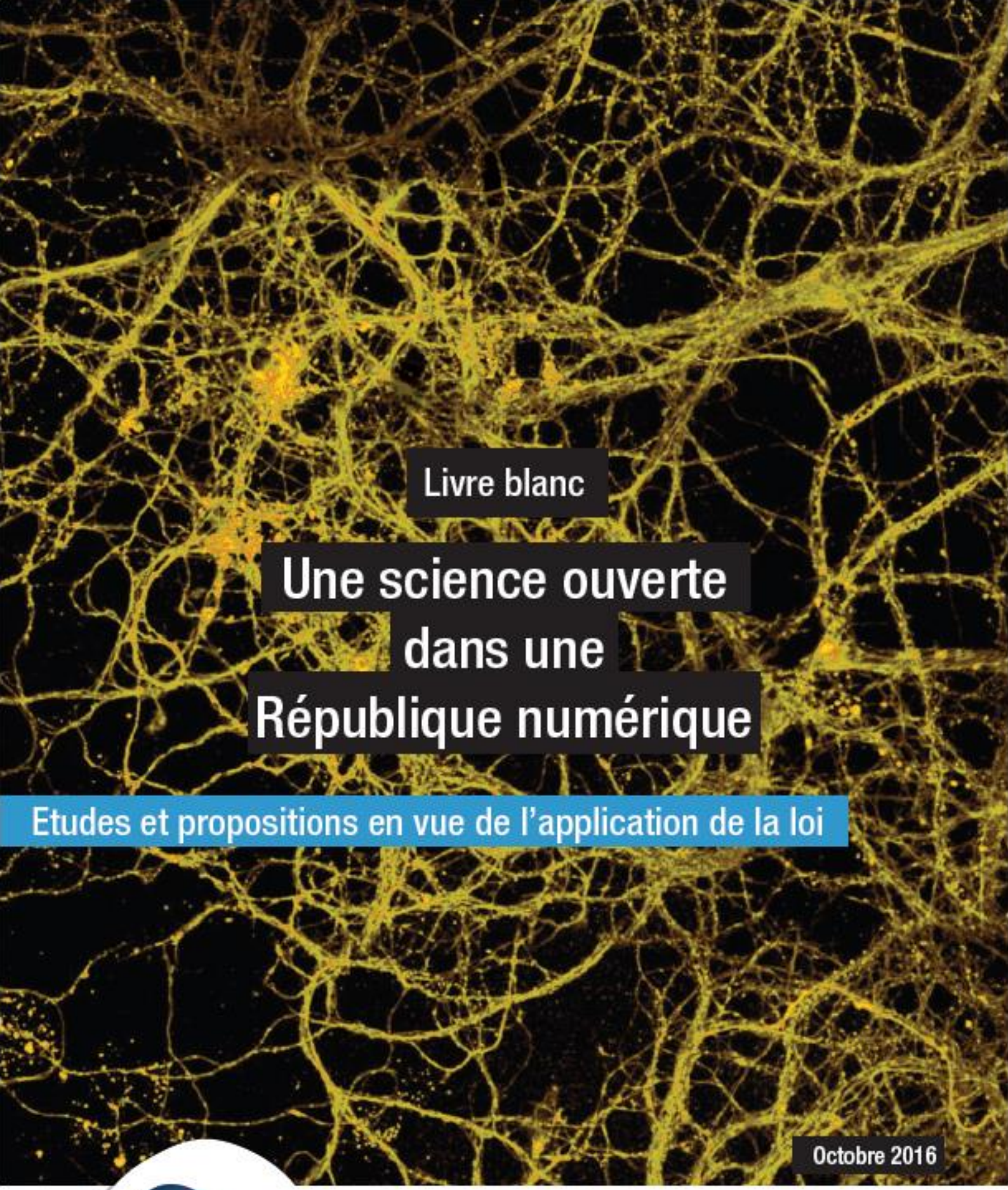


Livre blanc

Une science ouverte dans une République numérique

Etudes et propositions en vue de l'application de la loi

Octobre 2016

cnrs
dépasser les frontières

Direction de l'information scientifique et technique

Livre blanc

Une Science ouverte dans une République numérique

Études et propositions en vue de l'application de la loi

Guide stratégique d'applications

Direction de l'Information Scientifique et Technique

CNRS

Accompagnement, Expertise, Conseil : Cabinet ALAIN BENSOUSSAN

Table des matières

Préface

L'adoption par le Parlement de la loi « Pour une République numérique » ouvre une importante étape de mise en œuvre. Les décrets et modalités d'application vont pouvoir bénéficier du fort consensus qui a prévalu jusqu'ici dans la recherche publique. Toutes les communautés de chercheurs, d'usagers et de bénéficiaires des dispositifs numériques de science (Conseil scientifique du CNRS, ADBU, Couperin, EPRIST, CPU, CNum…) ont convergé pour soutenir intensément les dispositions législatives qui viennent d'aboutir.

Ce Guide stratégique a été construit pour jalonner et faciliter l'application de la loi : il est ainsi en phase avec les visées des initiateurs du Livre blanc « Une science ouverte dans une République numérique » qui ont tous souhaité associer clairement une vision nouvelle pour la recherche et une pratique qui la traduise dans les faits.

Rapport complémentaire du Livre blanc, ce Guide ambitionne un accompagnement des acteurs dans des démarches souvent complexes : chercheurs, techniciens, éditeurs, usagers des résultats de la science publique ont besoin aujourd'hui d'avancer dans l'esprit de découverte qui est celui des grands projets numériques pour la science, et d'abord pour l'Investissement d'Avenir ISTEX qui en est l'élément moteur.

Ce Guide stratégique contient des éléments comparatifs des pratiques étrangères et des propositions ou réflexions qui pourront être utiles à l'application de la loi. Celle-ci doit intervenir, selon les orientations données par le Premier ministre, pour la fin du mois de janvier 2017. Ce court délai justifie, à lui seul, que la DIST du CNRS ait eu à cœur de produire ce document immédiatement après le vote conclusif du Parlement, intervenu au Sénat le mercredi 28 septembre avec l'adoption du Projet de loi, à une forte majorité, dans des termes voisins de ceux de l'Assemblée nationale.

La genèse de la loi révèle la construction progressive d'un riche consensus au sein du Parlement, majorité et oppositions confondues, au sein du Gouvernement où les ministres en charge de la recherche et du numérique ont fait très tôt cause commune en faveur du TDM et de la libre circulation des publications scientifiques à des fins de recherche : Gouvernement et Parlement ont ainsi rejoint les attentes exprimées lors de la Consultation nationale lancée par le Premier ministre en septembre 2015 et soulignées ensuite par les Présidents des Universités et Organismes de recherche.

En liaison étroite avec la réflexion française, les choix de l'Europe se sont affinés dans le sens d'une liberté plus grande laissée à la libre circulation et à l'exploration intensive des résultats scientifiques. A l'heure où l'application de la loi s'engage en France, la

référence européenne est évidement autant un horizon qu'une force nouvelle : la compétitivité du travail scientifique est à la clef de cette nouvelle étape de Science ouverte engagée sous l'impulsion du Commissaire européen Carlos Moedas, en charge de la Recherche.

Qu'il soit permis sur ces bases de tirer quatre leçons de ce Guide stratégique :

- **Agir de manière globale** : de nombreuses interactions juridiques, scientifiques et techniques sont à prendre en considération, souvent de façon spécifique à chaque écosystème de la recherche. Des analyses systémiques précises permettront d'éviter des approches trop verticales et formelles, en se nourrissant bien entendu des exemples étrangers.
- **Se positionner au niveau européen** : notre pays entretient désormais sur ces sujets une relation dynamique avec la Commission européenne où l'expérience de la France a sa place. Les projets de révision en cours des directives européennes sont à suivre de près et donneront leur feuille de route aux grandes infrastructures numériques qui vont stimuler la compétitivité de la recherche communautaire.
- **Optimiser les interactions entre les dispositions de la loi** : la libre disposition des données publiques, le libre partage des publications scientifiques, la fouille de textes et de données sont trois approches indissociables, qui doivent faire l'objet d'une vision générale au service du chercheur mais aussi des usagers et des bénéficiaires de la recherche publique.
- **Ouvrir une phase d'expérimentation** : on ne saurait nier l'importance et la taille des changements qui se dessinent ainsi dans l'organisation numérique du travail de la Science, tout comme on ne saurait sous-estimer la complexité des interactions qui vont être à l'œuvre dans le système numérique de l'IST. Il faudra semble-t-il expérimenter, évaluer les formules nouvelles en évitant de les figer d'emblée. Cette tâche nécessitera sans doute de rédiger des rapports d'étape.

La nouvelle ingénierie des connaissances numériques poursuit ainsi sa mue. Un effort de rattrapage est amorcé : il conduit à une nouvelle étape de compétitivité pour la recherche, étape qu'ont déjà anticipée les grandes universités motrices en la matière regroupées dans la LERU.

Telles sont les idées clés de ce Guide stratégique, construit pour être un vecteur d'application de cette loi fondatrice pour le travail de la science et le partage de l'IST.

Renaud Fabre, Directeur de la DIST du CNRS

Les signataires du Livre blanc « Une science ouverte dans une République numérique » :

Les membres du Comité exécutif du projet d'Investissement d'Avenir ISTEX :

Grégory COLCANAP, Coordonnateur du Consortium Couperin

Renaud FABRE, Directeur de la DIST du CNRS

Jérôme KALFON, Directeur de l'ABES (jusqu'en septembre 2016)

Jean-Marie PIERREL, Professeur à l'Université de Lorraine

Laurent SCHMITT, Responsable du département Projets et Innovation, Inist-CNRS

Les Grands Témoins :

Alain BERETZ, Président de l'Université de Strasbourg

Jean CHAMBAZ, Président de l'UPMC

Bruno CHAUDRET, Président du Conseil Scientifique du CNRS

Bruno DAVID, Président du Muséum National d'Histoire Naturelle

Daniel EGRET, Astronome (PSL) ancien Président de l'Observatoire de Paris

Claude KIRCHNER, Conseiller du Président d'Inria, Directeur de Recherche

Benoit THIEULIN, Président du Conseil National du Numérique

Introduction

Partage et liberté d'analyse des textes et données scientifiques

Un guide stratégique et opérationnel…

L'objectif de ce Guide stratégique d'applications est de présenter à l'ensemble des communautés scientifiques, des parlementaires, des éditeurs scientifiques et de manière générale au grand public les applications pratiques des dispositions légales nouvelles introduites par la loi pour une République numérique dans le domaine des pratiques numériques de la Science.

… dans le prolongement des articles 30 et 38 de la loi pour une République numérique (Petite loi[1])…

Ce Guide constitue un rapport des commentaires et d'analyse des articles 30 et 38 de la loi pour une République numérique, qui introduit en droit français les bases légales d'une science ouverte en créant :

- un droit de mise à disposition des publications scientifiques après le respect d'une période d'embargo (article 30) ;
- un droit à l'exploration ou à la fouille de textes et de données aux fins de recherche publique par le biais d'une exception au droit d'auteur et au droit du producteur de base de données (article 38).

… dans le prolongement du Livre blanc « Une science ouverte dans une République numérique » …

Ce Guide s'inscrit dans le prolongement du Livre blanc « Une Science ouverte dans une République numérique » publié en mars 2016 par le CNRS pour le compte du projet ISTEX et qui a servi de fil conducteur pour les débats préparatoires à l'adoption de la loi.

Ce Livre blanc exprimait les pratiques et besoins des chercheurs quant à l'utilisation de l'information scientifique et technique et des outils numériques. Il présentait également une analyse comparée des textes relatifs au text and data mining (TDM) à l'étranger. Ces éléments démontraient la nécessité pour la recherche publique d'introduire dans notre législation un droit nouveau.

[1] Texte de la Petite loi de la loi pour une République numérique : http://www.senat.fr/petite-loi-ameli/2015-2016/744.html

Il était le fruit :

- d'un travail collectif initié dans le cadre d'ISTEX (ISTEX, Initiative d'excellence de l'Information Scientifique et Technique, est un projet de plateforme numérique multi-usages aux meilleurs standards internationaux, accessible à distance par l'ensemble des communautés scientifiques et offrant « tous les moyens accessibles de consultation et d'analyse aujourd'hui disponibles dans toutes les communautés de la Science »[2]) ;
- de témoignages forts de grands témoins du monde de la recherche : des universités, la League of European Research Universities (LERU), le Conseil scientifique du CNRS, l'Agence bibliographique de l'enseignement supérieur (ABES), Couperin et l'Université de Lorraine pour le compte de la conférence des présidents d'universités (CPU) en tant que membres du Comité exécutif ISTEX, le Comité d'éthique du CNRS, le Conseil national du numérique ;
- d'un travail d'expertise juridique mené par le cabinet Alain Bensoussan.

Le Livre blanc préconisait les lignes directrices suivantes[3] :

Directions principales :

- **Créer** : Créer un droit de l'Open science garantissant le libre accès et la libre réutilisation des données de la recherche publique.
- **Equilibrer** : Redéfinir les équilibres économiques de l'écosystème numérique de la science.
- **Sécuriser** : Adopter l'article 18 bis (nouveau) du projet de loi pour une République numérique tel qu'issu de la commission mixte paritaire [nouvel article 38 de la Petite loi] créant une exception au droit d'auteur et au droit du producteur de base de données en faveur du text and data mining sur les données de la recherche publique (articles et données de la recherche) afin de sécuriser les pratiques de traitement automatisé de données et réduire les risques de captation abusive.
- **Rivaliser** : Permettre à la recherche publique française de disposer des

[2] http://www.istex.fr/

[3] Livre blanc « Une Science ouverte dans une République numérique » Mars 2016, page 12

moyens légaux et techniques au moins équivalents à ceux de ses homologues européens et américains et en phase avec le mouvement international de l'Open science.

- **Protéger** : Protéger les intérêts légitimes : valorisation, secret, brevet, droit d'auteur, vie privée et données personnelles.

La loi pour une République numérique a transposé la majorité des propositions figurant dans le Livre blanc, ce dont se félicite l'ensemble des signataires du Livre blanc et du présent Guide.

Afin d'accompagner la rédaction des décrets d'application prévus notamment à l'article 38 de la loi pour une République numérique, le présent Guide propose une discussion de la notion de TDM et de ses enjeux ainsi qu'une analyse comparée des structures existantes en France et à l'étranger.

… proposant une analyse des implications du libre accès aux publications scientifiques …

Si l'article 30 de la loi pour une République numérique posant le principe de libre accès aux publications scientifiques ne fait pas référence à un décret d'application, des précisions pourraient être apportées, un cadre d'application et des valeurs pourraient être affirmés.

… proposant une analyse de la notion de TDM et de la chaine de valeur associée…

La notion même de text and data mining recouvre des réalités diverses selon l'angle d'analyse juridique, technique ou économique choisi. Si le TDM implique un chercheur, un sujet de recherche et des outils d'analyse automatique, d'autres acteurs interviennent dans la chaine de valeur :

- les éditeurs scientifiques ;
- les auteurs de publications scientifiques ;
- les chercheurs, laboratoires et instituts de recherche ;
- les correspondants IST ;
- les éditeurs d'outils numériques d'analyse ;
- les éditeurs de plateformes de dépôt et d'accès aux données scientifiques ;
- les hébergeurs de données ;
- les startups ou autres entreprises proposant des services innovants.

L'ensemble de ces acteurs ainsi que les nouveaux acteurs qui se développeront en marge du TDM forment l'écosystème complexe de la donnée scientifique. Par ailleurs, le monde de l'analyse de données numériques est aujourd'hui dominé par des grandes firmes américaines. Le développement et l'utilisation des outils numériques d'analyse sont des opportunités pour la recherche publique française qui entre dans l'ère nouvelle de la science numérique.

… proposant un benchmark des effets du TDM dans les pays ayant légalisé la pratique…

Ce Guide propose un benchmark et une analyse comparée des approches techniques, juridiques et économiques du TDM dans les pays ayant déjà légalisé la pratique ainsi que les enjeux et leviers du TDM constatés dans ces pays.

L'approche choisie est celle de l'observation des pratiques étrangères afin de proposer, au regard des besoins et spécificités françaises, le cadre juridique et organisationnel idéal d'application.

… formulant des propositions d'encadrement de l'application des dispositions légales relatives à la science ouverte.

Le Guide formule des propositions d'encadrement de l'application des dispositions légales introduites par la loi pour une République numérique notamment :

- un encadrement juridique par la définition des contours de la notion de text and data mining et de son périmètre d'application ;
- un encadrement technique par la création de plateformes interopérables, impliquant la définition de standards, permettant l'accès à l'ensemble de l'information scientifique et technique ainsi qu'à des outils de text and data mining. La plateforme ISTEX pourrait devenir précurseur de ce dispositif ;
- un encadrement structurel par la création d'un réseau de « conservateurs de données » agréés dont la mission serait de conserver les fichiers produits au terme des activités de recherche et d'organiser leur mise à disposition ;
- un encadrement éthique par la définition de bonnes pratiques de l'usage du TDM dans la recherche scientifique ;
- ce cadre juridique, éthique, organisationnel, structurel pourrait être chapeauté par une Agence nationale pour la Science ouverte chargée de la gouvernance de l'IST et garante de son efficience.

1. Le libre accès aux publications scientifiques

Data access management database © Coloures-pic - Fotolia.com.jpg

1. LIBRE ACCÈS AUX PUBLICATIONS SCIENTIFIQUES

1. Le Livre blanc « Une Science ouverte dans une République numérique » porte et argumente un double constat :

- l'état des lieux des usages de la recherche publique française dénote un fort besoin de rattrapage, là où aujourd'hui, les usages numériques de la science sont aujourd'hui en décalage avec les grandes pratiques émergentes et/ou installées dans les grands pays de science ; ce que la stratégie du CNRS « Mieux partager les connaissances »[4] a fait émerger ;
- les changements en cours doivent aller vers un « droit des ressources partagées et des usages protégés », vers la création d'un droit de l'Open science garantissant le libre accès et la libre réutilisation des données de la recherche publique.

2. La loi pour une République numérique a consacré par l'article 30 cette nécessité pour les chercheurs de disposer des travaux de leurs confrères et de créer un droit à l'accès et au partage de la connaissance, (1.2), répondant ainsi au risque de captation abusive de la connaissance (1.1) et s'inscrivant dans la tendance de nos voisins européens.

1.1 L'Open access : une réponse au risque de « captation abusive »

3. Les chercheurs ont besoin, pour mener à bien leurs travaux de recherche, de pouvoir accéder librement aux données scientifiques mais également aux publications de leurs pairs (en tant que résultat de recherche faisant l'objet d'une publication par un éditeur privé).

4. Les modèles économiques (auteur-payeur ou lecteur-payeur) et juridiques (cession exclusive de droits, contrat d'abonnement) de l'édition scientifique entraînent une forme de captation de la connaissance scientifique par les éditeurs privés. Si certains éditeurs autorisent après une période d'embargo le dépôt de l'article dans une archive institutionnelle, d'autres conservent l'intégralité des droits pendant toute la durée de protection du droit d'auteur (70 ans à compter de la mort de l'auteur).

[4] http://www.cnrs.fr/dist/strategie-ist.htm

5. Ces modèles de financement de la publication à l'heure du numérique ont conduit les institutions de recherche publiques vers un accès payant et restreint aux connaissances issues des programmes de travaux qu'elles financent.

6. L'étude menée par la Direction de l'Information Scientifique et Technique (DIST) du CNRS « Financer la publication scientifique - Le « Lecteur » et / ou « l'Auteur » ? » (janvier 2016) expose précisément cette nécessité de réformer les modèles « d'auteur payeur » (paiement d'Article Processing Charges), de « lecteur payeur » (la souscription d'abonnement) ainsi que le développement d'un modèle hybride, au regard notamment des impacts financiers et des risques de privatisation de la connaissance.

7. Ce modèle n'étant plus viable notamment économiquement, l'ouverture des publications scientifiques doit être organisée par la création d'un droit d'accès, pour la recherche publique, à l'ensemble des publications. Dans son étude, le CNRS précise que :

> « **L'objectif à atteindre est celui d'une sécurisation globale sur tous les paramètres d'évolution vers la science ouverte**. Faute de cette sécurisation globale, la juxtaposition des "négociations" nationales, dont les contenus et les résultats ne sont aujourd'hui pas révélés, place les éditeurs en situation d'arbitres de la circulation de l'IST numérique. Cette situation contient le risque, comme l'observe l'OCDE, du chacun pour soi, de la confusion et du morcellement des collaborations scientifiques internationales, sous le jeu d'intérêts éditoriaux étrangers au partage des résultats de la recherche publique. »

1.2 La consécration légale d'un droit à l'Open access

8. Ces dérives et risques de privatisation de la connaissance sont largement partagés et affirmés par l'ensemble des communautés scientifiques et notamment par les établissements d'enseignement supérieur dont les montants des abonnements aux plateformes des éditeurs augmentent de façon exponentielle.

9. Ces éléments ont été accueillis par le législateur qui a introduit dans la loi pour une République numérique le principe d'un Open access à la française aux publications scientifiques.

1.2.1 L'article 30 de la loi pour une République numérique : l'Open access

10. La loi pour une République numérique est venue consacrer ce droit d'accès aux publications scientifiques dans les termes suivants :

- Le chapitre III du titre III du livre V du code de la recherche est complété par un article L. 533-4 ainsi rédigé :

  « Art. L. 533-4. – I. – Lorsqu'un écrit scientifique issu d'une activité de recherche financée au moins pour moitié par des dotations de l'État, des collectivités territoriales ou des établissements publics, par des subventions d'agences de financement nationales ou par des fonds de l'Union européenne est publié dans un périodique paraissant au moins une fois par an, son auteur dispose, même après avoir accordé des droits exclusifs à un éditeur, du droit de mettre à disposition gratuitement dans un format ouvert, par voie numérique, sous réserve de l'accord des éventuels coauteurs, la version finale de son manuscrit acceptée pour publication, dès lors que l'éditeur met lui-même celle-ci gratuitement à disposition par voie numérique ou, à défaut, à l'expiration d'un délai courant à compter de la date de la première publication. Ce délai est au maximum de six mois pour une publication dans le domaine des sciences, de la technique et de la médecine et de douze mois dans celui des sciences humaines et sociales.

  La version mise à disposition en application du premier alinéa ne peut faire l'objet d'une exploitation dans le cadre d'une activité d'édition à caractère commercial.

  « II. – Dès lors que les données issues d'une activité de recherche financée au moins pour moitié par des dotations de l'État, des collectivités territoriales, des établissements publics, des subventions d'agences de financement nationales ou par des fonds de l'Union européenne ne sont pas protégées par un droit spécifique ou une réglementation particulière et qu'elles ont été rendues publiques par le chercheur, l'établissement ou l'organisme de recherche, leur réutilisation est libre.

  « III. – L'éditeur d'un écrit scientifique mentionné au I ne peut limiter la réutilisation des données de la recherche rendues publiques dans le cadre de sa publication.

  « IV. – Les dispositions du présent article sont d'ordre public et toute clause contraire à celles-ci est réputée non écrite. »

11. L'article organise de la manière suivante l'Open access :

- sur les publications :

  - l'article 30 prévoit un droit, pour l'auteur d'un écrit scientifique, de mise à disposition gratuite dans un format ouvert par voie numérique de la version finale du manuscrit acceptée pour publication ;

- cette version pourra être mise à disposition soit immédiatement si l'éditeur met la publication en ligne gratuitement, soit après le respect d'un délai d'embargo ;
- les délais d'embargo sont de six mois dans le domaine des sciences, de la technique et de la médecine et de douze mois dans celui des sciences humaines et sociales, et ce en conformité avec les recommandations européennes ;
- les clauses de cession exclusives de droit prévues dans les contrats d'édition n'entravent pas le droit de mise à disposition de l'auteur ;

- sur les données de la recherche :
  - les données de la recherche sont de libre réutilisation dès lors que l'établissement de recherche les a rendues publiques ;
  - l'éditeur ne peut pas réserver la propriété des données de recherche associées à une publication.
- les dispositions prévues dans cet article 30 sont d'ordre public et toute clause contraire est réputée non écrite.

1.2.2 La libre mise à disposition des écrits scientifiques

12. **Un besoin des chercheurs**. Il a été clairement exprimé et ce de manière consensuelle dans le cadre de la consultation publique sur la loi pour une République numérique, la nécessité de renforcer les droits des chercheurs à diffuser librement leurs travaux, lorsque ces travaux ont été financés par des fonds publics.

13. **Consécration légale**. Le législateur a introduit dans le code de la recherche française un droit pour l'auteur d'un écrit scientifique de mise à disposition gratuit de la **version finale du manuscrit accepté pour publication** lorsque cet écrit est issu d'une activité de recherche financée au moins pour moitié par des fonds publics.

14. Si l'intention des législateurs d'ouvrir l'accès et le partage des publications scientifiques ne peut être que saluée, certaines précisions doivent venir clarifier le texte et notamment la notion de « version finale du manuscrit accepté pour publication ».

15. **Précisions**. En effet, la loi (code de la propriété intellectuelle ou code du patrimoine) ne connait pas les notions de « manuscrit », « version auteur », « version éditeur », « prépublication », « postpublication » … Ces termes issus de la pratique doivent être

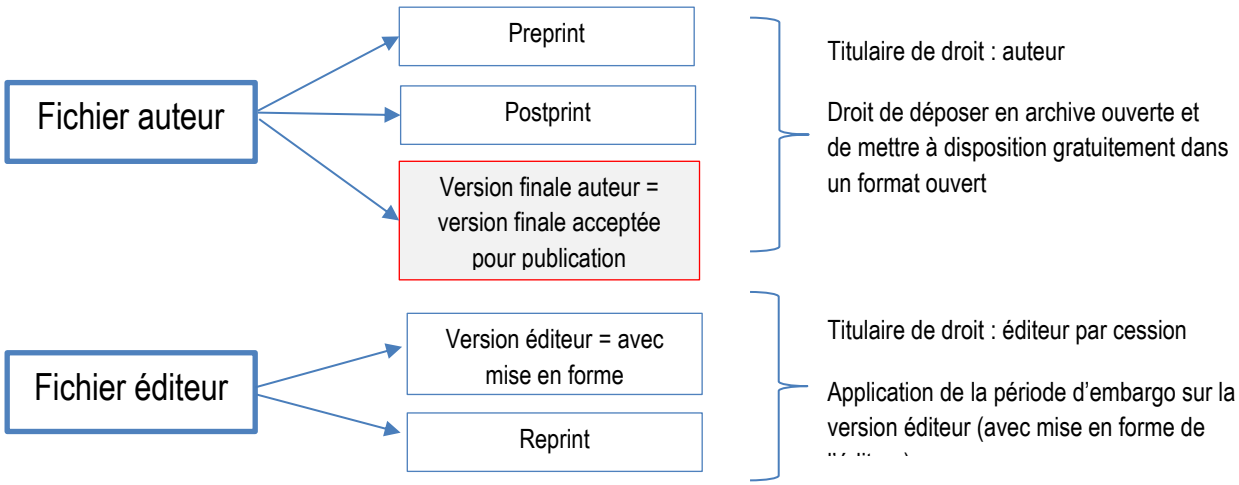
définis, qualifiés juridiquement et doivent être associés à un régime juridique (titularité des droits et droits d'exploitation associés).

16. La communauté scientifique et notamment les correspondants IST (CORIST) du CNRS se sont interrogés sur la définition des termes « manuscrit » et « version finale » au regard de la pratique. Dans son audition dans le cadre du Livre blanc, Claude Kirchner de l'INRIA résumait la problématique de la manière suivante :

> « Les éventuelles contraintes d'embargo ne peuvent porter que sur la ``version éditeur'' dans sa mise en forme finale et ce afin de respecter l'exploitation commerciale éventuelle. Elles ne sont acceptables que si la "version auteur" est effectivement libre de diffusion, et la durée de l'embargo devrait alors être fixée en cohérence avec les pratiques internationales. »
>
> (Audition INRIA, Claude Kirchner, 15 octobre 2015)[5]

17. Les décrets d'application sont l'occasion de proposer des définitions relevant de la pratique et de l'usage dans l'édition scientifique.

18. Pour ce faire, le présent Guide propose la création d'un **référentiel des usages**, référentiel qui peut contenir une nomenclature et une définition des termes utilisés en pratique, ainsi que le régime applicable pour les différentes versions de l'article. La typologie des différentes versions d'un article sont les suivantes :



[5] Audition Claude Kirchner, 15 octobre 2015, Livre blanc page 73

19. La notion de « version finale du manuscrit acceptée pour publication » semble vouloir désigner la dernière version de l'auteur avant publication et donc avant mise en forme par l'éditeur. Par conséquent, l'article 30 de la loi pour une République numérique pourrait être précisé par décret afin de clarifier la version objet de l'embargo.

⚠ Décret : création d'un référentiel des usages et précision quant à la version du manuscrit objet de l'embargo.

1.2.3 L'absence d'effet des clauses de cession exclusive de droits

20. Le texte de l'article 30 prévoit que le droit du chercheur de mettre à disposition gratuitement ses publications scientifiques s'applique « même après avoir accordé des droits exclusifs à un éditeur ».

21. Conscient que le contrat d'édition entre un chercheur et un éditeur prend le plus souvent la forme d'un contrat d'adhésion, le législateur a choisi de rendre inefficient la clause de cession exclusive de droits d'auteur pour les besoins de l'Open access.

22. **Proposition : contrat-type.** Afin de garantir les droits des chercheurs sur leur publication et de prendre en compte les risques d'asymétrie contractuelle, un décret pourrait également organiser un contrat type de cession de droits d'auteur destiné à la recherche publique.

23. Ce contrat définirait les règles du jeu entre les parties et la protection du chercheur dans sa relation avec l'éditeur. Il permettrait notamment de s'assurer de l'absence de cession à titre exclusif et de garantir les droits des chercheurs :

- d'autoriser le dépôt et la reproduction en archive ouverte de la publication dans la version auteur immédiatement et dans la version éditeur après le respect d'une période d'embargo ;
- de permettre l'exploration immédiate du contenu de l'article à partir d'outils numériques de traitement de données ;
- d'empêcher toutes formes de privatisation ou de réserve de propriété sur le contenu de l'article et les données associées.

24. Ce contrat pourrait faire l'objet d'un décret et ainsi avoir une valeur réglementaire qui s'imposerait à l'éditeur pour toute publication scientifique constituant un résultat de la recherche publique.

⚠ Décret : création d'un contrat type de cession de droits destiné aux publications scientifiques

1.2.4 Les recommandations de la communauté européenne sur les délais d'embargo

25. A la recherche d'un équilibre entre les positions des différents acteurs en présence à l'heure du numérique et de la société de la connaissance, le Gouvernement a introduit dans la loi :

- l'ouverture de la possibilité d'une diffusion en accès libre des travaux scientifiques financés sur fonds publics, au terme d'une durée dite « d'embargo » ;
- des délais « d'embargo » de 6 et 12 mois, au terme desquels l'auteur d'une publication financée sur fonds publics peut, au plus tard, mettre librement à disposition son écrit. Si l'article est mis à disposition gratuitement par l'éditeur en ligne, l'auteur pourra immédiatement faire usage de son droit.

26. **Recommandations CE**. Les délais d'embargo fixés par la loi correspondent aux délais maximaux prévus par la recommandation de la Commission européenne (C(2012) 4890)[6] :

Il est recommandé aux Etats membres :

- « de définir des politiques claires en matière de diffusion des publications scientifiques issues de la recherche financée par des fonds publics et du libre accès à ces dernières. Ces politiques devraient prévoir:
  - des objectifs et des indicateurs concrets permettant de mesurer les progrès accomplis,
  - des plans de mise en œuvre, incluant la répartition des responsabilités,
  - la programmation financière correspondante » ;
- de veiller à ce que « les publications issues de la recherche financée par des fonds publics soient librement accessibles dans les meilleurs délais, de préférence immédiatement et, dans tous les cas, au plus tard six mois après leur

6 https://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_fr.pdf

date de publication, et au plus tard douze mois pour les publications dans les domaines des sciences sociales et humaines ».

27. **Dispositions étrangères**. Les délais d'embargo français correspondent ou sont proches également de ceux prévus par les législations nationales de nos voisins européens :

- en Allemagne : délai d'embargo de 12 mois sans distinction entre les disciplines ;
- en Espagne : dépôt dans une archive institutionnelle le plus rapidement possible, sans dépasser 12 mois, sans distinction entre les disciplines.

1.2.5 L'interdiction de privatisation des données de la recherche

28. **Besoin de partage**. Le Code de la recherche définit parmi les missions de la recherche publique (article L.112-1 du Code de la recherche) :

- « le partage et la diffusion des connaissances scientifiques » ;
- « l'accès libre aux données scientifiques ».

Toutes les communautés scientifiques s'accordent à affirmer la nécessité d'avoir un accès libre et massif aux données de la recherche, au nom de l'intérêt supérieur de la recherche dont les enjeux sont multiples.

29. Dans un article « Préserver les données de la recherche à l'ère du Big Data »[7], la problématique de la conservation mais également du partage des données de la recherche est parfaitement et exhaustivement appréhendée.

> « Alors qu'on assiste à une explosion du volume des données produites par la recherche, la question de leur archivage est devenue cruciale, tant pour pérenniser notre héritage scientifique que pour permettre leur réutilisation par la communauté. (…) À mesure que les instruments et les outils d'analyse se perfectionnent, la quasi-totalité des disciplines fait face à une explosion du volume de données produites chaque année. Et ces données sont précieuses, car elles sont très souvent issues d'expériences complexes et coûteuses comme en physique des hautes énergies, ou sont le fruit d'observations ponctuelles sur une

[7] Préserver les données de la recherche à l'ère du Big Data du 9-9-2016 par Guillaume Garvanèse
https://lejournal.cnrs.fr/articles/preserver-les-donnees-de-la-recherche-a-lere-du-big-data

> longue période de temps à l'instar du suivi de la position des objets stellaires ou des relevés démographiques. »

30. L'article 30 de la loi pour une République numérique traduit cette nécessité de libérer et d'ouvrir l'accès « aux données issues d'une activité de recherche » mais également d'empêcher toute privatisation notamment par contrat d'édition de ces données.

> « II. – Dès lors que les données issues d'une activité de recherche financée au moins pour moitié par des dotations de l'État, des collectivités territoriales, des établissements publics, des subventions d'agences de financement nationales ou par des fonds de l'Union européenne ne sont pas protégées par un droit spécifique ou une réglementation particulière et qu'elles ont été rendues publiques par le chercheur, l'établissement ou l'organisme de recherche, leur réutilisation est libre.
>
> « III. – L'éditeur d'un écrit scientifique mentionné au I ne peut limiter la réutilisation des données de la recherche rendues publiques dans le cadre de sa publication. »

31. Le texte prévoit un principe de libre réutilisation des données de la recherche publique. Toutefois, le périmètre de « ces données issues d'une activité de recherche financée au moins pour moitié par des dotations » publiques n'est pas précisé et les modalités de partage et d'accès à ces données ne sont pas définies.

32. Ces précisions nécessaires à une bonne gouvernance des données de la recherche et de la science ouverte doivent figurer dans un décret d'application. En effet, valeur importante et fondement historique de la démarche scientifique, le partage des connaissances est le moteur même de la recherche. La transition numérique a bouleversé la pratique par un accès à une masse de données grandissante et globale, de manière instantanée et ce partout dans le monde. Ce big data des données scientifiques entraine le développement d'outils, de pratiques d'exploration intelligente par des services d'analyse et d'observation automatique des données.

33. L'utilisation de ces outils de text and data mining et l'avènement d'une pratique scientifique nouvelle, transverse et multidisciplinaire, sont le terrain d'enjeux multiples, scientifique mais également humain, économique, éthique. Le législateur a compris ces enjeux et la nécessité d'introduire ce droit au text and data mining dans la législation française, permettant à la recherche française de rivaliser avec ses homologues anglais, américains ou encore canadiens. La mise en œuvre de ces principes doit s'inscrire dans une organisation garante de son efficience.

2. La liberté d'analyse

Data mining © dizain - Fotolia.com.jpg

2. LIBERTÉ D'ANALYSE DES RÉSULTATS DE LA SCIENCE

34. Le Text and Data Mining (TDM) est un ensemble de techniques permettant d'explorer et de traiter de vastes corpus de textes et de données. Il ouvre des champs de recherche nouveaux et autorise de nouvelles approches méthodologiques de construction de connaissances. Outil au potentiel toujours à développer, le TDM répond à des enjeux scientifiques et économiques. D'un côté, il permet d'intensifier et de stimuler la recherche. De l'autre, il peut être valorisé économiquement et peut constituer un gain de coût et de temps pour l'économie de la recherche. Il est aussi un facteur d'amélioration des décisions publiques.

35. La notion de TDM recouvre des réalités diverses qu'il convient de clarifier d'un point de vue juridique, technique et économique (2.1). Les enjeux et leviers du TDM pour la recherche scientifique mais également de manière plus globale sont considérables et font entrer la science dans une nouvelle ère (2.2) impliquant un nécessaire et multiple encadrement des pratiques. Afin de définir une organisation structurante, il est proposé d'analyser des pratiques comparables (2.3) : les organisations relatives au TDM au Royaume Uni et aux Etats Unis (2.3.1) ; l'encadrement proposé dans les projets financés dans le cadre de H2020 (2.3.2) et enfin les organisations existant déjà en France (2.3.3).

2.1 Qu'est-ce que le text and data mining ?

36. Le data mining est un concept jeune qui apparaît en 1989 sous un premier nom de KDD (Knowledge Discovery in Databases, en français ECD pour Extraction de Connaissances à partir des Données).

37. Le terme de « text and data mining » est apparu pour la première fois dans le domaine du marketing au début des années 1990. Ce concept, tel qu'appliqué aux services marketing, est étroitement lié au concept du « one-to-one relationship » (Michael Berry et Gordon Linoff, créateurs du data mining dans le marketing), c'est-à-dire à la personnalisation des rapports entre l'entreprise et sa clientèle.

38. Si le domaine d'application du TDM qui intéresse le présent Guide est celui de la Science, la pratique du TDM est utilisée dans de nombreux secteurs d'activité comme par exemple[8] :

[8] http://www.rithme.eu/?m=resources&p=dmdomains&lang=fr

- le marketing direct : dans ce domaine des techniques de TDM sont par exemple utilisées pour segmenter les bases de données clients et pour prédire leur intention d'achat afin d'optimiser le discours marketing ;
- la communication : le filtrage anti-spam des courriers électroniques ou encore le système Echelon, système mondial d'interception des communications privées et publiques (SIGINT), élaboré par les États-Unis, le Royaume-Uni, le Canada, l'Australie et la Nouvelle-Zélande dans le cadre du traité UKUSA, sont des techniques de TDM ;
- le secteur bancaire et la finance ;
- l'assurance et la santé ;
- les secteurs médical et pharmaceutique.

39. Le développement des pratiques du TDM a vu le jour ses dernières années dans le domaine scientifique avec le développement des archives ouvertes de type arXiv ou HAL afin d'optimiser la recherche sur ses bases au volume croissant de données.

40. La notion de text and data mining ou encore d'exploration ou de fouille de textes et de données appliquée au domaine scientifique est aujourd'hui largement utilisée pour désigner des activités, des outils divers. Il est par conséquent proposé d'analyser la notion de TDM sous l'angle juridique, technique et économique afin de répondre aux questions suivantes :

- Qu'est-ce que le TDM ?
- Quelles sont les opérations que le TDM mobilise ?
- Dans quels domaines le TDM s'applique-t-il ?
- Comment mesurer l'efficience du TDM ?

2.1.1 Le droit du TDM

2.1.1.1 La consécration légale d'un droit au TDM par une exception

41. **Double exception**. L'article 38 de la loi (Petite loi) consacre un droit au text and data mining en introduisant une exception au droit d'auteur et au droit du producteur de base de données selon les termes suivant :

Le Code de la propriété intellectuelle est ainsi modifié :

1° Après le second alinéa du 9° de l'article L. 122-5, il est inséré un 10° ainsi rédigé :

« 10° Les copies ou reproductions numériques réalisées à partir d'une source licite, en vue de l'exploration de textes et de données incluses ou associées aux écrits scientifiques pour les besoins de la recherche publique, à l'exclusion de toute finalité commerciale. Un décret fixe les conditions dans lesquelles l'exploration des textes et des données est mise en œuvre, ainsi que les modalités de conservation et de communication des fichiers produits au terme des activités de recherche pour lesquelles elles ont été produites ; ces fichiers constituent des données de la recherche ; »

2° Après le 4° de l'article L. 342-3, il est inséré un 5° ainsi rédigé :

« 5° Les copies ou reproductions numériques de la base réalisées par une personne qui y a licitement accès, en vue de fouilles de textes et de données incluses ou associées aux écrits scientifiques dans un cadre de recherche, à l'exclusion de toute finalité commerciale. La conservation et la communication des copies techniques issues des traitements, au terme des activités de recherche pour lesquelles elles ont été produites, sont assurées par des organismes désignés par décret. Les autres copies ou reproductions sont détruites. »

42. **Absence de définition**. Le texte ne pose pas de définition de la notion même d'exploration ou de fouille de données. L'utilisation de ces deux termes dans un même texte appelle une remarque : l'utilisation du terme « exploration » de textes et de données dans la première partie du texte introduisant une exception au droit d'auteur et celui de « fouille » dans la seconde partie créant une exception au droit du producteur de base de données risque de soulever des problématiques d'interprétation. Le décret d'application pourrait à titre introductif préciser que les notions d'exploration et de fouille recouvrent les mêmes pratiques.

⚠ Le décret d'application pourrait à titre introductif préciser que les notions d'exploration et de fouille recouvrent les mêmes pratiques. Le décret doit à la manière des directives européennes comporter un article « Définition ».

43. **Encadrement de la notion**. Si les notions même de fouille et d'exploration de textes et de données ne sont pas définies, le texte pose des limites et un cadre à cette pratique:

Critères	Article 38
Fondement	Exception au droit d'auteur et au droit du producteur de base de données : droit de copie et de reproduction numérique aux fins de TDM
Périmètre du TDM	Fouille de textes et de données incluses ou associées aux écrits scientifiques
Bénéficiaire de l'exception	/
Limites	TDM limité aux besoins de la recherche scientifique / dans un cadre de recherche But non-commercial Source licite / accès licite aux textes et données objets du TDM

2.1.1.2 L'introduction d'une exception TDM dans le projet de directive Droit d'auteur dans le marché unique numérique

44. **Rapports préliminaires**. Le Livre blanc « Une Science ouverte dans une République numérique » relevait les nombreux rapports, dont certains commandés par la Commission européenne, qui préconisaient la révision de la directive 2001/29/CE « Droit d'auteur et droit voisin dans la société de l'information » et l'introduction d'un droit au TDM :

- le rapport Sirinelli pour le Conseil supérieur de la propriété littéraire et artistique (CSPLA) « Rapport de la mission sur la révision de la directive 2001/29/CE sur l'harmonisation de certains aspects du droit d'auteur et des droits voisins dans la société de l'information » de décembre 2014 demande la « création de nouvelles exceptions au droit d'auteur concernent notamment les activités dites de *text and data mining* (TDM) »[9] ;

[9] Rapport CSPLA page 8

- l'étude du cabinet Wolf & Partners de mars 2014, intitulée « Study on the legal framework of *text and data mining* »[10] pour la Commission européenne ;

- un groupe d'experts de la Commission européenne a également publié en avril 2014 un rapport intitulé « Standardisation in the area of innovation and technological development, notably in the field of *Text and data mining* »[11] ;

- le rapport Reda : ce rapport adopté par le Parlement européen le 9 juillet 2015 affirme « l'impératif d'évaluer avec soin la mise à disposition des techniques analytiques automatisées des textes et des données (par exemple la "fouille de textes et de données") à des fins de recherche. » ;

- le communiqué de presse de la Commission européenne du 9 décembre 2015 présentant les mesures pour améliorer l'accès aux contenus en ligne et présentant sa vision d'un droit d'auteur modernisé. Dans ce cadre, la Commission annonçait avoir « l'intention de travailler sur les exceptions au droit d'auteur » et notamment de réviser « les règles de l'Union afin de permettre aux chercheurs d'utiliser plus facilement les techniques de fouille data mining et de text mining pour analyser de grandes séries de données. »

45. **Projet de directive**. Le projet de directive Droit d'auteur dans le marché unique numérique (COM(2016) 593 final) a été publié par la Commission européenne le 14 septembre 2016.

46. Par ce projet de directive Droit d'auteur dans le marché unique numérique (COM(2016) 593 final)[12], la Commission a pour objectif « de moderniser les règles de l'UE sur le droit d'auteur pour favoriser l'essor et la diffusion de la culture européenne ». « Les propositions mettront également des outils permettant d'innover à la disposition de l'enseignement, de la recherche et des institutions du patrimoine culturel »[13]. L'objectif de cette directive est d'adapter les dispositions relatives au droit d'auteur à l'utilisation croissante des technologies numériques notamment dans le domaine de la recherche

[10] http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf

[11] http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf

[12] Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market, COM(2016) 593 final, 14-9-2016 http://ec.europa.eu/transparency/regdoc/rep/1/2016/EN/1-2016-593-EN-F1-1.PDF

[13] Communiqué de presse du 14 septembre 2016

scientifique, constatant l'application protéiforme des dispositions de la directive DADVSI et notamment des exceptions[14].

47. **Définition TDM**. L'article 2 de la directive propose une définition de la notion de text and data mining :

> "text and data mining means any automated analytical technique aiming to analyse text and data in digital form in order to generate information such as patterns, trends and correlations".

48. **Exception**. L'article 3 introduit une exception au droit d'auteur et au droit du producteur de base de données en faveur du text and data mining dans les termes suivants :

> Article 3 - Text and data mining
>
> 1. Member States shall provide for an exception to the rights provided for in Article 2 of Directive 2001/29/EC, Articles 5(a) and 7(1) of Directive 96/9/EC and Article 11(1) of this Directive for reproductions and extractions made by research organisations in order to carry out text and data mining of works or other subject-matter to which they have lawful access for the purposes of scientific research.
> 2. Any contractual provision contrary to the exception provided for in paragraph 1 shall be unenforceable.
> 3. Rightholders shall be allowed to apply measures to ensure the security and integrity of the networks and databases where the works or other subject-matter are hosted. Such measures shall not go beyond what is necessary to achieve that objective.
> 4. Member States shall encourage rightholders and research organisations to define commonly-agreed best practices concerning the application of the measures referred to in paragraph 3.

[14] Introduction d'une exception TDM au droit d'auteur dans la loi anglaise sur le fondement de l'article l'article 5, 3 a) de la Directive 2001/29/CE du 22 mai 2001 sur l'harmonisation de certains aspects du droit d'auteur et des droits voisins dans la société de l'information (Directive DADVSI). L'analyse des textes de la directive dans leur version anglaise et française ont permis de s'apercevoir que l'étendue de l'exception posée n'est pas la même selon la langue utilisée.

49. Cette exception peut être analysée selon les critères suivants (critères utilisés pour l'analyse du texte français) :

Critères	Projet de Directive
Fondement	Exception au droit d'auteur et au droit du producteur de base de données : droit de reproduction ou d'extraction aux fins de TDM
Périmètre du TDM	TDM sur des œuvres ou d'autres objets
Bénéficiaire de l'exception	Les organisations de recherche (la notion est définie de manière large à l'article 2 du projet de directive[15])
Limites	TDM limité aux besoins de la recherche scientifique But non-commercial Accès légal aux objets du TDM

50. La Commission justifie les lignes directrices de ce texte de la manière suivante :

> Text and data mining:
>
> - Option 1 consisted in self-regulation initiatives from the industry.
> - Other options consisted in the introduction of a mandatory exception covering text and data mining.
>   - In Option 2, the exception only covered uses pursuing a non-commercial scientific research purpose.
>   - Option 3 allowed uses for commercial scientific research purpose but limited the benefit of the exception to some beneficiaries.
>   - Option 4 went further as it did not restrict beneficiaries.

[15] Projet de Directive Art. 2 : "research organisation" : means a university, a research institute or any other organisation the primary goal of which is to conduct scientific research or to conduct scientific research and provide educational services:

(a) on a non-for-profit basis or by reinvesting all the profits in its scientific research; or

(b) pursuant to a public interest mission recognised by a Member State; in such a way that the access to the results generated by the scientific research cannot be enjoyed on a preferential basis by an undertaking exercising a decisive influence upon such organisation.

- Option 3 was deemed to be the most proportionate one.

51. La Commission précise que l'objectif de ce texte est de fournir une clarification juridique et un cadre de concurrence équitable afin que les chercheurs européens puissent utiliser des techniques innovantes d'analyse de données ; leur permettant de trouver plus rapidement des solutions novatrices en réponse aux défis majeurs tels que les épidémies mondiales et le changement climatique ; favorisant les collaborations transfrontalières et interdisciplinaires. Cette exception participe au soutien de la compétitivité européenne en favorisant l'Open science[16].

52. Carlos Moedas, Commissaire européen à la recherche, à l'innovation et à la science, a justifié la nécessité de cette exception de la manière suivante :

- "Science needs a copyright law that reflects the reality of the modern age. We must remove barriers that prevent scientists from digging deeper into the existing knowledge base. This proposed copyright exception will give researchers the freedom to pursue their work without fear of legal repercussions, and so allow our greatest minds to discover new solutions to major societal problems."

53. Si la France et l'Europe se dotent d'un arsenal législatif autorisant l'utilisation de techniques d'analyse automatiques, il est également intéressant de regarder les dispositions adoptées par d'autres pays.

2.1.1.3 Des périmètres mouvant de la notion de TDM à travers les textes nationaux

54. Le tableau ci-après présente une lecture analytique de la notion de TDM dans les législations anglaises, américaines, japonaises, autant de législations qui ont intégré légalement un droit au TDM.

[16] http://ec.europa.eu/research/index.cfm?pg=newsalert&year=2016&na=na-140916

PAYS	SOURCE	TEXTE	CARACTERISTIQUES DU TDM
Royaume-Uni	**Loi** Article 29 A introduit en 2014[16] dans le Copyright, Designs and Patents Act (1988)	"**29A Copies for text and data analysis for non-commercial research** (1) The making of a copy of a work by a person who has lawful access to the work does not infringe copyright in the work provided that : (a) the copy is made in order that a person who has lawful access to the work may carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose, and (b) the copy is accompanied by a sufficient acknowledgement (unless this would be impossible for reasons of practicality or otherwise). (2) Where a copy of a work has been made under this section, copyright in the work is infringed if : (a) the copy is transferred to any other person, except where the transfer is authorised by the copyright owner, or (b) the copy is used for any purpose other than that mentioned in subsection (1)(a), except where the use is authorised by the copyright owner. (3) If a copy made under this section is subsequently dealt	Fondement : Exception au droit d'auteur aux fins « d'analyse computationnelle » Périmètre : les œuvres et toutes données associées Bénéficiaire : / Limites : Accès licite. Fins non commerciales. Limité au seul but de la recherche. Mention de paternité. Interdiction de communication de la copie réalisée à un tiers / interdiction de conclure un contrat de cession ou licence de la copie réalisée

[16] http://www.legislation.gov.uk/uksi/2014/1372/pdfs/uksi_20141372_en.pdf

PAYS	SOURCE	TEXTE	CARACTERISTIQUES DU TDM
		with : (a) it is to be treated as an infringing copy for the purposes of that dealing, and (b) if that dealing infringes copyright, it is to be treated as an infringing copy for all subsequent purposes. (4) In subsection (3) "dealt with" means sold or let for hire, or offered or exposed for sale or hire. (5) To the extent that a term of a contract purports to prevent or restrict the making of a copy which, by virtue of this section, would not infringe copyright, that term is unenforceable."	
Etats-Unis	**Loi** Federal Agency Data Mining Reporting Act of 2007 : Article 42 U.S. Code § 2000ee-3	« **The term "data mining" means a program involving pattern-based queries, searches, or other analyses of 1 or more electronic databases**, where : (A) a department or agency of the Federal Government, or a non-Federal entity acting on behalf of the Federal Government, is conducting the queries, searches, or other analyses to discover or locate a predictive pattern or anomaly indicative of terrorist or criminal activity on the part of any individual or individuals; (B) the queries, searches, or other analyses are not subject-	Fondement : Droit positif de formuler des requêtes par un programme Périmètre : modèles, recherches, ou autres analyses d'une ou plusieurs bases de données électroniques Bénéficiaire : Agences fédérales Limites : Analyse prédictive dans le domaine d'activité terroriste ou criminelle Pas d'utilisation de données à caractère personnel

PAYS	SOURCE	TEXTE	CARACTERISTIQUES DU TDM
		based and do not use personal identifiers of a specific individual, or inputs associated with a specific individual or group of individuals, to retrieve information from the database or databases; and (C) the purpose of the queries, searches, or other analyses is not solely: (i) the detection of fraud, waste, or abuse in a Government agency or program; or (ii) the security of a Government computer system.»[17]	
Etats-Unis	**Décision de justice** Authors Guild v. HathiTrust, 755 F.3d 87 (2d Cir. 2014)	"The court held that the HDL's first use—creation of a full-text searchable database—was fair. It found that use "quintessentially transformative" because "the result of a word search is different in purpose, character, expression, meaning, and message from the page (and the book) from which it is drawn." The court further held that the copies were reasonably necessary to facilitate the HDL's services to the public and to mitigate the risk of disaster or data loss. In addition, it held that the full-text search posed no harm to	Fondement : La reproduction et l'utilisation de livres numériques à des fins de conservation, de recherche de texte et d'accessibilité pour les malvoyants ne constituent pas des atteintes au droit d'auteur car relèvent du « fair use ». Périmètre : Œuvre / base de données Bénéficiaire : Parties au litige

[17] https://www.law.cornell.edu/uscode/text/42/2000ee-3

PAYS	SOURCE	TEXTE	CARACTERISTIQUES DU TDM
		any existing or potential traditional market for the copyrighted works. The court also held that the second use—access for the print-disabled—was fair. It concluded that providing such access was a valid purpose under the first statutory factor, even though it was not transformative. The court held that it was reasonable for the defendants to retain both text and image copies because the text copies were required for text searching and text-to-speech capabilities, and the image copies provide an additional method by which many disabled patrons can access the works. Finally, the court held that the fourth factor favored fair use given the insignificance of the present-day market for books accessible to the handicapped."[18]	Limite : quatre critères du faire use - le but non lucratif - la nature de l'œuvre protégée par le droit d'auteur - la portion de l'œuvre utilisée - l'absence d'impact économique de l'usage
Etats-Unis	**Décision de justice** Authors Guild v. Google, Inc.[19]	"In sum, we conclude that: (1) Google's unauthorized digitizing of copyright-protected works, creation of a search functionality, and display of snippets from those works are non-infringing fair uses. The purpose of the copying is highly	Fondement : droit de mise à disposition par Google de certains passages de livres sous format numérique sur le fondement du fair use.

[18] http://copyright.gov/fair-use/summaries/authorsguild-hathitrust-2dcir2014.pdf

[19] https://www.authorsguild.org/wp-content/uploads/2015/10/CA2-Fair-Use-Ruling.pdf

PAYS	SOURCE	TEXTE	CARACTERISTIQUES DU TDM
	16 octobre 2015	transformative, the public display of text is limited, and the revelations do not provide a significant market substitute for the protected aspects of the originals. Google's commercial nature and profit motivation do not justify denial of fair use. (2) Google's provision of digitized copies to the libraries that supplied the books, on the understanding that the libraries will use the copies in a manner consistent with the copyright law, also does not constitute infringement. Nor, on this record, is Google a contributory infringer."	Périmètre : Œuvre Bénéficiaire : Parties au litige Limite : quatre critères du faire use - le but non lucratif - la nature de l'œuvre protégée par le droit d'auteur - la portion de l'œuvre utilisée - l'absence d'impact économique de l'usage. Portée : Avec cette jurisprudence, les Etats-Unis donnent un avantage significatif à ses chercheurs en leur ouvrant la possibilité de numériser des ensembles très larges de données accessibles licitement, de mutualiser les corpus et de développer des fonctionnalités de recherches et de traitements algorithmiques des données[20].

[20] Article « Comment l'affaire Google Books se termine en victoire pour le Text mining » 21-10-2015 http://scinfolex.com/2015/10/21/comment-laffaire-google-books-se-termine-en-victoire-pour-le-text-mining/

PAYS	SOURCE	TEXTE	CARACTERISTIQUES DU TDM
Japon	**Loi** Japan Copyright Act – Article 47 septies introduit en 2009	« For the purpose of information analysis ("information analysis" means to extract information, concerned with languages, sounds, images or other elements constituting such information, from many works or other much information, and to make a comparison, a classification or other statistical analysis of such information; the same shall apply hereinafter in this Article) by using a computer, it shall be permissible to make recording on a memory, or to make adaptation (including a recording of a derivative work created by such adaptation), of a work, to the extent deemed necessary. However, an exception is made of database works which are made for the use by a person who makes an information analysis."[21]	Fondement : Le texte prévoit une exception d'« analyse de l'information » aux fins de comparaison, classification, analyse statistique Périmètre : information de toute nature Bénéficiaire : / Limite : L'analyse de l'information n'est pas cantonnée à la recherche publique, ni à des fins non commerciales.

2.1.1.4 Analyse croisée des dispositions légales

55. L'analyse croisée des dispositions légales françaises, anglaises, américaines, japonaises ainsi que celles du projet de directive permettent de formuler les remarques suivantes au regard des quatre critères d'analyse utilisés (fondement légal, périmètre, bénéficiaires et limites du TDM)[17] :

Critères	Analyse croisée
Fondement	L'exception au droit d'auteur et au droit du producteur de base de données est privilégiée dans tous les pays ayant légalisé en faveur du TDM ; Le Royaume-Uni a préféré l'expression « d'analyse computationnelle » à celle de text and data mining.
Périmètre	Le périmètre du droit au TDM est plus ou moins large selon les législations : - la loi japonaise est la plus large en prévoyant que le TDM peut être pratiqué sur toute information ; - la loi anglaise et le projet de directive vise les œuvres et toutes données associées - la loi française vise les « textes et de données incluses ou associées aux écrits scientifiques »
Bénéficiaire	Les textes anglais et français ne prévoient pas de bénéficiaire expresse de la disposition. Toutefois, les bénéficiaires concernés sont indirectement les organismes de recherche publique ; le TDM étant limité à des fins de recherche publique. Le projet de directive limite de manière expresse l'utilisation du TDM aux organisations de recherche.
Limite	Trois limites sont principalement posées par l'ensemble des textes : - le TDM doit être limité aux besoins de la recherche scientifique ; - effectué dans un but non-commercial ; - sur les textes et données en accès licite.

[17] Le tableau d'analyse croisée figure en annexe 1.

56. Cette analyse croisée montre une certaine proximité des régimes du TDM dans les pays analysés.

57. Afin que le texte français ne soit pas en deçà des dispositions étrangères et du projet de directive, il est recommandé de préciser dans le décret d'application la notion de « textes et de données incluses ou associées aux écrits scientifiques » afin qu'elle soit la plus large possible.

⚠ Définition de la notion de TDM et son périmètre d'application. Clarification de la notion de « textes et de données incluses ou associées aux écrits scientifiques » afin qu'elle soit la plus large possible.

2.1.2 La technique du TDM

58. D'un point de vue technique, le TDM peut s'analyser au regard :

- des objets techniques nécessaires qu'il mobilise ;
- des opérations techniques qu'il nécessite ;
- de ses fonctionnalités générales.

2.1.2.1 Les objets techniques nécessaires au TDM

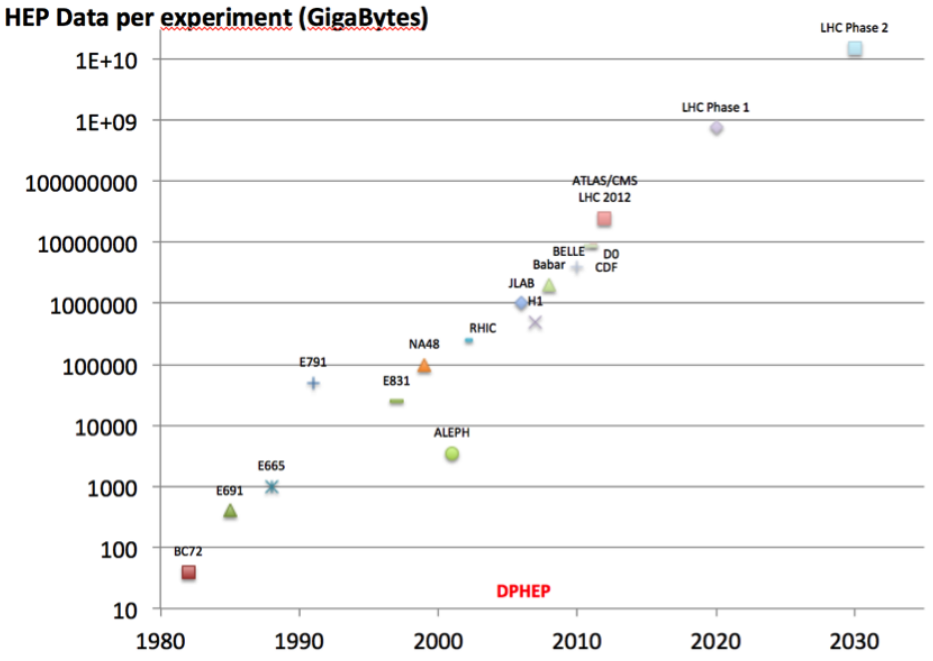
59. Pour que les pratiques de text and data mining puissent être mobilisées, le chercheur doit disposer :

- de données d'entrée ; les données input ;
- d'outils de traitement ;
- l'utilisation des outils de traitement sur les données d'entrée permettront de produire des données de sortie : les données output.

60. **La matière première : les données input**. Techniquement parlant, le TDM peut se décrire comme un processus d'extraction de connaissances à partir de textes et données sélectionnées, basé sur des mécanismes d'identification de structures jusque-là inconnues, scientifiquement valides et exploitables. Chaque champ scientifique (astrophysique, biologie moléculaire, sociologie des organisations, géologie marine, linguistique…) a développé son propre arsenal de moyens et techniques de collecte de données : capteurs, sondes, satellites, séquenceurs, cameras, numérisation, simulation,

analyses chimiques, … qui conduisent à l'accumulation de vastes ensembles de données (big data).

61. La production scientifique globale connaît également une croissance spectaculaire en raison de l'augmentation de l'activité scientifique au niveau mondial mais également d'une certaine pression à publier. La production annuelle globale de l'édition scientifique a largement progressé ; par exemple, elle est passée entre 1996 et 2012 de 1 134 000 à 2 250 000 articles/an sur la base Scopus d'Elsevier[18]. Toutefois, « malgré sa croissance rapide, l'effort de publication ne croît pas en proportion des données produites par la recherche » : plus de 90 % des données resteraient stockées sur des disques durs locaux et donc non partagés[19].



[20]

62. La démarche d'analyse scientifique nécessite l'utilisation des données de la recherche dans son sens le plus large, elle concerne tant les publications scientifiques que les données brutes ou encore les tableaux, images, données statistiques, sons,

[18] Schéma d'orientation stratégique de l'IST page 15 « les publications scientifiques : une augmentation continue et forte »

[19] Schéma d'orientation stratégique de l'IST page 15 « Données et publications : une course poursuite »

[20] http://informatique.in2p3.fr/li/spip.php?article327

autres corps de texte, de manière générale toutes données nécessaires au chercheur dans sa démarche.

63. La physique des hautes énergies et le LHC (Large Hadron Collider) fournit un exemple ultime de cette nouvelle pratique scientifique centrée sur les données : à un dispositif technique – le collisionneur – bâti par des ingénieurs sont associés des instruments – les détecteurs – imaginés par les chercheurs pour étudier les particules émises lors des milliards de collisions produites. La production de données est de l'ordre de 15 pétaoctets/an. Une telle production de données nécessite à la fois des moyens techniques de traitement colossaux mais également une infrastructure de gestion de données à même d'offrir une disponibilité maximale aux équipes de recherche.

64. Mais au préalable à cette mise à disposition, les données se doivent d'être préparées. Par exemple, on comprend aisément que les grands programmes d'observation de la terre et des océans, opérations menées le plus souvent en coopération internationale, réclament des processus de préparation des données préalable à la fouille de données elle-même. Ces activités, que l'on regroupe sous le terme de curation de données, englobent la sélection, vérification, normalisation, annotation, reformatage, enrichissement, structuration des données collectées ; le but ultime est de disposer de données qualifiées pour être soumises à un processus de fouille scientifiquement valide.

65. Les grands programmes de numérisation de livres anciens s'accompagnent ainsi de processus de création de métadonnées enrichies, facilitant l'extraction de connaissances.

66. Les entrepôts de données résultant de ces processus de curation sont alors mis à disposition des chercheurs à travers des portails d'accès thématiques tels l'observatoire virtuel astronomique international, l'observatoire mondial de la biodiversité, le programme mondial des données climatologiques…

67. Les capacités de traitement jouent évidemment un rôle essentiel dans le développement des usages du TDM. Les moyens des centres de calcul sont en progression constantes, et s'organisent en grilles de calcul à l'échelle internationale, pour répondre aux enjeux du traitement de flux de données massives en climatologie, environnement, santé, astronomie. En bout de chaîne de traitement interviennent les applications de représentation/visualisation de données complexes, fortement consommatrices en puissance informatique.

68. **Les outils techniques d'analyse.** Les communautés scientifiques assistent par ailleurs à une évolution des pratiques de recherche, vers une recherche plus collaborative impliquant des acteurs multiples et s'appuyant sur les technologies modernes pour traiter et exploiter des masses importantes de données produites dans un environnement partagé[21].

69. Les technologies de TDM visent à faire émerger des relations entre les données unitaires analysées, de détecter des liens de cause à effet, d'établir des modèles et de valider leur reproductibilité. Pour ce faire, en fonction des types de données et des objectifs proposés, seront utilisées de façon complémentaire des techniques issues de la statistique descriptive, de l'analyse de données (statistique exploratoire) ou de l'informatique (intelligence artificielle).

70. La fouille de textes utilise les mêmes techniques que la fouille de données mais il est nécessaire au préalable de traiter les données textuelles par des technologies linguistiques pour les rendre compatibles avec les méthodes de data mining. L'usage de structures de données adaptées aux propriétés des textes et d'algorithmes sémantiques est spécifique à la fouille de texte.

71. Les méthodes d'analyse de données (factorielle, discriminante, composantes principales, composantes multiples, correspondances...) conduisent à rendre compte des différentes dimensions internes des ensembles de données, permettant de mettre à jour les paramètres d'organisation des données. Les méthodes de classification (clustering, apprentissage non supervisé) permettent de faire ressortir des groupes d'éléments. Les méthodes de régression et d'apprentissage supervisé (intelligence artificielle) ont pour finalité de prédire l'évolution de certains comportements en fonction d'autres variables.

72. Les logiciels de TDM présentent des fonctionnalités regroupant tout ou partie de cette chaîne de traitement allant de l'accès à la préparation des données, l'application de calculs algorithmiques choisis (apprentissage), l'exploitation des sorties, l'exploitation des modèles, la visualisation des résultats.

73. Les logiciels spécifiques au text mining traitent le langage naturel par étiquetage grammatical, règles syntaxiques, ontologies, apprentissage à partir de corpus étiquetés… A partir des corpus de documents ainsi structurés, différents algorithmes d'analyse peuvent être utilisés : classification automatique, analyse de tendances, règles d'associations…

---

[21] Schéma d'orientation stratégique de l'IST page 16

74. De nombreux outils et techniques d'analyse automatique de textes et de données sont développés par les laboratoires de recherche français et déjà utilisés par les chercheurs.

75. **Au CEA**, la pratique du TDM est très courante tant en physique des particules, en physique nucléaire ou en astrophysique[22]. La mise en relation d'archives interopérables et d'outils intelligents pour leur analyse est à la base de projets européens, comme The European Virtual Observatory, http://www.euro-vo.org/, ou EUROPLANET, http://www.europlanet-eu.org/.

76. Dans d'autres domaines, l'usage du TDM est utilisé par des équipes du CEA pour le traitement automatique des langues, notamment pour la constitution, la visualisation et l'analyse de réseaux de citations (http://clair.eecs.umich.edu/aan/index.php) ou pour accroitre les performances de moteurs de recherche spécialisés (http://aclasb.dfki.de/ ou http://saffron.insight-centre.org/acl/).

77. **A l'INRA**, l'équipe de recherche Bibliome (unité MaIAGE-INRA) développe de nombreuses applications de TDM appliquées à des textes d'articles scientifiques dans des domaines intéressant l'Inra. Certains de ces projets utilisent également des références (PubMed), des brevets (EspaceNet) et des magazines professionnels (ex. Perspectives Agricoles).

78. **L'UMR Lisis** (INRA, ENPC, UPEM) (http://www.inra-ifris.org/), associée à l'**IFRIS**, développe une plateforme Cortext pour l'analyse de corpus textuels (http://www.cortext.net/) dans le cadre de recherches en sciences humaines et sociales.

79. **Le Cirad** est également concerné au travers de l'UMR IATE (CIRAD, INRA, SupAgro, Université de Montpellier II) et développe des méthodes et des outils innovants de traitement de données et de connaissances. L'objectif est de proposer des méthodes et outils d'aide à la décision pour le pilotage global de filières de transformation de la biomasse. Ces méthodes et outils doivent permettre de collecter, représenter et gérer différents types de données et de connaissances, incluant des données imparfaites (par exemple peu fiables, imprécises, ...), des connaissances à dire d'expert et des modèles de génie des procédés. D'autre part, les outils proposés doivent permettre la prise en compte de critères multiples, des préférences et des arguments des acteurs de la filière agricole.

---

[22] Voir les travaux du Centre de données astronomiques de Strasbourg : http://cdsweb.u-strasbg.fr/about

80. **A Irstea**, des scientifiques de l'UMR TETIS, utilisent des solutions innovantes de fouille de textes et de données dans le cadre de recherches conduites en collaboration entre IRSTEA, l'INRA et AgroParisTech : fouille d'articles scientifiques pour identifier des thématiques nouvelles de recherche par l'enrichissement de ressources sémantiques (ressources termino-ontologiques de spécialité, thésaurus, etc.) ou améliorer des outils de veille en épidémiologie animale. En outre, le TDM permet la mise en relation de données hétérogènes au sein de corpus très volumineux qui comprennent à la fois des textes scientifiques et non scientifiques, jeux et bases de données, images, etc.), ce qui a permis de découvrir des connaissances nouvelles et complémentaires (https://tetis.teledetection.fr/index.php/fr/).

81. **L'UMR GESTE** (IRSTEA, ENGEES), à Strasbourg (http://geste.engees.eu/), mobilise actuellement des scientifiques qui travaillent sur la problématique des polluants émergents dans l'eau et les changements de pratiques des particuliers et des artisans permettant d'en réduire l'émission. Il s'agit d'une question majeure en termes de santé-environnement, encore insuffisamment cernée sur le plan des enjeux et des solutions. L'usage du TDM permettra de réaliser une cartographie de l'enjeu, dans ses dimensions scientifiques et sociétales, d'en préciser la chronologie, d'identifier des sous-thématiques ainsi que des institutions et acteurs clés[23].

82. **L'Inria** a également développé GROBID (GeneRation Of BIbliographic Dataset)[24], outil d'apprentissage automatique (ou Machine Learning) disponible en Open source. Cet outil permet l'extraction, l'analyse et la restructuration de publications scientifiques d'un format bruts (ex : PDF) vers un format TEI.

83. **Autres outils**. Les logiciels Alceste[25] (développé par l'entreprise IMAGE et le CNRS) et Calliope[26] (par Astefo) sont spécialisés dans l'analyse lexicale. Le projet Gargantext offre un exemple d'une telle analyse. Il permet d'analyser les textes selon un autre paradigme que les mots complets :

---

[23] Ces exemples sont tirés d'une note EPRIST (Responsables IST des organismes de recherche) sur le text and data mining « Le TDM comme outil innovant de recherche scientifique »
http://www.cnrs.fr/dist/z-outils/documents/EPRIST%20text%20et%20data%20miningV3.pdf

[24] https://www.researchgate.net/publication/221176095_GROBID_Combining_Automatic_Bibliographic_Data_Recognition_and_Term_Extraction_for_Scholarship_Publications

[25] http://www.image-zafar.com/Logiciel.html

[26] https://www.calliope-textmining.com/

- « Il paraît que l'drore des lettres dans un mot n'a pas d'imtaorpnce. La preimère et la dernière lettre doveint être à la bonne plcae. Le reste peut être dans un désrodre total et on peut touojurs lire sans proèlbme. On ne lit donc pas chqaue lettre en elleê-mme, mais le mot comme un tout. Un chagemnent de réféenrtiel et nous transopsons ce réslutat au texte lui-êmme: l'rodre des mots est failbement imprtoant comparé au conetxte du texte qui, lui, est compté: comptexter avec Gargantext. »[27]

84. Parmi les logiciels de TDM largement utilisés dans le monde, Weka, développé à l'origine par l'université de Waikato en Nouvelle-Zélande, permet de visualiser et d'analyser rapidement les données. Enfin, de nombreux logiciels payants ont été produits par des entreprises, tels que RightsDirect[28], KNIME[29] ou RapidMiner[30].

85. Les données et articles scientifiques, qui constituent la source primaire d'information pour l'analyse par TDM, peuvent être importés depuis des plateformes de partage d'information scientifique et technique (tels que des bibliothèques numériques, bases de données, archives ouvertes, moteurs de recherche, etc.). Certaines plateformes œuvrent en faveur du libre accès à la production scientifique telles que, en France, **HAL, Gallica, NAKALA, ISIDORE, OpenEdition, Persée** [31].

86. **ISTEX**. Enfin, dans le cadre du projet ISTEX (Initiative d'excellence de l'Information Scientifique et Technique - ANR-10-IDEX-004-02), plateforme numérique d'accès, de partage et d'enrichissement de l'information scientifique et technique (IST) sont développés des « services à valeur ajoutée ». Ces services doivent permettre aux utilisateurs d'analyser automatiquement des masses importantes de données en proposant notamment les services suivants :

- l'enrichissement des données ;
- les jeux de données induites ;

---

[27] http://gargantext.org/

[28] http://www.rightsdirect.com/

[29] https://www.knime.org/

[30] https://rapidminer.com/

[31] Liste catégorisée de « Plateformes de partage d'information scientifique et technique » : http://www.cnrs.fr/dist/acces-ist.html

- les nano publications ;
- les schémas de collaboration ;
- les schémas d'influence ;
- l'analyse sémantique ;
- l'analyse d'impact ;
- le document automatique.

87. De nombreuses autres plateformes de partage de l'information scientifique et technique sont recensées par le CNRS et catégorisées afin de mettre en valeur les initiatives en faveur de la mutualisation des connaissances et du libre accès à la production scientifique[32].

88. **Les résultats.** A l'issu de l'application du processus de TDM sur des jeux de données, un résultat automatique est produit. Ce résultat se présente différemment selon la technique d'analyse utilisée, selon l'angle d'analyse choisi par le chercheur (« user generated contents »).

89. Ce résultat constitue des connaissances nouvelles. Celles-ci sont de plusieurs ordres et sont exploitées selon plusieurs axes :

- les systèmes de recommandation, par analyse et filtrage de l'information contenue dans les données, produisent des prédictions sur les comportements ;
- en science, les liens entre éléments de données mis en évidence conduisent à proposer de nouvelles hypothèses de recherche.

90. Le résultat produit s'intègre alors dans la démarche scientifique comme une nouvelle donnée d'analyse.

2.1.2.2 Les opérations techniques du TDM

91. Il convient d'appréhender le TDM également sous l'angle de la démarche technique nécessaire à toute opération de traitement.

92. **Présentation Atilf**. Dans une note datée du 14 mai 2014, M. Pierrel, représentant de l'Université de Lorraine (agissant pour le compte de la Conférences des Présidents

[32] http://www.cnrs.fr/dist/acces-ist.html

d'Université), en charge du développement des services à valeur ajoutée dans le cadre d'ISTEX, et Directeur de l'Atilf (Analyse et Traitement Informatique de la Langue Française) a précisé les opérations techniques réalisées par les laboratoires dans le cadre du développement du TDM. M. Pierrel distingue dans sa note deux temps :

1. Les travaux effectués sur les machines :

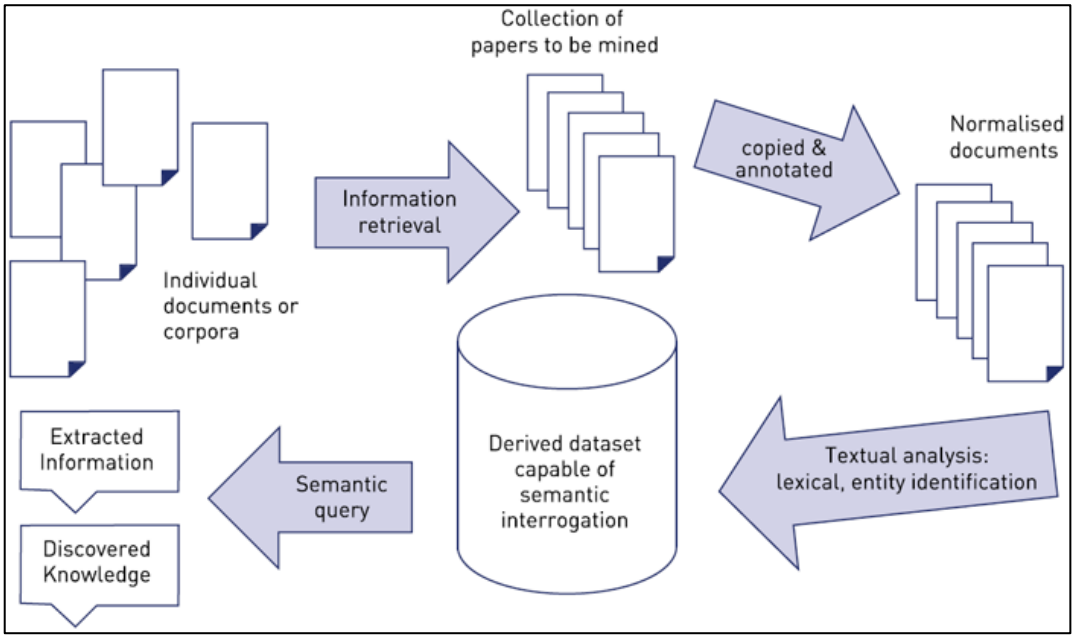
- copie de travail de sous-corpus sur des machines ;
- annotation linguistique sur les données :
- détection des termes et construction d'un référentiel terminologique ;
    - détection d'entités nommées et construction d'un référentiel d'entités nommées ;
    - annotation et lemmatisation des textes, construction et gestion de lexiques ;
    - balisage des références bibliographiques ;
- utilisation de ces annotations dans les procédures de sélection des articles suite à une demande utilisateur ;
- construction de cartographie à partir des éléments annotés.

2. La diffusion des résultats de recherche en TDM

- dans le cadre de publications scientifiques ;
- dans le cadre de diffusion des résultats de TDM.

93. **Schéma TDM**. Le schéma suivant illustre les étapes du traitement de l'information dans le cadre du TDM appliqué au marketing direct, dont les principes sont transposables au domaine scientifique[33] :

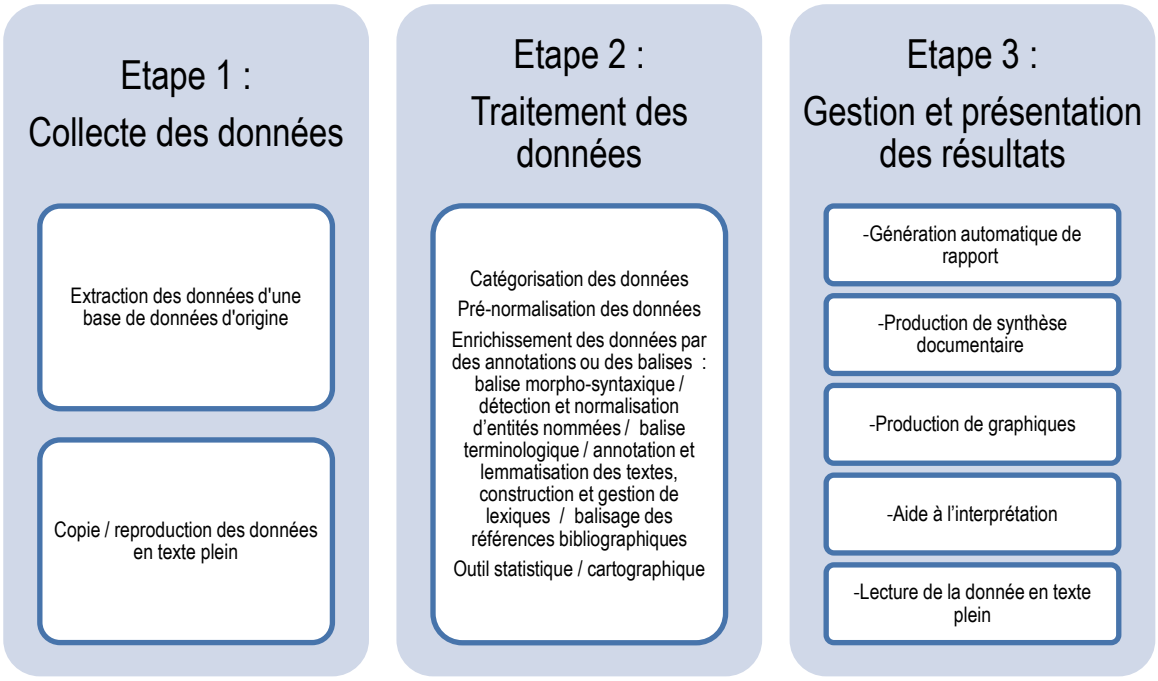
[33] http://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining



[34]

94. **Schéma de synthèse opérations TDM**. Eu égard à l'ensemble de ces éléments, le processus de traitement de l'information par TDM est complexe et peut être synthétisé en trois étapes ; les opérations techniques réalisées dans chacune des étapes sont précisées ci-après :

[34] ©Jisc CC BY-NC-ND



95. Chacune de ces étapes nécessite d'être encadrée, sécurisée, des standards et formats doivent être définis, un process de réalisation des opérations de TDM doit être rédigé voire normalisé.

2.1.2.3 Les fonctions macro du TDM

96. Une approche fonctionnelle macro du TDM permet également d'appréhender la notion au regard de ses objectifs.

97. Le développement de techniques de TDM ouvre de nouvelles perspectives dans l'analyse d'importants volumes de textes et de données (big data). Le TDM, grâce à une démarche inductive, analyse l'ensemble de ces informations formées en corpus. La fouille de textes regroupe un ensemble de pratiques et des méthodes diverses, qu'il est difficile d'unifier[35]. Il est pourtant possible de dégager de grands traits de compréhensions[36].

[35] Toussaint Yannick, « Extraction de connaissances à partir de textes structurés », *Document numérique,* Vol.8, 3/2004, pp. 11-34

[36] Fidelia Ibekwe-Sanjuan, *Fouille de textes : méthodes, outils et applications*, Coll. Systèmes d'information et organisations documentaires, Hermès, 2007, 352 p.

98. La multiplication des articles de sciences en format numérique a rendu accessible des informations diverses : tableaux, images, données statistiques, sons, corps de texte… Ces derniers sont quant à eux différemment rédigés (langue, expression…). L'ensemble des styles et des genres des textes à analyser invitent donc le chercheur qui utilise le TDM à former des corpus cohérents de données. L'objectif du TDM est un objectif ciblé. Une opération d'exploitation des données n'a pas comme finalité de donner des clés de lecture généraliste pour le corpus, mais de les explorer sur la base d'une question précise.

99. Le TDM analyse donc un ensemble de données selon un critère de nouveauté (qu'est-ce qui apparaît en croisant tous les textes du corpus, par exemple ?) ou un critère de similarité (qu'est-ce qui est récurrent dans tous les textes du corpus ?). C'est ensuite aux chercheurs de dégager le sens issu des croisements de données. Après avoir sélectionné puis transformé les données (selon leur codage notamment), le chercheur utilisera des outils de TDM pour aboutir à une interprétation et à une évaluation. Toutes les étapes du TDM sont donc essentielles, du prétraitement au formatage des données jusqu'aux conclusions.

100. Prenons un exemple pour mieux comprendre ces techniques. L'INRA mène actuellement une étude à l'aide de *text mining* afin d'identifier de nouvelles espèces de poissons domesticables, c'est-à-dire adaptables à l'aquaculture. De nombreuses données relatives aux poissons (reproduction, alimentation, milieu de vie) sont analysées à l'aide de technologies de *text mining*. Ces données proviennent notamment du corpus ISTEX. L'objectif est d'identifier les caractères types des poissons d'aquaculture afin de mieux comprendre le phénomène de la domestication ainsi que d'identifier les espèces de poissons qui ressemblent le plus à des espèces d'aquaculture (recherche fondamentale). Ces espèces identifiées pourront ainsi faire l'objet d'expérimentations de domestication (recherche appliquée). L'utilisation de leurs résultats pourrait ouvrir de nouvelles possibilités pour la pisciculture.

2.1.3 L'économie du TDM

101. Le TDM peut générer deux types de bénéfices économiques et sociaux :

- une baisse de coûts et une hausse de la productivité ;
- un accroissement des innovations de produits ou de services.

102. Il est par ailleurs important d'analyser également l'impact du TDM face au monde de l'édition, afin de dépassionner le débat tout en encadrant et sécurisant les pratiques.

2.1.3.1 Réduction des coûts de la recherche et accroissement des découvertes

103. Le potentiel économique du TDM tient d'abord à la réduction des coûts de la recherche et l'accroissement des découvertes envisageables.

104. En effet, l'usage du TDM, en réduisant les coûts de traitement, ouvre la possibilité de multiplier les nouveaux articles de recherche grâce à ce gain de temps, donc d'enrichir les bases de données, etc., dans une logique de cercle vertueux.

105. Le JISC[37], organisme britannique en charge des services numériques pour la recherche, a estimé que l'utilisation et le développement du TDM pour la recherche permettrait une hausse de la productivité de la recherche publique sans coûts supplémentaires. Le TDM équivaudrait à un gain de 4,7 millions d'heures de travail pour les chercheurs sur une année, sur l'ensemble du Royaume-Uni.

106. Dans le même sens, le TDM permet de développer les activités de recherche interdisciplinaire. Les corpus pourront être construits entre matières diverses, pour accéder aux avancées de la biochimie dans le cas d'une étude en biologie, ou encore pour comparer les évolutions de l'histoire dans les analyses sociologiques.

2.1.3.2 Accroissement de l'innovation

107. Une large part de ces nouvelles connaissances est convertible en innovation économique. Par exemple, dans le domaine de la santé, une analyse via PubMed[38] a permis de formuler de nouvelles hypothèses de recherche. Un groupe de chercheurs a procédé à une extraction de données sur un corpus d'articles de médecine. Ils cherchaient les liens entre « médicament – maladie » et « médicament – phénotype »[39]. En parcourant l'ensemble des articles, ils ont identifié des gènes pouvant être responsable de maladie dont l'origine était la toxicité d'un médicament. Cet enrichissement pourrait permettre aux chercheurs et aux médecins de diagnostiquer rapidement des liens nouveaux entre « médicament » et « maladie ». D'un point de vue économique, cette découverte pourrait faire avancer la recherche médicale, avec un coût de main d'œuvre limité.

---

[37] https://www.jisc.ac.uk/guides/text-and-data-mining-copyright-exception

[38] http://www.ncbi.nlm.nih.gov/pubmed

[39] A CTD–Pfizer collaboration: manual curation of 88 000 scientific articles text mined for drug–disease and drug–phenotype interactions", http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3842776/

108. L'intérêt économique qui se dégage de l'utilisation du TDM correspond donc essentiellement à une source d'innovation incrémentale. A cet égard, le JISC[40] souligne que le TDM est un « déverrouillage »[41] des connaissances dont les sphères économiques et sociales ne peuvent que bénéficier.

109. Par ailleurs, le TDM est une technique scientifique et numérique qui nécessite un capital humain conséquent pour être bien appliquée. La fouille de textes et de données ouvre donc des possibilités d'emplois spécialisés et techniques, d'ingénieurs et d'informaticiens essentiellement. Cette recherche utilisant le TDM demande aussi des infrastructures et des services dont la création et la production seraient des éléments stimulant de croissance.

2.1.3.3 Les enjeux économiques du TDM face au monde de l'édition

110. Les enjeux économiques du TDM sont à mettre en parallèle avec les attentes du monde de l'édition. Le TDM transforme en effet la conception classique du droit d'auteur. Si les interrogations sont normales, il convient de mettre en avant les bénéfices de cette pratique par rapport au modèle actuel de l'édition.

111. L'argument selon lequel l'activité de l'éditeur va purement et simplement disparaître a déjà été avancée par les éditeurs dans le cadre de l'analyse de l'article 17 du projet de loi pour une République numérique (article 30 de la Petite loi) organisant un droit d'accès aux publications scientifiques après le respect d'une période d'embargo. Cet argument n'a pas été retenu dans le cadre de l'Etude d'impact de cet article 17, publiée par l'Assemblée nationale le 9 décembre 2015[42].

---

[40] https://www.jisc.ac.uk/guides/text-and-data-mining-copyright-exception page 38

[41] JISC, p.38

[42] Projet de loi pour une République numérique NOR : EINI1524250L/Bleue ETUDE D'IMPACT 9 décembre 2015 : « Focus 1 : Impact économique sur l'édition scientifique institutionnelle en France

L'impact de cette mesure sur les équilibres économiques de l'édition scientifique institutionnelle française, essentiellement constituée d'éditeurs de sciences humaines et sociales, doit être relativisé dans la mesure où la majorité de leur chiffre d'affaires est aujourd'hui constitué de subventions apportées par des établissements ou des laboratoires. Les revues ne représentent en outre, en moyenne, que 18 % de leur production éditoriale, et entre 40 % et 60 % du chiffre global des ventes associées ces revues est réalisé grâce aux publications de l'année, qui demeureront sous embargo au terme de la mesure proposée, garantissant que ces acteurs ne devraient être touchés que marginalement. »

http://www.assemblee-nationale.fr/14/projets/pl3318-ei.asp

112. Cette affirmation n'a également pas été vérifiée au Royaume-Uni, pays qui a introduit une exception au droit d'auteur en faveur du text and data mining en 2014 ; ni aux Etats-Unis où la jurisprudence a également légalisé la pratique (Affaire Authors Guild v. Google - 14 novembre 2013).

113. Par ailleurs, les remises en question qu'introduit le TDM pour les bénéfices du monde de l'édition ne semblent en effet pas dépasser l'ensemble des bénéfices que la société pourrait tirer d'une exploitation intensive des données. La Commission européenne s'accorde ainsi sur les bénéfices à long terme du TDM en insérant une exception en faveur du TDM dans le projet de directive relative au droit d'auteur dans le marché unique numérique.

2.2 Quels enjeux pour le travail de la science ?

114. Le TDM fait entrer la Science et la recherche scientifique dans une nouvelle ère ; les vertus du TDM et de l'Open science sont multiples :

- le TDM permet certes dans le cadre de la démarche scientifique de procéder à de l'analyse automatique de la littérature scientifique et de l'ensemble des données ;
- il permet également d'organiser et de structurer les projets de recherche en France et de rationaliser les processus décisionnels qui régissent l'établissement des priorités et l'allocation des budgets ;
- il fait naître de nouveaux sujets de valorisation ;
- le TDM peut également constituer une aide à la décision publique.

2.2.1 Analyser les publications et données scientifiques

115. La littérature scientifique, articles de périodiques, monographies, thèses, rapports, fait l'objet d'un intérêt tout particulier pour les adeptes de la fouille de textes. Cette information en langage naturel, peu structurée comparée aux données chiffrées, recèle des gisements de connaissances nombreux et variés. Chaque article scientifique est le résultat d'un processus logique de recherche s'inscrivant dans un projet donné aux objectifs énoncés, et se positionne dans un contexte scientifique identifié et reflété par la bibliographie. L'analyse TDM de larges corpus de littérature scientifique couvrant, pour une problématique donnée ou limité à un champ disciplinaire spécifique, des périodes, méthodologies et contextes instrumentaux variés, permet de développer des produits et services au bénéfice des chercheurs :

- études longitudinales diachroniques sur l'évolution des pratiques scientifiques, l'apparition de nouveaux concepts, les fusions ou éclatements de champs scientifiques, les relations interdisciplinaires et leurs évolutions,

- la masse d'informations produites par les chercheurs (2,5 millions d'unités par an pour ce qui concerne les articles) conduit à rapprocher les technologies d'extraction de connaissances des besoins d'accès des chercheurs aux contenus des articles de leurs confrères. La lecture séquentielle est amenée à être remplacée au moins en partie par de l'approche navigationnelle dans des ensembles de connaissances issues des corpus initiaux. Ainsi, les activités de fouille de textes et de données ont des retombées directes dans les activités d'accès aux connaissances scientifiques par les chercheurs.

- l'amélioration de la pertinence des portails d'accès aux publications par enrichissement sémantique et annotations, l'offre de fonctionnalités de recommandations, sont autant d'évolutions requises pour mieux partager les connaissances contenues dans les articles

- catégorisation automatique, classification des publications, liens inter publications au-delà des simples citations,… autant d'usages d'analyses du langage naturel pour construire des savoirs dérivés

- bibliométrie rénovée et développement de nouvelles métriques et mesures d'impact des publications, conduisant à une évolution des processus d'évaluation des chercheurs, structures, programmes, institutions,…

- inflexion de la démarche scientométrique : analyse rétrospective de champs de recherche, mesures de la performance en recherche, fertilisation croisée entre disciplines

- recherche et gestion de réservoirs d'experts

- développement de la démarche prédictive par génération automatique d'hypothèses à partir de la fouille de littérature scientifique

116. Le TDM est par conséquent un facteur d'intensification de la recherche. Il permet de stimuler la recherche fondamentale mais aussi d'aiguiller la recherche appliquée. D'abord, le TDM ouvre des analyses quantitatives inédites dans l'histoire de la science. Jamais autant de corpus de textes et de données n'ont été si disponibles et possiblement analysables. L'interprétation scientifique du chercheur ne change pas en soi, mais son éventail de recherche s'élargit. L'intensification passe aussi par une

transformation de la manière dont les chercheurs perçoivent la science. Les paradigmes peuvent évoluer, comme le montre par exemple *Gargantext*.

117. Le TDM ouvre la possibilité de trouver de nouvelles corrélations ou des tendances émergentes. Outil de la recherche dérivée, la fouille de textes permet d'analyser des interactions souvent nouvelles, voire impossibles au niveau transdisciplinaire.

118. Le TDM, lié à la diffusion des données de la recherche et à l'accélération des connaissances, va faciliter le renouvellement des expériences et renforcer l'administration de la preuve. Dans certains cas, la réfutabilité devrait s'accroître, avec comme objectif final une pertinence scientifique renforcée. La vérification de la recherche est un corolaire du développement du TDM.

2.2.2 Construire des problématiques et des projets de recherche

119. L'évolution de la science s'inscrit dans une suite de ruptures méthodologiques concomitantes à l'enrichissement de l'arsenal méthodologique à la disposition des chercheurs. L'avènement de la science numérique, de l'omniprésence des instruments générateurs de données massives conduisent à la naissance d'une science guidée par les données (cf. Hey, Tolle, Tansley « The Fourth Paradigm »).

120. Les scientifiques confient à présent les tâches d'observation et de mesures soit à des instruments uniques et partagés comme les accélérateurs de particules, les télescopes embarqués, soit à des réseaux de nombreux petits instruments comme les bouées océanographiques, les stations de mesures météorologiques ou sismiques. Se développe alors en aval une science de l'interprétation des données dont le TDM est le principal pilier. Mais il est nécessaire d'accompagner ce développement des algorithmes d'analyse des données de fonctions moins visibles mais tout autant essentielles : les infrastructures de calcul et infrastructures de gestion/partage des données.

121. La mécanique du raisonnement scientifique est classiquement basée sur le raisonnement déductif, assis sur des théories servant de prémisses au raisonnement. De la connaissance théorique se construisent des hypothèses, les observations récoltées (données) servant à confirmer (ou infirmer en remettant en cause la théorie) ces hypothèses. Une grande partie de l'éducation occidentale est basée sur cette approche, dès « l'âge de raison ». Mais cette prééminence de la théorie sur l'expérimentation est discutée depuis quelques années, notamment avec la conjonction de l'avènement des big data et le développement du TDM.

122. La mécanique du raisonnement est alors inversée : la mise à jour de corrélations entre éléments des observations concentrées dans les jeux de données suggère des hypothèses permettant d'établir des modèles théoriques de comportement. Bien évidemment, ce raisonnement inductif ne garantit pas la conclusion et peut conduire à mener de nombreuses itérations successives. Le raisonnement abductif, inférence logique allant des observations vers la théorie, se focalise sur les hypothèses les plus simples et probables et peut être décrit comme « l'inférence de la meilleure explication »[43]. Les moteurs d'inférence en intelligence artificielle se basent fréquemment sur ces approches abductives pragmatiques et fécondes dans le développement des logiciels de TDM.

123. L'utilisation du TDM et l'organisation du partage de la connaissance ainsi que les changements de méthodes scientifiques vont permettre aux communautés de recherche de construire de nouveaux projets de recherche, de découvrir de nouveaux sujets.

124. Cette construction préalable par l'utilisation du TDM permet de gagner en maîtrise et en productivité par rapport à ce que sont aujourd'hui les bases décisionnelles d'un projet de recherche. La structuration des projets de recherche permettra une meilleure affectation des ressources.

2.2.3 Optimiser la gouvernance des systèmes scientifiques

125. Le TDM participera ainsi à une meilleure « gouvernance des systèmes scientifiques » en rationnalisant « les processus décisionnels », par « l'établissement des priorités, l'affectation des crédits et la gestion des ressources humaines de façon à répondre efficacement aux inquiétudes des différentes parties prenantes du système »[44].

126. En effet, l'analyse TDM des résultats des programmes de financement de la recherche (publications, brevets) permettent de mieux cerner les domaines les plus féconds ainsi que d'établir des liens entre disciplines. Des outils de visualisation / cartographie des coopérations disciplinaires ou géographiques nationales et internationales permettent de représenter ces informations essentielles au pilotage scientifique.

---

[43] https://en.wikipedia.org/wiki/Abductive_reasoning

[44] « Gouvernance de la recherche publique – Vers de meilleures pratiques », OCDE 2003

127. Cette exploitation d'informations à destination des décideurs de politique scientifique participe de la définition des stratégies scientifiques d'organismes, tant dans le domaine de la recherche fondamentale que dans sa valorisation économique.

2.2.4 Valoriser les données scientifiques

128. Le TDM et l'analyse de masse de données va non seulement permettre la valorisation des nouvelles connaissances issues de traitement automatique et le développement subséquent d'innovations et de découvertes, mais également permettre de valoriser les masses de données non utilisées, enregistrées sur les disques durs de chercheurs et non partagées, bien que constituant une valeur scientifique.

2.2.4.1 Valorisation de la lost science

129. Un « effet château d'eau »[45] est envisageable pour l'ensemble du monde de la recherche, avec le traitement de données oubliées.

130. Il y a aujourd'hui une forte déperdition des données recueillies par les expérimentateurs. On estime que les publications permettent d'accéder à environ 10 % de celles-ci, le reste restant disponible mais non utilisé sur les disques durs d'ordinateurs. Dans certaines disciplines, des résultats valables et importants restent non publiés et beaucoup de données sont sous-utilisées ou perdues (c'est en particulier le cas des données issues de résultats négatifs qui sont oubliées). Pour celles qui sont collectées par les grands instruments, les données brutes recueillies sont si massives qu'elles sont traitées directement en ligne sans être stockées, comme par exemple celles fournies par des observations spatiales[46].

131. Pour Cristinel Diaconu, la perte de données semble être le lot commun :

- « Quand on veut accéder aux données, ou on ne les trouve plus, ou on les trouve, mais on ne sait pas quoi faire parce qu'on ne comprend pas ce que c'est. Pire, certaines données ont parfois été détruites par les chercheurs qui les jugeaient inutiles à l'issue d'un projet. Sur le coup, on ne s'en rend pas compte mais, dix après, le projet en cours peut avoir une résonance avec le projet précédent et le

[45] L'expression « effet château d'eau » renvoie à une notion économique. C'est un écho à la théorie du ruissellement mais élargie aux innovations et non seulement au système fiscal. En somme, le TDM est un ensemble de techniques qui peut investir de nombreux secteurs, il « coule » sur eux et peut donc « irriguer » différentes activités économiques.

[46] COMETS, *Les enjeux éthiques du partage des connaissances*, 2015, p.4

potentiel de découverte est perdu car il n'y a plus de financement pour refaire ces manipulations. »

132. Pourtant, ces données archivées représentent une vraie manne. Cristinel Diaconu s'est livré à un calcul de rentabilité de ces données pour son domaine : « Nous nous sommes rendu compte avec mon équipe que le coût supplémentaire dédié à la préservation des données est de l'ordre de 1/1 000 du budget total. Or la publication de nouveaux articles issue de l'exploitation des archives dans les cinq années suivantes représente un bénéfice de 10 %. C'est de la recherche qui ne coûte presque rien ! Si on n'a pas de stratégie de préservation des données, on passe à côté de découvertes potentielles et de recherche à bas coût. Une fois qu'elles sont bien préservées, les données ne coûtent presque plus rien. »

2.2.4.2 Libre réutilisation des résultats de TDM

133. L'article 38 de la loi pour une République numérique prévoit que les « fichiers produits au terme des activités de recherche pour lesquelles elles ont été produites (…) constituent des données de la recherche ».

134. Le régime des « données de la recherche » est encadré :

- par l'article 30 de cette même loi : « Dès lors que les données issues d'une activité de recherche financée au moins pour moitié par des dotations de l'État, des collectivités territoriales, des établissements publics, des subventions d'agences de financement nationales ou par des fonds de l'Union européenne ne sont pas protégées par un droit spécifique ou une réglementation particulière et qu'elles ont été rendues publiques par le chercheur, l'établissement ou l'organisme de recherche, leur réutilisation est libre. »

- par la loi n° 2015-1779 du 28 décembre 2015 relative à la gratuité et aux modalités de la réutilisation des informations du secteur public (loi Valter) qui fait entrer les établissements et institutions d'enseignement et de recherche dans l'Open data et dans l'obligation de mise à disposition de leur données « sous forme électronique » « dans un standard ouvert, aisément réutilisable et exploitable par un système de traitement automatisé ».

135. Ces données de la recherche constituent par conséquent une source de connaissances qui doit être mise à disposition des communautés scientifiques aux fins d'enrichir de nouvelles connaissances et de nouveaux travaux.

136. Toutefois, l'Open science ne doit pas faire obstacle aux enjeux économiques de la recherche.

137. La mise à disposition des données scientifiques sur des plateformes Open science ne doit pas aller à l'encontre :

- de la valorisation des données notamment par brevet ; du respect des secrets et des dispositions spécifiques telles que les Zones à Régime Restrictif[47] ;
- du respect des règles contractuelles de confidentialité.

138. La mise à disposition des données de la recherche doit être encadrée également au regard des pratiques non homogènes des communautés scientifiques.

2.2.4.3 Développement de l'innovation

139. Les techniques de TDM s'appuient sur des outils innovants (logiciels, supercalculateur, technologies pour la collecte et le traitement massifs de données) que les organismes de recherche publique français ainsi que les entreprises françaises contribuent à développer.

140. Le TDM est un des grands secteurs d'innovation potentiel actuel de l'économie numérique et ces technologies sont centrales ; preuve en est des récompenses qui viennent consacrer les travaux sur le TDM. Xavier Tannier, du Laboratoire d'information pour la mécanique et les sciences de l'ingénieur à Orsay et Iona Manolescu de l'INRIA ont reçu un « Google Award »[48] pour leur algorithme de fouille de texte sur la presse.

---

[47] -Article R.413-5-1 du code pénal : « Sont dites " zones à régime restrictif " celles des zones, mentionnées à l'article R. 413-1, dont le besoin de protection tient à l'impératif qui s'attache à empêcher que des éléments essentiels du potentiel scientifique ou technique de la nation :

1° Fassent l'objet d'une captation de nature à affaiblir ses moyens de défense, à compromettre sa sécurité ou à porter préjudice à ses autres intérêts fondamentaux ;

2° Ou soient détournés à des fins de terrorisme, de prolifération d'armes de destruction massive et de leurs vecteurs ou de contribution à l'accroissement d'arsenaux militaires.

Les zones à régime restrictif peuvent inclure, dans leur périmètre, des locaux dont la protection renforcée est justifiée par l'entreposage de produits ou par l'exécution d'activités comportant des risques particuliers au regard des impératifs mentionnés aux trois premiers alinéas. »

[48] https://archives.limsi.fr/Actualites/GoogleAward.html

https://research.googleblog.com/2015/06/google-computational-journalism.html

https://www.amia.org/news-and-publications/press-release/19-fellows-inducted-american-college-medical-informatics

Dans le cas de l'informatique médicale, c'est Pierre Zweigenbaum[49] qui a été distingué par l'American College of medical Informatics en 2014.

141. Le TDM représente un potentiel de retombés économiques pour la France : plusieurs start-ups françaises ont vu le jour suite à des travaux de recherche ayant nécessité le développement d'outils de TDM (au CEA notamment) en lien avec des partenaires privés.

142. Dans son article « Mining external R&D »[50], Alan Porter précise que les entreprises sont « *désirables* » d'une telle innovation. Les entreprises ayant un pôle de recherche et développement ne peuvent en effet qu'attendre des bénéfices d'une recherche appliquée plus grande, sur des corpus plus importants. L'avantage est encore plus élevé pour les entreprises ne disposant pas de pôle R&D. La fouille de texte est à cet égard un facteur de croissance. Cela résulte des externalités positives de son développement, et ce sur trois points : l'innovation-produit, la productivité et l'augmentation du bien-être du consommateur. Cet impact économique est lié à la largeur juridique du TDM : plus les possibilités du TDM sont vastes, plus l'impact économique sera important.

2.2.5 Aider à la décision publique

143. Le TDM peut être un facteur d'amélioration de la prise de décision. Dans le secteur public, il faciliterait le développement des evidence-based policies (EBP). L'EBP est une approche politique basée sur l'analyse empirique de situations. La pratique consistant à analyser de larges bases de données factuelles à des fins décisionnelles s'est initialement développée dans le domaine médical (*evidence-based medicine*), au début des années 1990[51]. Cette approche s'est ensuite diffusée dans d'autres domaines de la décision publique tels que la protection de l'environnement et la sécurité.

144. Au Royaume-Uni, ces validations empiriques ont pris une place importante dans la vie publique sous le gouvernement de Tony Blair. Un guide d'utilisation a même été créé

---

[49] https://perso.limsi.fr/pz/

[50] http://www.sciencedirect.com/science/article/pii/S0166497211000113

[51] Laurent Catherine, Baudry Jacques, Berriet-Solliec Marielle, Kirsch Marc, Perraud Daniel, Tinel Bruno, Trouvé Aurélie, Allsopp Nicky, Bonnafous Patrick, Burel Françoise, Carneiro Maria José, Giraud Christophe, Labarthe Pierre, Mastose Frank, Ricroch Agnès, « Pourquoi s'intéresser à la notion d'evidence-based policy ? », Revue Tiers Monde, 4/2009 (n° 200).

http://www.cairn.info/revue-tiers-monde-2009-4-page-853.htm

par l'Overseas Development Institute (ODI) dans le but de diffuser ces pratiques[52]. Cette notion se décline en plusieurs branches : *evidence-based decision*, *evidence-informed decision, evidence-aware decision*. Le TDM permettrait d'optimiser les politiques économiques, sociales, environnementales, en ne se basant pas seulement sur les opinions et les modèles théoriques, mais également sur l'analyse factuelle.

145. Le TDM peut être utilisé pour d'autres catégories d'aide à la décision (pas nécessairement publique), par exemple en géographie (géographie économique, géographie sociale, géomarketing, etc.). L'idée est de croiser des données géo-référencés afin d'identifier des caractéristiques de zones géographiques. Pour ce faire, « les techniques descriptives du Data Mining et plus précisément la Classification Ascendante Hiérarchique (CAH) a été utilisée pour dégager le nombre de groupes homogènes basé sur le critère d'agrégation Ward et la métrique euclidienne et utilisant le logiciel de Data Mining R. » [53]. Les données traitées par TDM peuvent ainsi servir à une exploitation à l'aide de systèmes d'information géographiques (SIG).

146. L'économétrie est une science fondée sur les données (*data-driven*) qui pourrait également bénéficier des avancées du TDM. Cette discipline permet une observation du fonctionnement réel de l'économie, utile à l'analyse de situation et à la prise de décision publique et privée. Cependant, les liens entre l'économétrie et TDM/machine-learning/big data restent actuellement peu exploités. Une explication possible est la difficulté à établir des liens de causalité et à quantifier l'impact d'une variable sur un phénomène observé.[54]

147. Enfin, les activités de recherche et les publications qui en résultent impliquent une grande transparence et rigueur au niveau des méthodes, de la relecture par les pairs et des influences externes. Ces informations contribuent à éclairer la nature systémique du sujet étudié et à fournir un moyen d'évaluer la fiabilité des résultats. L'analyse de faits

---

[52] Overseas Development Institute, *Evidence-Based Policymaking: What is it? How does it work? What relevance for developing countries?,* November 2005.

[53] Marwa Chalgham, Abderrahmane Fadil, Abdelaziz Dammak. « Le Data Mining pour l'aide à la décision en géomarketing ». *ROADEF - 15ème congrès annuel de la Société française de recherche opérationnelle et d'aide à la décision*, Feb 2014, Bordeaux, France. <hal-00946452>

[54] Quora "Why is econometrics isolated from the big data revolution", 2013

https://www.quora.com/Why-is-econometrics-isolated-from-the-big-data-machine-learning-revolution

« Hal R. Varian, "Big Data: New Tricks for Econometrics", *Journal of Economic Perspectives*, vol. 28, no. 2, Spring 2014

empiriques grâce aux technologies TDM offre ainsi – dans certains cas – une possibilité d'analyser la véracité des affirmations politiques et d'apporter un appui d'aide à la décision. [55]

2.3 Quelle organisation pour le TDM ?

2.3.1 Les structures et centres de recherche à l'étranger sur le TDM

2.3.1.1 Au Royaume-Uni : le National Centre for Text Mining

148. Au Royaume Uni, le National Centre for Text Mining (NaCTeM) est une structure financée par l'Etat[56] et pilotée par l'Université de Manchester. Cette structure est née en 2004 pour répondre initialement aux besoins de la communauté universitaire.

149. Le NaCTeM vise au développement d'outils et de services relatifs au text mining[57]. Il propose des services de text mining, des logiciels développés par les équipes du NaCTeM ou par d'autres, des séminaires et ateliers sur le TDM, des tutoriels et démonstrations ainsi que des publications à partir de text mining.

150. Si ce site aspirait principalement à aider en data mining le milieu académique, il a désormais une portée et un public bien plus large y compris d'industriels[58].

151. Le NaCTeM propose de développer des outils et services sur mesures pour répondre aux besoins des chercheurs du monde académique ou industriel lorsque ceux proposés sur le site ne sont pas adaptés au contexte du projet[59]. Un exemple de services proposés par le NaCTeM est *Facta* + qui permet de trouver des associations entre des concepts biomédicaux[60] :

---

[55] NESTA, *Using Research Evidence. A Practice Guide*, Section A "What is evidence-informed decision-making, and why focus on research?" 2015

[56] http://www.nactem.ac.uk/

[57] http://openminted.eu/about/partners/univ-of-manchester-nactem/

[58] http://www.nactem.ac.uk/requestaccess.php

[59] http://www.nactem.ac.uk/customised.php

[60] http://www.nactem.ac.uk/facta/

152. Le NaCTeM veille à développer des outils logiciels qui permettent une interopérabilité, le manque d'interopérabilité étant l'un des plus grands défis auquel sont confrontés les chercheurs[61]. Le site du NaCTeM ne regroupe pas seulement des services et outils développés par ses soins. Des services et outils de text and data mining développés par des tiers sont également répertoriés.

153. Le NaCTeM pilote en outre des projets[62] relatifs au text and data mining et répertorie les évènements, nationaux et internationaux, en lien avec le text and data mining[63].

154. Le NaCTeM est le premier centre national dédié au text and data mining[64].

2.3.1.2 Aux Etats-Unis

155. Aux Etats-Unis, la National Science Foundation (NSF) est une agence indépendante du gouvernement des États-Unis, qui vise à soutenir financièrement la

[61] http://www.nactem.ac.uk/uima.php

[62] http://www.nactem.ac.uk/research.php

[63] http://www.nactem.ac.uk/news.php

[64] http://www.nactem.ac.uk/

recherche scientifique fondamentale. En 2002, 6 millions de dollars ont notamment été débloqués pour soutenir des recherches en data mining[65].

156. Quant aux agences fédérales, le Data Mining Reporting Act[66] invite celles qui utilisent le data mining ou le développent à établir un rapport annuel au Congrès[67].

157. Aucune structure de mutualisation des outils et de données aux fins de text and data mining ne semble exister aux Etats-Unis.

2.3.1.3 International Scientific Council (ICSU) et World Data System (WDS)

158. L'ICSU (International Scientific Council), créé en 1931, est le plus important organisme scientifique non gouvernemental au monde. Composé de 121 membres nationaux et 32 unions scientifiques internationales, il est chargé d'encourager les activités scientifiques et technologiques internationales et de défendre l'accès de tous à la science ainsi que l'accès universel à des données scientifiques de qualité et opérées dans une perspective de long terme.

159. L'ICSU a créé le World Data System en 2008[68] aux fins de promouvoir et faciliter l'échange de données entre les adhérents au WDS. L'ICSU World Data System vise à effectuer la transition de centres de données distincts vers un système de données global, interopérable qui incorporerait les technologies émergentes et des nouvelles activités scientifiques de données. La mission du ICSU WDS est de promouvoir un accès universel et équitable aux informations et données scientifiques fiables et de qualité, et ce dans toutes les disciplines[69].

160. En septembre 2016, l'ICSU WDS compte 98 membres[70]. Afin de devenir membre, les candidats doivent être certifiés[71], notamment sur des critères d'intégrité et de confidentialité[72] :

---

[65] https://www.nsf.gov/news/news_summ.jsp?cntn_id=103047

[66] Section 804, Implementing the Recommendations of the 9/11 Commission Act of 2007, intitulé The Federal Agency Data Mining Reporting Act of 2007 (Data Mining Reporting Act)

[67] https://www.dni.gov/index.php/newsroom/reports-and-publications/94-reports-publications-2011/619-data-mining-report

[68] https://www.icsu-wds.org/organization

[69] http://www.icsu-wds.org/services/data-sharing-principles

[70] http://www.icsu.org/what-we-do/interdisciplinary-bodies/wds/about

[71] http://www.icsu-wds.org/files/wds-certification-summary-11-june-2012.pdf

- « As part of the process of developing WDS, a certification procedure for evaluating candidates for membership was developed by the Scientific Committee to ensure the trustworthiness of WDS Members in terms of authenticity, integrity, confidentiality and availability of data and services. »

161. Des « data sharing principles », sorte de charte éthique, ont été établis, parmi lesquels se retrouve un principe de respect de la donnée et de son intégrité[73]:

- "Data, metadata, products, and information should be fully and openly shared, subject to national or international jurisdictional laws and policies, including respecting appropriate extant restrictions, and in accordance with international standards of ethical research conduct.

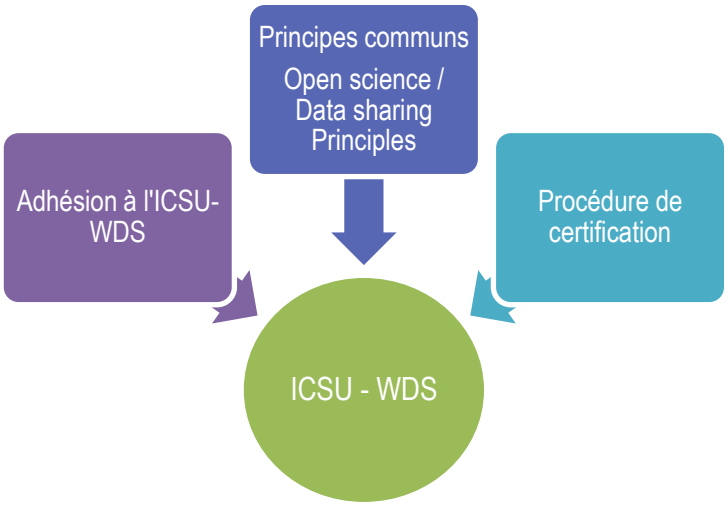
- Data, metadata, products, and information produced for research, education, and public-domain use will be made available with minimum time delay and free of charge, or for no more than the cost of dissemination, which may be waived for lower-income user communities to support equity in access.

- All who produce, share, and use data and metadata are stewards of those data, and have responsibility for ensuring that the authenticity, quality, and integrity of the data are preserved, and respect for the data source is maintained by ensuring privacy where appropriate, and encouraging appropriate citation of the dataset and original work and acknowledgement of the data repository.

- Data should be labelled 'sensitive' or 'restricted' only with appropriate justification and following clearly defined protocols, and should in any event be made available for use on the least restrictive basis possible."

[72] http://www.icsu-wds.org/services/certification

[73] http://www.icsu-wds.org/files/WDS_Data_Sharing_Principles_2015.pdf



2.3.1.4 L'Alliance pour les Données de Recherche (RDA-Research Data Alliance)

162. La RDA est une initiative conjointe de la Commission européenne, la NSF (National Science Foundation) et le NIST (National Institute of Standards and Technology) américains ainsi que le Département de l'Innovation du gouvernement Australien. L'objectif est de construire une infrastructure sociale et technique facilitant l'ouverture et le partage des données de recherche. C'est une initiative pilotée par les communautés de recherche elles-mêmes, qui compte aujourd'hui plus de 4300 membres venant de 111 pays. Les membres de RDA se répartissent dans des groupes de travail ou groupes d'intérêt thématiques[74], chargés de développer et de valider une infrastructure et accompagner la croissance d'une communauté des données intégrant des contributeurs de toute discipline scientifique et de toute origine géographique.

163. RDA, lancée en 2012, a d'ores et déjà produit des recommandations, testé et harmonisé des standards dans différents aspects des données. Ces premiers résultats concernent des éléments d'infrastructure en lien avec : la reproductibilité de la science, la préservation à long terme, les bonne pratiques de maintenance des réservoirs de données, les formations de spécialistes de données, la citation des données, etc.

2.3.2 L'encadrement proposé dans les projets européens H2020

164. Une prise de conscience politique de la nécessité d'encadrer légalement les opérations de TDM dans le domaine de la recherche a émergé :

[74] https://www.rd-alliance.org/groups/interest-groups

- la Commission européenne avait évoqué dans sa stratégie *Digital Single Market* la nécessité d'adapter la directive sur le droit d'auteur (Directive DADVSI) aux avancées technologiques afin d'éviter des disparités juridiques au sein du marché unique ;

- un rapport de la Commission européenne datant de juillet 2016 affirmait que l'absence de disposition juridique sur le TDM à des fins de recherche, au niveau de l'UE, entraîne des incertitudes qui sont nuisibles à la recherche et à l'innovation (R&I).[75]

165. Ces prises de position favorables au TDM ont contribué à orienter Horizon 2020 vers l'Open access et le TDM et le projet de directive Droit d'auteur dans le marché unique numérique introduit une exception en faveur du TDM.

166. Horizon 2020 contribue indirectement au développement du TDM en favorisant le libre accès aux publications et aux données de la recherche. En effet, ce programme européen pour la recherche et l'innovation impose la mise en libre accès, après un certain délai, de toutes les publications et de certaines données de recherche issues des projets financés. Une gestion transparente et pérenne des données de recherche doit être assurée à travers la rédaction d'un Data Management Plan (pour les projets participant au pilote). Afin d'accompagner cette démarche, la Commission européenne fournit des indications, rassemblées dans un guide, pour une bonne gestion des données de recherche[76]. La stratégie de la Commission européenne pour un libre accès aux données de la recherche est explicitée dans un rapport (Open Research Data Pilot[77]), publié en décembre 2013. D'après ce document, les projets Horizon 2020 doivent permettre à des tiers – autant que possible – d'accéder, de fouiller, d'exploiter, de reproduire et de disséminer (gratuitement pour les utilisateurs) les données issues de leurs recherches.

167. Horizon 2020 contribue également directement au développement du TDM à travers le financement de projets européens sur ce sujet. La Commission européenne investit dans la recherche sur le TDM avec par exemple l'appel à projet H2020 GARRI-3-

---

[75] European Commission, *Towards a modern, more European copyright framework*, Brussels, 9.12.2015, COM(2015) 626 final.

[76] European Commission, *Guidelines on FAIR Data Management in Horizon 2020*, 26 July 2016

[77] https://www.openaire.eu/opendatapilot

2014 « Scientific Information in the Digital Age: Text and Data Mining (TDM) » lancé en 2014. Cet appel a donné naissance à plusieurs projets, pour la période 2015-2018, dont :

- FutureTDM[78], "the Future of Text and Data Mining", vise à diffuser l'utilisation du TDM en Europe. Le projet organise notamment des forums afin d'obtenir des avis et retours d'expérience sur le TDM de la part de chercheurs, développeurs, éditeurs et autres acteurs liés au TDM.

- OpenMinTeD[79], "Open Mining INfrastructure for TExt and Data", vise à faciliter l'utilisation de technologies de TDM sur des publications scientifiques, en rendant interopérables des logiciels et plateformes, grâce à une standardisation.

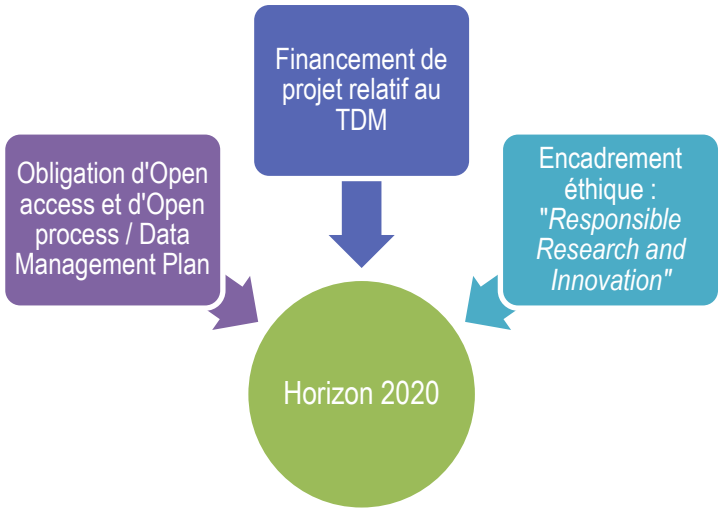
168. La part des projets de recherche sur le TDM parmi les projets financés par la Commission européenne peut se chiffrer grâce à une analyse par TDM. Parmi plus de trente mille projets financés par la Commission européenne pendant la période 2007-2016 (FP7 et H2020), environ 2,90% sont liés au TDM (soit 885 projets)[80] . Certains projets (0,38%, soit 115 projets) traitent spécifiquement du TDM. Ils comportent dans leur description les termes "Text Mining", "Data Mining" ou "Text and Data Mining". Les autres (2,53%, soit 770 projets) contiennent des termes proches tels que "Big Data", "Data Analysis" ou "Machine Learning". Cette analyse a également fait ressortir un glissement sémantique, avec l'expression "Text and Data Mining" utilisée de façon croissante.

169. Horizon 2020 propose également un encadrement éthique de la recherche. Les bonnes pratiques de recherche sont regroupées sous l'appelation *Responsible Research and Innovation* (RRI). L'objectif est de prendre en compte l'impact potentiel de chaque innovation sur la société, dans une logique d'anticipation, de développement durable et partagé. En particulier, le traitement des données personnelles doit s'effectuer dans le respect du droit à la vie privée.

[78] http://www.futuretdm.eu/

[79] http://openminted.eu/

[80] Toutes ces données sont issues de : FutureTDM, *Deliverable D4.1, European Landscape of TDM, Applications Report*, May 2016, p. 38.



2.3.3 Les réservoirs de données en France : quelques exemples

2.3.3.1 L'encadrement dans une discipline pionnière : l'astronomie

170. Dans l'article « Préserver les données de la recherche à l'ère du Big Data » précité, il est pris en exemple l'organisation de l'astronomie, discipline pionnière dans le domaine de la préservation et de partage des données.

171. L'article cite Françoise Genova, chercheuse à l'Institut national des sciences de l'Univers du CNRS et au Centre de données astronomiques de Strasbourg (CDS) qui mentionne que :

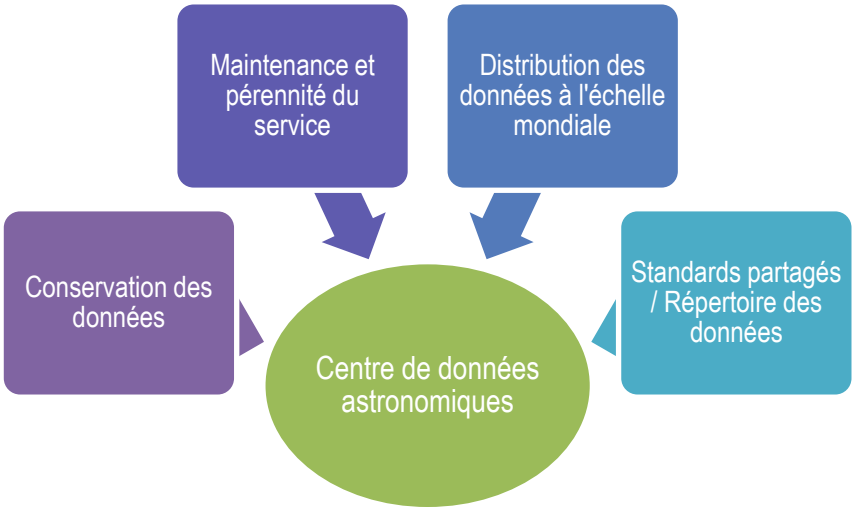
- « Pour comprendre les phénomènes physiques à l'œuvre en astronomie, nous avons besoin de rassembler des observations obtenues par différentes techniques et de travailler à partir de données obtenues par d'autres instruments et d'autres équipes ».

- « Pour répondre à ce besoin d'échange et de préservation des données, la communauté astronomique internationale s'est structurée autour de l'Observatoire virtuel, un ensemble de services qui permet de retrouver l'information utile parmi toutes les données astronomiques ouvertes aux chercheurs grâce à un grand répertoire et à des standards partagés. »[81]

172. Le Centre de Données astronomiques de Strasbourg (CDS) accueille la SIMBAD base de données astronomiques de référence mondiale pour l'identification des objets astronomiques. Il a pour mission la collecte, la valorisation et la distribution à l'échelle

[81] https://lejournal.cnrs.fr/articles/preserver-les-donnees-de-la-recherche-a-lere-du-big-data

mondiale de données astronomiques et des informations connexes. Il a la responsabilité de la conservation des données et de la maintenance du service ainsi que sa viabilité à long terme.



2.3.3.2 Huma-Num, dans le domaine des sciences humaines et sociales

173. Huma-Num est une très grande infrastructure de recherche (TGIR) visant à faciliter le tournant numérique de la recherche en sciences humaines et sociales.

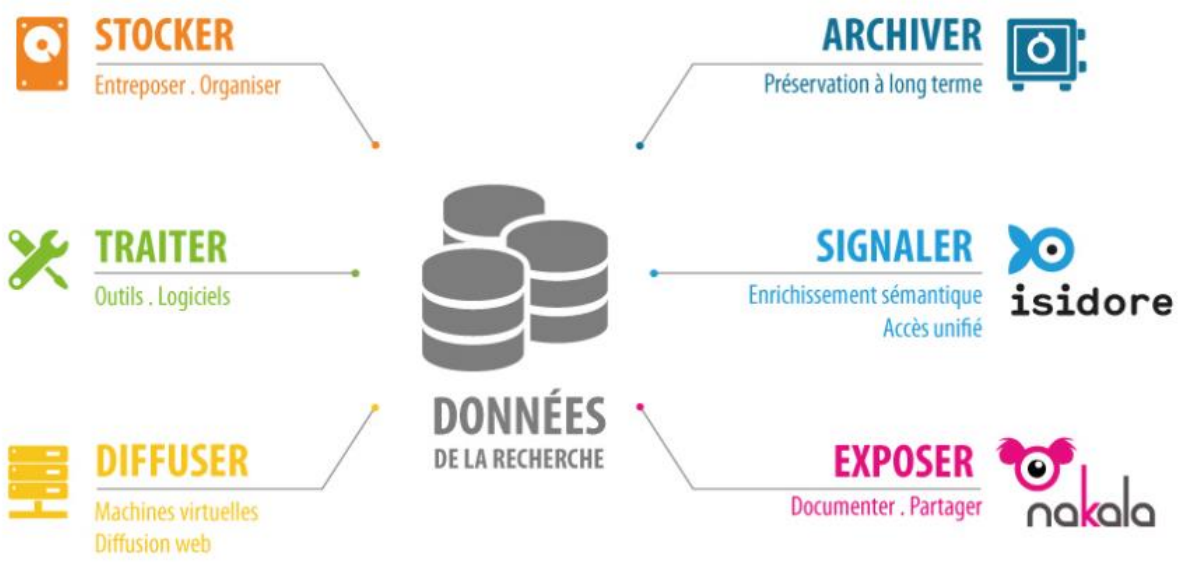
174. « Pour remplir cette mission, la TGIR Huma-Num est bâtie sur une organisation originale consistant à mettre en œuvre un dispositif humain (concertation collective) et technologique (services numériques pérennes) à l'échelle nationale et européenne en s'appuyant sur un important réseau de partenaires et d'opérateurs.

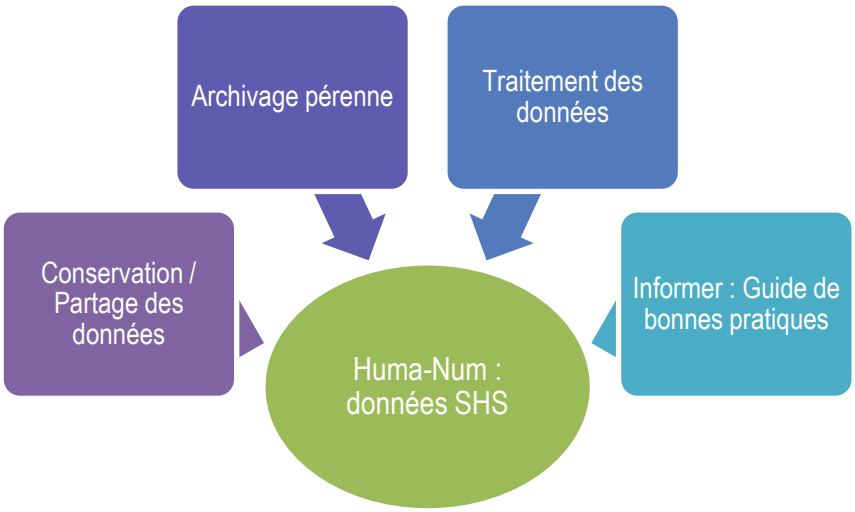
175. La TGIR Huma-Num favorise ainsi, par l'intermédiaire de consortiums regroupant des acteurs des communautés scientifiques, la coordination de la production raisonnée et collective de corpus de sources (recommandations scientifiques, bonnes pratiques technologiques). Elle développe également un dispositif technologique unique permettant le traitement, la conservation, l'accès et l'interopérabilité des données de la recherche. Ce dispositif est composé d'une grille de services dédiés, d'une plateforme d'accès unifié (ISIDORE) et d'une procédure d'archivage à long terme.

176. La TGIR Huma-Num propose en outre des guides de bonnes pratiques technologiques généralistes à destination des chercheurs. Elle peut mener

ponctuellement des actions d'expertise et de formation. Elle porte la participation de la France dans le projet DARIAH en coordonnant les contributions nationales.

177. La TGIR Huma-Num est portée par l'Unité Mixte de Services 3598 associant le CNRS, l'Université d'Aix-Marseille et le Campus Condorcet. »[82]





[82] http://www.huma-num.fr/la-tgir-en-bref

2.3.3.3 Le projet PREDON (PREservation des DONnées)

178. Le projet PREDON, issu du programme MASTODONS relatif à la coopération de différentes disciplines autour du concept « big data », est à l'initiative d'un petit groupe constitué de chercheurs de l'IN2P3[83]. Il a la mission de fédérer les initiatives au niveau national dans le domaine de la préservation des données scientifiques. Le projet propose une approche nouvelle basée sur les capacités scientifique, technique et organisationnelle d'unités de recherche, collaborations internationales et grands centres de calcul.

179. Le groupe PREDON est en liaison étroite avec des initiatives similaires au niveau national et international, notamment avec le panel de l'ICFA pour la préservation des données dans la physique des hautes énergies.

180. Le groupe de travail PREDON a produit en 2014 un document de synthèse (« Scientific Data Preservation 2014 ») qui résume les contributions des participants à des ateliers de travail. Le document est structuré en trois parties qui reflètent les aspects complémentaires de la préservation des données scientifiques : potentiel scientifique, méthodologie et technologies.

2.3.3.4 L'expertise du Cines

181. Le Centre Informatique National de l'Enseignement Supérieur (Cines) est un établissement public national placé sous la tutelle du Ministère de l'Enseignement Supérieur et de la Recherche. Le Cines a trois missions stratégiques nationales :

- le calcul numérique intensif,
- l'archivage pérenne de données électroniques,
- l'hébergement de plates-formes informatiques d'envergure nationale.

182. Dans le cadre de sa deuxième mission, le CINES assure l'archivage des données et documents numériques produits par la communauté de l'enseignement supérieur et la recherche française, y compris les données scientifiques d'observation, d'expérimentation, les résultats de calcul.

---

[83] Institut national de physique nucléaire et de physique des particules

183. « Il propose des solutions d'archivage numérique sur le moyen et long terme, mutualisées, économiques et personnalisables »[84] et « fort de sa double expérience en calcul intensif et en archivage pérenne et de ses infrastructures informatiques d'excellence », le Cines « accompagne les producteurs et gestionnaires de données dans leurs problématiques d'archivage »[85].

184. L'objectif principal est de garantir l'accessibilité, l'intégrité, la lisibilité et la compréhension des données aussi longtemps que nécessaire en adaptant le niveau d'exigences requis par l'archivage en fonction de cette durée.

185. Dans ce cadre, le Cines entreprend des actions de lobbying auprès des acteurs du marché afin que leurs formats de fichiers soient encore lisibles dans plusieurs années. Une équipe d'ingénieurs est engagée dans une course permanente contre l'obsolescence en veillant à ce que les logiciels ou lecteurs matériels puissent en permanence accéder aux données. La plateforme Facile[86] recense la palette de formats actuellement pris en charge par le Cines.

2.3.3.5 Un parallèle avec l'organisation de l'hébergement des données de santé

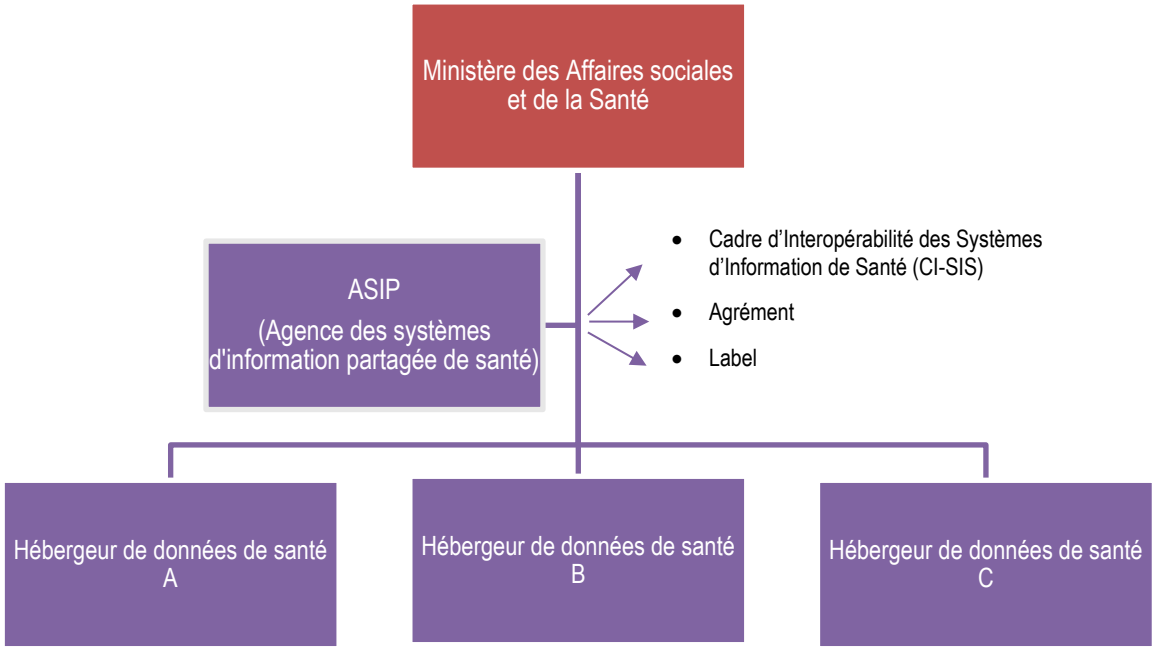
186. Il est intéressant d'observer l'organisation du système d'hébergement des données de santé en France afin de définir une organisation de la « conservation et la communication » des fichiers issus des traitements de données de la recherche.

187. L'organisation est la suivante :

[84] https://www.cines.fr/archivage/

[85] https://www.cines.fr/archivage/typologies/donnees-scientifiques/

[86] https://facile.cines.fr/



188. **Création.** L'ASIP Santé est un groupement d'intérêt public, établi entre l'Etat, la Caisse nationale d'assurance maladie (CNAM) et la CNSA (Caisse nationale de solidarité pour l'autonomie)[87]. C'est en 2009 que le groupement d'intérêt public dossier médical personnel (GIP-DMP) s'est transformé en Agence des Systèmes d'Information Partagés de santé (ASIP Santé)[88]. L'ASIP Santé est placée sous la tutelle du Ministère en charge de la santé[89].

189. **Rôle.** Le rôle de L'ASIP Santé est de favoriser et encadrer le développement des systèmes partagés dans les domaines de la santé et du secteur médico-social.

190. **Nécessité d'interopérabilité.** Afin de faciliter et d'accroitre les échanges entre les différents professionnels, des domaines de la santé et du secteur médico-social, une interopérabilité est nécessaire au niveau sémantique, syntaxique et technique[90]. La dématérialisation des données nécessite « la définition de langages communs aux

[87] http://esante.gouv.fr/partenaires/nationaux/339

[88] Arrêté du 8 septembre 2009 portant approbation de la convention constitutive d'un groupement d'intérêt public https://www.legifrance.gouv.fr/eli/arrete/2009/9/8/SASC0917305A/jo

http://esante.gouv.fr/asip-sante/espace-presse/communiques-de-presse/le-gip-dmp-devient-l-asip-sante

[89] http://esante.gouv.fr/actus/politique-publique/l-asip-sante-publie-un-etat-des-lieux-des-maitrises-d-ouvrage-regionales

[90] http://esante.gouv.fr/services/referentiels/referentiels-d-interoperabilite/cadre-d-interoperabilite-des-systemes-d

systèmes d'information amenés à les manipuler de manière à éviter la définition de nouveaux langages et donc de nouveaux développements à chaque fois que deux systèmes d'information veulent échanger ou partager des données ».

191. Conformément à sa convention constitutive, l'ASIP Santé établit des référentiels, standards, produits ou services contribuant à l'interopérabilité, à la sécurité et à l'usage des systèmes d'information de santé et de la télésanté, et veille de leur bonne application. Le Cadre d'Interopérabilité des Systèmes d'Information de Santé (CI-SIS) « fixe les règles d'une informatique de santé communicante »[91].

192. **Agrément.** L'hébergement de données de santé à caractère personnel nécessite, au vu de leur caractère sensible, l'obtention d'un agrément[92]. Un référentiel de constitution de dossier de demande d'agrément a été élaboré par l'ASIP Santé, qui est en charge de la pré-instruction de ces demandes[93]. Lors de ce pré-examen l'ASIP Santé va analyser trois axes, détaillés dans les formulaires que le demandeur est invité à remplir :

- un axe éthique et juridique (notamment les moyens mis en œuvre pour obtenir le consentement des personnes intéressées et les conditions relatives aux demandes de rectifications des données) ;
- un axe sécurité et technique (notamment des mesures offrant la garantie de la sécurité des accès et de la transmission des données, des mesures de contrôle des droits d'accès et traçabilité des accès et des traitements, des conditions de vérification des tentatives d'effractions et d'accès non autorisés, des modalités de vérification des registres des personnes habilitées et de leurs mises à jour) ;
- un axe économique et financier.

193. L'agrément est délivré par le Ministère en charge de la Santé pour une durée de trois ans et est renouvelable[94].

---

[91] http://esante.gouv.fr/services/referentiels/referentiels-d-interoperabilite/cadre-d-interoperabilite-des-systemes-d

[92] Article L. 1111-8 du Code de la Santé Publique

https://www.legifrance.gouv.fr/affichCodeArticle.do?idArticle=LEGIARTI000021941353&cidTexte=LEGITEXT000006072665

[93] http://esante.gouv.fr/services/reperes-juridiques/le-role-de-l-agence-des-systemes-d-information-partages-de-sante-dans-la

[94] Article R. 1111-15 du Code la Santé Publique

194. **Labellisation.** L'ASIP Santé propose en outre la délivrance de labels[95]. Les étapes de la labellisation prennent la forme suivante[96] :

195. **Observations.** L'arrivée d'une gouvernance par l'ASIP-Santé a été saluée par les professionnels du secteur des systèmes d'information sanitaires et sociaux: « Enfin un pilote clairement identifié qui œuvre dans un esprit de concertation »[97]. Les opérateurs et acteurs du secteur sont invités à participer à l'évolution du CI-SIS en exprimant leurs besoins, ce qui permet de légitimer les référentiels établis par l'ASIP Santé.

---

https://www.legifrance.gouv.fr/affichCodeArticle.do;jsessionid=60C3A828598DFE6EA728751C24A32AF5.tpdjo15v_2?cidTexte=LEGITEXT000006072665&idArticle=LEGIARTI000006908152&dateTexte=20130118&categorieLien=id#vig

[95] http://esante.gouv.fr/services/label-e-sante-logiciel-maisons-et-centres-de-sante

[96] http://esante.gouv.fr/services/labellisation/editeurs-comment-obtenir-le-label#P1

[97] Rapport d'activité 2009 de l'ASIP-Santé, p.29
http://esante.gouv.fr/sites/default/files/ASIP_Sante_Rapport_d_activite_2009.pdf

196. Quant à la procédure d'agrément en vue d'héberger des données de santé, si elle peut paraitre lourde en toute première approche, d'une part il est nécessaire d'assurer la sécurité des données conservées au vu de leur caractère sensible et d'autre part, l'ASIP Santé, par la constitution des formulaires, permet d'assister le demandeur. Il est à souligner que cette procédure est essentiellement déclarative puisque l'instruction se fait uniquement sur pièces. A noter, que la loi de modernisation de notre système de santé[98] prévoit à terme une évaluation de conformité technique par un organisme certificateur accrédité par le COFRAC en lieu et place de la procédure d'agrément. En conséquence, l'ASIP-Santé a publié le 16 septembre 2016 un référentiel de certification pour concertation[99]. Elle invite ainsi les opérateurs du secteur à faire des remarques et propositions. Elle pointe du doigt[100] :

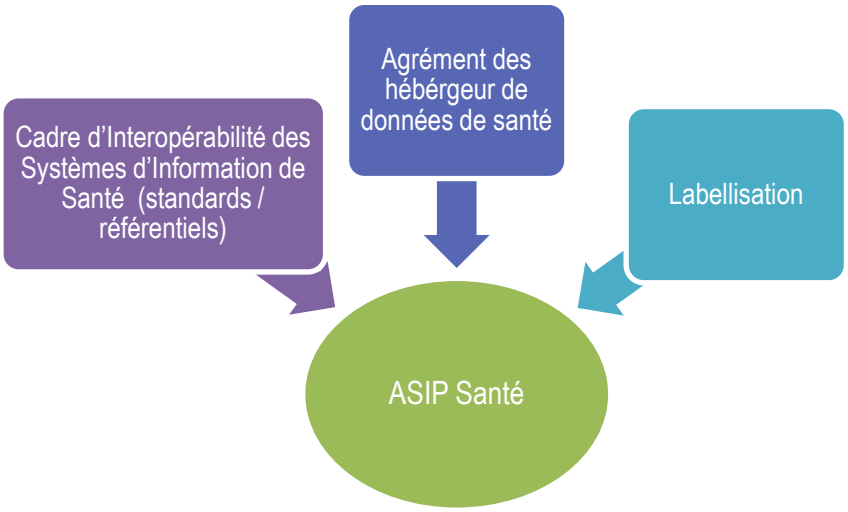
- l'absence de mise en œuvre de la possibilité de contrôle a posteriori donnée par la loi à l'Inspection générale des affaires sociales ;
- la nécessité de définir dans le référentiel des moyens à mettre en œuvre et non uniquement des buts à atteindre ;
- l'absence de visibilité, en conséquence, pour les demandeurs quant à leurs chances d'obtenir l'agrément ;
- la nécessité de moyens et ressources techniques pour l'instruction des dossiers ;
- l'absence d'audits externes par des auditeurs qualifiés ;
- la nécessité de prendre en compte l'évolution technique et l'offre commerciale des services d'hébergement.

---

[98] LOI n° 2016-41 du 26 janvier 2016 de modernisation de notre système de santé (1), article 204, 5e c : « I. - Dans les conditions prévues à l'article 38 de la Constitution et dans un délai de douze mois à compter de la promulgation de la présente loi, le Gouvernement est autorisé à prendre par ordonnances les mesures d'amélioration et de simplification du système de santé relevant du domaine de la loi visant à : (…) c) Remplacer l'agrément prévu au même article L. 1111-8 par une évaluation de conformité technique réalisée par un organisme certificateur accrédité par l'instance nationale d'accréditation mentionnée à l'article 137 de la loi n° 2008-776 du 4 août 2008 de modernisation de l'économie ou par l'organisme compétent d'un autre Etat membre de l'Union européenne. Cette certification de conformité porte notamment sur le contrôle des procédures, de l'organisation et des moyens matériels et humains ainsi que sur les modalités de qualification des applications hébergées »

[99] http://esante.gouv.fr/actus/services/agrement-des-hebergeurs-de-donnees-de-sante-publication-du-referentiel-de

[100] http://esante.gouv.fr/sites/default/files/asset/document/asip_sante_-_vue_densemble_referentiel_hds_-_v0.3.0.pdf



197. L'analyse de l'ensemble des cadres, acteurs existants et de leur mission et organisation permet d'esquisser et de proposer un cadre organisationnel de la conservation et du partage des données de la recherche scientifique aux fins notamment de TDM.

3. Une proposition d'encadrement global

3. PROPOSITIONS POUR L'APPLICATION DE LA LOI

198. Le projet de directive sur le droit d'auteur dans le marché unique numérique encourage les titulaires de droits et les organismes de recherche à définir ensemble de pratiques exemplaires relatives à la sécurité et à l'intégrité des réseaux et des bases de données où les travaux sont hébergés.

199. Par ailleurs, l'article 38 de la loi « pour une République numérique » prévoit que « un décret fixe les conditions dans lesquelles l'exploration des textes et des données est mise en œuvre, ainsi que les modalités de conservation et de communication des fichiers produits au terme des activités de recherche pour lesquelles elles ont été produites ».

200. L'analyse ci-dessus de la notion de TDM et de ses enjeux ainsi que l'observation des modèles organisationnels existants déjà en France, dans le programme H2020 ainsi qu'à l'étranger permet d'affiner une proposition d'encadrement de la pratique du TDM et de manière plus globale de l'Open science.

3.1 La définition de standards

3.1.1 Référentiel d'interopérabilité spécifique à l'Open science

201. Pour que les données et réseaux communiquent entre eux, il est nécessaire que des méthodologies, des standards soient rédigés afin d'assurer une interopérabilité entre les données de tout domaine scientifique.

202. Le réseau de conservateurs doit former un ensemble de réservoirs de données interopérables et parlant le même langage.

203. A la manière du Référentiel Général d'Interopérabilité (RGI) - la version 2.0 du RGI est officialisée par l'arrêté en date du 20 avril 2016 (JORF n°0095 du 22 avril 2016) - la science ouverte pourrait se doter d'un standard favorisant l'interopérabilité entre les conservateurs de données.

204. Le RGI est un cadre de recommandations référençant des normes et standards qui favorisent l'interopérabilité au sein des systèmes d'information de l'administration. Ces recommandations constituent les objectifs à atteindre pour favoriser l'interopérabilité. Elles permettent aux acteurs cherchant à interagir et donc à favoriser l'interopérabilité de leur système d'information d'aller au-delà de simples arrangements bilatéraux.

205. Le RGI est défini dans l'ordonnance n° 2005-1516 du 8 décembre 2005 relative aux échanges électroniques entre les usagers et les autorités administratives et entre les autorités administratives. Dans l'article 11 de cette ordonnance, le « RGI fixe les règles techniques permettant d'assurer l'interopérabilité des systèmes d'information. Il détermine notamment les répertoires de données, les normes et les standards qui doivent être utilisés par les autorités administratives. »

3.1.2 Procédure de certification ou d'agrément

206. Une procédure de certification ou d'agrément des conservateurs de données peut être définie.

207. Cette procédure d'évaluation des conservateurs de données doit permettre de s'assurer du respect des engagements des conservateurs de données, notamment en matière de :

- sécurité des données ;
- respect des standards et des formats de données ;
- respect des engagements relatifs aux infrastructures techniques et en réseau (maintien des plateformes de collecte et de mise à disposition des données, maintien des hubs pour la mise en réseau) ;
- mise à disposition des guides de bonnes conduites et charte éthique.

3.2 La création d'un réseau de conservateurs des données

3.2.1.1 Modélisation d'un conservateur type de données

208. Des structures de conservation et de mise à disposition des données existent déjà en France, le Centre de Données astronomiques de Strasbourg pour les données astronomiques ou encore le Huma-Num pour les données en sciences humaines et sociales.

209. Afin de modéliser un conservateur de données type, il est proposé de procéder à l'analyse précise de ces réservoirs de données existants notamment en termes de structure, de sécurité, de puissance, de capacité, de personnel, de fonctionnalités, d'utilisation par la communauté scientifique, d'indicateurs.

3.2.1.2 Mise en réseau des conservateurs

210. Il est proposé de s'appuyer sur ces structures existantes et de construire des structures proches dans chacune des disciplines scientifiques.

211. Ces structures doivent avoir comme fonctions communes :

- la conservation des données scientifiques de la discipline ;
- la mise à disposition et l'accessibilité des données scientifiques de la discipline par une plateforme ;
- l'enrichissement des données et la fourniture d'outils de traitement ;
- le maintien en conditions opérationnelles ;
- la tenue d'un répertoire des données de la discipline.

212. Par ailleurs, afin de permettre les recherches et l'analyse trans et multidisciplinaire, il convient de privilégier la mise en place d'une science en réseau[101]. S'il existe déjà certains échanges théoriques et de corpus, la création de réservoirs de données par discipline scientifique et leurs mises en réseau facilitera le partage de la connaissance.

213. Enfin, l'archivage des données et leur préservation dans le temps pourra être assurée par ces conservateurs de données avec l'appui du Cines.

214. Cette mise en réseau et la construction de hubs informatiques entre les conservateurs de données de chaque discipline scientifique implique la définition de formats standards et d'une réflexion sur l'interopérabilité des données.

3.3 Un encadrement éthique du TDM par une « charte éthique »

215. Les résultats obtenus par TDM doivent évidemment être interprétés et analysés avec un esprit critique. Un risque serait, par exemple, d'établir des liens de causalité abusifs pour expliquer les statistiques obtenues par TDM, menant à des corrélations fallacieuses. Le TDM ouvre ainsi des nouvelles questions méthodologiques.

216. L'intégrité de la recherche par TDM passe également par un respect de règles d'éthique. En particulier, le traitement des données personnelles doit s'effectuer dans le

[101] Florence Millerand, « La science en réseau. Les gestionnaires d'information « invisibles » dans la production d'une base de données scientifiques », *Revue d'anthropologie des connaissances*, Vol. 6, 2012/1, pp.163-190

respect du droit à la vie privée. Plus largement, la Commission européenne a regroupé des bonnes pratiques de recherche sous l'appellation Responsible Research and Innovation (RRI). L'objectif est de prendre en compte l'impact potentiel de chaque innovation sur la société, dans une logique d'anticipation, de développement durable et partagé. Par exemple, l' « évaluation des choix scientifiques et technologiques » (plus connue sous l'expression technology assessment) est un processus scientifique interactif dont l'objectif est de contribuer à faire émerger des opinions publiques et politiques sur de nouvelles technologies. Ainsi, les implications du TDM doivent être anticipées afin d'avoir un impact social bénéfique.

217. Il convient donc d'engager une réflexion profonde sur l'utilisation de ces outils. Comme le souligne le Comité d'éthique du CNRS à ce propos :

> « Il est devenu difficile d'appliquer dans tous les cas les principes de base du traitement des données personnelles, tels qu'informer les personnes sur le devenir et l'utilisation des données, ou obtenir leur consentement. Il peut arriver que la démarche du chercheur impose d'obtenir des informations à l'insu de la personne objet de son enquête. Il serait alors nécessaire de prévoir des principes à respecter s'il n'y a pas consentement, comme l'engagement à informer a posteriori cette personne. De même la question du consentement se pose quand les recherches portent sur les informations issues de la fouille de données sur des réseaux sociaux. Ces données, publiquement disponibles, sont considérées par la CNIL comme des données personnelles. »[102]

218. Le Comité d'éthique a également fait valoir que « face à cette dynamique de circulation des données relayée par leurs autorités de tutelle et par leur communauté, les chercheurs doivent :

- prendre conscience de leur responsabilité individuelle, déontologique[103] et éthique, vis-à-vis de la communauté à laquelle ils appartiennent ;
- avoir connaissance des engagements internationaux des institutions dont ils dépendent ;

[102] COMETS, *Les enjeux éthiques du partage des connaissances*, 2015, p.6

[103] "Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences", *Committee on Responsibilities of Authorship in the Biological Sciences, National Research Council*. National Academy of Sciences.

- participer à la définition de principes éthiques propres à leur discipline dans le domaine du data sharing et du big data en général. »[104]

219. Ce besoin de régulation par l'éthique a été exprimé par les chercheurs eux-mêmes lors de l'enquête sur les usages et besoins d'IST réalisée par le CNRS en mars 2015.

220. La mise en place d'une « charte d'éthique de l'IST » qui pose des « principes éthiques qui soient de nature à transcender les catégories instrumentales et à affirmer les buts de la recherche publique dans un contexte global de science ouverte » peut répondre en partie à ce besoin[105].

3.4 La formation des chercheurs et personnels de recherche aux pratiques de TDM

221. Face à cette utilisation des outils du TDM, c'est aussi la formation des chercheurs à la pratique de la fouille de textes et de données qui est en jeu. Actuellement essentiellement dispensés dans les cursus d'informatique, les enseignements du TDM sont un atout évident qui reste à développer. Pour que toutes les sciences et tous les domaines de recherche puissent en bénéficier, il convient d'élargir son étude et son apprentissage à un maximum de cursus.

222. Le TDM est un vecteur de dynamisme pour le capital humain d'un pays. L'ensemble des techniques qu'il représente sont demandeuses d'une main-d'œuvre qualifiée, voire très qualifiée.

3.4.1 La formation des métiers techniques

223. D'un point de vue infrastructurel, le TDM implique la multiplication de *data centers* et de tous les personnels techniques nécessaires à ce développement : métiers de la construction, métiers de la maintenance informatique et technique, métiers de l'ingénierie numérique.

[104] Autosaisine du COMETS « Les enjeux éthiques du partage des données scientifiques » Groupe Data Sharing 12-12-2014

[105] Résultats de l'enquête sur les usages et besoins d'IST des Unités de recherche, réalisées auprès des Directrices et Directeurs d'Unités du CNRS - mars 2015 page 59

3.4.2 L'émergence de nouveaux métiers et qualifications

224. Au-delà des infrastructures, c'est l'ensemble des métiers techniques spécifiques au TDM qu'il convient aussi de développer. Ces professions, spécialistes du *big data* ou du traitement des bases de données, forment un capital humain essentiel pour le dynamisme économique national. Leur formation est donc centrale.

3.4.3 La formation initiale des chercheurs

225. Actuellement, trop peu de formations spécialisées sont orientées vers les connaissances issues du TDM. On notera dans le cas des universités un master mention informatique spécialisé en *data mining* à l'Université Lyon 2[106] ou encore à l'Université Paris 8[107], spécialisé en *Big data et fouille de textes*. Ces masters reviennent à former des ingénieurs informatiques compétents en *deep learning*, analyse et management des données, modélisation par graphes des résultats de fouille de textes… En plus des possibilités pour devenir chercheur ou enseignant-chercheur dans ces disciplines informatiques, les formations donnent aussi accès aux métiers de *data scientist* ou d'ingénieur d'étude, voire de chargé d'étude statistique. L'Université Paris 6 propose aussi un parcours de spécialisation en master, ainsi que celle de Nice Sophia-Antipolis ou Paris 13. Toutefois, ces formations sont restreintes (on compte 20 places en M2 à l'Université Paris 8 par exemple).

226. Des spécialités d'analyse des données sont enfin dispensées dans des cursus économiques, comme à l'ESSEC (*Data Science and Business Analytics*) ou encore à la *Toulouse Business School* (*Digital Intelligence and Marketing Intelligence*).

227. Or, les besoins se multipliant, la généralisation de tels parcours en fouille de données paraît nécessaire.

228. On retrouve, à côté de ces formations universitaires ou spécialisées, des spécialisations en écoles d'ingénieurs. Polytechnique a par exemple lancé en 2014 une chaire de *data scientist*, en partenariat avec des entreprises comme Keyrus, Orange ou Thales[108]. Télécom Paritech est dans une même dynamique de formation, ainsi que

---

106 http://master-datamining.univ-lyon2.fr/

107 http://www.univ-paris8.fr/Master-MIASHS-Big-Data-et-fouille-de-donnees

108 https://www.polytechnique.edu/fr/polytechnique-keyrus-orange-thales-creent-une-chaire-data-scientists

d'autres écoles, réparties trop inégalement sur le territoire pour offrir des parcours aussi riches et valorisés qu'aux Etats-Unis, en Allemagne ou au Royaume-Uni.

229. La centralité du TDM dans le marché de l'emploi a mieux été prise en compte dans d'autres pays, comme les pays anglo-saxons. C'est sur ce point que la France doit rattraper un retard de formation, malgré l'excellence des structures existantes. En Allemagne, des masters en *compututional and data science* ou en *data and knowledge engineering* sont plus développés (Dortmund, Berlin notamment). Le Royaume-Uni et les Etats-Unis possèdent néanmoins les infrastructures et centres de formation les plus avancées dans ces domaines. On ne compte pas moins d'une vingtaine de formations différentes au Royaume-Uni, à Coventry, Leicester, Londres, Edinburgh, Lancaster, Nottingham, Oxford, Bristol, Norwich, Glasgow ou Sheffield. Une telle répartition géographique, éclatée sur le territoire national, a aussi comme corollaire une répartition technique : formation théorique et généraliste (*Data Science, Machine Learning*), formation appliquée (Applied Data Analytics), formation spécialisée (Business Intelligence Systems)…

230. Dans le cas des Etats-Unis, la formation est aussi riche des centres de recherche spécialisés en TDM et en data science. C'est le cas de l'Université de New York et de son Center for Data Science notamment, ou de l'Université d'Harvard et le Berkman Center for Internet and Society. Un ensemble d'autres universités forment aussi au TDM, de façon plus développée qu'en France : Stanford, MIT, Northwestern University, Penn State, Indiana, University of Maryland, Carnegie Mellon, Berkeley…

Ces formations sont essentielles pour dynamiser le capital humain mais aussi pour enrichir l'accroissement des connaissances.

3.4.4 La formation continue et les actions de sensibilisation

231. Il convient également de sensibiliser et de former les doctorants, chercheurs et enseignants-chercheurs au dépôt, au partage et aux techniques de traitement des données.

232. Cette sensibilisation peut passer par la rédaction d'un guide bonnes pratiques de l'Open science, par des tutoriels et des offres de formation en ligne (e-learning).

233. Cette offre de formation globale doit être structurée et s'articuler avec les acteurs nationaux déjà présents (Huma-Num, Inist, Persée, CCSD, URFIST, OST…) afin de proposer une offre nationale de formation en IST pour l'ensemble des communautés scientifiques.

3.5 La création d'une agence nationale de la science ouverte

234. A l'étranger et notamment dans les pays anglo-saxons il n'a pas été identifié d'agence nationale ou autre autorité qui organise la collecte et la mise à disposition des données scientifiques. La tradition de droit coutumier des pays anglo-saxons explique cette absence de structure.

235. Compte-tenu de la structure administrative française et de la multitude des missions nécessaire à la mise en œuvre de l'Open science, il est proposé la création d'une Agence nationale de la Science ouverte. Cette Agence pourrait prendre la forme d'une Autorité Administrative Indépendante. Elle constituerait un lieu de convergence des points de vue entre éditeurs scientifiques et communautés scientifiques.

236. Les missions suivantes pourraient être dévolues à cette Agence nationale de la Science ouverte:

- la rédaction et le respect de référentiel d'interopérabilité spécifique aux données scientifiques ;
- la création d'un réseau de conservateurs de données par disciplines scientifiques et le respect de leurs missions ;
- la formation des chercheurs ;
- la rédaction et le respect d'un guide de bonnes pratiques ;
- la rédaction et le respect d'une charte éthique ;
- la définition et le suivi de la procédure de certification ou d'agrément des conservateurs de données ;
- la tenue d'un répertoire des données scientifiques.

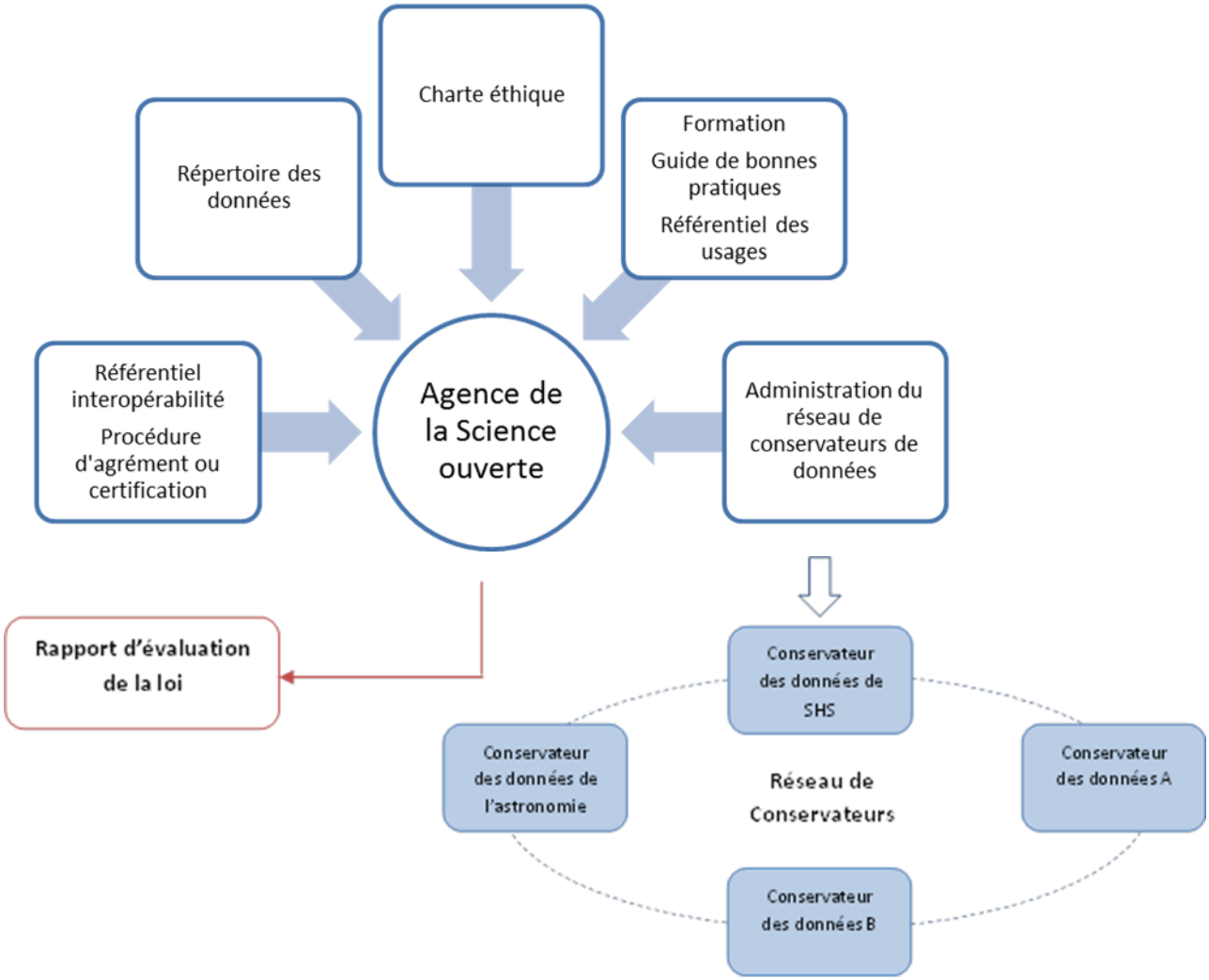
237. Cette agence nationale sera garante de l'efficience de la Science ouverte et de sa mesure. Des indicateurs relatifs par exemple au partage des données par discipline, à l'utilisation des outils d'analyse, à l'émergence de nouvelles connaissances par l'utilisation de ces outils, l'impact de ces outils sur le monde de l'édition scientifique pourront être développés.

238. En effet, cette agence nationale pourra apporter son concours à l'évaluation de l'application progressive des dispositions légales dans le cadre d'un rapport d'évaluation

des dispositions relatives à la science ouverte de la loi pour une République numérique et ses décrets d'application.

3.6 Schéma de synthèse de l'encadrement global

239. Le schéma ci-dessous résume l'ensemble des propositions d'encadrement formulées :



3.7 Des lignes directrices pour le décret d'application de l'article 38

3.7.1 L'affirmation des principes de la science ouverte

240. Le Code de la recherche organise institutionnellement les organismes qui participent en France à la recherche scientifique mais aucun texte ne définit les principes et les valeurs de la communauté scientifique.

241. Un droit de la Science, défini par consensus par les chercheurs pour la recherche publique, serait la retranscription des valeurs des communautés scientifiques telles que :

- le partage de la connaissance ;
- le libre accès aux données scientifiques ;
- la liberté de traitement des données scientifiques.

242. Un texte posant les principes d'une Science ouverte permettrait à la France d'être pionnière dans ce domaine.

243. Ce décret d'application pourrait en article 1 affirmer ces valeurs et proposer des définitions (article 2) des termes structurants du secteur comme les notions :

- d'exploration et de fouille de textes et de données ;
- de la version du manuscrit objet de l'embargo ;
- du périmètre de la notion de « textes et de données incluses ou associées aux écrits scientifiques ».

3.7.2 La création d'une science en réseau

244. Le décret pourrait prévoir le principe d'une architecture en réseau à partir de conservateurs de données par discipline scientifiques.

245. Ces conservateurs de données pourraient faire l'objet d'une procédure d'agrément.

246. Ils auraient en charge la collecte, la conservation et la mise à disposition des données scientifiques, d'assurer la sécurité de leur système d'information, de respecter les standards et référentiel d'interopérabilité permettant une communication en réseau.

3.7.3 Un contrat type de cession de droits entre les auteurs et éditeurs

247. Afin de garantir les droits des chercheurs sur leurs publications et de prendre en compte les risques d'asymétrie contractuelle, le décret pourrait organiser un contrat type de cession de droit d'auteur destiné à la recherche publique.

248. Ce contrat définirait les règles du jeu entre les parties et la protection du chercheur dans sa relation avec l'éditeur. Il permettrait notamment de s'assurer de l'absence de cession à titre exclusif et de garantir les droits des chercheurs en :

- autorisant le dépôt et la reproduction en archive ouverte de la publication dans la version auteur immédiatement et dans la version éditeur après le respect d'une période d'embargo ;
- permettant l'exploration immédiate du contenu de l'article à partir d'outils numériques de traitement de données ;
- empêchant toutes formes de privatisation ou de réserve de propriété sur le contenu de l'article.

249. Ce contrat pourrait faire l'objet d'un décret et ainsi avoir une valeur réglementaire qui s'imposerait à l'éditeur pour toute publication scientifique constituant un résultat de la recherche publique.

3.7.4 La création d'un référentiel d'interopérabilité et de standards

250. Le principe d'un référentiel d'interopérabilité spécifique aux données scientifiques pourrait être prévu dans ce décret.

251. La mise en réseau des conservateurs de données par discipline scientifique nécessite une communication entre les systèmes d'information et le respect de formats de données et de standards.

3.7.5 Une charte éthique de la Science numérique

252. Le Comité d'éthique du CNRS soutenu par le Conseil Scientifique du CNRS soutient l'introduction d'une charte éthique de la Science numérique. Cette charte définirait les valeurs d'accès et de partage des données scientifiques ainsi que les bonnes pratiques des chercheurs telles que :

- le dépôt des données scientifiques sur des plateformes Open science ;
- le respect des mentions de paternité.

253. Un comité d'éthique serait garant du respect de cette charte en veillant notamment à :

- la diffusion et à la compréhension de son contenu ;
- la sensibilisation des chercheurs à l'importance de l'éthique: « Les chercheurs et les personnels du monde de la recherche doivent être formés aux dimensions éthiques de la gestion des données, en particulier au respect de la vie privée, de

la propriété intellectuelle, de la qualité et de l'intégrité des données. Ils doivent être informés de l'état actuel et de l'évolution des règles juridiques concernant le partage responsable de données utilisées »[109] ;

- formuler des avis assortis de recommandations afin de préciser les lignes de conduite définies dans la charte.

3.7.6 La création d'une agence nationale de la science ouverte

254. Le décret pourrait prévoir la création d'une agence nationale de la science ouverte, lieu de convergence des points de vue, de sécurisation et de contrôle des pratiques, de doléances des acteurs de l'Open science.

255. Elle aurait notamment pour rôles :

- le contrôle du respect des principes et valeurs de la science ouverte ;
- le contrôle du respect des règles éthiques posées dans la Charte éthique ;
- l'administration du réseau de conservateurs de données ;
- la création d'un référentiel d'interopérabilité des données posant des standards et des formats de données ;
- la procédure d'agrément des conservateurs de données ;
- la rédaction d'un guide de bonnes pratiques ;
- la tenue du répertoire des données ;
- le suivi de la formation des chercheurs ;
- le suivi des évolutions techniques et des usages ainsi que la gestion du référentiel des usages ;
- un rôle consultatif ;
- des propositions d'adaptation des cadres légaux existants au regard de l'évolution des pratiques et de besoins.

[109] Avis COMETS « Les enjeux éthiques du partage des données scientifiques » 7-6-2015

256. L'Agence pourrait également être en charge de la rédaction du rapport sur l'impact du principe de libre accès aux données scientifiques sur le marché de l'édition scientifique et sur la circulation des idées et des données scientifiques.

257. Elle pourrait également être à l'initiative d'une Convention internationale de la Science ouverte universelle.

3.7.7 La création d'une agence européenne de la Science ouverte

258. Dans le prolongement de l'article 3 du projet de la directive sur le droit d'auteur dans le marché unique numérique introduisant une exception au droit d'auteur et au droit du producteur de base de données, il pourrait être créé une agence européenne de la Science ouverte.

259. Le modèle de l'Agence française pourrait être étendu au niveau européen et la France pourrait être à l'initiative de cette création.

260. En effet, les enjeux de l'Open science sont universels et appellent un traitement harmonisé des valeurs de partage et d'accès à la connaissance au moins au niveau communautaire.

261. Cette agence européenne pourrait également être le vecteur d'un discours universel auprès de l'Organisation de Coopération et de Développement Économiques (OCDE) par exemple.

Annexes

Tableau d'analyse croisée des législations TDM

	France	Union européenne	Royaume-Uni	Etats-Unis	Japon
Fondement	Exception au droit d'auteur et au droit du producteur de base de données : droit de copie et de reproduction numérique aux fins de TDM	Exception au droit d'auteur et au droit du producteur de base de données : droit de reproduction ou d'extraction aux fins de TDM	Exception au droit d'auteur aux fins « d'analyse computationnelle »	Décision de justice Exception du fair use	Exception au droit d'auteur aux fins d'« analyse de l'information » aux fins de comparaison, classification, analyse statistique.
Périmètre	Fouille de textes et de données incluses ou associées aux écrits scientifiques	TDM sur des œuvres ou d'autres objets	Les œuvres et toutes données associées	Œuvres	Information de toute nature
Bénéficiaire	/	Les organisations de recherche (la notion est définie de manière large à l'article 2 du projet de Directive)	/	Parties au litige	/
Limite	TDM limité aux besoins de la recherche scientifique / dans un cadre de recherche But non-commercial Source licite / accès licite aux textes et données objets du TDM	TDM limité aux besoins de la recherche scientifique But non-commercial Accès légal aux objets du TDM	Limité au seul but de la recherche. Fins non commerciales. Accès licite. Mention de paternité. Interdiction de communication de la copie réalisée à un tiers / interdiction de conclure un contrat de cession ou licence de la copie réalisée	But non lucratif Nature de l'œuvre protégée par le droit d'auteur Portion de l'œuvre utilisée Absence d'impact économique de l'usage.	Pas de limitation à la recherche publique no à des fins non commerciales

Références bibliographiques

1. Textes et législations sur le TDM

HM GOVERNMENT, *Modernising Copyright. A modern, robust and flexible framework*, 2012

EUROPE, *Rapport Reda*, 2015

EPRS, Tambiama Madiega, *EU copyright reform: Revisiting the principle of territoriality*, 2015

Loi pour une République numérique, dont l'étude d'impact du 9 décembre 2015[110]

RAPPORT SIRINELLI pour le CSPLA « Rapport de la mission sur la révision de la directive 2001/29/CE sur l'harmonisation de certains aspects du droit d'auteur et des droits voisins dans la société de l'information » de décembre 2014

ETUDE DU CABINET WOLF & PARTNERS, *Study on the legal framework of text and data mining*, pour la Commission européenne, mars 2014

RAPPORT DU GROUPE D'EXPERTS de la Commission européenne, *Standardisation in the area of innovation and technological development, notably in the field of Text and data mining*, avril 2014

PROPOSAL FOR A DIRECTIVE of the European Parliament and of the Council on copyright in the Digital Single Market, COM(2016) 593 final, 14-9-2016

2. Analyses institutionnelles du TDM

S. ANANIADOU, The National Centre for Text Mining: A Vision for the Future, 2007

J. CLARK, Text Mining and Scholarly Publishing, Study Commissioned by the Publishing Research Consortium (PRC), Amsterdam, 2013

CNRS, Livre blanc, Une science ouverte dans une République numérique, 2015

COUPERIN et ADBU, Mission relative au data mining : l'analyse de Couperin et de l'ADBU, 2014

[110] http://www.assemblee-nationale.fr/14/projets/pl3318-ei.asp

DE WOLF & PARTNERS, Study on the Legal Framework of TDM, 2014

U.K. IPO (Intellectual Property Office), Impact Assessment (IA), Exception for copying of works for use by text and data analytics, 2012

JISC, « The Value and Benefits of Text Mining », Digital Infrastructure Directions Report, Doc#811, 2012

J. KELLY, "The Text and data mining copyright exception: benefits and implications for UK higher education", JISC Publications, 2016

NESTA, Alliance for useful Evidence, *Using Research Evidence. A Practice Guide*, 2015

OUTSELL, *Text and Data Mining: Technologies Under Construction*, 2016

SCIENCE EUROPE, *Text and Data Mining and the Need for a Science-friendly EU Copyright Reform*, Briefing Paper, 2015

Guillaume GARVANESE *Préserver les données de la recherche à l'ère du Big Data,* 9-9-2016 https://lejournal.cnrs.fr/articles/preserver-les-donnees-de-la-recherche-a-lere-du-big-data

EPRIST (Responsables IST des organismes de recherche) note sur le text and data mining « Le TDM comme outil innovant de recherche scientifique »

3. Travaux de recherche sur le TDM

M. BORGHI and S. KARAPAPA, *Copyright and Mass Digitization: a Cross-Jurisdictional Perspective*, Oxford University Press, 2013

W. FAN, L. WALLACE, S. RICH and Z. ZHANG, *Tapping into the Power of Text Mining,* 2005

U. FAYYAD and R. UTHURUSAMY, "Data mining and knowledge discovery in databases: Introduction to the special issue", *Communications of the ACM,* 39(11), 1999

W. J. FRAWLEY, G. PIATETSKY-SHAPIRO and C. J. MATHEUS, *Knowledge Discovery in Databases: An Overview,* 1992

HANDKE, GUIBAULT et VALLBE, "Is Europe Falling Behind in Data Mining? Copyright's impact on Data Mining in Academic Research", Juin 2015

M.A. HEARST, *Untangling Text Data Mining*, School of Information Management & Systems, University of California, Berkeley, 1999

M.A. HEARST, *What is Text Mining?*, SIMS, UC Berkeley, October 17, 2003

F. IBEKWE-SANJUAN, *Fouille de textes : méthodes, outils et applications*, Coll. Systèmes d'information et organisations documentaires, Hermès, 2007, 352 p.

S. JUSOH and H. M. ALFAWAREH, *Techniques, Applications and Challenging Issue in Text Mining,* November 2012

C. LAURENT, J. BAUDRY et al., « Pourquoi s'intéresser à la notion d' « evidence-based policy » ? », Revue Tiers Monde 4/2009 (n° 200), p. 853-873

F. MILLERAND, « La science en réseau. Les gestionnaires d'information « invisibles » dans la production d'une base de données scientifiques », *Revue d'anthropologie des connaissances*, Vol. 6, 2012/1, pp.163-190

Y. TOUSSAINT, « Extraction de connaissances à partir de textes structurés », *Document numérique, Vol.8*, 3/2004, pp. 11-34

Remerciements

Nos remerciements aux rédacteurs de ce Guide stratégique :

- Le cabinet Alain Bensoussan :
    - Maître Alain BENSOUSSAN
    - Maître Sarah LENOIR
- Les collaborateurs de la DIST du CNRS :
    - Renaud FABRE, Directeur de la DIST du CNRS
    - Laurence EL KHOURI, Directrice adjointe de la DIST du CNRS
    - Francis ANDRÉ, Expert national en données de recherches
    - Stéphanie DOS SANTOS, Chargée d'études
    - Quentin MESSERSCHMIDT-MARIET, Chargé d'études
    - Marc ROUX, Chargé de projets