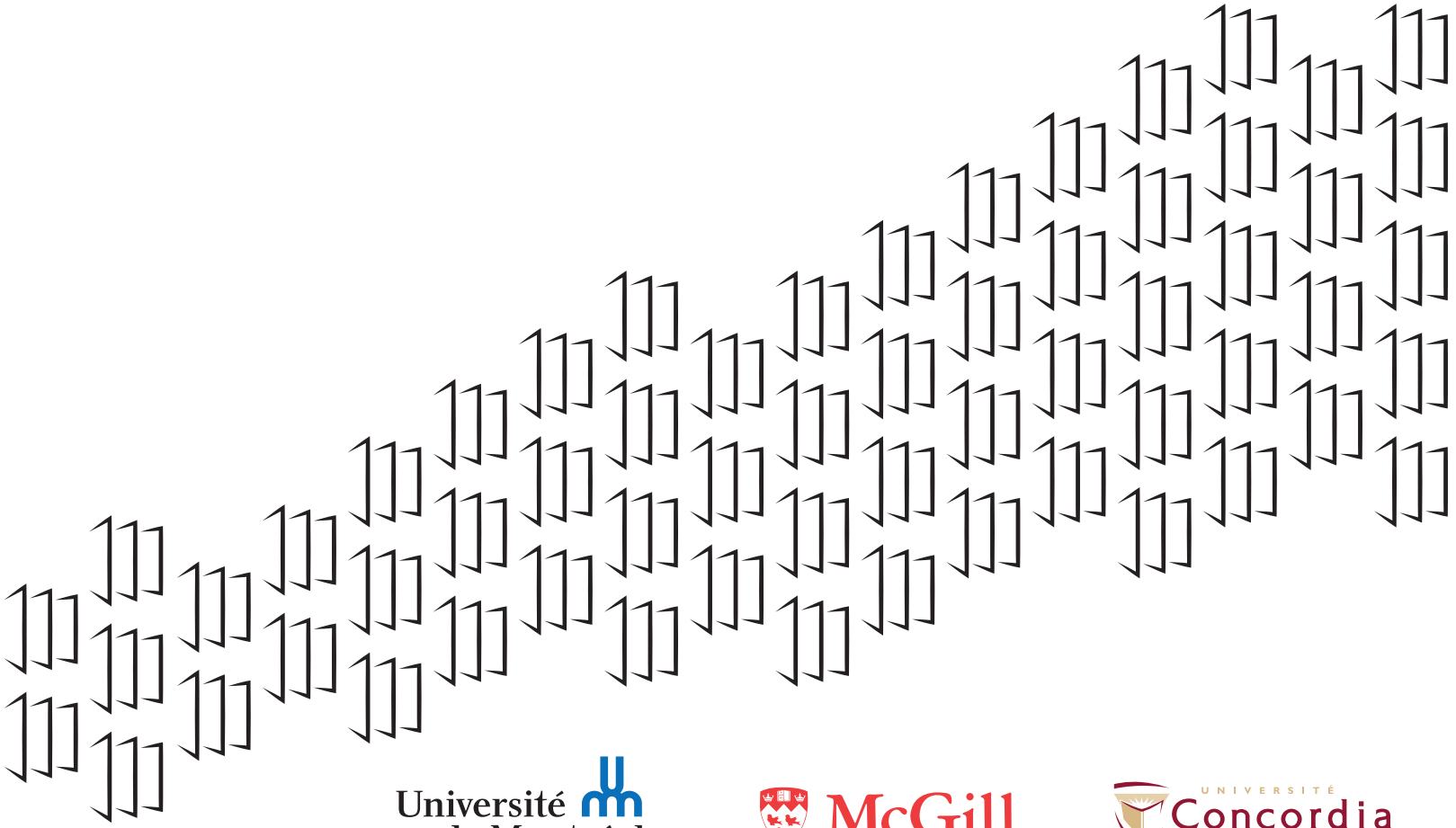
CRIHN
www.crihn.org

Centre de recherche interuniversitaire
sur les humanités numériques

Développer une infrastructure de services numériques pour les Humanités numériques canadiennes

Livre blanc pour la NOIRN, décembre 2020



Université de Montréal

McGill

Université Concordia University

Développer une infrastructure de services numériques pour les Humanités numériques canadiennes

Livre Blanc du Centre de recherche interuniversitaire sur les humanités numériques (CRIHN) sur le futur écosystème d'IRN du Canada, consultation du NOIRN, décembre 2020

Personnes contact :

- **Michael Eberle-Sinatra**, michael.eberle.sinatra[@/umontreal.ca
- **Emmanuel Chateau-Dutier**, emmanuel.chateau.dutier[@/umontreal.ca

Au cours des dernières années, une concentration d'expertises et de compétences dans le domaine des humanités numériques a émergé au Québec. Basé à l'Université de Montréal, le **Centre de recherche interuniversitaire sur les humanités numériques** (CRIHN) regroupe 45 membres réguliers dont 4 chaires de recherche du Canada et 3 chaires institutionnelles, 35 collaborateurs et 5 stagiaires de recherche postdoctorale provenant de huit universités, CÉGEP et établissements de recherche québécois. https://www.crihn.org/

Les propositions ci-dessous essaient de dresser un portrait qui découle de la diversité des situations rencontrées par les membres du Centre et de l'expérience acquise dans le travail avec des moyens numériques en sciences humaines et sociales.

Enjeux actuels

Quels sont les principaux outils, services et/ou ressources d'IRN que vous utilisez actuellement dans votre recherche?

Alors que les chercheurs de notre centre de recherche devraient largement avoir recours à une infrastructure de recherche numérique en raison de leur implication dans des projets mobilisant largement l'informatique, nous sommes contraints de constater qu'un très faible recours à l'infrastructure de recherche actuelle et notamment aux ressources offertes par Calcul Canada.

De notre point de vue, cette situation s'explique sans doute principalement par une prise en compte insatisfaisante des besoins spécifiques des chercheurs en sciences humaines et sociales dans les services actuellement offerts aux chercheurs mais peut-être aussi par une méconnaissance de l'existence de cette infrastructure et des modalités d'accès pour pouvoir en bénéficier.

Avez-vous accès à tous les outils, services et/ou ressources dont vous avez besoin pour mener votre recherche? Quels sont-ils? Qu'est-ce qui manque?

Pour autant, nombre de nos équipes de recherche ont souvent besoin d'entretenir des serveurs pour héberger des applications web, de gérer des données de recherche ou de faire fonctionner des logiciels de recherche. Les solutions adoptées sont souvent hébergées sur des serveurs acquis individuellement dans le cadre de projets financés par le Fonds canadien pour l'innovation (FCI) ou bien en ayant recours à l'offre commerciale. La diversité des situations rencontrées s'explique par l'inégalité de l'offre de services numériques proposée par les établissements universitaires.

Cet état de fait nous semble particulièrement insatisfaisant. En effet, celui-ci ne permet pas toujours de répondre aux besoins spécifiques de la recherche, l'offre commerciale ou des établissements n'étant pas toujours adaptée ou facile d'accès. Il est souvent difficile de pouvoir disposer de bacs-à-sable pour prototyper des applications ou des outils. De plus, cette situation ne favorise pas la mutualisation des moyens et ne permet pas de garantir la maintenance de l'infrastructure de recherche sur la durée. Enfin, elle constitue un frein notable à l'intégration et aux développements de pratiques qui adressent le cycle de vie des données de recherche depuis leur création jusqu'à leur archivage à long terme en passant par leur curation.

Quels sont les plus grands défis auxquels vous êtes confrontés pour accéder aux outils, aux services et/ou aux ressources d'IRN actuels à votre disposition, et pour les utiliser?

Une des difficultés principale également rencontrée par nos disciplines concerne le financement de personnel technique qualifié dédié au développement d'outils de recherche et d'applications web, à l'entretien et à la maintenance des infrastructures numériques. Les financements du CRSH favorisent l'emploi d'étudiants, ce qui présente un réel bénéfice pour la formation mais ne permet pas de capitaliser sur les développements réalisés et l'expérience acquise. Les financements du FCI sont concentrés sur les trois premières années pour la mise en place de l'infrastructure et sont plutôt conçus pour l'acquisition de matériel, ils ne permettent donc pas réellement de financer de manière durable un personnel technique. En outre, les projets conduits dans le domaine des humanités numériques sont d'ampleur très diverse et n'ont pas toujours la taille critique nécessaire pour disposer d'un personnel technique stable. Lorsque par chance ils en possèdent, l'entretien et la maintenance des serveurs, les mises à jour de sécurité, détournent les maigres ressources techniques disponibles du travail de recherche proprement dit qui porte souvent principalement sur le traitement des données ou les logiciels et les applications de recherche.

Il nous semble donc que les modalités d'accès à l'infrastructure numérique en général et la nature des services actuellement offerts constituent un réel frein au développement de projets en humanités numériques qui présentent une forte dimension technique. Les difficultés de financement en matière de personnel qualifié menacent la pérennité de nombreux projets numériques, y compris ceux qui connaissent un succès important ou un rayonnement international. L'absence de mutualisation entraîne une déperdition des dépenses sans réel bénéfice en matière de services disponibles pour la communauté dans son ensemble.

État futur de l'IRN

Quelle est votre vision d'un écosystème d'IRN canadien qui répondrait à vos besoins en matière de recherche?

La création de la nouvelle organisation d'infrastructure numérique qui doit réunir les services précédemment répartis entre Calcul Canada, CANARIE et RDC/RDC nous paraît offrir une réelle opportunité pour mieux répondre aux besoins des chercheurs dans le domaine des humanités numériques. Si certains projets ont parfois besoin d'importantes ressources en calcul de pointe (CPU/GPU), le plus souvent les recherches conduites dans notre champ impliquent des besoins en matière d'hébergement web d'applications spécialisées, de développements logiciels, et de services dédiés à la création, la gestion, la dissémination et la préservation à long terme de données de recherche. Nos travaux s'inscrivent le plus souvent dans une dynamique circulaire associant étroitement **production, circulation** et **validation** des connaissances à laquelle doit directement répondre l'infrastructure de recherche.

L'émergence des Humanités numériques répond directement à la question fondamentale de la transformation du rapport au savoir provoqué par le numérique non seulement en termes de méthode, mais plus généralement comme fait culturel. Ces changements affectent plus largement toutes nos disciplines et il convient de **devancer et de soutenir la transformation générale des pratiques en sciences humaines et sociales**. Cela implique la **création de services spécifiques ainsi que la mise à**

disposition de personnels dédiés pour accompagner cette transformation en répondant aux besoins de formation.

Nous sommes convaincus qu'un socle de service étendu doit être offert aux chercheurs sans qu'ils aient à dépendre des logiques de demandes de financement de la recherche par projet.

Quels sont les types d'outils, de services et/ou de ressources d'IRN que vous aimeriez utiliser ou que vous envisagez d'utiliser à l'avenir ?

Le développement de services infonuagiques modulaires et spécialisés

Il nous semble crucial que la nouvelle infrastructure propose à tous les membres de la communauté académique de recherche **un accès universel à des services infonuagiques** pour l'hébergement d'applications web. Nous parlons d'accès universel car ces services doivent être facilement accessibles pour tous les étudiants en recherche, les enseignants-chercheurs académiques et les groupes ou équipes de recherche.

Deux types d'hébergement doivent être offerts : d'une part **des bacs-à-sable qui facilitent la modélisation des données et des outils, l'expérimentation, et le prototypage,** d'autre part des services dédiés à la publication web pour l'hébergement d'outils et de logiciels de recherche et pour la dissémination des résultats de recherche. Cela implique une infrastructure robuste et distribuée mais surtout la création d'**une offre modulaire hautement personnalisable** pour répondre le plus parfaitement possibles aux besoins des chercheurs, en particuliers pour les environnements de développement qui ne sont pas toujours offerts dans le secteur commercial.

La nouvelle infrastructure prendrait en charge la gestion de l'identité des utilisateurs et les accès sécurisés au serveur ainsi que le déploiement, la maintenance et l'entretien de ces environnements numériques. Il s'agit de **dégager les chercheurs des tâches d'administration système** qui sont souvent hors de leur compétence ou lorsqu'ils ne disposent pas du personnel dédié afin de pouvoir se concentrer sur le développement de leurs applications de recherche. Les besoins en recherche ne pouvant être standardisés, cela implique de développer **un modèle de plate-forme en tant que service (PaaS) hautement personnalisable** à partir de technologies libres et ouvertes. Il s'agit aussi d'inventer une gestion flexible et décentralisée qui réponde aux besoins spécifiques et évolutifs de la recherche tant sur le plan administratif que financier.

La création de services dédiés au traitement des données textuelles et audiovisuelles

Les chercheurs en sciences humaines travaillent souvent sur des données culturelles qui peuvent aussi bien être de nature textuelle, visuelle (image fixe ou vidéo), sonore, géographique ou tridimensionnelle. La nouvelle infrastructure de recherche doit donc être en mesure de fournir des services dédiés pour faciliter le travail et le traitement de ce type de données. Plusieurs logiciels de recherche ont été développés internationalement pour travailler sur ce genre de données. Nous pensons que ces développements peuvent plus largement intéresser les chercheurs des autres disciplines. Par ailleurs, l'infrastructure doit permettre de nouer d'intenses collaborations avec des institutions culturelles et patrimoniales qui sont actuellement confrontées à ces enjeux avec la numérisation massive du patrimoine culturel.

Plusieurs services qui s'appuient sur des développements de logiciels de recherche en cours peuvent être ainsi imaginés :

- un service de reconnaissance optique de caractères OCR multilingue et adapté au traitement de données patrimoniales
- un service de reconnaissance optique de manuscrits HCR
- un service de transcription automatique
- un service d'annotation de média temporel (vidéo/son)
- un système de description et de catalogage de contenus audiovisuel
- un service pour la diffusion de ressources visuelles et audiovisuelles utilisant le protocole IIIF

- des bancs de travail et des boîtes-à-outils dans une grande diversité de langages informatiques (Julia, Python, XQuery, R, etc.)
- des services pour faciliter l'utilisation d'algorithmes apprenants (machine learning) ou d'apprentissage profond (deep learning) pour enrichir ou exploiter des corpus

La mise en place d'une chaîne de traitement pour la dissémination et l'archivage à long terme des données de recherche

Toutes les disciplines sont actuellement confrontées à des enjeux en matière de gestion des données de recherche, de dissémination et d'archivage à long terme. Dès lors que la nouvelle infrastructure de recherche héberge les applications de traitement de données, elle va être amenée à occuper un rôle crucial dans la création de chaînes de traitement des données pour faciliter la diffusion des données de recherche et leur archivage à long terme, conformément aux politiques des grands conseils.

Nos disciplines disposent d'une riche expérience en matière de description des contenus et de production de riches jeux de métadonnées. Cependant un travail important reste à faire pour mettre en œuvre plus systématiquement des plans de gestion de données et adopter des principes FAIR. Ici, les choix d'infrastructure peuvent jouer un grand rôle dans l'adoption de ces logiques. Dans ce domaine, des collaborations avec d'autres acteurs qui ont développé des solutions mutualisées pour l'archivage à long terme des contenus numériques comme les bibliothèques doivent être envisagées.

Ces solutions doivent valoriser l'emploi de riches standards de métadonnées pour la diffusion et la documentation des jeux de données. En privilégiant des technologies sémantiques, en favorisant l'interconnexion des données et leur dissémination, il s'agit de faire émerger de véritables bases de connaissances conçues de manière à renouveler l'ensemble de l'écosystème de recherche, d'écriture et de diffusion de la recherche en sciences humaines.

La création d'une cellule d'innovation pour faire évoluer les services

Si une infrastructure doit proposer une architecture durable et fiable pour supporter le travail des chercheurs, les services proposés doivent également accompagner l'évolution des pratiques des chercheurs et de leurs besoins. Il nous semble ainsi que la nouvelle organisation doit adopter un fonctionnement fondé sur une logique de recherche et développement. Cela pourrait se matérialiser par la création d'un ou plusieurs laboratoires d'innovation destinés à l'expérimentation de nouveaux services numériques qui seraient proposés par les utilisateurs ou les employés de l'infrastructure.

Quels sont les défis que vous anticipez pendant que vous utilisez les outils, les services et/ou les ressources d'IRN intégrés?

Plusieurs avantages notables peuvent être attendus d'une telle organisation :

- **un accès facilité aux ressources et moins dépendant des cycles de financements de la recherche par projet pour soutenir l'innovation et l'expérimentation**
- **une mutualisation des moyens techniques pour l'hébergement des applications web et des outils**
- **une économie d'échelle qui permettrait aux chercheurs de se consacrer aux seuls travaux qui concernent directement leur recherche**
- **une facilitation du travail sur des données culturelles qui pourrait bénéficier à l'ensemble des autres disciplines et au développement de collaboration avec le secteur patrimonial**
- **une meilleure implémentation des politiques de sciences ouverte et une articulation avec les politiques des bibliothèques ou d'autres acteurs**
- **une capacité d'adaptation de l'infrastructure à l'évolution des besoins des chercheurs**
- **un cercle vertueux où les développements des logiciels de recherche bénéficient aux services offerts à l'ensemble de la communauté à travers l'infrastructure**

Combler l'écart

Quels sont les outils, les services et/ou les ressources que la NOIRN devrait exploiter pour créer votre état futur désiré?

La création des nouveaux services proposés devrait résolument **s'appuyer sur l'utilisation des logiciels libres et ouverts**. Il s'agit de pouvoir construire une infrastructure maîtrisable et personnalisable qui puisse évoluer en fonction des besoins de l'infrastructure. De telles solutions permettraient également de développer des collaborations internationales pour pouvoir bénéficier de services déjà développés avec succès ailleurs. L'ensemble des solutions développées doivent également être documentées et mises à dispositions sous licences libres et ouvertes pour bénéficier plus largement à la communauté internationale et servir de base à une coopération internationale, en particulier avec les pays dont le système éducatif supérieur est moins doté économiquement.

Comment percevez-vous le rôle de la NOIRN dans les initiatives visant à combler les écarts actuels dans l'écosystème national d'IRN?

Plus généralement, la nouvelle infrastructure doit jouer **un rôle dans la promotion de l'utilisation de modèles de métadonnées ouverts et documentés**. L'utilisation de certains services pourrait être conditionnée à l'adoption de politiques claires d'ouverture de données par les chercheurs avec des exceptions pour certaines données de recherche qui présentent des enjeux de confidentialité ou de droits particuliers, conformément aux politiques des grands organismes subventionnaires canadiens. Quoiqu'il en soit, l'ensemble des données de recherches hébergées par la cyberinfrastructure devraient être signalées et exposées sur le web en utilisant des protocoles pour garantir leur découvrabilité.

Comme il est difficile de séparer la production des données de recherche, de leur gestion et de leur archivage à long terme, des solutions doivent être développées pour collaborer avec les acteurs du domaine notamment dans le secteur des bibliothèques (réseau Portage) pour identifier une répartition des rôles adéquate. Des modèles doivent aussi être développés pour favoriser la durabilité des projets en Humanités numériques.

Le financement de l'infrastructure de recherche devrait être directement lié à l'obtention du financement d'une recherche. Lorsqu'un projet de recherche est sélectionné, celui-ci devrait nécessairement pouvoir bénéficier des moyens techniques nécessaires à sa réalisation. Un tel processus implique sans doute une évaluation des aspects techniques de la demande et de sa faisabilité par la cyber-infrastructure de recherche au cours du processus d'évaluation.

Quelles sont vos autres suggestions?

Compte tenu de l'hétérogénéité des situations rencontrées dans le secteur académique au Canada, en particulier dans les communautés autochtones et certaines régions éloignées, nous pensons qu'il y a un réel intérêt à proposer des services au niveau fédéral même lorsque ceux-ci peuvent être offerts dans certains établissements universitaires. Afin de répondre le mieux possible aux besoins locaux des chercheurs et pour nouer des collaborations de proximité, nous pensons qu'**une fédération d'organismes provinciaux** pourrait être très adaptée à la condition que les services développés régionalement soient systématiquement offerts à l'échelle canadienne.

Certaines universités ont financé des services numériques de la recherche ou des Centres dédiés aux humanités numériques. L'organisation aura intérêt à s'appuyer sur leur existence pour soutenir la formation et rejoindre les communautés. Elle pourrait également avoir intérêt à faire émerger **des communautés de pratiques thématiques pour mettre au point des bonnes pratiques et des solutions techniques adaptées** (gestion de données tridimensionnelles, utilisation des ressources informatiques de pointe pour du deep learning, etc.).

Nous sommes convaincus que mieux couvrir les besoins spécifiques des chercheurs en sciences humaines et sociales permettrait de développer un nouveau modèle de services qui bénéficierait plus largement à toutes les disciplines en diversifiant l'offre et l'accessibilité à la nouvelle infrastructure de recherche numérique.