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

Le principe d'ouverture des données de la recherche scientifique. Réflexions autour du croisement de l'informatique et du droit

Agnès ROBIN *Maître de conférences HDR*
Droit
LICeM
University of Montpellier

Édition électronique :
URL :
https://intelligibilite-numerique.numerev.com/articles/revue-1/2160-le-principe-douverture-des-donnees-de-la-recherche-scientifique-reflexions-autour-du-croisement-de-linformatique-et-du-droit
DOI : numerev_1690
Date de publication : 13/05/2020

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

Pour **citer cette publication** : ROBIN, A. (2020) Le principe d'ouverture des données de la recherche scientifique. Réflexions autour du croisement de l'informatique et du droit. *Intelligibilité du numérique*, (1). https://doi.org/10.34745/numerev_1690

Les données de la recherche scientifique sont actuellement soumises à un programme de standardisation technique (FAIR) dont l'objectif est d'en permettre la diffusion aux fins de réutilisation par le public (entreprises privées ou autre). Cette politique, qui sans se confondre avec elle, converge avec celle dite de « science ouverte », s'articule autour d'un principe normatif conflictuel, selon lequel le résultats de la recherche (et donc les données) doivent être « aussi ouverts que possibles et pas plus fermés que nécessaire », obligeant alors les chercheurs, ingénieurs et documentalistes, éventuellement chargés de la gestion des données de la recherche, à procéder à une qualification juridique délicate des données.

Scientific research data is currently subject to a technical standardization program (FAIR), the objective of which is to allow its dissemination for re-use by the public (private companies or other). This policy, which, without being confused with it, converges with that known as "open science", revolves around a conflicting normative principle, according to which the results of research (and therefore the data) must be "as open as possible and no more closed than necessary ", thus forcing researchers, engineers and documentalists, possibly responsible for the management of research data, to carry out a delicate legal qualification of the data.

Mots-clés :
Informatique, Données, Droit, Principe d'ouverture, FAIR, Data, Computer science, Law, Opening principle, FAIR

Pierre Catala, professeur à la Faculté de Droit de Montpellier de 1963 à 1968, avait publié en 1998 un ouvrage qui présentait les travaux menés en informatique juridique dans les années 1970-1980[1]. Il y écrivait que « l'informatisation du droit est une chose trop sérieuse pour la laisser aux informaticiens » (Catala, 1998, p. 1). L'auteur poursuivait en observant que les juristes avaient un rôle majeur à jouer et initiait une recherche interdisciplinaire à une époque où la découverte des potentialités de l'informatique nourrissait des fantasmes de type techno-scientifiste. La recherche avait été l'occasion de reprendre les travaux menés antérieurement, dès les années 1970, sur la confrontation entre la logique informatique et la logique juridique, ou plus exactement entre la rationalité du droit et la « logique des ordinateurs ». Et, si l'on

trouve des tentatives de modélisation de la rédaction des contrats, c'est surtout les travaux de la doctrine (informatisation de l'appareil documentaire) et des juges (jurisprudence) qui retient l'attention des pionniers de l'informatique juridique. Il s'agit à l'époque d'assurer l'accès au droit. De cette recherche interdisciplinaire entre universitaires (juristes et informaticiens) est née, sous l'impulsion de Pierre Catala, la base de données JurisData, très prisée des juristes, et qui s'avère être un bel exemple de valorisation de la recherche publique scientifique menée en interdisciplinarité. Dans ces différentes déclinaisons, l'informatique juridique mettait alors déjà à l'épreuve la rationalité de l'expression juridique.

Mais des espoirs étaient également formulés à l'endroit de la Justice : l'informatique devait permettre d'éclairer le degré de rationalité qui préside à la décision judiciaire (*ibid.*, p. 8-9). Pierre Catala prédisait à cet égard qu'au XXIe siècle, l'analyse automatique du discours deviendrait une technique courante (*ibid.*, p. 51). C'est aujourd'hui une chose en train de se faire : en témoignent les multiples projets de recherche sur la justice prédictive[2]. Si l'objectif poursuivi est encore bien sûr d'analyser le langage juridique et de concevoir des outils d'aide à la décision, il s'agit également de concevoir des systèmes-experts capables de mesurer l'aléa judiciaire, et donc de mesurer le risque juridique. Or, ces analyses ne peuvent être menées sans qu'aient été collectées toutes les décisions rendues. Ceci explique aujourd'hui la mise en application de la politique d'ouverture des données dans le domaine judiciaire. Etant, en vertu du Code des relations entre le public et l'administration (CRPA), considérées comme des données publiques, les données judiciaires sont soumises à une obligation de diffusion[3].
Les questionnements de notre contribution sont directement liés aux réflexions produites par le professeur P. Catala. Il s'agit en effet toujours de questionner l'interaction entre la règle de droit et l'informatique, et précisément de comprendre si, et comment, la loi (au sens large) peut être implémentée dans le code informatique, dans l'écriture algorithmique. L'informatique et le droit entretiennent un rapport étroit en raison de ce qu'ils produisent tous deux du langage. La notion de code elle-même permet d'entrevoir cette similarité : qu'on l'entende comme le code secret servant à échanger des informations en toute confidentialité, comme système d'informations et de symboles permettant de représenter une information dans un système technique ou scientifique (ex. code génétique ou code informatique), c'est-à-dire finalement un système permettant de transcrire un message, qu'on l'entende comme un recueil ou une compilation de textes de loi ou au contraire, comme une norme non écrite (coutume), ou enfin comme un support permettant de tenir un registre d'informations[4], le terme « code » s'avère être tout à la fois une forme d'expression du droit et de l'informatique.

Il s'agit plus précisément de comprendre si, et surtout comment, le droit et l'informatique peuvent dialoguer dans la mise en œuvre de la politique de science ouverte. A cet effet, nous retiendrons une définition large de la notion de donnée de la recherche scientifique qui mérite cependant d'être explicitée. Tout d'abord, la notion de donnée elle-même est officiellement définie comme la représentation de l'information sous une forme conventionnelle ou codifiée, cette forme permettant d'en faciliter le

traitement (informatique)[5]. Ensuite, la donnée n'est scientifique que parce qu'elle est intégrée dans un processus scientifique d'exploration et de compréhension de la réalité, lui-même ne pouvant, à notre sens, être qualifié de tel, que sur la base d'un critère intellectuel (compétences cognitives) et moral (valeurs morales). Les données de la recherche scientifique sont donc tout à la fois le matériau de base et le résultat d'une production intellectuelle. La définition est très large dans la mesure où elle a pour effet d'englober des éléments qui ne sauraient de prime abord être qualifiés de tel. Ainsi, par exemple, en est-il des documents d'archives qui constituent les sources principales de recherche en histoire. Dès lors qu'ils prennent une forme numérique et qu'ils sont contenus dans un ensemble plus vaste, les documents deviennent des données. De même, bien que constituée de documents d'archives, une compilation d'archives est, au sens juridique et au sens informatique du terme, une base de données. Ainsi, vu sous l'angle du traitement de l'information — quel que soit le but de ce traitement par ailleurs (gestion, stockage, archivage, partage, diffusion ou réutilisation) —, tout document ou toute ressource devient donc une donnée.

Les données de la recherche scientifique figurent aujourd'hui au cœur de la politique européenne et nationale d'ouverture des données. Cette politique par laquelle un organisme public met à la disposition de tous des données numériques, dans un objectif de transparence ou afin de permettre leur réutilisation, notamment à des fins économiques[6], est mise en œuvre à travers plusieurs types de normes. Il s'agira donc de comprendre comment cette politique s'exprime au plan juridique (I), mais aussi comment elle prend forme au plan technique (II), tout en montrant à chaque fois l'interaction entre l'informatique et le droit.

1. L'ouverture des données de la recherche scientifique et le droit

Les données de la recherche scientifique font l'objet d'une double normativité, la première est purement juridique (A) et trouve son corollaire dans un processus de normalisation (ou de standardisation) purement technique (B).

1.1. Droit de l'ouverture des données de la recherche scientifique

Les chercheurs sont aujourd'hui dans l'obligation d'ouvrir leurs mémoires et les données scientifiques qu'elles contiennent afin d'en permettre la diffusion et l'accès. Plusieurs dispositifs expriment aujourd'hui cette injonction. Tout d'abord, l'obligation trouve sa source dans la volonté de la communauté scientifique de remplir au mieux l'une des missions cardinales qui lui incombe, à savoir la diffusion des connaissances scientifiques, mission par ailleurs mise à mal par la politique commerciale des puissants groupes d'édition scientifique[7]. Aussi, face aux moyens qu'offre désormais l'internet

pour diffuser les résultats de la recherche scientifique, les stratégies monopolistiques et non coopératives de l'édition traditionnelle dont la déontologie et la qualité scientifique ont, par ailleurs, été souvent critiquées, sont clairement apparues comme étant archaïques. L'on comprend, à cet égard, que le discours et la philosophie open access aient été largement relayés par les bibliothécaires[8] qui encouragent activement les chercheurs à publier leurs productions (articles, mémoires et thèses) via l'internet. C'est officiellement à travers les déclarations des « trois B » (Budapest, 2002[9] ; Bethesda, 2003[10] et Berlin, 2003[11]) que se sont dessinées les véritables injonctions de mettre la ressource scientifique en accès libre et ouvert (Robin, 2013). Cette mission est aujourd'hui inscrite dans le code français de la recherche qui prévoit que les chercheurs doivent organiser l'accès libre aux données scientifiques[12] et qu'ils doivent assurer le partage et la diffusion des connaissances en donnant priorité aux formats libres d'accès (art. L. 112-1 C. rech.)[13]. Cette politique fut renforcée par les dispositions de la loi n° 2016-1321 pour une République numérique du 7 octobre 2016[14] qui prévoit plus précisément le principe et les conditions de diffusion des données de la recherche scientifique publique (art. L. 533-4 C. rech.).

Cette injonction repose également sur les principes d'ouverture des données publiques énoncés par les dispositions légales du CRPA, auxquelles les données de la recherche publique scientifique n'échappent plus[15]. En effet, à se fonder sur ces dispositions (art. L. 300-2 CRPA), les données scientifiques peuvent recevoir la qualification de données publiques dans la mesure où elles figurent dans des documents produits ou reçus par l'administration (Robin, 2017). En tant que telles, elles sont donc soumises aux règles de communication, de diffusion et de réutilisation gouvernant l'ensemble des données publiques. Le principe d'ouverture des données qui est fondé sur un principe de transparence des données de l'administration, n'est cependant pas étranger à la volonté d'offrir aux entreprises la possibilité d'exploiter les gisements de données produites ou reçues par les administrations. Les établissements de recherche scientifique publique échappent d'autant moins à cette volonté qu'ils organisent de longue date non seulement la diffusion des connaissances (puisqu'il s'agit là, comme on l'a rappelé, de l'une de leurs missions), mais également la valorisation des résultats de leurs recherches en opérant des transferts de technologie en direction du secteur privé. Dans la continuité de cette politique, mais désormais sur le fondement des dispositions du CRPA, les institutions publiques de recherche doivent pouvoir assurer cette diffusion de leurs données afin de permettre aux entreprises de les exploiter. Cette volonté s'appuie elle-même sur l'idée que les données peuvent être réutilisées par d'autres personnes (entreprises ou communautés de chercheurs) que celles qui les ont produites afin de servir des analyses en big data, et d'en déduire éventuellement d'autres vérités. Le partage de données massives suppose toutefois, comme on le verra plus loin, qu'elles soient lisibles par les machines (« données lisibles en machine »).

Enfin, le principe de libre diffusion des données se trouve en résonnance avec la volonté de permettre l'administration de la preuve de la validité des résultats scientifiques. La transparence qui sous-tend ainsi la « fairisation » des données permettrait en effet de vérifier la validité des données (données non falsifiées, non fabriquées) elles-mêmes,

mais également de vérifier la méthode scientifique utilisée et ainsi, de répondre aux exigences résultant de l'application des principes d'intégrité scientifique résultant, au plan européen, du Code de conduite européen pour l'intégrité en recherche, et, au plan international de la Convention de Singapour (2010). Ces principes qui orientent les chercheurs dans leurs travaux ainsi que dans leur engagement, sont la fiabilité, l'honnêteté, le respect envers les collègues, les participants à la recherche, la société, les écosystèmes, l'héritage culturel et l'environnement et enfin, la responsabilité assumée pour les activités de recherche, de l'idée à la publication, leur gestion et leur organisation, pour la formation, la supervisions et le mentorat, et pour les implications plus générales de la recherche[16]. Le respect de ces principes passe par la nécessité de fournir, avec les données, les métadonnées qui permettent de contextualiser les données et de comprendre l'environnement et les conditions dans lesquelles elles ont été collectées ou produites[17], en précisant l'objectif de la recherche, les conditions de l'expérimentation, les hypothèses scientifiques, le protocole suivi, etc. ; processus cognitif que certains chercheurs en sciences de l'information et de la communication dénomment le « constructivisme numérique » (Verlaet, 2015).

1.2. Normalisation de l'ouverture des données de la recherche scientifique

Pour pouvoir alimenter l'analyse de données massives, les données scientifiques doivent pouvoir être encodées selon des formats, protocoles et schémas qui assurent une compatibilité et permettent le dialogue entre machines. Or, l'une des plus grandes difficultés à laquelle l'analyse en big data se trouve confrontée est l'hétérogénéité des données. Ainsi, afin de les rendre homogènes et ainsi d'en faciliter la lecture, puis l'analyse, la communauté scientifique s'est mobilisée autour de la conception d'un standard technique[18]. C'est ainsi que la norme FAIR fut conçue[19]. La norme FAIR exprime le fait que les données doivent être techniquement trouvables, accessibles, interopérables et réutilisables. Il est intéressant, à cet égard, d'observer qu'un standard technique porte le nom, en forme d'acronyme, d'une valeur, le fair, dont la traduction renvoie aux notions d'équité et de justice. Le génie provient, à notre sens, de là : la fair attitude (ou parfois utilisée en mot-valise la « fairitude ») consiste ainsi à assurer la diffusion et la réutilisation des données de la recherche scientifique, dont on estime qu'elles traduisent avec objectivité une réalité observable[20]. La norme FAIR est ainsi le fruit d'une politique d'innovation industrielle qu'une recommandation de la Commission européenne du 25 avril 2018 résume fort bien en posant le principe selon lequel les résultats de la recherche doivent être « aussi ouverts que possible et pas plus fermés que nécessaire » (« *as open as possible, as closed as necessary* »)[21]. Elle répond également à la volonté de faire dialoguer les machines entre elles : les données de la recherche présentées sous la forme numérique sont ainsi perçues, voire fantasmées, comme une lingua franca, langue véhiculaire des chercheurs.

Pour le juriste, la norme FAIR qui est donc en passe de devenir le standard (ou

condition) de fonctionnement du langage scientifique, est un phénomène intéressant à observer sous l'angle de la normativité de l'activité scientifique. L'adoption de la norme FAIR conditionne en effet le financement des projets de recherche (avec l'exigence corrélative de fournir un plan de gestion des données). Ceci n'est cependant pas étonnant dans le contexte de la recherche scientifique. D'une part, l'on sait depuis les travaux de Léon Duguit, et d'Emile Durkheim, que l'estampe formelle de l'Etat n'est pas un critère nécessaire pour assurer l'existence d'une norme. La normativité infra-légale (usages) est, en outre, très caractéristique de la communauté scientifique, dont la particularité est précisément d'adopter ses propres règles de structuration et de fonctionnement (Encinas de Munagorri, 1998, p. 247). La norme FAIR s'impose-t-elle toutefois véritablement comme une norme infra-légale, c'est-à-dire comme un usage résultant d'un consensus de la communauté scientifique ? La norme FAIR ferait-elle partie de ces valeurs partagées par la communauté scientifique telles que les avait étudiées Merton ? Rien n'est moins sûr. D'une part, malgré l'acronyme qui la désigne, la norme est clairement le résultat d'une politique publique plus générale, ainsi qu'on l'a montré plus haut, dans la mesure où elle est formellement exprimée par les pouvoirs publics tant au niveau national qu'européen. D'autre part, elle trouve son terrain d'élection dans les sciences de l'univers (astronomie, astrophysique) et dans les sciences du vivant (génomique végétale, animale, humaine), mais également dans les sciences du langage (linguistique computationnelle), mais pourrait s'avérer peu adaptée à d'autres disciplines, y compris celles travaillant à partir d'analyses quantitatives (histoire, économie, sociologie, etc.). Enfin, l'ouverture des données en mode FAIR suppose la mobilisation d'un certain nombre de ressources, essentiellement humaines. L'application de ce standard implique en effet une organisation logistique importante, comme de recruter des personnes chargées de la gestion des données, d'organiser le stockage et l'archivage des données, de réfléchir au périmètre de leur accès et de leur diffusion — question relevant de l'analyse juridique —, d'organiser leur lisibilité scientifique par le travail de conception et d'indexation des métadonnées[22], etc. L'« ingénierie des données » qui se met ainsi progressivement en place nécessite donc l'allocation de budgets conséquents.

2. L'ouverture des données de la recherche scientifique et l'informatique

Sur un plan technique, la normalisation des données scientifiques implique la mise en place de procédures informatiques complexes permettant leur lecture, leur compréhension et surtout leur articulation avec d'autres. La science informatique développe ainsi des systèmes de présentation standardisée des données reposant nécessairement sur une réduction de la polysémie (A). La standardisation technique des données pose également la question de leur qualification juridique : pour être en mesure de décider si les données peuvent faire l'objet d'une diffusion ou d'un partage, il y a lieu de savoir à quelle(s) catégorie(s) juridique(s) elles appartiennent (B).

2.1. Standardisation scientifique et réduction sémantique ?

La communauté scientifique s'est structurée autour de l'idée du FAIR, notamment avec la création de la Research Data Alliance (RDA). L'organisation, qui est un réseau international de chercheurs[23], propose, de façon très opérationnelle, de réfléchir aux protocoles de standardisation des données. La RDA organise par exemple également une réflexion sur les indicateurs permettant de rendre compte du niveau de maturité d'un modèle d'ouverture de données tels qu'adoptés par certaines communautés et du degré d'ouverture opérationnelle des données (fairness), ainsi que sur les modèles de licence — plutôt des modèles de licences libres — autorisant la réutilisation des données. Les préconisations de la RDA[24] méritent d'être présentées car elles traduisent en langage technique le sens qu'il convient de donner à la norme FAIR :

- pour être trouvables, les données doivent être identifiées grâce à un identifiant unique et permanent, elles doivent être décrites par de nombreuses métadonnées qui intègrent clairement l'identifiant des données, et elles doivent être indexées dans une base de données interrogeable ;
- pour être accessibles, les données doivent pouvoir être accessibles par un protocole standardisé de communication qui est libre, ouvert et qui peut ainsi être universellement implémenté (idée sous-jacente de neutralité technologique). Il est également précisé que le protocole qui permet d'accéder aux données et métadonnées peut donner lieu à une procédure d'authentification et d'autorisation si nécessaire, et enfin que les métadonnées doivent pouvoir être accessibles quand bien même les données ne seraient plus disponibles ;
- pour être interopérables, les données doivent être présentées selon un langage compatible. Ainsi, le format des données doit être suffisamment général pour en permettre le partage, les ontologies et le vocabulaire utilisé doivent suivre les principes FAIR et les données ou métadonnées doivent inclure des références qualifiées à d'autres données ou métadonnées ;
- enfin, pour être réutilisables, les données doivent être décrites de façon précise et exacte grâce à une pluralité de caractéristiques et d'attributs, qu'elles soient accompagnées d'une licence d'utilisation claire et accessible, que leur provenance soit présentée de façon détaillée et enfin, que les standards de réutilisation choisis correspondent de préférence à ceux utilisés par la communauté scientifique en question. Le détail des principes FAIR indique clairement comment les données font et feront de plus en plus l'objet d'une standardisation technique.

Si l'objectif assigné est aujourd'hui de mettre en œuvre le partage et la diffusion des données, il faut que l'acquisition et la production de données soient dès le départ conditionnées par celui-ci. C'est d'ailleurs ce qui est explicitement exigé : pour que les données soient accessibles et réutilisables, il faut qu'elles soient d'emblée, c'est-à-dire lors de leur collecte, conditionnées selon les préconisations techniques mentionnées plus haut, et en particulier qu'elles soient interopérables[25]. Le partage des connaissances suppose ainsi la mise au point d'ontologies qui soient conceptuelles, normatives, cohérentes, partageables et réutilisables. Mark Musen, professeur à

l'Université de Stanford et spécialiste des questions d'informatique biomédicale, insiste également sur la nécessité, dans l'activité d'ingénierie des connaissances, de standardiser les ontologies présidant à l'indexation des données, mais aussi des métadonnées. Avant de conclure que « le problème du big data est un problème sémantique »[26], il préconise ainsi de fournir les définitions retenues pour les termes figurant dans les ontologies. De même, Pierre Catala constatait, à propos du vocabulaire et de la polysémie des mots du droit par rapport au langage commun, ou de la polysémie des mots du droit tout court, la nécessité de « réduire le champ sémantique pour éliminer la polysémie »[27]. Le croisement disciplinaire par le prisme de l'informatique impose-t-elle la réduction du champ sémantique ? Autre question plus vaste que l'on peut se poser : la standardisation technique impliquera-t-elle à terme une standardisation scientifique ? L'administration et la gestion des données de la recherche scientifique dans l'optique de leur ouverture et donc de leur croisement interdisciplinaire, suppose en tout état de cause qu'un langage commun soit adopté, ou à tout le moins, qu'un socle commun de définitions, lui-même évolutif, soit établi et fourni avec les données.

2.2. Bonnes pratiques et qualification juridique

La mise en application de ces « bonnes pratiques » suppose, pour être menée à bien, que les données soient préalablement qualifiées au plan juridique. Il nous semble en effet que l'ouverture des données ne peut être mise en œuvre sans que certaines règles ne soient observées. Toutes les données ne peuvent donner lieu à une diffusion ouverte, en particulier lorsqu'elles sont couvertes par un secret. Et même à supposer qu'elles soient juridiquement diffusables, la diffusion des données implique que soient respectées un certain nombre de règles, notamment celles applicables en matière de données personnelles (Loi Informatique et Libertés). Ainsi en est-il, par exemple, des données de santé dont l'ouverture a été organisée par la loi n° 2016-41 du 26 janvier 2016 de modernisation de notre système de santé[28]. Afin que ces donnés puissent donner lieu à une réutilisation conforme à la loi, des protocoles d'anonymisation, qu'on appelle des méthodologies de référence, ont été mis en place en collaboration avec la CNIL. Ces protocoles ont pour effet d'intégrer la norme dans la technique elle-même, c'est-à-dire d'utiliser la technique pour faire respecter la norme. Dans un domaine différent, l'installation des Digital Right Management sur les œuvres de l'esprit (œuvres musicales, audiovisuelles, logicielles, jeux-vidéos) a été, par exemple, une façon innovante d'empêcher la circulation gratuite des œuvres et de faire respecter le droit d'auteur. C'est ce que l'on pourrait appeler l'ethic by design. La technique informatique est en mesure, grâce à l'ensemble des codes qu'elle utilise, de proposer un système variable d'accès ou, au contraire, de verrouillage de l'information.

La qualification juridique des données est un exercice qui ne se laisse cependant pas enfermer dans une binarité. Si l'opération juridique qui consiste à affirmer que tel ou tel jeu de données renferme des données personnelles, par exemple, doit aboutir à une qualification définitive, elle passe néanmoins par un processus d'appréciation des faits

parfois délicat. Cette opération de qualification, qui permet de rattacher un objet à telle ou telle catégorie d'objets et donc à telle ou telle catégorie de règles, ne peut en effet être le résultat que d'une analyse casuistique. Le traitement informatique des données de la recherche scientifique ne peut ainsi éluder la phase de qualification juridique qui permet ensuite de décider du sort de celles-ci. Ainsi, de façon tout à fait concrète, les données, objets du traitement, seront-elles qualifiables de données personnelles ou non personnelles, de données publiques ou de données privées, de données mixtes (si elles sont produites dans le cadre de partenariats public-privé et, si tel est le cas, il conviendra alors de savoir quel sort leur réserve le contrat de partenariat), de données confidentielles ou non, etc. ? La « fairisation » des données dans l'optique de leur ouverture ne peut donc faire l'économie d'un travail préalable de qualification juridique qui peut, dans certains cas, s'avérer conséquent. La mise en œuvre d'une politique de gestion des données de la recherche implique donc qu'un processus, pas toujours simple, de travail de qualification juridique soit mis en place ; processus dont la notion de « bonnes pratiques » peine à rendre compte. Ce n'est pas parce que l'on s'engage à respecter la loi que l'on répond à la question de la qualification juridique des faits. Si la notion de « bonnes pratiques » renvoie à l'idée d'une volonté de bien faire au regard de la loi ou d'une norme, elle laisse toutefois en suspens la question pratique de la mise en œuvre concrète. Cette question peut être simplement exprimée comme suit : « telle règle s'applique-t-elle à tel fait ? », supposant alors de comprendre et de caractériser le fait lui-même. L'opération de qualification juridique, et le raisonnement qui en résulte, permettent ainsi d'insérer le fait dans le droit dans le dessein de le mener à la règle. Or, l'axiome sur lequel la politique de science ouverte est fondée, et qui rappelons-le est contenu dans la formule « *as open as possible, as closed as necessary* », contient un renvoi implicite mais très clair à la règle de droit qui, seule, peut dire ce qui est possible et ce qui est nécessaire. L'axiome pose cependant, et c'est là son principal intérêt, une règle conflictuelle hiérarchisante : la liberté doit être le principe, l'interdiction doit demeurer l'exception. Si l'on prend l'exemple du secret d'affaires, il y a lieu, pour savoir si un jeu de données peut entrer dans le périmètre de l'obligation de diffusion, de savoir si le jeu de données est couvert par un secret. A suivre les dispositions légales sur le secret d'affaires (art. L. 151-1 s. C. com.), l'information n'est considérée comme confidentielle qu'à une triple condition : l'information ne doit pas être connue ou aisément accessible pour les personnes familières de ce type d'informations en raison de leur secteur d'activité, elle doit revêtir une valeur commerciale effective ou potentielle du fait de son caractère secret, et elle doit avoir fait l'objet de mesures de protection raisonnables, compte tenu des circonstances, pour en conserver le caractère secret. Ainsi, le fait que le principe du secret soit prévu dans le contrat ne suffit pas toujours… Autrement dit, la connaissance du caractère nécessaire de la non diffusion des données scientifiques ne peut résulter que de l'appréciation des circonstances de la mise en place du secret. D'autres exemples pourraient être donnés à l'aune de l'application des droits de propriété intellectuelle.

3. Conclusion

L'intégration de la norme dans le processus de traitement informatique de l'information relève à notre sens de la gageure si elle était réalisée sans tenir compte de décisions produites en amont par les personnes chargées de faire appliquer la loi (juges ou autorités de régulation), mais aussi par toutes celles qui l'appliquent au quotidien en se fondant sur la connaissance qu'elles ont de la règle de droit (juristes). Une ingénierie juridique paraît donc dans ce domaine, comme dans d'autres, inéluctable si l'on veut intégrer la norme à la technique et ainsi bénéficier des avantages offerts par la diffusion des données de la recherche scientifique publique. Il ne s'agit pas de dire que le raisonnement juridique humain ne soit pas substituable, ni qu'à terme l'analyse des données en masse ne permette pas de modéliser les décisions issues de ces processus de qualification, mais pour l'heure, ce travail ne peut se résumer ni à des bonnes intentions, ni à des raisonnements binaires. C'est ce que nous enseignait déjà Pierre Catala à l'époque de ses travaux de recherche. Par ailleurs, s'il est absolument nécessaire et utile, l'énoncé de la politique générale de science ouverte ne saurait se suffire à lui-même. La diffusion des connaissances, ici des données, nécessite que soient mis en œuvre des mesures importantes d'accompagnement des chercheurs, leur permettant non seulement de savoir à quel régime juridique sont soumises leurs données, mais également de choisir la voie la plus appropriée de diffusion et/ou d'exploitation pour les données qu'ils ont contribué à produire.

4. Bibliographie

Bulletin des bibliothèques de France (BBF), 1982, n° 6.
Catala, P. (1998). *Le droit à l'épreuve du numérique. Jus ex Machina*. Paris : PUF.
Robin, A. (2013). Créations immatérielles et technologies numériques : la recherche en mode open science. *Propriété intellectuelle*, n° 48, 260-270.
Robin, A. (2017). Les données scientifiques au prisme du dispositif open data. *Communication - commerce électronique*, étude 14, 7-14.
Westeel, A. & Claerr, T. (dir.) (2010). *Numériser et mettre en ligne*. Villeurbanne, Presses de l'Enssib, n° 19. http://books.openedition.org/pressesenssib/414.
Verlaet, L. (2015). La deuxième révolution des systèmes d'information : vers le constructivisme numérique. *Hermès La Revue*, vol. 2, n° 71, 249-254.
Encinas de Munagorri, R. (1998). La communauté scientifique est-elle un ordre juridique ? *Revue trimestrielle de droit civil*, 2, 247-283.

5. Notes

[1] En 1967, il avait fondé l'Institut de Recherche et d'Etude pour le Traitement de l'information juridique (IRETIJ).
[2] Ainsi par exemple de la société Lex Machina qui opère dans le domaine de la propriété intellectuelle et qui résulte d'un programme de recherche lancé par l'Université de Stanford en 2008 ou en France Doctrine.fr.
[3] A l'heure où s'écrivent ces lignes, les décrets d'application de la loi du 7 octobre 2016 (JORF 8 oct. 2016) n'ayant pas été pris, le gouvernement réfléchit encore au moyen technique le plus sûr d'anonymiser les décisions de manière à empêcher le ré-identification des personnes. V. aussi, Rapport Cadiet, « L'open data des décisions de justice », Mission d'étude et de préfiguration sur l'ouverture au public des décisions de justice, Rapport à Madame la garde des Sceaux, ministre de la Justice, nov. 2017 ; D. 2018, p. 232, obs. L. Cadiet et entretiens in JCP G 2018, 290 ; JCP G 2018, 282, obs. N. Fricero ; JCP G 2018, 1096, obs. B. Mathis.
[4] Comme des formules médicales approuvées par exemple (codex).
[5] C'est la définition retenue par l'Enrichissement du vocabulaire de l'informatique, Bulletin des bibliothèques de France (BBF), 1982, n° 6, p. 355-358. Disponible en ligne : http://bbf.enssib.fr/consulter/bbf-1982-06-0355-009.
[6] Commission générale de terminologie et de néologie, Vocabulaire de l'informatique et du droit, Avis, JORF 3 mai 2014.
[7] Il s'agit ici de lutter contre les tentatives de monopolisation des jeux de données accompagnant les publications scientifiques par les éditeurs. Cette politique éditoriale avait été dénoncée lors des déclarations des « 3 B » (Berlin, Bethesda, Budapest) au début des années 2000.
[8] V. le site de l'École nationale supérieure des sciences de l'information et des bibliothèques : www.enssib.fr.
[9] Initiative de Budapest pour l'accès ouvert, 14 févr. 2002.
[10] Déclaration de Bethesda pour l'édition en libre accès (recherche biomédicale), 11 avr. 2003.
[11] Déclaration de Berlin sur le libre accès à la connaissance en sciences exactes, Sciences de la vie, Sciences humaines et sociales, 22 oct. 2003.
[12] Depuis l'adoption de la loi du 22 juillet 2013, JORF 23 juill. 2013.
[13] V. égal. Recommandation de la Commission européenne, 17 juill. 2012 relative à l'accès aux informations scientifiques et à leur conservation, C(2012)4890 final.
[14] Préc.
[15] Les établissements de recherche sont restés longtemps en dehors de l'application du CRPA, du fait de l'exception dont ils bénéficiaient en particulier quant au principe de réutilisation des données. Cette exception a été retirée par la loi n° 2015-1779 du 28 décembre 2015, dite « loi Valter », relative à la gratuité et aux modalités de la réutilisation des informations du secteur public (JORF, 29 déc. 2015).
[16] Code de conduite européen pour l'intégrité en recherche, All European Academies (ALLEA), Berlin, 2018, www.allea.org.
[17] La métadonnée est une donnée à propos d'un ensemble signifiant de données. Il s'agit une information structurée qui décrit, explique, localise la ressource et en facilite la recherche, l'usage et la gestion (Westeel et Claerr, 2010).
[18] De nombreux standards techniques sont conçus dans d'autres domaines, ou par

secteurs. Ainsi par exemple de la description des données géographiques qui est soumise à la norme ISO 19115 (2003), déclinée en 2007 pour donner la norme ISO 19139, compatible XML.

[19] Findable, accessible, interoperable, reusable.

[20] Puisque par hypothèse, elles été collectées ou produites selon une méthode scientifique rigoureuse.

[21] Recommandation de la Commission européenne du 25 avril 2018 (C(2018) 2375 final), n° 3 : « research data that results from publicy funded research becomes and stays findable, accessible, interoperable and re-usable (« FAIR principles ») within a secure and trusted envrironment, through digital infrastructures (including those federated winthin the European Open science Cloud (EOSC), where relevant), unless this is not possible or is incompatible with the further exploitation of the research resultts (« as open as possible, as closed as necessary »). This could be for reasons, in particular, of privacy, trade secrets, national security, legitimate commercial interests and to intellectual property rights of third parties. Any data, know how and/or information whatever its form or nature which is held by private parties in a joint public/private partnership prior to the research action should not be affected by theses policies or national action plans ».

[22] L'idéal est de pouvoir utiliser un standard commun pour décrire les données. Le schéma général Dublin Core a ainsi été conçu pour remplir cette fonction (v. Westeel et Claerr, op. cit.).

[23] L'alliance compte aujourd'hui 8 418 membres à travers le monde.

[24] https://www.rd-alliance.org/system/files/documents/20190603_FAIR_WG_slides_v0.06_0.pdf. Il y a d'autres réseaux qui se mobilisent et proposent les protocoles et principes de « fairisation » des données, comme GoFair ou Fairsharingdata, etc.

[25] L'interopérabilité peut être définie comme la capacité que possède un produit ou un système, dont les interfaces sont intégralement connues, à fonctionner avec d'autres produits ou systèmes existants ou futurs et ce, sans restriction d'accès ou de mise en œuvre, v. site de l'AFUL.

[26] Conférence donnée à l'Université de Montpellier, 29 mai 2019 (LIRMM).

[27] Op. cit., p. 26.

[28] JORF, 27 janv. 2016.