

Ouvrir un chantier commun

Réconcilier la logique d'archive et la recherche active autour d'un dépôt numérique thématique en littérature et culture québécoises

René Audet

CRILCQ, Université Laval
Directeur, Laboratoire Ex situ

Ce texte est la version écrite de la communication prononcée lors du congrès annuel de la [Society for Digital Humanities](#), dans le cadre du Congrès de la Fédération canadienne des sciences humaines et sociales (Montréal, Concordia University, 1^{er} juin 2010).

I. Des retailles

Comme chercheurs en études littéraires, en sciences humaines, que produisons-nous ? Des communications, des articles, des monographies, évidemment. Mais aussi, nous produisons une masse de documentation qui se sédimente, qui s'enfouit dans des classeurs, qui meurt dans des caisses.

Comme chercheurs, nous avons constamment à nous questionner sur la portée de notre travail, sur la valeur relative de nos résultats de recherche — qui constituent, à n'en point douter, notre principale contribution à l'avancement du savoir. Nous oublions pourtant trop souvent les restes de l'exercice de la recherche, ce qui persiste en dehors des conclusions que nous avançons. Ces restes, quels sont-ils ? Du papier, bien du papier... Des documents pêle-mêle : photocopies, courriels, documents de traitement de texte ou pdf. Plus concrètement encore : des bibliographies, des notes de lecture, des tables statistiques, des plans, des ébauches de texte, des banques de données, des schémas, des articles savants, de la documentation diverse.

Que faire de ces traces de la recherche ? (si l'on exclut les feux de joie auxquels recourent nos étudiants à la fin des trimestres...) En fait, y a-t-il quelque chose à faire avec les brouillons, les signaux de tâtonnement, les listes interminables ? Ce n'est pourtant pas une question que l'on se pose explicitement, sinon lorsque des considérations concrètes (classement, ménage, débordements sur nos tables de travail) nous y poussent.

D'autres se sont posé la question : pensons la valorisation des manuscrits d'écrivains, qui a cours depuis quelques décennies, dans le sillage d'une approche plus processuelle de l'écriture littéraire (et, selon certains autres plus cyniques, conformément à la mythification de la figure de l'écrivain depuis la deuxième moitié du dix-neuvième siècle). On voit un phénomène similaire en études cinématographiques, où les *rush* et les chutes de film sont considérés comme

traces de la création artistique à part entière. En regard de ces domaines, les restes de la recherche scientifique en arts et lettres, en sciences humaines demeurent encore sous-estimés : nous consultons et lisons le produit fini, où rien ne dépasse, où tout est bien ficelé... Mais persistent les brouillons, les ensembles documentaires, les notes, les fiches, les grilles d'analyse... Persiste donc un ensemble de retailles produites par l'exercice de la recherche en sciences humaines.

Des retailles de la recherche ? Chaque chercheur en produit, assurément, en marge de ses réflexions et travaux. Le cas qui m'intéressera ici est plus spécifique, à savoir les retailles produits par de grands projets de recherche et des unités plus larges, que sont les centres de recherche. Un centre en l'occurrence, le Centre de recherche interuniversitaire sur la littérature et la culture québécoises ([CRILCO](#)), qui compte trois composantes (à l'Université Laval, l'Université de Montréal et l'Université du Québec à Montréal).

Ce centre (et ses composantes antérieures : le CRELIQ à l'Université Laval et le CÉTUQ à l'Université de Montréal) est le lieu où ont été menés nombre de projets sur la littérature et la culture québécoises. Pensons au *Dictionnaire des œuvres littéraires du Québec* (qui en est à son huitième tome), à *l'Histoire de la vie littéraire au Québec* (dont le sixième tome paraîtra sous peu), au projet *Penser l'histoire de la vie culturelle au Québec* — sans compter d'anciens projets sur l'enseignement de la littérature au collégial, sur les best-sellers, sur la littérature par fascicule, des projets d'édition critique (Hubert Aquin, Philippe-Aubert de Gaspé, pour ne nommer que ces deux exemples), des projets sur la littérature contemporaine, sur la littérature dans les revues populaires, sur le théâtre de Robert Lepage, sur l'essai dans les années 1960...

Ces projets ont légué un impressionnant héritage, d'abord en termes de publications scientifiques, mais aussi en documentation d'appoint : des dizaines et des dizaines de mètres linéaires d'archives subsistent, comportant :

- des listes de best-sellers ;
- des bibliographies sur fiches carton ;
- de la documentation iconographique ;
- des notes de lecture ;
- des schémas d'analyse ;
- des enregistrements de téléthéâtres, leurs retranscriptions, leur édition critique, leur analyse ;
- des entrevues audio avec des éditeurs ;
- des banques de données ;
- des demandes de subvention ;
- des données statistiques...

Des archives de tous les formats :

- papier, papier fax, affiches, fiches de carton, affiches, liasses de papier fleuries de post-it qui décollent ;
- des documents numériques de différentes générations (banques de données en pro-cite, des documents wordperfect) ;
- des supports variés : microfilms, bandes audio, photos, diapositives ;
- des supports informatiques divers : cd, zip, bernouilli, disquettes 3 1/2 pouces, 5 1/4 pouces, 8 pouces, des cartes perforées...

L'enjeu, on commence à le pressentir (ici comme dans la réalité concrète du centre et de ses composantes) : que faire de ces retailles de la recherche ? Comment gérer ces retailles ? Tenter de répondre à cette question, c'est se confronter à des logiques variées et en relation tendue, tension qu'il faut arriver à réduire.

2. Des logiques

Dans le contexte d'une réflexion sur les retailles de la recherche s'opposent d'emblée des logiques de gestion documentaire numérique. Ce qui vient spontanément en tête, ce sont tous les discours sur l'*open access* qui inondent le champ discursif documentaire en contexte numérique. On évoque ses formes (que [Peter Suber](#) résume bien), à savoir le *green open access* (qui consiste à déposer les documents scientifiques dans des dépôts numériques ouverts) et le *gold open access* (qui appelle la publication d'articles dans des périodiques savants en ligne et ouverts). Ces discours glosent également les modalités de l'*open access*, les uns visant à changer les mentalités (en convaincant de l'impératif de diffuser le savoir par des lieux ouverts), les autres travaillant plutôt à changer les cadres (ainsi l'[open access mandate](#) de l'université Harvard qui oblige les chercheurs, sauf exceptions, à déposer une copie de leurs publications dans un dépôt institutionnel ouvert ; [Bernard Rentier](#), recteur de l'Université de Liège, qui prône une contrainte forte d'utilisation des dépôts institutionnels ; ou encore le dépôt centralisé [HAL](#) en France, qui fonctionne relativement bien actuellement).

Ces propos sur l'*open access*, il ne faut pas l'oublier, portent sur des publications (ou, plus précisément, sur des *pre-prints*, des versions pré-finales des articles en cours de publication). Il importe donc ici d'établir une distinction claire entre les publications, l'édition d'une part, et d'autres part les données, la documentation d'accompagnement de la recherche. Ces deux catégories de documentation numérique appellent des types de projet distincts. Les publications en *open access* trouvent leur niche dans des dépôts numériques, qu'ils soient thématiques, institutionnels ou centralisés. Pour leur part, les ensembles de données, davantage liés à l'esprit des *digital humanities*, commandent ce qui est au cœur du projet qui m'occupe, à savoir une instance de diffusion des sources, des archives documentaires.

Ces sources, quelles sont-elles ? En études littéraires, en histoire de la culture, ce ne sont pas tant des données quantitatives que de la documentation brute, en fait plutôt des données à demi-traitées. Nous nous butons ici au mythe du *raw data*, qui joue encore de façon plus forte dans le champ des arts et lettres. La donnée brute n'existe pas, toute donnée étant déjà travaillée, ne serait-ce que par sa sélection (qui lui confère un champ de pertinence). L'inscription d'une telle donnée, d'un tel document dans un ensemble numérique de données vient répercuter cette contextualisation, voire l'amplifier (par un travail d'annotation ou d'étiquetage). Mais à quel point faut-il pousser cet accompagnement contextuel ? [Dan Cohen](#), dès 2004, montrait que l'on se trompe souvent dans la prédiction des usages précis des documents d'archives (l'étude des *logs* illustrant bien la consultation directe de certains documents, découverts grâce aux moteurs d'indexation du web ou par hasard, par *sérendipité*). À quel point la définition de la donnée inclut-elle sa mise en contexte ?

Poser cette question, c'est ouvrir la problématique de la gestion des données — problématique fondamentale et coûteuse, qui dépasse les limites de la présente réflexion. Néanmoins, elle mérite qu'on lui accorde une attention particulière. C'est d'ailleurs ce que faisait Marin Dacos, directeur du Centre pour l'édition électronique ouverte, dans l'atelier qu'il animait récemment au [THATcamp](#) de Paris (18-19 mai 2010) : il y a évoqué les nécessités, les risques, les contraintes et les enjeux de reconnaissance posés par la diffusion des sources en sciences humaines et sociales, [exercice](#) fort ambitieux mais nécessaire.

Une telle réflexion nous conduit sur le terrain des fonds documentaires et met en relief deux logiques possibles de prise en charge. D'une part apparaît la nécessité de saisir ce qui croupit dans les classeurs et les boîtes : il faut opérer une manœuvre de stabilisation, d'arrêt de la documentation, pour classer celle-ci, l'identifier et la décrire. C'est une logique proprement archivistique qui intervient. D'autre part surgit également le fort sentiment qu'il faut *rentabiliser* cette documentation : puisqu'elle a déjà été utile, peut-être peut-elle l'être à d'autres ? À tout le moins faut-il la rendre disponible, pour que d'autres chercheurs puissent faire cette évaluation... La manœuvre est alors celle de l'extraction et de la dissémination, qui correspond à une logique de diffusion.

Y a-t-il là une contradiction fondamentale ? Ou ne s'agirait-il pas plutôt d'une opposition rhétorique, ces logiques correspondant à des étapes formant une séquence ? Si ces logiques appellent des conceptualisations et des opérations bien distinctes, elles demeurent toutes deux inévitables, voire fondatrices de la prise en charge nécessaire des fonds documentaires scientifiques. La logique archivistique est déjà fortement balisée et maîtrisée ; son application ne pose pas de problèmes majeurs. Il en est autrement de la logique de diffusion, qui renvoie pour sa part à une idée plus vague, et à des processus encore plus fortement indéterminés, en ce qu'ils peuvent répondre à des conceptions variées de la diffusion. Cette portion de la prise en

charge, on le sent, oblige à un positionnement clair dans un champ très peu balisé. C'est là certainement un défi tout à fait corollaire aux problématiques concrètes liées à la constitution d'un tel ensemble de données — pensons aux coûts reliés à l'opération et aux spécialisations techniques impliquées, pour résumer caricaturalement cette part pragmatique du projet.

Lorsque la détermination du type de projet et l'application des logiques de prise en charge surviennent, l'épreuve de la réalité constitue le meilleur guide pour orienter les réflexions et les décisions. C'est alors que se dessinent concrètement divers usages au sein de la mise en place d'un dépôt numérique de documentation.

3. Des usages

Le projet qui se met en place actuellement est nommé le DÉCALCQ (Dépôt électronique et vitrine de consultation des archives en littérature et culture québécoises). Il s'agit du principal projet ayant justifié la mise en place du Laboratoire Ex situ, en réponse à une commande du CRILCQ portant sur la prise en charge des archives scientifiques du centre. Ce projet a été planifié et construit en amont, avec les risques que cela comporte, à partir d'estimations des usages possibles ou souhaités d'un tel dépôt numérique. Le projet DÉCALCQ est d'une ampleur intermédiaire : il est plus important, par exemple, que les projets menés au [Center for History and New Media](#) que dirige Dan Cohen à George Mason University ou que les projets fédérés par le portail [NINES](#) sur la littérature britannique du dix-neuvième siècle, mais plus modeste qu'un dépôt institutionnel, à plus forte raison qu'un dépôt centralisé. Pour parvenir à sa réalisation, le projet a appelé la mise en place d'une petite [équipe](#) (directeur, archiviste, technicien en informatique) qui supervise le travail de sous-équipes de « production » dans les trois sites universitaires du CRILCQ.

L'esprit du projet repose sur la dualité des usages souhaités pour le dépôt : il se caractérise par la cohabitation des archives et d'un volet collaboratif rattaché à l'actualité de certains projets de recherche.

Le volet « archives » comporte une série d'interventions attendues dans une telle approche. À partir d'une documentation papier, média et numérique, il faut opérer un tri, un élagage et un ordonnancement du fonds, qui mène au final à une description (et l'établissement d'un plan de classification). C'est là la base pour saisir les fonds et permettre une publicité, si minime soit-elle, des fonds disponibles dans chacun des sites du centre. De cette façon est établi ce qui constitue le patrimoine documentaire scientifique du CRILCQ. Suit l'étape de la numérisation (PDF/A, OCR lorsque possible) d'une large part des fonds, qui conduit à l'intégration des métadonnées (une interprétation des champs-types du Dublin Core et la description des relations de parenté entre les documents et les projets de recherche qui les ont générés) ainsi

que l'établissement des droits d'accès (en fonction des sensibilités autant personnelles que juridiques). Cette documentation numérisée prend ensuite place dans un dépôt numérique. Plutôt que de recourir à des logiciels spécialisés pour dépôts institutionnels (Fedora, DSpace, Eprints, etc.), nous avons comme équipe opté pour [Alfresco](#) (un *Entreprise Content Management system*), un outil *open source* très robuste mais flexible, qui constitue la voûte documentaire au cœur du projet. À cette infrastructure se greffent des outils de consultation : navigation arborescente pour retrouver les documents selon leur rattachement originel, recherche par mots-clés et facettes pour favoriser le repérage sémantique de la documentation, ainsi que modules de mise en valeur de certains contenus (à partir de résultats stabilisés ou fondés sur des indicateurs de recherches transversales ou d'ensembles déjà constitués).

En soi, jusqu'ici, ce projet correspond à la pratique générale de la numérisation d'archives. Toutefois, là où DÉCALCQ se différencie, c'est au niveau de la nature et de la stabilité des fonds : les fonds de projets de recherche ne sont pas (tous) inactifs comme le sont celui des manuscrits de Jane Austen ou celui des brouillons de *Madame Bovary*... Il y a toujours de nouveaux projets ; plusieurs sont en opération depuis plusieurs années et ont déjà produit des masses documentaires à prendre en charge. Déjà, sur ce point, des enjeux distincts se profilent.

À ce volet « archives » se greffe un volet « collaboration », qui prend acte des caractéristiques fondamentales du milieu d'où émergent ces fonds documentaires. Très tôt, nous en sommes venus à envisager comment cet outil du dépôt numérique pouvait être mis à profit pour les recherches en cours et à venir, et non pas seulement pour les projets terminés. Profitant notamment d'une couche logicielle collaborative venant avec Alfresco, nous planifions faire de ce dépôt numérique une infrastructure pour la recherche, par l'utilisation d'outils de collaboration qui ont pour effet de transformer le dépôt en zone de travail partagé. En lui arrimant des wikis, des banques de données, des sites web, des outils de gestion de projets ou des calendriers, nous refusons de faire du dépôt numérique un lieu statique. Les documents, les données n'y meurent pas : ils sont remobilisés par de nouveaux projets ; les nouveaux documents numérisés ou créés numériquement sont utilisables et utilisés, puis survient une bascule des données en statut d'archives et de données accessibles au moment de la conclusion des projets de recherche (ce qui évite d'avoir à planifier à l'avenir un aussi lourd traitement des archives que pour les projets actuellement terminés).

Le projet DÉCALCQ n'est donc pas un dépôt thématique au sens strict, puisqu'il rassemble des données et qu'il est une interface de collaboration. Il n'est pas une seule plateforme de travail, comme il a pour fonction de stabiliser les documents, de leur associer des métadonnées contrôlées. Il n'est pas plus une plateforme de publication — il vise d'abord et avant tout à rassembler des sources documentaires. S'il fallait le définir, ce projet constituerait une *plateforme hybride de diffusion d'archives et de collaboration pour équipes de chercheurs*.

En ce sens, il s'oriente clairement sur plusieurs des dimensions majeures de la recherche en sciences humaines et sociales (telles que les définissent [Caverni et Dacos, 2009](#)) : sont ainsi favorisés « l'accès, la stabilisation et la mise en relation des données numériques entre elles », « la vie des communautés scientifiques (débat scientifique, identité numérique) », « les données sur lesquelles s'appuient les chercheurs (archives historiques, enquêtes orales, statistiques diverses, données archéologiques...) mais aussi les méthodes et outils qui permettent d'en extraire des découvertes scientifiques », méthodes qui se trouvent partagées par l'expertise développée au centre de recherche ; il n'y a que « l'édition des résultats de la recherche (livres, revues, archives ouvertes) » qui est un objectif indirectement rencontré pour son volet « publication ».

Cette réflexion sur les usages ne peut faire l'économie de l'identification de certains enjeux sous-jacents. Ils sont d'abord et avant tout techniques. Toute mise en place d'un projet de ce type doit interroger les possibilités et limites des logiciels disponibles, et veiller à ne pas s'isoler en choisissant un logiciel fermé sur lui-même ou en développant une solution de toutes pièces (choix souvent coûteux, surtout dans un monde où bien des applications sont développées de façon semi-confidentielle). Il en est de même des protocoles (choix d'encodage, types de métadonnées, séquences de traitement pour la numérisation...) dont l'issue est souvent la stabilité et le caractère opératoire du fonds numérique, ainsi que la possibilité de s'ouvrir sur d'autres ensembles de données — d'où la préoccupation grandissante pour une interopérabilité des métadonnées des fonds numériques. Un dernier enjeu technique, à la frontière d'enjeux plus politiques (au sens strict), concerne les arrimages institutionnels : il importe de faire en sorte que ce type de projet soit institutionnellement supporté, autant pour partager les coûts rattachés à l'infrastructure que pour assurer la pérennisation du travail réalisé.

La réflexion sur les enjeux engage également la question de la valeur. Celle-ci se pose de façon forte pour les publications en *open access* : c'est toute la problématique de la validation des contenus et de la reconnaissance de ces publications. Dan Cohen, dans son [texte préparatoire](#) à « The Shape of Things to Come Conference » de mars 2010, résumait que « we need to work on the demand side of the social contract of scholarly publishing », signalant que le « supply side » est déjà bien doté avec une variété d'expérimentations offrant une diversité de possibilités aux usagers. Il y a des mentalités à faire évoluer, afin de mieux asseoir la recevabilité, la reconnaissance de cette documentation. Mais qu'en est-il pour les données ? Ce rapport avec la validation et la reconnaissance paraît également problématique, et plus encore que pour les publications en *open access*. Christine L. Borgman, autant dans une conférence prononcée à Columbia University en 2009 que dans le [texte](#) qui a suivi cette réflexion, montre que le travail reste encore à faire, comme l'entrée des données dans le processus de valorisation est plus complexe et moins cadrée : « The key to "better" data — that is, data suitable for curation, reuse, and sharing — is capturing data as cleanly as possible and as early as possible

in its life cycle. Agreements about data sources, structures, and formats will further the development of information infrastructure for digital humanities scholarship. » Des pistes existent déjà, notamment lorsqu'on s'appuie sur les acquis en sciences de l'information et en archivistique.

Enfin, les enjeux sont clairement politiques, au sens large cette fois. Il importe de conduire les gens à recourir à de telles banques de sources, pour qu'ensuite ils puissent y contribuer. De nombreux projets scientifiques ont été menés en sciences humaines, en arts et lettres, et il paraît primordial de ne pas constamment recommencer le même travail ; s'alimenter à des projets antérieurs peut souvent se révéler une approche rentable pour de nouveaux projets. Évidemment, bien des freins au partage et à la collaboration émergent — Christine L. Borgman fait état de la récompense symbolique liée à la publication plutôt qu'au partage de données, de l'effort lié à ce partage, de l'avantage stratégique lié à la possession de sources et du (faux) sentiment de propriété des sources. On le sent bien, les mentalités sont fortes et le milieu académique est féroce. Bien du travail reste à faire sur les perceptions pour favoriser la mise en place et l'intégration dans les usages de telles initiatives pourtant fondées sur la collaboration, la collégialité et l'absolu de l'avancement de la science, toutes valeurs profondément associées à la recherche universitaire.

([Présentation électronique accompagnant ce texte](#))

René Audet

Courriel : rene.audet@lit.ulaval.ca

Twitter : [@reneaudet](https://twitter.com/reneaudet)

Carnet web : <http://contemporain.info/audet>

Site du Laboratoire Ex situ : <http://ex-situ.info>

Twitter du Laboratoire : [@labexsitu](https://twitter.com/labexsitu)