ailleurs, l'emploi généralisé de méthodes quantitatives soulève un problème de reproductibilité. Pour pouvoir être évalué, le processus de recherche devrait être retracé dans ses étapes successives, ce qui suppose d'avoir communication au moins partielle du corpus ou de la base de données utilisée. À nouveau, ce qui a été une source licite pour les participants du projet risque de ne pas l'être pour les évaluateurs potentiels.

Quel est le statut juridique des « résultats » ?

Le *text & data mining* ne se résume par à la réutilisation de données ou de textes existants : il produit des résultats inédits. Le nombre d'occurrences de mots par documents (ou *term document matrix*) constitue un exemple très simple de production dérivée, ensuite utilisée pour de nombreuses tâches (notamment pour la classification automatisée des textes). Les droits de protection associés aux ressources analysées pourraient-ils « contaminer » ces résultats ?

Pour l'instant, les techniques employées ne permettent pas de réidentifier le texte d'origine et, à ce titre, épuisent l'originalité préexistante (même si, comme nous le verrons, certains acteurs émettent des revendications en ce sens). La bibliothèque numérique Hathi Trust (l'équivalent américain de Gallica) a ainsi pu republier le décompte des occurrences page par page de millions d'ouvrages encore soumis au droit d'auteur. Ces données suffisent pour mener un grand nombre d'analyses statistiques (notamment de la classification automatisée), tout en ne suffisant pas pour reconstruire le texte d'origine : si tous les mots sont là, ils le sont dans le désordre.

Les limites du règlement contractuel

Les différents problèmes que nous listons ici sont devenus patents vers la fin de la décennie 2000, alors que le *text & data mining* suscitait de plus en plus l'intérêt de plusieurs communautés de recherche. L'essentiel de ces recherches portait sur les publications scientifiques en sciences techniques et médicales. La concentration considérable de l'édition académique laissait alors augurer de la possibilité d'un arrangement contractuel ou « organique » entre les grands acteurs économiques (Elsevier, Springer, Informa, etc.) et les communautés scientifiques.

En 2012 et 2013, l'Union européenne organisa un « dialogue structuré entre parties intéressées » visant à aboutir à l'élaboration de licences contractuelles par les détenteurs de droits (« Licences for Europe »). Dans ce cadre, les éditeurs scientifiques ont développé des processus d'accès normalisés pour extraire leurs ressources[20]. Ces initiatives ont rapidement suscité des critiques dans les milieux

20 Voir à titre d'exemple, la présentation de la licence de Springer : https://ec.europa.eu/licences-for-europe-dialogue/sites/licences-for-europe-dialogue/files/Publishers-Perspective-Initiatives_0.pdf.

scientifiques. Elles correspondent pour partie à une privatisation du droit de lire et à la substitution de la loi par des régulations privées [21].

1. L'extraction n'épuise pas les droits sur la ressource : la licence normalisée d'Elsevier implique que les « résultats » (« *output* ») soient diffusés sous une licence Creative Commons non commerciale. La notion d'*output* est volontairement indéfinie et concerne visiblement la totalité des données/informations recueillies au terme de l'extraction (par exemple sur les entités) : « *You are able to distribute the findings of your text mining,* in *line with the following conditions* [22] ». Concrètement, le projet Text2Genome devrait adopter une licence non commerciale pour toutes les « fiches d'identité » sur les gènes constitués à partir de corpus d'Elsevier. En effet, en dehors des *faits* et *résultats*, les *outputs* peuvent inclure des brèves citations (« *snippets* ») de moins de 200 caractères. Cette exigence contrevient au droit de courte citation qui n'a pas de limite de taille définie. Elle constitue une limite problématique pour certains champs de recherche : des noms de composants chimiques ont parfois plus de 200 caractères [23]… En l'état, la licence normalisée d'Elsevier revient sur plusieurs concepts fondamentaux de la propriété intellectuelle, en redéfinissant les frontières entre faits et expression originale, ou entre courte citation et extraits disproportionnés : elle revient à « écrire sa propre loi ».
2. L'extraction passe exclusivement par une interface normalisée, l'API : bien que les acteurs revendiquent un accès au texte « complet », l'API ne permet pas d'accéder à certains éléments importants d'une publication scientifique (comme les images ou les schémas) et inclut des commandes prédéfinies qui « norment » par anticipation le processus de recherche.
3. L'extraction est plafonnée : elle ne peut excéder un nombre prédéfini d'appels à l'API (10 000 par semaines). Pour un projet de l'ampleur de Text2Genome, l'extraction aurait pris quasiment une année entière.
4. L'extraction doit être « déclarée » en détail : chaque projet doit indiquer précisément ses objectifs et l'identité de ses équipes – alors que par définition ils ont déjà un accès licite à la ressource. Cette étape rajoute une couche bureaucratique supplémentaire et contribue à ficher les projets de *text & data mining*.

L'évolution des usages a graduellement rendu les licences contractuelles obsolètes : les techniques de *text & data mining* sont employées bien au-delà des seuls corpus scientifiques détenus par de grands éditeurs, notamment à la suite de leur appropriation par des chercheurs en sciences humaines et sociales. Les projets se sont multipliés sur des sources de nature très variées : réseaux sociaux, presse, illustrations, romans du XIX^e siècle…

Or, le degré de concentration de l'édition scientifique a peu d'équivalents : dans de nombreux domaines concernés, les « parties intéressées » sont trop nombreuses et dispersées pour qu'un « dialogue » puisse aboutir à des normes communes. Il n'est clairement pas envisageable de développer des licences contractuelles pour chaque

21 Pour un compte-rendu détaillé voir Langlais (2014) : https://scoms.hypotheses.org/98.

22 https://www.elsevier.com/__data/assets/pdf_file/0012/102234/TDM-sign-up-short-form.pdf.

23 https://blogs.ch.cam.ac.uk/pmr/2014/01/31/content-mining-why-you-and-i-should-not-sign-up-for-elseviers-tdm-service/.

détenteur des droits (ce qui était l'objectif sous-jacent du programme « Licences for Europe » : les accords contractuels dispensant de repenser la loi).

L'avènement des exceptions

Plusieurs pays disposaient déjà d'un cadre légal autorisant en principe le *text & data mining* à des fins de recherche : l'évolution a pu être jurisprudentielle, sans qu'une réforme ne soit nécessaire.

Aux États-Unis, le « *fair use*[24] » autorise anciennement la reproduction d'œuvres protégées sous certaines conditions (notamment à des fins de recherche et d'enseignement). Le procès Authors Guild v. Hathi Trust a confirmé l'assimilation des nouvelles pratiques d'extraction au « *fair use* » : la bibliothèque numérique Hathi Trust peut légitimement proposer une recherche en plein texte de livres soumis au droit d'auteur. Incidemment, le jugement exclut la possibilité d'un maintien des droits : la recherche plein texte (tout comme, l'extraction d'entités où la reconnaissance automatisée des genres) ne maintient pas le texte en l'état ; elle introduit « une fonction entièrement distincte » qui relève d'un « usage transformatif[25] ». Un autre procès parallèle visant Google Books aboutit à des conclusions convergentes.

Le compromis actuel prévalant dans la jurisprudence américaine ne permet pas de rendre public les bases de données textuelles mais leur stockage privé pour générer des résultats de recherche. Il a eu une incidence immédiate sur les pratiques de recherche. En janvier 2017, Hathi Trust a pu proposer un décompte complet des occurrences de mots dans chaque page de 13 millions d'ouvrages (dont 8,4 millions d'ouvrages soumis au droit d'auteur)[26]. Jusqu'à présent, les initiatives de ce type se limitaient au domaine public et/ou n'incluaient aucune information permettant la réidentification des ouvrages et des métadonnées (comme *Ngram Viewer*).

Les États-Unis ne constituent pas un cas à part. Plusieurs pays ont procédé à une autorisation tacite ou explicite du *text & data mining* (Chine, Canada, Japon, Corée, Israël, Taïwan, etc.), généralement en procédant à une interprétation élargie de libertés préexistantes (« *fair use* », « *fair dealing* », etc.) (Handke, Guibault et Vallbé 2015[27]).

En Europe, cette évolution est actuellement bloquée : elle passe nécessairement par une réforme. Le cadre légal commun créé par la directive de 2001 sur l'harmonisation des droits d'auteur n'autorise en effet pas une évolution

---

24 Le terme « *fair use* » (ou « usage juste ») est apparu dans la jurisprudence anglo-saxonne au XVIII[e] siècle et est couramment utilisé aux États-Unis pour désigner tout un ensemble d'exceptions qui ne nuisent pas aux intérêts du détenteur du droit et/ou favorables aux intérêts du public. Le droit de parodie constitue un exemple de « *fair use* » également présent en droit français. Le droit américain reconnaît également des exceptions pour usages scientifiques et pédagogiques. Par exemple, la version de Wikipédia en anglais intègre de nombreux contenus sous droit d'auteur (comme des images de films ou de jeux vidéo) au titre du « *fair use* » ; ces contenus ne sont pas présents dans la version francophone.

25 Décision du juge Harold Baer à l'issue du procès Authors Guild v. HathiTrust, p. 16 : https://cases.justia.com/federal/district-courts/new-york/nysdce/1:2011cv06351/384619/156/0.pdf.

26 https://wiki.htrc.illinois.edu/display/COM/Extracted+Features+Dataset.

27 http://dx.doi.org/10.2139/ssrn.2608513.

purement jurisprudentielle. Ni l'exception pédagogique et scientifique, ni le droit de reproduction « transitoires » à des fins techniques ne sont suffisamment extensibles pour pérenniser les projets de *text & data mining* (Caspers et Guibault 2016 : 23). Ces blocages pénalisent la recherche européenne. Une étude empirique menée sur les données bibliométriques montre que « là où le *data mining* requiert le consentement explicite des ayants droit, il représente une part significativement inférieure des projets de recherche » (Handke, Guibault et Vallbé 2015 : 3).

Au Royaume-Uni, le rapport Hargreaves appelle dès 2011 à introduire une exception : « *Text mining is one current example of a new technology which copyright should not inhibit, but does* » (Ian Hargreaves 2011 : 48[28]). La réforme se concrétise en 2014. Elle prend la forme d'un « *fair dealing* » ou « utilisation équitable » clairement délimitée : les projets de recherche à visée non commerciale peuvent mener des « analyses computationnelles » de sources auxquelles ils ont accès[29]. Si la conformité de l'exception anglaise à la directive de 2001 n'est pas totalement acquise[30], elle constitue néanmoins une inspiration forte pour toutes les réformes ultérieures, qu'elles soient discutées au niveau national ou européen.

L'exception française : un projet inachevé

En France, après de premières consultations peu concluantes[31], le débat sur l'exception prend de l'ampleur à la faveur d'une grande loi sur les nouveaux usages du numériques : la loi pour une République numérique. La loi est dotée d'emblée d'un important volet scientifique en prévoyant l'introduction d'un droit de republication des publications scientifiques, indépendamment des conditions contractuelles, sur le modèle allemand et italien. Les premières tractations menées auprès de représentants des communautés scientifiques entraînent l'ajout d'une première exception autorisant « l'exploration de textes et de données pour les besoins de la recherche publique » dans un préprojet de juillet 2016[32]. Le compromis retenu est le même que celui de l'exception anglaise : autoriser « l'exploration des textes de données » provenant d'une « source licite » pour les projets de recherche à des fins de commerciales.

L'exception disparaît dès septembre : les acteurs de l'édition se sont mobilisés contre la mesure. Une brochure de l'avocat du Syndicat national de l'édition dénonce une menace majeure contre l'économie du livre : « Si de tels investissements pouvaient être légalement pillés, aucun éditeur n'engagerait désormais le moindre

---

28 https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/32563/ipreview-finalreport.pdf.

29 Voir la traduction et la présentation dans Lionel Maurel (2014) : https://scinfolex.com/2014/04/01/le-royaume-uni-sanctuarise-les-pratiques-de-data-mining-par-le-biais-dune-exception-au-droit-dauteur/.

30 Les exceptions prévues par la directive de 2001 sont limitatives et n'admettent en principe aucun aménagement au titre de la subsidiarité.

31 Voir notamment le rapport commandité par le CSPLA pour lequel j'ai été auditionné au nom de Savoirscom1 avec Lionel Maurel. Voir Martin et Carvalho (2014). Dans ses conclusions, le rapport estime que l'introduction d'une exception n'est pas nécessaire.

32 Voir l'analyse de ce texte sur *Sciences communes* : https://scoms.hypotheses.org/473.

financement pour créer de tels outils. » (Malka 2015) Alors que les éditeurs scientifiques étaient assez isolés sur l'autre grande réforme concédée aux communautés scientifiques, le droit de « republication » en libre accès, ils parviennent ici à fédérer une coalition plus large. Dans la mesure où elle vise des contenus non produits par des chercheurs, l'exception requiert en effet une évolution du Code de la propriété intellectuelle (et non simplement du Code de la recherche).

L'exception revient dès octobre : la consultation ouverte par Axelle Lemaire autour de la loi sur le numérique se traduit par une mobilisation significative des institutions et des communautés scientifiques. La proposition d'exception « de fouille de texte de données » portée par le consortium Couperin est la 4e mesure la plus soutenue dans la section 2 de la loi (sur les « travaux de recherche »). Ce texte introduit une innovation qui va demeurer par rapport au précédent compromis : les « copies techniques issues des traitements » ne sont pas systématiquement détruites mais peuvent être conservées par des « organismes désignés par décret »[33].

En raison de l'émergence d'un clivage fort entre deux coalitions, le débat parlementaire autour de l'exception est très animé. Sept amendements similaires proposent le rétablissement de l'exception ; ils sont déposés par plusieurs dizaines de députés, de presque toutes tendances politiques[34]. Un amendement de synthèse est définitivement adopté par la commission mixte parlementaire[35]. Suite à ce soutien inattendu, le « gouvernement s'est trouvé très embarrassé »[36]. Plusieurs voies alternatives sont encore étudiées, telles que l'intégration d'une clause impérative dans les contrats conclus par les universités avec les éditeurs. Cette dernière configuration excluait toutes les ressources non encadrées par un contrat : par exemple, une exploration des textes d'un forum consultable en ligne vise manifestement une source « licite », sans jamais donner lieu à un accord écrit entre l'utilisateur, l'hébergeur et les auteurs.

Produit d'une évolution complexe et contrariée, le texte final de l'exception en « porte les stigmates » (Maurel 2016[37]). Par contraste avec la simplicité du « *fair use* » américain, elle est assortie de plusieurs réserves qui compliquent sa mise en œuvre :

1. Comme la loi anglaise de 2014, *l'exception ne peut être invoquée que dans le cadre de recherches publiques non commerciales* (« à l'exclusion de toute finalité commerciale »). La notion de « finalité non commerciale » reste imprécise. Elle ne s'étend probablement pas jusqu'à la publication d'article (qui constitue une forme de finalité commerciale pour l'éditeur, mais dans ce cas l'exception

33 Proposition du consortium COUPERIN : http://www.republique-numerique.fr/projects/projet-de-loi-numerique/consultation/consultation/opinions/section-2-travaux-de-recherche-et-de-statistique/exception-de-fouille-de-texte-et-de-donnees.

34 Amendements AC11, AC16, CL84, CL344 et CL463, CE29 et CE78 de la loi numérique. Ils émanaient aussi bien de parlementaires de gauche (CE29 pour les communistes, CL344 pour des écologistes et indépendants [Attard], CL463 pour les socialistes) que de parlementaires de droite (le seul CL84, mais adopté par un grand nombre de signataires).

35 Voir l'amendement n° 180 : http://www.assemblee-nationale.fr/14/amendements/3399/AN/180.asp.

36 http://www.eprist.fr/wp-content/uploads/2016/01/I-IST_12_LoiNum%C3%A9rique-ExceptionTDM.pdf.

37 https://scinfolex.com/2016/11/09/lexception-tdm-dans-la-loi-numerique-merites-limites-et-perspectives. Nous nous sommes appuyés sur cette excellente analyse pour décrire les avancées et limites de l'exception.

serait en grande partie inexploitable) mais couvre très certainement le dépôt de brevets ou les partenariats avec des acteurs privés.

2. *L'exception est pour l'essentiel limitée aux textes* (« en vue de l'exploration de textes »). L'analyse computationnelle des images ou des vidéos, qui suscite de plus en plus d'intérêt suite au développement de nouvelles techniques de *deep learning* se trouve de fait exclue du champ de l'exception. En cela, l'exception française se distingue de ses homologues anglais ou allemands, qui portent sur l'ensemble des « sources » sans distinction.
3. *L'exception porte également sur les données, mais uniquement si elles sont « incluses ou associées aux écrits scientifiques ».* En pratique, cela couvre non seulement les données intégrées dans le texte d'un article de recherche (par exemple sous forme de tableaux), mais aussi les fichiers « additionnels » (déposés dans un entrepôt de données associé, comme *Zenodo* ou *Figshare*). Cette restriction peut incidemment compliquer l'extraction des textes : celle-ci se limite rarement au texte « brut », mais intègre aussi des métadonnées (par exemple, la date de publication, l'auteur, etc.).
4. *La conservation des données et des textes n'est pas laissée à la libre appréciation du projet de recherche.* L'encadrement est plus restrictif s'agissant du droit des bases de données que du droit d'auteur. L'exception ajoutée à l'article L342-3 indique que les « copies techniques issues des traitements » sont remises à des « organismes désignés par décret » au terme du projet de recherche ; toutes les autres copies ou reproductions « sont détruites ». Inversement, l'exception ajoutée à l'article L122-5 ne fait qu'une mention vague des « modalités de conservation et de communication » destinées à être fixées par décret. En l'état, ces dispositions ont pour effet de rendre l'exploration de textes « bruts » plus simple que l'exploration de ressources protégées au titre du droit des bases de données.

Au-delà de ces limites, l'exception règle définitivement deux points contentieux majeurs.

1. *Les sources licites peuvent être copiées et analysées pour les besoins du projet*, sans que cette copie ne puisse être assimilée à une diffusion « publique » illégale. Les membres d'un projet de *text & data mining* constituent, en quelque sorte, une incarnation nouvelle du « cercle de famille ».
2. *Les droits sur la ressource sont « épuisés » au terme du traitement*. Les résultats constituent en effet des « données de la recherche » et, en accord avec la définition donnée par l'art. 30 de la loi sur le numérique leur « réutilisation est libre ».

L'exception française apporte ainsi une sécurité juridique tangible, tout en maintenant plusieurs zones d'ombre et d'incertitudes. Le compromis actuel intègre plusieurs restrictions non négligeables pour la recherche (limitation aux textes et données scientifiques, cadre non commercial, encadrement de la conservation des « copies techniques »), voire des complications pas totalement motivées (telles que la divergence entre droit d'auteur et droit des bases de données, s'agissant de la conservation des copies).

Le décret d'application constitue une pièce importante du dispositif. Si toutes les mesures de l'exception n'en dépendent pas (l'accès aux copies et l'épuisement des

droits sur la ressource constituent *a priori* des points déjà réglés), c'est notamment le cas des modalités de conservation.

La version initialement envisagée ouvrait des pistes intéressantes. La plupart des institutions scientifiques françaises et certains acteurs externes comme la BnF auraient le statut de « tiers de confiance » : ils seraient habilités à conserver les copies techniques et à les rendre disponibles pour d'autres projets tant que ces derniers disposeraient d'un « accès licite ». Au-delà de la sécurité juridique garantie par l'exception, ce mécanisme permettrait à terme de contourner certaines limitations techniques qui entravent la pratique du *text & data mining* : la récupération de la ressource étudiée se fait en une seule fois. Au lieu d'être tributaire des « plafonds » imposés par certains acteurs (tels que les 10 000 articles par semaine d'Elsevier), l'extraction pourra être directement effectuée sur les corpus conservés par les tiers de confiance.

Le statut de tiers de confiance présente un autre aspect bénéfique : faire émerger un « écosystème » structuré du *text & data mining*. La diffusion restreinte des copies techniques auprès d'autres projets ayant également un accès technique contribue à l'émergence de « bonnes pratiques » (puisque l'enjeu n'est plus seulement d'élaborer des méthodes *ad hoc*, le temps d'un seul projet, mais de pouvoir les communiquer, si possible en reprenant des normes mutuellement compréhensibles) ainsi que le développement d'infrastructures adaptées (les volumes de données concernés peuvent être assez considérable).

Suite à l'opposition toujours soutenue de plusieurs représentants d'ayants droit (dont des éditeurs scientifiques et des éditeurs de presse écrite), la proposition de décret d'application a été soumise en mai 2017 au Conseil d'État. Elle n'a pas été jugée conforme à la législation européenne : son périmètre excèderait le champ des exceptions autorisées par la directive de 2001. Ce rejet suscite des incertitudes non seulement sur les modalités du décret d'application, mais sur son principe même.

En l'état, l'exception française demeure un projet inachevé : elle apporte une sécurité juridique limitée, sans résoudre les problèmes d'accès techniques et sans donner réellement à la recherche française sur le *text & data mining* les moyens de se structurer durablement.

En dépit de ces imperfections, ce dispositif a commencé à inspirer des initiatives similaires. L'Allemagne a voté en juin 2017 une exception qui reprend certains éléments du texte français : il est possible d'envoyer les copies à des « institutions » dont la liste est prédéfinie dans la loi (et non dans un décret risquant d'être invalidé) [38]. Paradoxalement, tout en étant en partie dérivée de l'exception française, l'exception allemande sera appliquée plus tôt.

Le futur des exceptions : une construction européenne ?

Les déboires des décrets d'applications français montrent que l'encadrement strict des exceptions prévues dans la directive européenne de 2001 fragilise les réformes nationales : en principe aucune évolution ne doit « excéder » cette liste limitative.

38 http://www.bibliothequescientifiquenumerique.fr/exception-au-tdm-en-allemagne/.

Dans le cas du *text & data mining*, il est nécessaire de faire preuve d'une grande créativité juridique pour affilier l'exception à un des cas de figure envisagés, comme l'exception pour la recherche ou pour les copies transitoires.

Ce cadre commun est actuellement en cours de refonte, sous le titre de « réforme européenne du droit d'auteur ». Le thème du *text & data mining* apparaît initialement lors d'une consultation préalable en 2013, alors que la perspective d'une résolution contractuelle (« licences for Europe ») s'éloignait de plus en plus. Deux ans plus tard, le rapport Reda émet une préconisation très large en appelant à « clarifier que l'accès licite aux données inclut le droit de les extraire avec des techniques d'analyse automatisées[39] ».

Dans les débats au Parlement, ce droit d'extraction est finalement limité à la recherche scientifique non commerciale. Ce périmètre reste toujours discuté alors que le principe même d'une exception semble faire l'objet d'une acceptation consensuelle. La proposition de directive soumise par la commission en septembre 2016 appelle à « répondre à cette insécurité juridique en introduisant une exception » dont le périmètre serait limité à la recherche scientifique tout en admettant (contrairement aux exceptions anglaises et françaises) des usages commerciaux[40].

Le rapport de Theresa Comodini pour la Commission aux affaires juridiques publié en mars 2017[41] plaide pour plusieurs élargissements. L'exception n'est plus seulement utilisable dans un contexte de recherche scientifique mais aussi d'« innovation » (ce qui inclut tacitement des acteurs commerciaux et non commerciaux situés en dehors du monde académique). De manière plus inédite, Comodoni s'intéresse au cas des sources « non licites » : l'un des amendements proposés vise à permettre à des chercheurs d'extraire, dans un format standardisé, des informations auxquels ils n'ont pas accès moyennant une compensation « proportionnée ». Ce dispositif s'apparente à une ébauche de résolution de l'une des principales limites des exceptions actuellement en vigueur : les arrangements contractuels sont toujours nécessaires pour les sources non licites. Par contraste avec la démarche d'ouverture du rapport Comodini, la commission du marché intérieur s'est finalement ralliée début juin à une interprétation restrictive (avec une exception limitée à la recherche).

Si l'introduction à terme d'une exception européenne paraît acquise, ses modalités sont encore très incertaines. Les procédures européennes déjà complexes sont « doublonnées » par un processus parallèle : les différentes lois nationales passées successivement en France, au Royaume-Uni et, tout dernièrement, en Allemagne s'apparentent à un exercice d'écriture collaborative à distance. L'article 38 de la loi pour une République numérique est un emprunt avoué au compromis anglais défini par le rapport Hargreaves (notamment pour la définition de la « source licite ») ; l'exception allemande reprend plusieurs innovations introduites en France (comme le statut d'intermédiaire de confiance, chargé de conserver les copies techniques). Tous ces échanges irriguent les débats européens et constituent pour partie des expérimentations grandeur nature de telle ou telle version de l'exception.

---

39 https://juliareda.eu/copyright-evaluation-report-explained/#tdm.

40 Proposition pour une directive sur le droit d'auteur dans le marché commun de la Commission européenne, par. 8-13, http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=17200.

41 http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2f%2fEP%2f%2fNONSGML%2bCOMPARL%2bPE-601.094%2b01%2bDOC%2bPDF%2bV0%2f%2fEN.

Au-delà des mobilisations explicites, les communautés scientifiques sont appelées à jouer un rôle considérable dans ce processus juridique. L'écosystème du *text & data mining* reste pour l'essentiel à construire et l'irruption de nouveaux acteurs a suscité et suscite encore des évolutions significatives. Le développement de projets et d'outils spécifiques en sciences sociales et dans les humanités a fortement contribué à rendre les licences contractuelles obsolètes : dès lors que le champ de l'extraction automatisé se déporte des corpus scientifiques vers la totalité des productions écrites, la régulation ne peut plus passer que par la loi. Les enjeux émergents de conservation des « copies techniques » ou d'accès à des sources non licites vont nécessiter le développement de cadres juridiques, d'infrastructures techniques et d'usages scientifiques concordants : si elle reste un critère déterminant pour qu'un projet soit possible ou non, la loi n'est que l'un des lieux où se joue ce processus de normalisation.

Table des matières

La diffusion numérique des données en SHS

Guide des bonnes pratiques éthiques et juridiques

DIGITALES

La collection « Digitales » s'intéresse aux rapports entre les sciences humaines et le monde numérique, qu'il fournisse des outils critiques ou qu'il soit un domaine de création.

Produire, exploiter, éditer, publier ou valoriser des données numériques fait partie du travail quotidien des chercheurs en sciences humaines et sociales (SHS). Ces données sont aujourd'hui disséminées sous de multiples formats dans le monde de la recherche et, au-delà, auprès de citoyens de plus en plus curieux et intéressés par les documents produits par les scientifiques. Dans un contexte de mutation fulgurante des méthodes de travail, ce guide aborde avec simplicité des questions et des enjeux complexes auxquels se confronte quotidiennement la communauté des SHS. De leur collecte à leur réutilisation, les données de la recherche sont manipulées, éditorialisées, interrogées, mises en ligne… par tous les acteurs du monde académique qui ne savent pas toujours répondre aux questions juridiques et éthiques ou même, ne parviennent pas à les poser clairement. C'est à eux que s'adresse cet ouvrage, fondé sur des réflexions et des retours d'expériences qui présentent les bonnes pratiques pour accompagner celles et ceux qui s'inscrivent dans la dynamique de la science ouverte.

conception graphique
et illustration de couverture
J.-B. Cholbi

Véronique Ginouvès est responsable des archives sonores et audiovisuelles à la Maison méditerranéenne des sciences de l'homme (AMU-CNRS) à Aix-en-Provence.

Isabelle Gras est conservatrice des bibliothèques au Service commun de la documentation de l'université d'Aix-Marseille (SCD AMU).

Presses Universitaires de Provence
Aix*Marseille université initiative d'excellence
Bibliothèques universitaires
cnrs
Maison méditerranéenne des sciences de l'homme USR 3125
Huma-Num

9 791032 001790

20 €