Sciences technologiques et sciences humaines et sociales : les enjeux d'une gestion mutualisée des données de la recherche.

Note pour l'audition OPECST(Office parlementaire d'évaluation des choix scientifiques et technologique)

Assemblée nationale. Séance du 21 janvier 2016 consacrée aux sciences humaines et sciences technologiques

Mon intervention portera sur un point précis de convergence entre les sciences humaines et les sciences technologiques : la gestion des données de la recherche.

Jusqu'à une époque récente, cette question pouvait sembler relativement déconnectée de l'activité principale du chercheur. Certes, le chercheur en SHS dépendait hier comme aujourd'hui des données accessibles et il lui fallait parfois, hier comme aujourd'hui, faire preuve d'ingéniosité et d'inventivité pour recueillir des données, susciter leur production et, ensuite, les faire parler, par des problématisations ou des outils originaux. Si je prends l'exemple de l'histoire, il y avait en amont les sources conservées par les archives et les bibliothèques et, en aval, la publication des résultats de la recherche au format papier, sous forme d'édition critique de sources, d'article ou de livre.

Depuis le 20^{e} siècle, de plus en plus de chercheurs en SHS sont également producteurs de données sur des supports de haute technologie et de faible pérennité : photographies, films, enregistrements sonores, bandes magnétiques, fiches perforées, disquettes, disques durs sont autant de supports qui posent la problématique de la conservation et de la transmission des contenus.

Le tournant numérique en SHS a accéléré encore la nécessité d'une politique de gestion de nos données, de leur traitement, de leur interopérabilité avec d'autres jeux de données et, enfin, de leur conservation à long terme.

La fragilité des données numériques est en effet très préoccupante car elle tient à deux facteurs : fragilité des supports, évolution des standards de lecture

1/ Fragilité des supports

Nous sommes confrontés à un paradoxe, qui est aussi un défi technologique, qui fait que la durée de vie des supports de mémoire ne cesse de s'écourter à mesure que l'on avance dans le temps. En grossière approximation :

- une inscription sur une pierre durera 10 000 ans
- un parchemin en moyenne 1 000 ans
- une pellicule = 100 ans
- un disque vinyl, 50 ans
- un support informatique ? On ne sait pas.

Dans les années 80, on a cru que le support CD constituait un support inusable, en raison de l'absence de contact entre le support physique et le lecteur de l'information. On sait depuis que leur fiabilité est incertaine, et ne peut être garantie au-delà de 20 ans

Les disques durs sont garantie 5 ans, la mémoire flash ne dépasse guère 10 années si on ne la sollicite pas au-delà du seuil des 100 000 réécritures.

2/ Deuxième facteur de fragilisation : l'évolution des standards :

Évolution des standards de lecture des supports. Perte des lecteurs Betacam, ¾ Umatic en vidéo, perte des lecteurs de fiches perforées, de disquette 5 ¼ en informatique etc.

Pire, dans le numérique, la lecture de l'information évolue en fonction des logiciels permettant l'interprétation du code. Nous avons tous fait un jour l'expérience d'une difficulté de lecture d'un « vieux » fichier de traitement de texte, devenu incompatible avec un logiciel plus récent. L'encapsulage des données dans des logiciels propriétaires pose également la question du libre accès de ces données, de leur interconnexion possible à d'autres corpus, de leur disponibilité future.

A ces questions, qui sont à la fois des défis scientifiques et technologiques, il n'y a pas aujourd'hui de réponse simple et définitive, mais on peut en tirer un constat, qui doit guider une politique rationnelle de gestion des données de la recherche en sciences humaines et sociales :

1 – l'enjeu porte sur notre capacité à préserver l'accumulation des savoirs, et à la transmettre

2 – cette transmission passe par une politique de conservation coûteuse en temps (parce qu'elle suppose une veille technologique coordonnée) et en matériel (datacenter sécurisés)

3 – cette politique de gestion des données ne peut être prise en charge au niveau du chercheur ou d'un laboratoire

4 – elle doit relever d'une stratégie coordonnée, à l'échelle nationale (MSH, Equipex, Idex, UMS…) et internationale (européenne, avec les ERIC), qui passe par la prise de conscience des chercheurs, et qui offre des solutions technologiques et humaines de concertation humaine et de mutualisation technologique

Un exemple de structure contribuant à la mise en œuvre de cette politique de gestion des données numériques de la recherche :

TGIR Huma-Num, créée en mars 2013 par la fusion du TGE-Adonis et de l'IR Corpus, est une très grande infrastructure de recherche (TGIR) visant à faciliter le tournant numérique de la recherche en sciences humaines et sociales.

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.

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.

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. Huma-Num a élaboré des services qui correspondent à chaque étape du cycle de vie des données de la recherche :

Stocker – Traiter – Exposer – Signaler – Diffuser – Archiver

Ex de service : Nakala. Entrepôt de données numériques avec leur description, qui permet d'assurer leur conservation, leur citabilité (url pérennes), leur signalement (moissonnage distant avec le protocole OAI-PMH), leur archivage à long terme, en partenariat avec le CINES

Ces institutions sont elles-mêmes fragiles, et il me paraît indispensable d'assurer leur développement pour construire des points de repères forts permettant de rallier les chercheurs à une politique nationale coordonnée

Pour conclure, je voudrais revenir à la recherche, pour souligner combien le développement de structure permettant une gestion mutualisée des données peut aussi avoir un effet bénéfique à l'évolution de nos savoirs et au décloisonnnement de nos disciplines.

L'incidence de la diffusion de ces solutions de gestion co-élaborée avec les chercheurs, c'est qu'elle stimule le travail collectif et nous permet de nous concentrer sur l'élaboration de nouveaux outils de publications, de visualisation et de diffusion de l'information scientifique.

Nous sommes ici plusieurs à porter depuis de nombreuses années des plateformes de publication originale et évolutive, qui ont permis peu à peu de constituer des communautés d'intérêt scientifique qui expérimentent la voie de la convergence des SHS avec les sciences technologiques. Je pense au site Fabula.org par ex, mais Alexandre Geffen ici présent en parlera mieux que moi, et pour ma part, je peux témoigner du fait que ces expériences produisent des effets inattendus et des perspectives nouvelles sur le terrain émergent de la recherche participative.

Dans le domaine de l'histoire de la justice, qui intéresse à la fois les historiens, les juristes, les professionnels de la justice, les journalistes, les documentaristes et un public non négligeable, il existe la plateforme Criminocorpus, que j'anime avec un collectif depuis 2003. On trouve sur cette plateforme des outils de recherche assez classiques, comme des bases de données relationnelles, des chronologies juridiques, une revue, un blog d'actualités, une bibliothèque numérique, mais aussi des réalisations plus spécifiques et plus expérimentales, comme des visites de lieux de justice, combinant plan de situation et vidéos ; mais aussi un outil de consultation de corpus permettant de comparer des corpus évolutifs, comme des textes de lois (on peut ainsi suivre toutes les versions de l'Ordonnance du 2 février 1945 sur la justice des mineurs, de 45 à nos jours, ou encore le Code civil, de 1804 à 2004)

Parmi nos projets en cours, nous souhaitons initier un grand recueil du patrimoine judiciaire de la France (tribunaux, lieux d'exécution des peines), qui soit à la fois coordonné scientifiquement, et ouvert à des contributions citoyennes.

Il s'agit là, parmi bien d'autres, d'exemples qui démontrent que le tournant numérique offre au SHS de nouvelles opportunités d'ouverture de nos savoirs vers la société.

Marc Renneville, DR CNRS

Directeur du CLAMOR. Centre pour les humanités numériques et l'histoire de la justice (UMS 3726)

Ancien directeur de la TGIR Huma-Num (UMS 3598)