La loi sur le numérique et le libre accès

septembre 26, 2015

Data-mining, domaine public, Libre accès

Pierre-Carl Langlais

Je profite de l'annonce officielle de la proposition de loi sur le numérique pour relancer ce carnet de recherche (temporairement interrompu pour cause de… fin de thèse). Au terme d'une gestation tortueuse, un texte stabilisé vient d'atterrir sur une forme de "wikipédia" législatif. La plate-forme contributive République numérique permet à tout un chacun de proposer des articles ou d'amender les articles proposés. En tant que wikipédien averti, l'opération m'est plutôt sympathique. En tant que wikipédien averti, j'espère juste qu'elle ne finira pas comme de nombreuses prises de décision de Wikipédia (i. e. l'équivalent d'un roman de deux cent pages de discussion pour aboutir à pas grand chose). La démocratie collaborative est sans doute le pire des systèmes à l'exclusion de tous les autres (dont la démocratie représentative)…

La loi touche à de nombreux aspects des usages numériques (notamment l'open data, qui se taille la part du lion). Je vais ici me focaliser sur mon thème de prédilection : le libre accès. En apparence, tout est réglé dans l'art. 9 sur le libre accès aux publications scientifiques de la recherche publique. En réalité, ce n'est là que le morceau d'un puzzle plus vaste.

L'art. 9 sur le libre accès dans la plate-forme Republique Numérique

L'affirmation d'un grand principe : le domaine public informationnel

Par-delà les nombreuses pressions qui se sont agitées en coulisse sur le projet de loi, et qui ont tout particulièrement ciblé l'art. 9, un acquis capital a été préservé : la définition positive d'un domaine public informationnel. Elle est formalisée en ces termes dans l'art. 8 :

Les choses qui composent le domaine commun informationnel sont des choses communes au sens de l'article 714 du Code civil. Elles ne peuvent, en tant que tels, faire l'objet d'une exclusivité, ni d'une restriction de l'usage commun à tous, autre que l'exercice du droit moral.

Le périmètre de ce domaine commun informationnel intègre les "objets" suivants. Il s'apparente à un miroir inversé du code de la propriété intellectuelle en explicitant les conditions à partir desquelles les productions intellectuelles sont exemptées de toute protection.

1. **Les informations, faits ou idées** : il s'agit ici des "significations" distinctes d'une expression originale (par exemple, on pourrait reformuler toutes les idées exprimées dans ce billet, sans attenter à aucun droit d'auteur) ou exprimées sous une forme a priori non originale (par exemple, un énoncé procédural du style "L'action de telle entreprise est cotée à telle valeur"). En apparence, tout ceci ne relève évidemment pas du droit d'auteur. En pratique, c'est beaucoup moins simple depuis une vingtaine d'années. L'Union européenne a reconnu un "droit des bases de données", intégré depuis dans le code du droit d'auteur français. Tout "producteur" d'une base de donnée peut

restreindre les reprises “substantielles”, même lorsqu’elle ne peut être considérée comme une création originale. Ce droit, qui s’intègre difficilement dans la terminologie générale du droit d’auteur à la française (il est question d’un “producteur” et non plus d’un “auteur” et le critère fondamental d’originalité passe à la trappe) a permis de justifier de nombreuses dérives. À l’image d’Elsevier, les grands éditeurs scientifiques tentent de revendiquer un “droit d’auteur” sur l’information brute en arguant d’une interprétation extensive du droit des bases de données (il est très difficile de déterminer où commence une reprise “substantielle”). Ces prétentions sont consolidées par l’absence de statut juridique clair et explicite pour les informations. Les projets scientifiques ou communautaires compilant des données sont confrontées à une grande insécurité juridique.

2. **Les créations de l’esprit dont toutes les formes de protections ont expirées**. À nouveau, il s’agit ici d’asséner une évidence, voire une tautologie (tout ce qui n’est pas protégé… n’est pas protégé). À nouveau, ce rappel est loin d’être inutile. Les protections indues apposées au domaine public des œuvres (aussi qualifiées de copyfraud) sont monnaie courantes. Des institutions publiques prestigieuses s’y livrent sans ménagement : comme le rappelais dans un article pour Rue89, le Musée d’Orsay multiplie les mentions “© Orsay” sous des reproductions d’œuvres du XIXe siècle ; Gallica requiert l’acceptation de conditions de réutilisation non commerciales avant tout téléchargement. Les utilisateurs expérimentés ne se laissent pas impressionner : Wikisource ou Commons reprennent couramment des publications diffusées par Gallica ou Orsay. Il n’en va pas de même pour les lecteurs ou amateurs occasionnels, qui en viennent à restreindre leurs usages faute de déceler les chausses-trappes de ces incantations juridiques. L’invocation de l’art. 8 permettrait d’expliciter l’illégalité des copyfraud. Même si aucune mesure de rétorsion précises ne sont prévues, une institution pourrait sans doute, sur cette base, être contrainte de revoir sa politique.
3. **Les informations issues des documents administratifs** au titre de la loi sur la réutilisation des informations publiques de 1970 ne peuvent faire l’objet d’une protection. En effet, le premier grand dispositif juridique organisant un “open data” à la française a permis de justifier paradoxalement de nombreuses dérives : des organismes publics apposent fréquemment des restrictions sur la reprise sur d’informations ou, s’agissant de bibliothèques, d’œuvres du domaines public, en invoquant le fait qu’ils sont libres de déterminer les conditions de réutilisation.

L’article s’achève sur l’évocation d’un mécanisme de “contrôle” de l’application de ces droits : une association reconnue “ayant pour objet la diffusion des savoirs ou la défense des choses communes” peut tenter de “de faire cesser toute atteinte au domaine commun informationnel”. Apparemment, dans ce cadre, une association de chercheurs ou de bibliothécaires pourrait remettre en cause la licence de text and data mining d’Elsevier.

Comme on le voit, la codification de ce grand principe n’est pas une simple proclamation abstraite. Elle va faciliter et sécuriser de nombreux travaux de recherche. Les projets scientifiques sont généralement très soucieux de la légalité de leur activité. Si l’usage des technologies numériques ouvre de nombreuses perspectives (par exemple, en terme d’analyse statistique de grands corpus), leur mise en œuvre effective reste limitée de facto par l’absence d’un cadre juridique clair (peut-on ou non importer automatiquement le contenu de telle bibliothèque numérique ou la base de données de telle institution publique ?)

Un libre accès en demi-teinte

Revenons maintenant l’article ciblant précisément le libre accès : l’art. 9. J’ai suivi à distance

l'histoire de son élaboration. Et elle est plutôt mitigée (même si il y a aussi de bonnes surprises imprévues en bout de course).

Initialement, la loi dérive d'une recommandation non contraignante de l'Union Européenne prise en 2012. D'après son premier article,

il devrait y avoir un libre accès à toute publication résultant d'une recherche sur fonds publics aussi rapidement que possible, de préférence immédiatement et, de toute manière, par plus tard que 6 mois après publication et 12 mois pour les sciences sociales et les humanités.

En janvier dernier, j'avais conçu une proposition de loi adaptée au droit français. Elle reprenait dans ses grandes lignes le projet de loi européen, en y joignant des provisions supplémentaires sur les nouvelles formes de publications. L'article scientifique est en effet en pleine métamorphose. Il se présente sous la forme de plusieurs versions (preprint, postprint) ; l'écrit rédigé ne résume plus à lui seul tout l'article, il convient d'y adjoindre d'autres productions permettant d'évaluer sa reproductibilité (les données, les algorithmes statistiques). D'où une interprétation élargie :

Cette exemption porte sur la totalité des formes prises par les travaux de recherche. Elle couvre notamment les versions finales acceptées des articles de recherche parus dans des revues périodiques. Les formes non rédigées donnant lieu à des droits de propriété intellectuelle, telles que les compilations de données, sont également concernées.

Cette disposition (qui visait à anticiper des transformations en cours) n'a finalement pas eu d'écho — même si j'ai eu confirmation par ailleurs que mon texte a été pris en considération lors de la phase d'élaboration de l'art. 9.

L'une des premières moutures du texte reprend les conditions temporelles de la recommandation européennes (6 mois pour les STM, 12 mois pour les SHS), à ceci près qu'il les rend impératives (dans la recommandation, il était seulement question d'un "au plus tard"). Il y adjoint une provision, introduite par une adaptation allemande de la recommandation européenne : toute fin commerciale est exclue. Concrètement, les chercheurs ne peuvent republier leurs travaux que sous une licence non commerciale. Par contre, il est clairement explicité que tous les droits exclusifs des éditeurs, quels que soient les contrats signés, sont abolis :

L'auteur d'une contribution scientifique, issue d'une activité de recherche financée au moins pour moitié par des fonds publics et publiée dans le cadre d'une collection paraissant au moins deux fois par an, dispose du droit, même s'il a cédé un droit d'exploitation exclusif à l'éditeur, de rendre publiquement accessible la version acceptée de son manuscrit aux termes d'un délai de six mois pour les Sciences et de douze mois pour les Sciences humaines et sociales à compter de la première publication, toute fin commerciale étant exclue. L'alinéa précédent est d'ordre public((Première version de l'art. 9 de la Loi sur le numérique))

En juillet, une version de travail du projet de loi "fuite" sur Internet. Il est beaucoup plus étendu que le projet actuel (il comprend 80 articles, vs. seulement 30). La disposition sur le libre accès se trouve dans l'art. 39. Elle n'a pas fondamentalement bougé par rapport à la version de mars :

Il est créé dans le code de la propriété intellectuelle un article L. 132-8-1 ainsi rédigé : Art. L132-8-1. – L'auteur d'une contribution scientifique, issue d'une activité de recherche financée au moins pour moitié par des fonds publics et publiée dans le cadre d'une collection paraissant au moins une fois par an, dispose du droit, même s'il a cédé un droit d'exploitation exclusif à

l'éditeur, de rendre publiquement accessible la version acceptée de son manuscrit, au terme d'un délai de six mois pour les sciences et de douze mois pour les Sciences humaines et sociales à compter de la première publication, toute fin commerciale étant exclue.
Il s'agissait explicitement d'amender le code de la propriété intellectuelle. Je ne suis pas certain que c'était là une excellente méthode. Dans ma proposition de loi de janvier, je préconisai, tactiquement, de cibler plutôt le code de l'éducation nationale, afin de présenter le texte comme une application "particulière" du droit d'auteur plutôt que comme une exception "générale".
Le choix du code de l'éducation nationale ne relève pas non plus du hasard. Cela permet de centrer le sujet sur une situation bien précise. Une réforme du code de la propriété intellectuelle risque a contrario de susciter des réactions ou des réticences annexes (bien qu'elles ne soient pas du tout concernées, les industries culturelles ou les société de gestion de droit y verraient une tentative de détricotage du droit d'auteur).
Mes appréhensions ne relevaient pas d'un pur fantasme. Sous l'effet de diverses pressions des éditeurs scientifiques, et plus largement par les institutions et organisations défendant les intérêts des ayant-droit, les durées d'exclusivité sont prolongées : elles passent à 12 mois pour les STM et 24 mois pour les SHS. Par ailleurs, la republication ne peut pas porter sur les version finalement publiée par l'éditeur, mais sur celle finalement transmise par l'auteur au terme du processus de peer-review :
Lorsque un écrit scientifique, issu d'une activité de recherche financée au moins pour moitié par des fonds publics, est publié dans un périodique, un ouvrage paraissant au moins une fois par an, des actes de congrès ou de colloques ou des recueils de mélanges, son auteur, même en cas de cession exclusive à un éditeur, dispose du droit de mettre à disposition gratuitement sous une forme numérique, sous réserve des droits des éventuels coauteurs, la dernière version acceptée de son manuscrit par son éditeur et à l'exclusion du travail de mise en forme qui incombe à ce dernier, au terme d'un délai de douze mois pour les sciences, la technique et la médecine et de vingt-quatre mois pour les sciences humaines et sociales, à compter de date de la première publication. Cette mise à disposition ne peut donner lieu à aucune exploitation commerciale[1] .
Par contre, le périmètre des publications est élargi. Il n'est plus seulement question des revues, mais aussi des actes de colloques et de congrès et des compilations d'articles réalisées par l'auteur (le "recueil de mélanges"). À la limite, la jurisprudence pourrait élargir cette notion de "recueil de mélanges" aux formes de publications annexes à l'article : les données ou algorithmes associées feraient parties d'une forme de compilation.

Le grand oublié : le Text and Data Mining

C'était l'autre grand attendu de la loi sur le numérique pour les communautés scientifiques : la reconnaissance d'une exception pour la fouille automatisée de données et de textes (ou Text and Data Mining). La version "fuitée" de juillet mentionne la modification suivante apportée à l'art. 122-5 du code de la propriété intellectuelle (art. 40).

« f) Les copies ou reproductions numériques réalisées à partir d'une source licite, en vue de l'exploration de textes et de données 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 destruction des fichiers au terme des activités de recherche pour lesquelles elles ont été produites; »

En janvier 2014, j'avais publié sur ce blog une synthèse détaillée sur les problématiques légales du Text and Data Mining. L'analyse automatisée de grands corpus requiert en effet une copie substantielle d'une base de données ou d'un ensemble de texte. Pour des raisons de performances et de limitations techniques, il n'est souvent pas envisageable d'effectuer

directement des requêtes sur des publications en ligne. Au titre du droit des bases de données, ces recopies massives sont clairement illégales. Chaque projet de recherche doit obtenir l'acquiescement des détenteurs des droit, ce qui peut prendre des années. Paradoxalement, il serait possible de réaliser strictement le même travail de manière "manuelle" (en consultant les publications visées, sans faire appel à des techniques d'extraction). Il y aurait ainsi une disjonction entre le droit d'extraire et le droit de lire.

Les exceptions au text and data mining se sont multipliées au cours de ces dernières années. C'est notamment le cas au Royaume Uni qui a mis en place un fair dealing depuis un an : toute recherche à des visées non commerciale peut réaliser légalement une copie substantielle d'une base de donnée mis en ligne (sous réserve, évidemment, de ne pas la divulguer). Aux États-Unis, la jurisprudence évolue également en ce sens : le procès Google Books reconnaît ainsi que la fouille automatisée pourrait relever du *fair use*.

Le projet français de la loi sur le numérique se situait déjà en retrait par rapport à ces évolutions. Il met en place une procédure de "destruction" des fichiers obtenus : un projet de recherche ne pourrait pas conserver durablement une copie ; dès que certains objectifs sont accomplis, la base de donnée disparaît, sans espoir de revenir en arrière.

Finalement, même cette adoption timide n'a pas été retenue. Entre temps, le text and data mining est devenu l'un des "épouvantails" du lobby de l'édition. D'après la brochure La Gratuité c'est le vol (ironiquement diffusée… gratuitement) de l'avocat du Syndicat National de L'Édition, Richard Malka, l'adoption d'une exception matérialiserait une forme d'apocalypse communiste :

Si de tels investissements pouvaient être légalement pillés, aucun éditeur n'engagerait désormais le moindre financement pour créer de tels outils. Il n'existe en réalité, aucune activité économique au monde dont les productions peuvent être librement expropriées pour cause d'utilité privée et sans aucun dédommagement.

Par-delà les fantasmes d'une industrie culturelle aux aboies, l'absence de cadre légal pour la fouille de texte de données risque d'acter un "déclin" bien réel : celui de la recherche française, désavantagée sur ce point par rapport à ce que peuvent réaliser les pays anglo-saxon. L'élaboration de vastes de base de donnée en libre accès représente un tournant majeur de ces dernières années. Il serait notamment impossible de réaliser en France l'équivalent du projet Content Mine, qui ambitionne de créer une collection de 100 millions de "faits".

En l'état, ce nouveau monde de la donnée scientifique a peu de chance de parler français…

1. Art. 9 de la Loi sur le Numérique