Contribution du Conseil Scientifique au livre blanc du CNRS sur le

« projet de loi sur le numérique »

Introduction

La place du "numérique" dans l'activité scientifique est devenue capitale aujourd'hui, même si bien entendu, il faut avoir conscience qu'il ne fait en partie que figer ou "photographier" le réel et la recherche dans un état donné. L'activité scientifique comporte bien d'autres facettes que la gestion de données. Cependant, la numérisation des données utilisées par les scientifiques et de leurs publications permet un traitement automatisé, un transfert rapide, une harmonisation des méthodes d'accès et des descriptions ; tout cela est susceptible de mettre à la portée du chercheur une immense matière, riche et diversifiée, en des temps singulièrement raccourcis. Dégageant ainsi les scientifiques d'une partie des tâches répétitives et grandes consommatrices de temps, le numérique peut libérer leurs capacités réflexives et créatrices. On peut sans doute à cet égard, comme cela a été maintes fois écrit, comparer les possibilités ouvertes à la recherche par le numérique à celles qu'ont connues les savants des 16 e et 17 e siècles avec l'invention de l'imprimerie et l'accélération des échanges de savoir qui en a résulté.

A travers les disciplines

Les SHS:

En sciences humaines et sociales, en ce qui concerne les publications scientifiques récentes, alors que de plus en plus de revues francophones sont en libre accès immédiat ou après quelques années (notamment grâce à HumaNum, BSN, OpenEdition), les revues anglophones, sont souvent cantonnées à des plateformes assez chères ; le dépôt parallèle d'articles dans des archives ouvertes est peu développé. La numérisation des sources imprimées utilisées par nombre de disciplines, qu'il s'agisse de publications scientifiques plus anciennes ou de romans, journaux, traités juridiques, etc. est, elle, en bonne voie ; elle se fait souvent en libre accès, même si certaines grandes entreprises (comme Gale) produisent également des bases de données à des prix prohibitifs, pratiquement inaccessibles en Europe. Des plateformes d'échange de données chiffrées, qu'il s'agisse des échelles les plus détaillées de la statistique publique ou de données produites par la recherche, ont également été mises en place (Réseau Quetelet, DIMESHS, etc.) : elles assurent à la fois une meilleure circulation des données, le respect des contraintes nécessaires, d'anonymisation par exemple, et la documentation ("métadonnées") sans laquelle les chiffres seraient inutilisables.

Cela dit, les données des sciences humaines et sociales, extrêmement variées selon les disciplines (de l'histoire de l'art à l'économie en passant par la linguistique), sont loin de se limiter à des imprimés libres de droits ou des chiffres. Les plateformes de partage restent à inventer pour, par exemple, les photographies d'archives ou d'oeuvres d'art prises dans un but scientifique (ce qui se heurte au droit de reproduction) ; elles sont encore peu développées pour les données issues d'enquêtes qualitatives de terrain (qui posent de délicats problèmes

d’anonymisation, de mise en forme et de documentation). Le problème est ici que nombre de données utilisées par les scientifiques en SHS ne sont pas produites par eux (qu’il s’agisse d’une chanson, d’un rapport d’activité d’entreprise ou de l’architecture d’un monument) : d’autres personnes physiques ou morales disposent de droits sur elles. Partage de données et techniques de “fouille de données et de textes” sont ainsi inégalement répandus selon les types de données, pour des raisons principalement d’obstacles juridiques, et de manque de moyens en personnel pour la production et le maintien de métadonnées de qualité. Du fait de ces contraintes, pour beaucoup de types de données en SHS, il semble difficile d’imaginer un libre partage qui irait au delà d’un partage pour usage scientifique , avec toutes les difficultés que présente la définition de ce périmètre.

Au delà, il y aurait en effet des dangers de captation de données qui peuvent être très sensibles. Par ailleurs, pour certains types de données, le temps d’exploitation pouvant mener à des publications est assez long, ce qui plaide pour des délais avant mise en partage ajustés à ces particularités.

Sciences de l'univers:

En astronomie, et plus généralement dans une partie des sciences de l'univers ou des sciences d'observation, se généralise le paradigme d'observatoire virtuel. Les données sont libres d'accès en astronomie pour toute la communauté après la fin d'une période propriétaire. C'est l'approche du réemploi maximalisé des données qui est privilégiée. Pour cela, il est nécessaire d'harmoniser et de standardiser les formats, les descriptions et les modes d'accès aux données d'archives, aux métadonnées et aux applications susceptibles de les traiter, de manière à réaliser l'interopérabilité. Cette interopérabilité s'étend à la liaison des données de la recherche avec les publications en ligne. Les dangers de captation à des fins commerciales ont été peu prégnants jusqu'à ce jour, même si les choses pourraient changer à l'avenir (par exemple avec la “météorologie spatiale” et l’observation détaillée des éruptions solaires).

Biologie

En biologie, l’édition numérique est généralisée et la recherche d’articles se fait via des plateformes thématiques développées par des institutions académiques et permettant d’accéder librement aux résumés d’articles. La plus importante (PubMed) est proposée par le National Institute of Health américain (NIH). L’accès à la totalité des articles est lui généralement payant, le transfert du copyright à l’éditeur étant la pratique la plus courante. Il est à noter que le NIH, s’est opposé à cette pratique et propose en accès libre, via PubMed, une version non formatée de tout article publié par un éditeur décrivant des travaux financés par le NIH. Depuis une dizaine d’années, l’Open Access se développe. Le coût de publication est alors généralement payé par les auteurs lors de la publication.

Si les techniques de "fouille de données"_sur les textes ne sont pas prioritaires pour la plupart des domaines de la biologie en termes d'outil de découverte, (mais plutôt en termes de collecte documentaire) la "fouille de données" proprement dite prend une place de plus en plus importante. Le libre accès à ces données est répandu, à l'instar de ce qui s'est passé pour le génome humain. De nombreux éditeurs dont Nature, conditionnent d’ailleurs la publication d’un article au dépôt des données en masse, associées à une publication, sur une plateforme accessible gratuitement par tous. Il est à noter que cette exigence va au-delà des données numériques et concerne également le matériel produit dans le cadre d’une publication. Lorsqu’un papier décrit un matériel particulier (lignée de cellules, microorganismes ou souris génétiquement modifiées, virus, anticorps, ...), l’éditeur (*Nature*…) demande un engagement

de l’auteur pour la donation de ce matériel aux autres chercheurs académiques. Des plateformes internationales existent pour la conservation et la distribution de ce matériel.

Au delà de la question des données en masse, plusieurs éditeurs dont *Nature* songent à mettre en place un système permettant d’accéder, via leur site, aux données brutes ayant conduit à l’élaboration des figures d’un article. Si cela permettra au lecteur de s’assurer de la bonne interprétation des résultats, la question se pose de la propriété de ces données et de leur cession éventuelle.

Physique

En ce qui concerne la physique hors « grands instruments », l’accès libre aux données brutes n’est pas encore très répandu. En revanche, de nombreuses bibliothèques numériques ont été constituées et mises en libre accès par des groupes de chercheurs ; régulièrement mises à jour, elles portent aussi bien sur la modélisation théorique de problèmes génériques (conduction électrique, dynamique moléculaire) que sur la mise au point et le pilotage d’expériences (interfaçage d’appareils, bibliothèques pour le traitement des données). Le numérique joue également un rôle essentiel dans le domaine de la diffusion des résultats, avec l’utilisation quasi-systématique des serveurs de pré-publications. Le dépôt sur ces serveurs se fait de manière simultanée avec l’envoi à une revue scientifique avec comité de lecture, et permet de « prendre date » immédiatement.

Chimie:

Le domaine de la chimie est en fait vraiment intermédiaire entre les pratiques des sciences de la biologie et celles de la physique. La règle, ce sont des publications dans des journaux payants de sociétés savantes (American Chemical Society, Royal Society) ou sociétés commerciales (Wiley, Elsevier, ...) et un développement timide de l'"open access" type "gold" payé par les auteurs. Il y a en fait peu de différences entre les deux, les négociations avec l'ACS ont été à un moment donné plus dures qu'avec Elsevier. Il n'y a pas de pré-publication comme ArXiv. Des bases de données libres d'accès se développent, notamment la base de Cambridge qui contient toutes les structures moléculaires publiées.

Mathématiques et informatique:

En mathématiques, les bases de données relatives aux publications sont très importantes pour le travail individuel et communautaire. Une particularité de cette discipline est l'importance de l’accès facile aux publications "anciennes" (plusieurs années, décennies voire siècles). La pérennité quant à l'accès à ces publications est donc capitale pour la recherche. Les plateformes d'archivage des publications comme HAL ou Arxiv répondent donc en partie à cette problématique et sont à soutenir, ainsi que les plateformes de métadonnées (MathSciNet, Zentralblatt,…). En ce qui concerne les données numériques, il est nécessaire que pour des questions de reproductibilité, comparaison et interprétation de méthodes de simulation et calcul, elles soient librement accessibles et elles-aussi pérennisées (archivage, catalogues de jeux de données, …) et ceci concerne aussi bien les logiciels que les codes de calcul. Par ailleurs les mathématiques jouent un rôle important dans l’analyse, la gestion et l’exploitation des masses des données (questions autour du Big Data). Il est certainement très important que les données soient accessibles, mais quand elles deviennent de plus en plus massives, il faut aussi pouvoir les exploiter de manière efficace. Dans ce domaine il y a d’importants défis à relever pour la recherche mathématique.

La "publication des données":

Une question importante transversale aux disciplines est celle de la "publication" des données. L'exigence du libre accès est claire dans le cas de données associées à des publications dûment validées par les revues à comité de lecture. Mais qu'en est il de données qui seraient mise en ligne avant publication, par exemple pour analyse et interprétation dans le cadre de large colllaborations? C'est une réalité qui monte dans nombre de disciplines. Ce problème est d'autant plus aigu que la définition même de ce qu'est une donnée publiée est parfois floue.

Dangers et garde-fous juridiques

On constate donc, que comme toute avancée, la numérisation des données et des résultats de la science peut avoir des contre-effets. Les résultats scientifiques jouent un rôle capital dans la concurrence économique mondiale en conférant à leurs détenteurs des avantages compétitifs parfois considérables. En retour, la science moderne a besoin, pour se développer et expérimenter, de technologies que va souvent lui offrir le monde de la production régi par le marché. C'est particulièrement le cas des fonctionnalités qui peuvent être fournies par les éditeurs scientifiques

Pourtant chacun sait que le développement de la connaissance se fait à travers les échanges d'idées, de résultats et de données entre scientifiques. Il est donc impératif de limiter la captation du travail de ces derniers par les intérêts privés et en même temps de fournir un cadre légal pour permettre de libérer autant que faire se peut l'échange des données d'usage scientifique.

Trois principes:

Le conseil scientifique soutient trois principes importants qui permettraient de répondre à ces objectifs :

- La liberté complète de circulation et d'usage des données scientifiques pour réutilisation dans le cadre de la science, sous réserve qu'un temps minimal de réserve permettant aux producteurs de données de les interpréter et de publier soit garanti par la loi. Cette exigence de libre circulation des données s'étend d'une part aux publications et d'autre part aux données et textes non scientifiques à l'origine mais constituant la matière première de beaucoup de recherches, notamment en sciences humaines et sociales.

- Cette exigence de la mise à disposition des données s'étend aux services à valeur ajoutée (traitement massif de type "Big data", fouilles de données, lien aux métadonnées, interopérabilité) qui doivent également être publics et libres d'accès pour éviter toute captation abusive. Ceci supposerait, dans le cas de création de services et de plateformes par les éditeurs et plus généralement le secteur privé, des garanties légales de juste prix non discriminants.

- Elle suppose aussi la clarification des droits à disposer de l'usage de leurs productions scientifiques et de leurs publications par les auteurs face aux éditeurs et aux autres acteurs privés. Les droits de propriété intellectuelle des scientifiques ne doivent en aucun cas être concédés gratuitement aux éditeurs, de manière à favoriser la libre circulation des résultats scientifiques.

Le conseil scientifique tient également à saluer le travail réalisé par le COMETS dans son rapport intitulé “les enjeux éthiques du partage des données scientifiques” et il souscrit aux recommandations contenues dans ce texte.

Annexe:

Commentaires sur le "tableau de propositions" du livre blanc du CNRS concernant la loi "l'ambition numérique française ».

Nous commentons ici brièvement le document émanant de la DIST et du cabinet d'avocats "Alain Bensoussan" daté du 18/05/2015. Ce tableau nous semble globalement apte à promouvoir une conception ouverte et libre de l'usage des données numériques pour la science. Nous avons néanmoins quelques remarques et questions

Une première question est relative au paragraphe "Approche générale" : les thèmes et sous thèmes ou sujets sont-ils le reflet de l'organisation du projet de loi ou sont-ils une initiative des rédacteurs du document que nous commentons ici ? Ce découpage ne nous paraît pas forcément le plus pertinent du point de vue de la science.

Au sujet de la sous-section "Tirer parti de l'économie de la donnée", le point clef de la réflexion sous-jacente aux cinq propositions qui y sont déclinées semble être dans cette affirmation "le système de biens communs ne doit pas empêcher le transfert et la valorisation de la recherche et un équilibre doit être trouvé entre les deux systèmes". ***La notion de bien commun avancée dans le texte nous parait être un présupposé incontournable*** *et les propositions faites nous semblent saines, en particulier la possibilité d'avoir des périodes "propriétaires" sur les publications qui devraient selon nous être étendues aux données (il faut laisser aux auteurs le temps de "chercher"), et la mise en place d'un statut de données d'intérêt général étendu aux données de la science. Il conviendrait cependant de mieux valoriser que ne le fait le texte dans ces propositions la facilitation de l'accès aux données et publications scientifiques pour les scientifiques comme condition préalable au développement de la science elle-même. Les droits particuliers de propriété intellectuelle et les règles de réutilisation ne doivent pas faire obstacle aux échanges nécessaires aux chercheurs.*

Par ailleurs création d'un statut juridique pour les données scientifiques pourrait aussi permettre de définir plus précisément ce qu'est une donnée scientifique publiée et donc publique.

Concernant les sections 2.1.2 et 2.1.3 et la proposition 6, le point soulevé est fondamental. L'accès libre aux services permettant d'optimiser l'usage des données est aussi important que l'accès aux données elles-mêmes. Les données et les plateformes ne doivent pas seulement être distinguées du point de vue de leur "nature" même mais aussi de l'usage que l'on en fait. C'est sans doute la notion d' "usage équitable" qu'il faut promouvoir. Les chartes sont-elles cependant un outil suffisant pour promouvoir les "valeurs de partage" et l' "esprit de collaboration "inhérents à la science » ?

Concernant la section 2.2 et la proposition 7 consistant à promouvoir les formats et standards ouverts et interopérables, il convient de noter que c'est également un point fondamental. Mais

il ne s'agit pas que d'une question de "loyauté entre acteurs économiques" ; c'est aussi une condition "sine qua non" de l'interopérabilité et de l'existence même de la "science ouverte".

Concernant la section 2.3 et la proposition 8, nous estimons que l'existence d'un comité général d'éthique et aussi de comités par domaine d'activité serait sans doute utile pour donner des avis sur les nouvelles pratiques qui ne manqueront pas d'émerger ; cependant, le socle fourni par la loi sur le numérique doit être suffisamment clair et précis dès le départ sur ce point pour encadrer les pratiques actuelles.