Data Mining : quand Elsevier écrit sa propre loi…

Posted on février 8, 2014 by Pierre-Carl Langlais

Le leader mondial incontesté de l'édition scientifique, Elsevier s'engage en faveur d'une simplification du data mining. De nouvelles conditions d'accès, dévoilées le mois dernier, vont grandement simplifier l'accès à l'un des principaux corpus de publications scientifiques. D'autres éditeurs devraient prochainement adopter un modèle similaire. C'est notamment le cas du principal concurrent d'Elsevier, Springer.

En apparence ce pourrait être une bonne nouvelle. La recension de Nature met ainsi en évidence l'engouement de certains chercheurs. Max Hauessler, l'instigateur d'un immense projet d'extraction des articles scientifiques sur le génome humain, a salué l'initiative : « Finalement, tout ceci montre qu'il n'y a plus aucune raison d'être effrayé par le text-mining ». Les membres du Human Brain Project (le projet européen d'étude du cerveau humain, doté d'un budget d'un milliard d'euros) semblent également emballés par l'affaire : « Nous sommes enchanté par tout ceci. Cela résout d'importantes questions techniques ».

En France, le consortium Couperin vient de conclure un accord avec Elsevier, qui comprend une clause spécifique sur le data mining :

> *Tous les contenus accessibles et souscrits sur ScienceDirect dans le cadre de cet accord seront utilisables à des fins de data et text mining via une interrogation des données par une API connectée à la plateforme ScienceDirect. Les modalités appliquées seront celle du cadre juridique défini par Elsevier pour ce type de service.*

Selon les termes de l'accord, plus de 600 établissements publics français pourront accéder au corpus Elsevier. Ce qui constitue autant de point d'accès pour faire du data-mining à une vaste échelle.

Alors, une bonne nouvelle ? En apparence seulement… Le choc de simplification d'Elsevier n'a qu'un mérite : il met fin à une situation relativement chaotique où, faute de cadre légal défini, les chercheurs passaient plus de temps à négocier avec les éditeurs qu'à extraire et analyser le corpus. Pour le reste, tout est mauvais…

API obligatoire…

Les termes de la licence sont profondément inadaptés. Elsevier oblige tout-le-monde à passer par son API. Les robots et les algorithmes de recueil automatiques ne peuvent accéder au site lui-même afin de « garantir la performance et la disponibilité du site pour tous les usagers ». Ce prétexte ne saurait légitimer une interdiction complète. Seul l'usage excessif de requêtes automatisées peut entraîner une baisse de performance du site — raison pour laquelle Wikipédia plafonne les requêtes sur sa base de données (je n'ai plus les chiffres en tête, mais cela devait tourner autour de quelques milliers de requêtes par heure par utilisateur).

L'API n'est obligatoire que parce que Elsevier tient à contrôler directement les termes et les modalités de la recherche. Pour reprendre la terminologie introduite par la thèse (en cours…) de Samuel Goyet, l'API s'apparente étroitement à un texte de loi : elle détermine par avance les commandes qu'il est possible de réaliser.

Les requêtes à l'API sont elles-mêmes plafonnées à 10 000 articles par semaine. Pour un projet ambitieux, c'est peu. En réutilisant ce système pour analyser 3 millions d'articles, Text2Genome y aurait passé près de 300 semaines, soit 6 ans. Et en plus, il lui aurait fallu découper son corpus en rondelle, sans avoir la possibilité de faire une requête unique à l'ensemble du corpus.

Le passage par l'API limite considérablement les possibilités d'analyse. Il est nécessaire de passer par le réseau

Internet. Là où une simple base de donnée MySQL permet de survoler plusieurs millions d'entrées en quelques dizaines de seconde, le recours à l'API génère des délais d'attente bien plus longs, dépendant qui plus est des aléas de la connexion (le wifi parfait, ça n'existe pas…).

Vers une privatisation de l'information

L'emprise d'Elsevier ne s'arrête pas là. L'éditeur ne cherche pas seulement à contrôler les projets de data mining en amont, mais aussi en aval. La licence Elsevier comprend trois conditions. Tout élément (*output*) issu de l'extraction :

1. peut comprendre des extraits de 200 caractères au maximum du texte original.
2. doit être publié sous une licence non commercial (CC-BY-NC)
3. doit inclure un lien DOI vers le contenu original.

La dernière condition est clairement la moins problématique. La traçabilité des données nécessite certes un équipement technique adéquat (pour reprendre le jargon des data scientists, il faut passer du triplet sujet-prédicat-objet au quadruplet sujet-prédicat-objet-source). Mais elle constitue une pratique saine, d'ailleurs appelée à se généraliser. Wikidata inclut ainsi des notes de base de donnée : chaque donnée étant appelée à être étayée par une source fiable.

La première condition contredit ouvertement l'exception de courte citation. En France, le législateur a volontairement adopté une définition floue afin de laisser le juge statuer sur chaque cas selon ses spécificités. Les exceptions sont beaucoup trop nombreuses pour envisager de faire une règle fixe. L'application du droit de citation sera beaucoup plus rigoureuse pour une œuvre courte (comme par exemple un poème) que pour un roman de mille pages. Or Elsevier tente de contourner le droit existant pour imposer sa propre règle : une limite de 200 caractères, purement arbitraire. Cela hypothèque quantité de projet comme le souligne bien Peter Murray-Rust : les noms de composés chimiques excèdent fréquemment la barre des 200 caractères.

Là n'est pas encore le plus grave. La deuxième condition introduit une dérive beaucoup plus problématique : elle requiert une forme de privatisation de l'information. Nul ne peut en effet réclamer de droit de propriété sur une information, une opinion ou une idée. Tous ces éléments relèvent du domaine public de l'information. Pour reprendre l'excellente définition de l'UNESCO, il existe un « indivis mondial de l'information » composé de toutes les informations publiquement accessibles. Le fonctionnement de la recherche repose explicitement sur cet indivis : le format classique de l'article de recherche, quadrillé de notes de base de page, suppose de pouvoir légitimement reprendre une information préalablement publiée.

Les données « solitaires » font globalement partie de ce domaine public de l'information. À moins de reprendre une expression textuelle originale, elles ne sont pas concernées par le droit d'auteur. Le droit des bases de données ne porte qu'au niveau structurel : chaque donnée individuelle est disponible pour n'importe quelle reprise, au même titre que n'importe quelle information publiquement accessible.

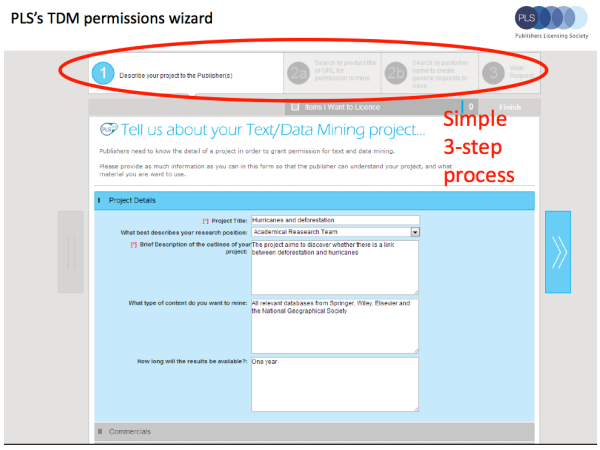
Or Elsevier impose discrètement l'idée que les informations pourraient être protégées. L'utilisation du terme ambigu *output* permet d'opérer ce glissement lourd de conséquence : tout ce qui sort des publications d'Elsevier reste la propriété d'Elsevier et l'éditeur est libre d'imposer ses propres conditions pour leur réutilisation. Ni les informations, ni les données « solitaires » n'échappent à cette appropriation globale : toute reprise du contenu d'Elsevier à des fins de data mining devra être publié sous une licence non-commerciale.

Cette dérive est lourde de conséquence. Une fois que Elsevier est parvenu à imposer l'idée que tout *output* donne potentiellement lieu à une protection, rien ne l'empêche d'aller beaucoup plus loin. Cette règle peut aisément être généralisée à l'ensemble des pratiques de reprises de l'information. Nous faisons, tous, au quotidien du

data-mining en récupérant des informations dans des textes préalablement publiés. L'API et les algorithmes d'extraction ne sont que des outils supplémentaires visant à simplifier cette activité fondamentale. Un projet de l'ampleur de Text2Genome aurait très bien pu être réalisé avec un papier et un crayon : son élaboration aurait simplement pris un ou deux siècles plutôt que quelques années.

Les données sur le data mining

Il ne suffit pas de payer un abonnement à Elsevier pour avoir accès à l'API. Il est nécessaire d'enregistrer son projet sur developers.elsevier.com. Je n'ai pas pu accéder au formulaire (cela requiert un enregistrement préalable). Par contre, Springer a dévoilé un système assez similaire dans une présentation publiée sur le site de la Commission européenne.



Le système de licences de Springer

L'éventail des informations nécessaires est large. Les chercheurs doivent décrire le projet et spécifier sa durée, délimiter le contenu à extraire, indiquer si un tiers aurait accès aux informations, etc. Le data-mining selon Springer et Elsevier s'apparente ainsi à une transaction commerciale non monétaire : les projets doivent céder leurs métadonnées pour obtenir les données du corpus éditorial. L'éditeur est ensuite libre d'exploiter ces métadonnées à des fins de marketing ou de les revendre à d'autres organisations.

On aboutit ainsi à un véritable paradoxe. Alors que l'essentiel de la recherche française est financée par l'argent public, seuls quelques grands éditeurs privés ont une vue globale sur l'ensemble des projets réalisés. Et les licences sur le data-mining vont aggraver une situation déjà déséquilibrée. Les éditeurs n'auront pas seulement des informations sur les projets réalisés, mais sur ceux en train de se faire.

Elsevier vs. La loi

Au début de l'année 2013, les grands éditeurs scientifiques marquent un grand coup. Ils parviennent à s'approprier entièrement le seul processus européen visant à doter le data mining scientifique d'un cadre légal : le text and data mining working group. En dépit de nombreuses protestations, les licences éditoriales au cas par cas s'imposent comme la seule solution envisageable : la Commission européenne a explicitement refusé de considérer des solutions alternatives (comme une exception générale).

Dès lors, chaque éditeur pourrait écrire sa propre loi et s'arranger librement avec la législation existante. Elsevier s'affranchit ainsi à la foi du droit de courte citation et du domaine public de l'information.

Depuis quelques mois, cette issue est devenue beaucoup moins certaine. Boycotté par de nombreux interlocuteurs essentiels (comme l'OKFN ou la Ligue des bibliothèques européennes), le Text and Data Mining Working Group a succombé à une certaine léthargie. Plusieurs pays européens envisagent ouvertement une exception (c'est notamment le cas du Royaume-Uni et de l'Irlande). L'exception a d'ailleurs été actée de facto dans la loi américaine à l'issue de la dernière décision du procès Google Books : elle est couverte par le fair use.

La recension de Nature fait ainsi état d'un revirement progressif de la Commission Européenne (peut-être lié à l'approche des prochaines élections européennes). Un nouveau processus de réflexion aurait été confié au juriste britannique Ian Hargreaves. Hargreaves a joué un rôle essentiel dans la conception du programme britannique de modernisation du copyright. Il a également fortement encouragé le développement d'une exception pour le data-mining. En juin dernier, il critiquait ouvertement la politique actuelle de l'Union européenne en matière de droit des bases de données.

Hargreaves devrait remettre son rapport à l'Union Européenne à la fin du mois. La Commission semble désormais assez sensible à un argument économique : les éditeurs ne sont pas, et de loin, les principaux bénéficiaires d'un data-mining libéralisé ; un tel cadre légal profiterait à l'ensemble de la société.

En France, le collectif Savoirscom1 a été récemment auditionné sur le sujet par le Conseil Supérieur de la Propriété Littéraire et Artistique (CSPLA). Les propositions du collectif ont été exposées dans une petite synthèse, que j'ai rédigé avec Lionel Maurel, Quel statut légal pour le content-mining ?.

Le rapport formule pour l'essentiel deux propositions :

- Reconnaître l'existence d'un domaine public de l'information, qui permet d'extraire toutes les informations et les données solitaires ne répondant pas à un critère d'originalité.
- Mettre en place une exception limitative, autorisant notamment l'importation substantielle d'une base de données ou de textes soumis au droit d'auteur sous réserve que la copie ne reste accessible qu'au seul projet de recherche.

Bien que nous ayons reçu un assez bon accueil du CSPLA, ces deux propositions sont très loin d'être acquises. Au cours de la semaine passé j'ai été auditionné au sénat par la Mission commune d'information sur l'accès aux documents administratifs et aux données publiques. Je représentais le collectif Savoirscom1 avec Silvère Mercier et Thomas Fourmeux. J'ai commencé à évoquer l'intégration des données de la recherche au régime général des données publiques et la nécessité de formaliser le domaine public de l'information : le président de la mission n'a pu réprimer un certain étonnement tant ces suggestions paraissaient exotiques…

Imprimer ce billet

This entry was posted in Billets, Data-mining, Industries académiques, Licences, Ouverture des données scientifiques. Bookmark the permalink.