Une plateforme de recherche et d'expérimentation pour l'édition ouverte

Colloque réalisé dans le cadre du Congrès de l'Acfas 2015
à l'Université du Québec à Rimouski (UQAR).
Lundi 25 mai 2015 de 09h15 à 17h.

érudit

Chaire de recherche du Canada sur les transformations de la communication savante

Université de Montréal

83e CONGRÈS DE L'ACFAS DU 25 AU 29 MAI 2015

SORTIR DES SENTIERS BATTUS

Présentation du colloque

Alors que les capacités de stockage et de calcul s'accroissent de façon exponentielle et que les outils de fouille, d'analyse et de visualisation des collections numériques se multiplient, les propriétés du corpus d'Érudit (erudit.org) offrent des perspectives de recherche exceptionnelle en bibliométrie, en linguistique informatique, en web sémantique, ainsi qu'en histoire et en sociologie des sciences.

L'exploration automatisée d'un corpus numérique enrichi comme celui d'Érudit, regroupant une quantité croissante d'archives et de numéros courants de revues scientifiques et culturelles, mais aussi de thèses et de documents et données divers, permet en effet d'extraire, de compiler et d'analyser quantités de données autrement dispersées sur de nombreuses plateformes ou dont l'accès était limité au format papier.

Mais qu'en est-il concrètement de ces nouvelles possibilités de recherche? Il s'agira ainsi de se demander, à partir de l'exemple d'Érudit, quelles questions inédites le traitement automatisé d'un corpus numérique permet de poser et comment ce corpus devrait idéalement évoluer (collections, structuration, sémantisation) afin de répondre aux besoins émergents des chercheurs; d'explorer, en somme, les possibilités de recherche présentes et futures que recèle une plateforme comme Érudit.

Informations

Colloque 427 - Une plateforme de recherche et d'expérimentation pour l'édition ouverte
Organisé par Vincent LARIVIÈRE (Professeur adjoint à l'École de bibliothéconomie et des sciences de l'information, Université de Montréal, Titulaire de la Chaire de recherche du Canada sur les transformations de la communication savante) dans le cadre du congrès de l'Acfas 2015.
Date: Lundi 25 mai de 09h15 à 17h00.
Lieu: Bâtiment – Local: IMQ – A356, Université du Québec à Rimouski (UQAR), Québec, Canada
Mot-clic: #AcfasC427

Le programme est également consultable en ligne depuis le site internet de l'Acfas à l'adresse suivante :
http://www.acfas.ca/evenements/congres/programme/83/400/427/c

Programme de la journée

9h15-9h30 Mot de bienvenue **Vincent LARIVIÈRE**	**12h00-13h30: Dîner**
9h30 - 10h30: Plateformes	**13 h30 - 14h30: Traitement et outil 1**
9h30 Les projets de données liées chez Canadiana.org - Daniel VELARDE (Canadiana), Julienne PASCOE	**13h30** Valorisation du fonds documentaire numérique pour la recherche - Marc BERTIN (UQAM - Université du Québec à Montréal)
9h50 Huma-Num, une infrastructure de recherche au service des données de la recherche et des savoirs en sciences humaines et sociales - Stéphane POUYLLAU (CNRS)	**13h50** Contraintes et atouts du corpus Érudit pour le traitement automatique de la langue - Lyne DA SYLVA (Université de Montréal)
10h10 Calcul Québec : une plateforme pour vos projets en humanités numériques - Suzanne TALON (Université de Montréal), Félix-Antoine FORTIN	**14h10** Fouille de textes et cartographie thématique des corpus numériques - Dominic FOREST (Université de Montréal), Marcela BAIOCCHI (Université de Montréal)
10h30-10h40: Pause	**14h30-14h45: Pause**
10 h 40 - 12 h 00: Usages et pratiques	**14h45 - 15h55: Traitement et outil 2**
10h40 L'archive et l'analyse réseautique des revues d'idées au Québec: entre le bricolage et les promesses du numérique Jean-Pierre COUTURE (Université d'Ottawa)	**14h45** Extraction et structuration de connaissances pour une plateforme interactive dédiée à Érudit: état de l'art et points de réflexion - Philippe LANGLAIS
11h00 Altmetrics: opportunités et défis associés à l'usage des médias sociaux dans la communication savante - Stefanie HAUSTEIN (Université de Montréal)	**15h05** Utilisation des citations pour le résumé automatique de la contribution des articles scientifiques - Bruno MALENFANT (Université de Montréal)
11h20 Comment mutualiser les données documentaires et les potentialités d'exploration sans concentrer le pouvoir de décision sur les systèmes et architextes ? - Jérôme VALLUY (Université Panthéon-Sorbonne (Paris 1) / Université de Technologie de Compiègne, France)	**15h25** La recherche sémantique dans le domaine juridique: résultats et réflexions - Martin BOUCHER (Université de Montréal)
11h40 Fouille textuelle de revues intellectuelles québécoises - Iana ATANASSOVA (Université de Franche-Comté)	**15h55-16h05: Mot de clôture - Tanja NIEMANN (Université de Montréal)**
	16h05 - 17h: Cocktail (Local : IMQ – A235)

Résumé des présentations

9h15 - 9h30 : Mot de bienvenue

Vincent LARIVIÈRE (Université de Montréal)

9 h 30 - 10 h 30 Plateformes

9h30 - Les projets de données liées chez Canadiana.org

Daniel VELARDE (Canadiana), Julienne PASCOE

En tant qu'organisation dirigée par ses membres, dont le but est de fournir un large accès au patrimoine documentaire canadien, Canadiana.org s'est donné pour mission de relier les ressources patrimoniales culturelles du Canada au monde en faisant appel aux principes des données liées. Cette communication vise à analyser les défis et les possibilités qu'offrent les données liées pour la description et l'exploration des ressources patrimoniales du Canada. Elle comprendra un résumé des principes du Web sémantique tels qu'ils s'appliquent au patrimoine culturel, à la vision et à la stratégie de Canadiana concernant les données liées, et aux approches expérimentales sur le développement et l'enrichissement des métadonnées en utilisant le modèle et les technologies des données liées.

Les données liées sont un ensemble de meilleurs pratiques et de spécifications techniques pour la publication et la liaison de données structurées sur le Web. Il s'agit de l'une des plus récentes initiatives se rapportant aux données à être développées par Canadiana.org. Les principes du Web sémantique permettent aux institutions de la mémoire de relier des collections hétérogènes dans les divers domaines des bibliothèques, archives et musées, en exploitant les capacités de distribution sur le Web pour faciliter la description et l'accès au patrimoine documentaire à travers des silos institutionnels.

9h50 - Huma-Num, une infrastructure de recherche au service des données de la recherche et des savoirs en sciences humaines et sociales

Stéphane POUYLLAU (CNRS)

Huma-Num est une très grande infrastructure de recherche visant à faciliter le tournant numérique de la recherche en sciences humaines et sociales (SHS). Elle est bâtie sur une organisation originale consistant à mettre en oeuvre 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 des humanités numériques. Elle favorise la coordination de la production raisonnée et collective de corpus de données ouvertes et interopérables.
Elle développe pour cela un dispositif technologique unique permettant le traitement, la conservation, l'accès et l'interopérabilité des données de la recherche des SHS. Ce dispositif est composé d'unegrillede services dédiés pour les données de la recherche, d'une plateforme d'accès unifié (ISIDORE) et d'une procédure d'archivage à long terme. Elle propose en outre desguides debonnes pratiquestechnologiques généralistes à destination des chercheurs. Huma-Num coordonne la participation française àDARIAH(Digital Research Infrastructure for the Arts and Humanities), dont l'objectif est de développer l'échange de données, d'expertises et de services au niveau européen.

Huma-Num est portée par le CNRS, l'Université Aix Marseille et le Campus Condorcet. La communication présentera la stratégie et la position d'Huma-num dans le cycle de vie des données numériques en SHS.

10h10 - Calcul Québec : une plateforme pour vos projets en humanités numériques

Suzanne TALON (Université de Montréal), Félix-Antoine FORTIN

Calcul Québec (CQ) est un regroupement d'universités québécoises réunies autour du calcul informatique de pointe (CIP). Il a pour mission de fournir au milieu de la recherche des infrastructures matérielles et logicielles de pointe en CIP ainsi que des services d'expert-conseil, afin de contribuer à l'avancement des connaissances dans toutes les branches du savoir et à la formation de personnel hautement qualifié (PHQ) en CIP, capable d'exploiter efficacement le parallélisme des systèmes informatiques modernes. Ces dernières années, les humanités numériques ont connu une forte croissance. Les analystes en CIP de Calcul Québec travaillent en collaboration avec divers groupes de recherche afin de les aider à mieux tirer profit des ressources matérielles qui sont à leur disposition.

Dans cette présentation, nous discuterons de l'offre de services adaptés aux humanités numérqieus ainsi que quelques exemples de travaux réalisés auprès de groupes de recherche.

10h30 - 10h40: Pause

10 h 40 - 12 h 00: Usages et pratiques

10h40 L'archive et l'analyse réseautique des revues d'idées au Québec: entre le bricolage et les promesses du numérique

Jean-Pierre COUTURE (Université d'Ottawa)

L'Action nationale, Relations, Liberté, Parti pris, Chroniques, Possibles, Temps fou, À bâbord. Depuis un siècle, les revues d'idées servent de véhicule pour l'animation du débat idéologique et politique au Québec. Les nombreux intellectuels qui les investissent choisissent cet espace de ommunication intermédiaire entre le journalisme et l'écriture savante pour diffuser leur pensée au sein d'un cercle plus restreint que celui des médias de masse, mais dans un format qui se veut accessible au public lettré et cultivé. Cette hybridité des publics et des registres discursifs (en sus de la courte existence de la plupart des titres) fait en sorte que les revues d'idées ne sont pas toutes indexées par la banque d'Érudit. Il s'agit pourtant d'un terreau éminemment fertile pour faire valoir la portée heuristique de la bibliométrie et de l'analyse de

réseau dans la documentation des interactions entre le champ politique et parlementaire et le champ intellectuel et idéologique. Or en l'absence d'outil adéquat et standardisé, la recherche contemporaine doit bricoler ses propres index à partir du fastidieux dépouillement des revues du corpus (près de 200 revues d'idées distinctes ont paru depuis 1917 au Québec). La communication fera état de résultats récents quant au potentiel de ces analyses de réseau et souhaitera contribuer à la réflexion commune quant à la d'Érudit pour la sociologie des idées.

11h00 - Altmetrics: opportunités et défis associés à l'usage des médias sociaux dans la communication savante

Stefanie HAUSTEIN (Université de Montréal)

350 ans après sa création, la revue savante demeure le principal moyen de diffusion des connaissances savantes, et les citations reçues par les articles constituent la mesure principale de leur impact scientifique. Les médias sociaux et leur introduction dans un contexte académique ont généré de nouvelles opportunités pour capturer l'impact sur un public potentiellement plus large—pas simplement les auteurs qui citent—et plus rapide, compte tenu de la vitesse avec laquelle l'activité dans les médias sociaux peut être mesurée. Le nombre de tweets, de publications Facebook, de lecteurs sur Mendeley, d'évaluations d'experts sur F1000, et de vues sur Slideshare sont des exemples d'indicateurs considérés comme des «altmetrics». De nombreuses revues fournissent également les «altmetrics» associées à chacun de leurs articles, certains chercheurs les présentent sur leurs CVs, et certains organismes subventionnaires commencent à envisager leur utilisation. Même s'il est devenu évident que ces nouvelles mesures sont très hétérogènes et ne peuvent remplacer les citations, on sait encore peu de choses sur leur signification et le type d'impact qu'ils reflètent. Cette communication fera un tour d'horizon des opportunités et des défis associés à l'utilisation de médias sociaux dans la communication savante.

11h20 - Comment mutualiser les données documentaires et les potentialités d'exploration sans concentrer le pouvoir de décision sur les systèmes et architextes ?

Jérôme VALLUY (Université Panthéon-Sorbonne (Paris 1) / Université de Technologie de Compiègne , France)

L'édition numérique de textes scientifiques et pédagogiques est écartelée entre un mouvement de dilution dans l'océan des autoéditions numériques en libre accès et, en sens inverse, un mouvement de concentration politique ou commerciale de la décision éditoriale. Les plateformes francophones s'inscrivent dans la deuxième tendance par emprise étatique française (Hal, Persee, OpenEdition) ou gestion commerciale (Cairn). Par ailleurs, trop étroitement académiques ou commercialement contraintes, elles ne tirent pas parti des gigantesques ressources (écrits, sons, images) de l'océan numérique en libre accès et ne parviennent pas à suivre l'évolution vers les ouvrages numériques dynamiques («enrichis», «augmentés») faisant appel à ces ressources. Dans ce contexte Erudit pourrait accroître son volume de publications et le potentiel collectif d'exploration et de réutilisation des contenus, en créant un dispositif ouvert intégrant, par duplication, les bases de données de sites sous SPIP ou Wordpress, en libre accès intégral, de collectifs d'auteurs (revues, laboratoires, associations, réseaux, équipes...) demeurant autonomes tant pour leurs contenus que pour les modalités d'affichages sur leurs sites. Si le système est accessible aux collectifs sans contrôle a priori (mais avec contrôle a posteriori des contenus publiés au regard des lois canadiennes), il offrira une alternative précieuse à toute la francophonie.

11h40 - Fouille textuelle de revues intellectuelles québécoises

Iana ATANASSOVA (Université de Franche-Comté)

Nous présentons une étude logométrique de revues intellectuelles québécoises du début du XXe siècle, avec pour objectif d'identifier des tendances et des phénomènes dans le discours intellectuel et de fournir des outils de fouille textuelle permettant d'observer les concepts et leurs usages dans les textes afin d'en dégager des tendances. Le développement des nouvelles technologies et la numérisation des revues intellectuelles rendent possibles aujourd'hui des analyses automatisées de textes par des méthodes issues du traitement automatique des langues. À travers l'étude des cooccurrences, des fréquences des termes et du lexique, sur une période donnée, nous pouvons révéler l'évolution des thématiques traitées, identifier l'émergence de nouveaux concepts dans les textes. De plus, la mise en relation entre les sujets traités dans les revues intellectuelles québécoises et le contexte historique apporte un éclairage lors d'études sociologiques. Les politiques d'accès de la plateforme Érudit et de la Bibliothèque et Archives Nationales du Québec ont permis la création d'un corpus textuel numérise, constitué d'une part de « La Revue d'histoire de l'Amérique française » et de la revue « Recherches sociographiques », et d'autre part, des revues : « L'action française », « L'action canadienne-française » et « L'action nationale ». Pour ce corpus de cinq revues, nous avons traité l'intégralité des publications sur la période de 1917 à 2005.

12h00 - 13h30: Dîner

13 h 30 - 14 h 30: Traitement et outils (partie 1)

13h30 - Valorisation du fonds documentaire numérique pour la recherche

Marc BERTIN UQAM (Université du Québec à Montréal)

Si la numérisation des fonds documentaires et l'accès au texte intégral offrent de nombreuses possibilités, ces dernières reposent avant tout sur le traitement de grands ensembles de textuelle. À la frontière de l'informatique et de la linguistique, le champ du traitement automatique des langues propose des méthodes et des outils pour l'exploitation de l'information textuelle en la segmentant, la classant, la catégorisant ou même en l'annotant sémantiquement et automatiquement. À partir des résultats obtenus, nous disposons de nouvelles données apportant un éclairage nouveau sur le fonds avec une information textuelle organiser, condition nécessaire à la production de nouvelles connaissances. Les métadonnées se voient enrichies, les textes annotés, les formes désambiguïsées et les index optimisés : l'accès au document se voit ainsi revisité.

Nous donnerons des exemples concrets autour du traitement de corpus et de solution informatisés autour des index et des bibliographies. Nous présenterons des outils de lexicométrie comme Iramuteq ainsi que des solutions de production d'interface web dédiée au monde de la recherche comme RShiny. Nous conclurons par une réflexion autour des services qui peuvent être proposés au monde de la recherche à partir de ces nouvelles pratiques.

13h50 - Contraintes et atouts du corpus Érudit pour le traitement automatique de la langue

Lyne DA SYLVA (Université de Montréal)

La présentation procédera à un examen systématique des caractéristiques du corpus d'Érudit d'un point de vue de traitement automatique de la langue (TAL). Celles-ci incluent les suivantes :

(1) ses caractéristiques informatiques, dont principalement le format des documents, la présence de métadonnées explicites et l'existence de balisage sémantique étendu; (2) les caractéristiques linguistiques du corpus, notamment le degré de multilinguisme des textes, le vocabulaire utilisé, étudié à la fois d'un point de vue terminologique et de sémantique lexicale, ainsi que quelques éléments de linguistique textuelle telle qu'observée dans un échantillon du corpus; (3) un certain nombre de critères pragmatiques, incluant les distinctions entre revues scientifiques et culturelles ainsi que les propriétés de cette bibliothèque numérique comparée aux corpus normalement utilisés en TAL.

Ceci sera suivi d'une analyse, d'une part, des atouts que ces caractéristiques identifiées présentent, et d'autre part des contraintes qu'elles imposent au traitement.

14h10 - Fouille de textes et cartographie thématique des corpus numériques

Dominic FOREST (Université de Montréal), Marcela BAIOCCHI (Université de Montréal)

On observe depuis une dizaine d'années une hausse du nombre d'initiatives visant à numériser et à diffuser le patrimoine informationnel des différentes branches du savoir. Dans certains domaines, les conséquences des initiatives de numérisation ont des répercussions sur le développement d'applications visant à assister la recherche, l'analyse, la structuration et la gestion des informations. Lors de cette communication, nous exposerons comment certaines techniques de fouille de textes peuvent être exploitées afin d'assister l'extraction et l'organisation et la visualisation d'informations présentes dans des corpus de documents scientifiques en sciences humaines. Les données que nous avons traitées dans le cadre de nos recherches sont issues de la plate-forme Érudit. La démarche que nous avons menée repose sur une méthodologie inspirée de travaux dans le domaine de la fouille de données. Cette démarche est composé de 4 principales étapes : 1. Le pré-traitement, 2. La transformation numérique, 3. L'application des algorithmes de fouille et l'extraction des termes caractéristiques et 4. L'évaluation et la visualisation. Dans cette démarche, nous avons principalement mis à contribution les propriétés structurantes des algorithmes de classification que nous avons couplées à des modalités devisualisation de l'information qui permettent de présenterde manière conviviale les résultats obtenus.

14h30 - 14h45: Pause

14 h 45 - 15 h 55: Traitement et outil 2

14h45 Extraction et structuration de connaissances pour une plateforme interactive dédiée à Érudit: état de l'art et points de réflexion

Philippe LANGLAIS

L'extraction automatique de connaissances à partir de données textuelles en partie structurées trouve un nombre croissant d'applications comme l'aide interactive au furetage de grande collections de documents, le recensement d'informations implicites dans les textes ou encore la réponse à des questions complexes. Dans cette présentation je compte décrire des chaînes de traitement développées par la communauté du traitement des langues et proposer des scénarios possibles de leur intégration dans une plateforme de furetage interactive dédiée à Érudit. La première étape de cette réalisation consiste à construire une base de connaissances sous la forme d'une (large) collection de triplets <sujet,relation,prédicat> à la façon des triplets RDF qui constituent le socle du Web sémantique. J'illustrerai les sorties des extracteurs de triplets actuels sur quelques documents d'Érudit. La seconde étape consiste à structurer ces triplets. Je dresserai une cartographie des principaux niveaux de structuration que l'on peut obtenir automatiquement. La troisième et dernière étape consiste à mettre à l'usage une telle base de connaissances. Je décrirai à cet effet 2 applications qui selon moi auraient un intérêt dans une plateforme dédiée à Érudit: le furetage interactif d'une large collection de documents et la réponse automatique à des questions complexes.

15h05 - Utilisation des citations pour le résumé automatique de la contribution des articles scientifiques

Bruno MALENFANT Université de Montréal

Une des tâches d'un chercheur est la lecture d'articles scientifiques, que ce soit pour les comparer, pour identifier de nouveaux problèmes, pour situer son travail dans la littérature courante ou pour définir des propositions de recherche. Nous voulons construire un résumé d'un article scientifique qui permettrait à un chercheur de décider rapidement s'il doit lire un article ou non. À l'intérieur d'une citation, il y a une description des liens entre plusieurs articles. Ces articles sont comparés, commentés et combinés. Cette information n'était pas disponible lors de la rédaction de l'article, cela lui ajoute donc un niveau d'interprétation et un indice sur son apport à la communauté scientifique, les citations reflétant l'opinion de la communauté scientifique. Nous proposons un système de résumé à base de citances, terme proposé par Nakov et al (2004) désignant l'ensemble des phrases entourant une citation expliquant ce qu'un autre article a réalisé dans un domaine lié à l'article citant. Notre système identifie le rôle attribuable à une citance : hypothèse, discussion, méthode, résultat ou implication. Nous travaillons aussi sur le balisage automatique d'un article et la construction d'une base de données RDF contenant la méta-information des articles.

15h25 - La recherche sémantique dans le domaine juridique : résultats et réflexions

Martin BOUCHER (Université de Montréal)

Depuis quelques années, les techniques de traitement automatique de la langue naturelle (TALN) ont été appliquées à la recherche d'information dans le but, entre autres, d'améliorer la précision de la recherche par une meilleure compréhension des termes de requête dans leur contexte d'utilisation, de proposer des suggestions de requêtes apparentées ou encore de faire ressortir des documents dont le contenu partage des similarités conceptuelles avec la requête. De telles techniques de recherche dites « sémantiques » sont à la base de produits innovateurs développés par une équipe de chercheurs et de développeurs de la compagnie québécoise KeaText.

Dans cette communication, nous ferons d'abord un bref survol des techniques de recherche sémantique utilisées par l'équipe de KeaText. Par la suite, nous présenterons le résultat de la mise oeuvre de ces techniques, notamment dans le domaine juridique, ouvrant ainsi la voie à de nouvelles possibilités pour le repérage et la recherche d'information basés sur le contexte sémantique. Nous discuterons des avantages et des limites associées à ce type d'approche et enfin, nous formulerons des réflexions concernant l'application de telles techniques de recherche sémantique sur des corpus documentaires savants comme Érudit.

15h55 -16h05: Mot de clôture

Tanja NIEMANN (Université de Montréal)

16h05 - 17h: Cocktail

Bâtiment – Local: IMQ – A235 Salon des anciens