Données de recherche Canada – sous-comité de l'infrastructure

Activités 2 à 5 : Déceler les lacunes de l'infrastructure de données de recherche au Canada et déterminer comment y remédier.

Données de recherche Canada a créé un sous-comité de l'infrastructure (DRC-I) chargé de se pencher sur les enjeux relatifs à l'infrastructure de gestion des données de recherche. Ce DRC-I a établi un plan de travail en mars 2013.

Le présent document constitue le deuxième rapport du sous-comité de l'infrastructure. Un premier rapport intitulé *Activité 1 : Déterminer l'infrastructure pour les données en vigueur au Canada* a été soumis en septembre 2013 (voir la note 2). Le présent document tient lieu de complément à ce premier rapport en synthétisant les constats du sous-comité relativement à l'activité 2 ↖ déceler les lacunes de l'infrastructure de soutien aux données au Canada, à l'activité 3 ↖ déterminer le niveau de sensibilisation à l'importance des données de recherche dans le cadre des efforts visant à aborder les enjeux plus vastes liés à l'infrastructure de recherche, à l'activité 4 ↖ regrouper les capacités, les hypothèses, les leçons apprises et les plans d'avenir en matière de conservation des données, et à l'activité 5 ↖ évaluer les stratégies de rechange pour l'infrastructure de données, comme les nuages commerciaux.

Ce document présente donc les résultats de l'examen effectué par l'équipe du DRC-I au meilleur de sa connaissance. Il ne prétend pas fournir une liste exhaustive d'installations et d'infrastructure, mais les efforts ont été multipliés en vue de traiter du sujet en profondeur.

Les membres de l'équipe du DRC-I qui ont participé à cet exercice sont : Jim Ghadbane (CANARIE), David Moorman (Fondation canadienne pour l'innovation), John Morton (Calcul Canada), Francis Ouellette (Santé, Ontario), Benoît Pirenne (Sciences naturelles, président d'Ocean Networks Canada), David Schade (astronomie, CNRC), Ray Siemens (sciences humaines et sociales, Université de Victoria), Scott Tomlinson (Affaires autochtones et Développement du Nord) et Mark Wolff (CANARIE).

Résumé des constats et recommandations

Les lignes qui suivent présentent un résumé des grands constats et des principales recommandations du sous-comité. Veuillez consulter le texte pour de plus amples renseignements.

- La plupart des utilisateurs du pays sont d'avis que les capacités de réseau sont excellentes.

- Il est relativement facile d'acquérir des capacités de stockage, ce qui n'est pas (encore) le cas des capacités d'archivage.

- Dans bon nombre de cas, les modules logiciels de gestion des données de recherche font défaut. Ces modules, collectivement désignés sous le nom d'« intergiciels », représentent l'élément clé qui permet d'optimiser l'infrastructure brute de stockage et de réseautage pour constituer ce qu'on appelle des « installations de gérance des données ».

- Les installations de gérance des données (IGD) devraient jouer les rôles suivants :
    - recommander, mettre en œuvre et soutenir des normes relatives à la définition des données disciplinaires et mettre en œuvre des politiques sur l'archivage à l'échelle des territoires de compétence (pays);
    - offrir des services de stockage à long terme et d'extraction des données de recherche soutenus par des systèmes logiciels à jour et du personnel qualifié;
    - tenir lieu de réceptacle des produits de données issus des programmes de recherche en cours ou terminés, en conformité avec les exigences des organismes de financement;
    - gérer les transitions technologiques découlant de l'évolution à court terme du matériel informatique, des logiciels et des normes sur la définition des données, comme l'exige la réalité des technologies numériques.

- La numérisation du matériel analogique demeure une tâche inachevée, plus particulièrement dans le domaine des sciences humaines et sociales.

Le concept d'installations de gérance des données pourrait apporter des solutions à des problèmes comme la multiplicité des dépôts de données, les données orphelines, les résultats faisant l'objet de piètres descriptions, les sources impossibles à retracer, ainsi que les supports numériques illisibles et les dossiers inaccessibles, pour n'en nommer que quelques-uns. Il existe déjà au Canada des IGD disciplinaires qui sont exploitées avec succès. Les éventuels modèles de gouvernance, de gestion et de mise en œuvre d'IGD feront l'objet d'un prochain rapport de Données de recherche Canada.

I. Contexte et définitions

Le présent rapport sur les lacunes et les stratégies de rechange tient pour acquis que les cinq préoccupations suivantes ont déjà été résolues par les parties intéressées du domaine de la conservation des données de recherche :

- organiser les données de sorte qu'elles soient utiles au projet et les présenter sous une forme qui se prête bien à la gestion informatisée;
- cerner les données qui devront être conservées;
- cerner les données qui doivent être rendues accessibles (et déterminer à quel moment) (politique sur les données);
- organiser convenablement les données qui ont été ciblées aux fins de conservation;
- convenir de la ou des méthodes à privilégier en vue de rendre les données accessibles.

Le présent document applique les définitions établies dans le Glossaire de Données de recherche Canada (voir la note 4). Les autres termes utilisés dans ce document sont définis dans le tableau ci-dessous.

Terme	Définition dans le contexte de DRC
Intergiciel	D'après Wikipédia : Un intergiciel est un logiciel qui crée un réseau d'échange d'informations entre différentes applications informatiques, quels que soient les systèmes d'exploitation impliqués. Les intergiciels sont communément décrits comme un « ciment logiciel ». Les intergiciels facilitent la tâche des concepteurs de logiciels qui doivent assurer la communication et les entrées-sorties entre les logiciels, de sorte qu'ils puissent se concentrer sur l'objet particulier de leurs applications.
Données	D'après la National Science Foundation des États-Unis : Éléments factuels consignés couramment considérés, au sein de la collectivité scientifique, comme étant nécessaires à la validation des constats de recherche, y compris les données originales et les « métadonnées » (p. ex. protocoles expérimentaux, codes rédigés aux fins d'analyse statistique, etc.). On reconnaît que de nombreuses variables contribuent à définir ce qui constitue des « données » et la gestion de celles-ci, et chaque discipline scientifique a sa propre culture en ce qui concerne les données.

Tableau 1 : Définitions des termes utilisés dans le présent document.

II. Lacunes

Les lignes qui suivent décrivent quelques-unes des lacunes les plus importantes cernées par l'équipe en ce qui a trait à la capacité d'offrir des services de gestion des données de recherche, y compris la conservation de ces données et l'accès à celles-ci. Avant d'entrer dans le vif du sujet toutefois, nous souhaitons rectifier certaines idées fausses couramment répandues à propos de l'archivage des données scientifiques. En effet, contrairement à ce que les fournisseurs de matériel informatique peuvent laisser entendre, l'achat de disques, de bandes magnétiques, d'ordinateurs et de services d'accès réseau ne suffit manifestement pas à assurer la conservation des données. Acheter quelques disques durs pour y garder des données n'équivaut qu'à déplacer une entropie existante. De même, stocker des données dans des nuages commerciaux sans se doter de capacités de gestion, de conservation et d'accès n'a pour effet que de déplacer le siège de l'entropie des données. Comme nous le verrons plus loin dans le présent rapport, les principales lacunes ont davantage trait aux logiciels appropriés et à l'expertise du personnel.

Profusion de systèmes de gestion des données

Tout chercheur menant un projet qui produit un volume quelconque de données prend rapidement conscience de la nécessité de procéder à une certaine forme de gestion de ces données à l'appui des objectifs de la recherche et qu'une forme quelconque d'infrastructure est également nécessaire de manière à relier les données à leurs utilisateurs. Étant donné que les données de recherche proviennent de diverses sources, sont de diverses natures et se présentent sous différents formats, et comme il existe peu de systèmes capables d'y mettre de l'ordre, on constate une pléthore de systèmes de gestion des données, comme en faisait clairement état le rapport précédent du sous-comité (Activité 1). Il semble en effet que la gestion des données n'est qu'une arrière-pensée dans le cadre de nombreux projets de recherche scientifique financés ou que les responsables de ces projets ont jugé que leurs besoins étaient si différents des autres qu'ils ont bricolé leurs propres systèmes maison. De tels systèmes ont donc été conçus pour répondre à des besoins spécifiques et immédiats. Qui plus est, le degré de perfectionnement de ces systèmes est proportionnel au niveau de confiance et de ressources que les chercheurs principaux accordent à leur personnel affecté aux données et aux logiciels.

Ce phénomène entraîne d'ailleurs un certain nombre de problèmes :

- l'existence d'un grand nombre de petits systèmes distincts qui ont chacun leurs propres objectifs, un niveau d'extensibilité mal défini et, vraisemblablement, leurs propres idiosyncrasies et méthodes d'accès;
- le manque d'uniformité de l'accès aux données, ou le caractère privé de cet accès;
- les coûts de mise en œuvre des normes, le cas échéant, ce qui signifie que ces petits systèmes renferment des données qui sont incompatibles avec les autres données provenant de sources similaires et ne peuvent donc pas être utilisées avec ces dernières.

Bon nombre des lacunes décrites ci-dessus résultent du **manque d'installations disciplinaires communes, prêtes à l'emploi et capables de soutenir les projets de recherche de petite et de moyenne envergure et leurs extrants. Si elles existaient dans davantage de disciplines, de telles installations communes constitueraient un moyen fiable et sécuritaire de recueillir, de stocker, de diffuser et de gérer les données**. La situation force donc les chercheurs à concevoir un nouveau système pour chaque nouveau projet dans chaque discipline scientifique.

Le Centre canadien de données astronomiques (CCDA) constitue toutefois un contre-exemple encourageant dans le domaine de l'astronomie au Canada. Au cours des deux dernières décennies, le CCDA a en effet démontré sa capacité à fournir les services clés, ainsi qu'à garantir la fiabilité, la sécurité, l'interopérabilité et l'accessibilité souhaitées, en plus de convaincre un nombre toujours plus grand de petits observatoires et projets de recherche de lui confier leurs données. Il s'agit toutefois d'un cas isolé d'initiative locale qui a connu un succès progressif sans s'imposer à qui que ce soit ni bénéficier d'un soutien approprié des organismes de financement.

Les installations communes telles que le CCDA seront donc désignées sous le terme d'installations de gérance de données dans le reste du présent document.

Enjeux relatifs à l'infrastructure

Afin de s'acquitter de leur mandat de manière fiable, les installations de gérance de données doivent s'appuyer sur une infrastructure. Une telle infrastructure est essentielle à la réalisation de quatre types d'activités, soit *la collecte ou l'acquisition de données*, *le stockage de ces données*, *leur diffusion* et *leur gérance*.

L'acquisition et la diffusion des données requièrent évidemment des **réseaux**. L'acquisition a en outre besoin de **logiciels** s'exprimant dans le langage particulier des producteurs de données (instruments, etc.). Ces outils évoluent au fil du temps et les logiciels, plus particulièrement, doivent être adaptés.

La diffusion de données nécessite non seulement des interfaces utilisateurs et des interfaces de programmation d'applications particulières, mais également des protocoles de transport et des mécanismes d'encodage des données. Ces éléments se fondent généralement sur des **logiciels**. Les interfaces utilisateurs et les interfaces de programmation d'applications ont **des durées de vie utile plus courtes que celles des données qu'elles permettent d'utiliser** (p. ex., les terminaux à base de caractères qui prévalaient jusqu'au début des années 1990 ont tous été supplantés par les technologies Web utilisées aujourd'hui, lesquelles finiront également par être remplacées par des approches qu'on n'arrive pas encore à envisager). Encore une fois, les logiciels habilitants devront évoluer au même rythme que la technologie afin de demeurer pertinents et exploitables.

Le stockage aussi doit nécessairement s'appuyer sur la technologie, mais comme les dernières décennies l'ont démontré, la durée de vie habituelle des **outils de stockage** est généralement inférieure à celle des données qu'ils servent à stocker. Or, continuer à utiliser

des technologies devenues désuètes est le meilleur moyen de faire grimper les coûts de gestion. Il convient donc de faire preuve d'une vision à long terme et de prévoir les ressources nécessaires à l'évolution de ce volet de l'infrastructure.

L'archivage des données est lié non seulement au simple stockage de celles-ci, mais a également trait à des aspects tels que les pratiques et les **protocoles de conservation**, la **planification** de la gestion des données et le **régime réglementaire** qui encadre le versement obligatoire des données, c'est-à-dire la mise en œuvre de **politiques d'archivage des données**.

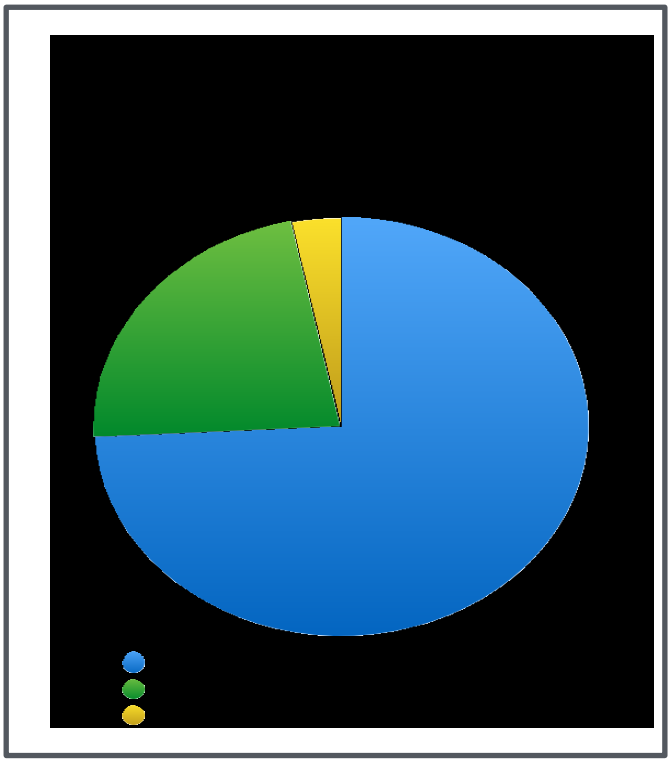
La gérance des données est une activité qui permet de faire en sorte qu'un jeu de données demeure utilisable en s'assurant continuellement de la qualité des données et des métadonnées connexes par le truchement d'agents (**systèmes logiciels** et **spécialistes humains**) dont le rôle consiste à vérifier l'uniformité, l'intégralité, la qualité et la description des données. Le soutien aux utilisateurs constitue également un aspect crucial si l'on souhaite que des non-spécialistes puissent comprendre et utiliser des jeux de données spécifiques.

Indice encourageant donnant à penser qu'on tire effectivement des leçons dans ce domaine, une proposition de nouvelle installation de gestion des données a récemment été soumise relativement à la Station canadienne de recherche dans l'Extrême-Arctique (SCREA) (voir la note 5). La proposition mentionne clairement que les promoteurs sont au fait de tous les enjeux soulevés par Données de recherche Canada. En fait, la plupart des volets de cette proposition sont conformes aux idées actuelles concernant les installations de gestion des données scientifiques. Elle pourrait toutefois s'articuler autour d'une vision à plus long terme plutôt que de se concentrer aussi étroitement sur les modalités de la mise en œuvre de l'infrastructure qui dépendent des technologies modernes.

En somme, la majorité des défis associés aux installations de gérance des données ont trait aux réseaux, à la technologie de stockage, à l'omniprésence des logiciels et à l'expertise relative aux données. Les logiciels étant particulièrement spécialisés, ils sont plus susceptibles de faire l'objet d'évolutions et d'adaptations périodiques, devenant ainsi au fil du temps le poste de dépenses le plus important de l'installation. Les coûts de stockage des données sont liés au volume, mais ils sont vraisemblablement inférieurs aux coûts du personnel et des opérations.

Équipement et ressources humaines

L'expérience acquise dans le domaine des sciences naturelles (voir la note 7) révèle qu'au fil du temps, les coûts de gérance et de gestion des données tendent à se répartir selon les proportions suivantes :



Personnel
Matériel, logiciels et services de réseau
Autres opérations

Au chapitre des particularités de l'infrastructure fondamentale dans le contexte canadien, nous sommes privilégiés de pouvoir compter sur des organisations nationales unifiées qui sont présentes et représentées d'un océan à l'autre. En collaboration avec ses partenaires provinciaux, le Réseau canadien pour l'avancement de la recherche, de l'industrie et de l'enseignement (CANARIE) fournit les capacités de communications numériques qui permettent aux chercheurs d'utiliser des données et d'établir des liens avec leurs pairs. Calcul Canada met des capacités de stockage et de calcul informatique à la disposition de tous les chercheurs du pays en s'appuyant sur des regroupements régionaux existants et en administrant un processus concurrentiel unifié et centralisé aux fins d'accès à toutes les ressources participantes du pays. Le modèle établi par CANARIE répond à la plupart des besoins en matière de transport des données de recherche, mais Calcul Canada n'est pas en mesure d'offrir un tel service à l'heure actuelle, bien que l'organisation ait été envisagée comme éventuelle solution aux besoins en matière de plateforme de stockage à long terme au pays. Les principales raisons en sont que les scientifiques doivent justifier (à nouveau) chaque année leur utilisation des capacités de stockage de Calcul Canada pour lesquelles ils se livrent concurrence et que cette justification doit être étayée par un flot constant de résultats, de publications et d'activités scientifiques. Pour l'heure, Calcul Canada ne représente donc pas une solution de stockage à long terme pour les données qui sont susceptibles de cesser d'être productives au terme d'un projet de recherche. L'approche de Calcul Canada dans ce domaine est compréhensible, l'organisation ne souhaitant pas être responsables de données qui seraient mal gérées ou qui ne seraient pas décrites

conformément à des normes reconnues. Il serait toutefois possible de remédier à ce problème en intégrant au portefeuille de Calcul Canada un mandat explicite en vertu duquel l'organisation offrirait des services de stockage à long terme destinés aux collections de données faisant l'objet d'une gestion, de descriptions et d'un accès adéquats, même si elles ne donnent pas nécessairement lieu à des publications scientifiques. Évidemment, les services nuagiques commerciaux peuvent constituer une solution de rechange à cet égard, mais le caractère flou de l'emplacement des données pose souvent problème, notamment si l'on tient compte des obligations juridiques associées, par exemple, à la confidentialité des données sur la santé, qui doivent ainsi demeurer dans une province donnée ou, à tout le moins, au pays.

Dans le domaine des sciences humaines et sociales, il pourrait être utile d'envisager les solutions préconisées par le rapport sur la cyberinfrastructure *Our Cultural Commonwealth: Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences* (voir la note 8). Les réponses que les collectivités des sciences humaines et sociales ont fournies aux documents de consultation des trois Conseils s'inspirent largement de ce rapport, ainsi que d'un rapport similaire de la Fondation européenne de la science intitulé *Humanities Researchers and Digital Technologies: Building Infrastructures for a New Age* (voir la note 9). Des travaux tels que le projet du Collaboratoire scientifique des écrits du Canada (voir la note 10) financé par la FCI s'en sont également grandement inspirés.

Interopérabilité

Même dans un monde idéal où il n'existe que quelques installations de gérance de données disciplinaires ↖ une installation particulière pour chaque grande discipline scientifique, par exemple, l'astronomie, les sciences sociales, la biologie, l'océanographie, etc. ↖, la nécessité de l'interopérabilité demeure, car les données produites par les différents groupes actifs sur la scène internationale doivent être compatibles les unes avec les autres au chapitre de l'accès, du transport et des formats.

En d'autres mots, on parle d'interopérabilité lorsque deux organisations ou plus conviennent des méthodes à utiliser en vue d'échanger des données. Les parties doivent s'entendre sur deux grands aspects : le conditionnement des données (p. ex. formats et description sémantique du contenu des fichiers) et les protocoles d'accès et de transport (comment rechercher, demander et transporter des jeux de données). Or, il s'avère que l'interopérabilité des données est généralement très spécifique, bon nombre de groupes ayant passé de nombreuses années à **définir des structures appropriées** répondant à une portion suffisamment importante de leurs exigences. Par la suite, il incombe aux différents participants de **mettre en place les protocoles et les formats** établis, habituellement par le truchement de logiciels qui traduisent les données de leur format d'origine à la norme établie d'un commun accord. Ces normes évoluent généralement en fonction des nouveaux besoins et des changements technologiques.

Rareté des données numériques

Dans le domaine des sciences humaines, la rareté des données conservées et décrites adéquatement constitue un obstacle, plus particulièrement la rareté des données issues des sciences humaines qui peuvent être retracées par des moyens informatiques ou fusionnées d'une façon que bon nombre de disciplines scientifiques fondamentales et appliquées tiennent aujourd'hui pour acquise. On constate donc des « **vides » de données numériques sur les sciences humaines**, ce qui devrait peut-être faire l'objet d'une enquête. Le Réseau canadien de documentation pour la recherche (RCDR), les projets Synergies et TAPoR financés par la FCI, le programme Les textes, les documents visuels, le son et la technologie du CRSH, la Consultation sur les archives nationales de données et d'autres initiatives ont été très utiles à cet égard, mais des investissements (intellectuels et

autres) seraient indiqués pour faire en sorte que les données sur les sciences humaines atteignent le même niveau que celles des données produites dans des disciplines telles que l'astronomie ou l'océanologie, par exemple (voir également la note 1).

Dans le même ordre d'idées, le tollé récemment soulevé par le démantèlement des bibliothèques scientifiques du MPO (voir, par exemple, l'article publié dans *The Tyee* mentionné à la note 3) a mis en lumière les préoccupations relatives à l'exhaustivité et à la qualité du processus de numérisation préalable à la destruction des vieux dossiers analogiques.

III. Résumé

Bien que le Canada dispose aujourd'hui d'une vaste infrastructure de réseautage numérique et d'importantes capacités de stockage, sa plus grande lacune semble être le manque d'« intergiciels » (ou modules logiciels de gestion des données de recherche). Le terme intergiciel est une appellation fourre-tout désignant des types de logiciels qui remplissent une très large gamme de fonctions, de la gestion du flux des données aux systèmes de visualisation. Le manque d'intergiciels découle du peu de soutien offert aux ingénieurs en logiciels et aux spécialistes des données compétents dont la tâche consiste à mettre en œuvre les formats de données spécifiques et les normes de diffusion, à adapter les fonctions d'acquisition, de stockage, de recherche et de diffusion des données, ainsi qu'à satisfaire aux exigences en matière de qualité des données et de soutien aux utilisateurs. À cet égard, il convient de souligner le travail de CANARIE, qui a cerné ces enjeux il y a déjà plusieurs années et qui a financé des groupes de chercheurs et d'ingénieurs en vue d'investir dans les intergiciels (voir la note 6).

Il semble être nécessaire d'accorder davantage d'attention à la gestion des activités de numérisation des dossiers de données relatives aux sciences humaines et sociales afin d'en garantir l'exhaustivité et la qualité supérieure. À défaut d'autre chose, le degré de complétude des données devrait être qualifié ou, mieux, quantifié.

Au chapitre des capacités de stockage, l'élargissement du mandat de Calcul Canada devrait prévoir le soutien à la conservation à long terme de données numériques adéquatement décrites au moyen d'installations financées de gérance de données, même en l'absence de recherche continue.

Les options relatives à la mise en œuvre des installations de gérance de données feront l'objet d'un prochain rapport.

Annexe I : Notes et documents de référence

1. History of Sciences Society (2013). Report on Data Management and Data-Management Plans for the History of Science Society Committee on Research and the Profession (6 septembre 2013). *The History of Science Society Newsletter*. Volume 42, numéro 4, octobre 2013.

2. Données de recherche Canada. *Rapport du sous-comité de l'infrastructure sur l'activité 1 : Déterminer l'infrastructure pour les données en vigueur au Canada*.

3. The Tyee (2013). *What's Driving Chaotic Dismantling of Canada's Science Libraries?* http://thetyee.ca/News/2013/12/23/Canadian-Science-Libraries/.

4. Données de recherche Canada (2013). *Glossaire*. Ébauche 3B, novembre 2013.

5. Réseau canadien de données polaires (2013). *Data and Information Management for Arctic Science and Technology: A Proposed Approach for the Canadian High Arctic Research Station*. Juillet 2013.

6. Les programmes de plateformes sur réseau et d'interfaces pour plateformes de recherche de CANARIE soutiennent les travaux d'élaboration d'intergiciels, tel qu'il est mentionné dans la section Résumé. Consulter le http://www.canarie.ca/fr/reseau/.

7. Les données comparatives sur les coûts associés au matériel de l'infrastructure numérique et au personnel d'exploitation proviennent d'Ocean Networks Canada. Le ratio est similaire dans le cas du Centre canadien de données astronomiques. Il serait intéressant de comparer ces données avec celles d'autres disciplines telles que la biologie computationnelle.

8. American Council of Learned Societies. *Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*.

9. Fondation européenne de la science. *Humanities Researchers and Digital Technologies: Building Infrastructures for a New Age*.

10. Le Collaboratoire scientifique des écrits du Canada (CSEC, http://www.cwrc.ca/fr/) est un projet financé par la Fondation canadienne pour l'innovation, entre autres subventionnaires.