10 octobre 2018

Journée d’étude
CO.SHS 2018

Compte-rendu

CO.SHS

Journée d’étude **CO.SHS**

Le mercredi 10 octobre 2018, Université de Montréal

La deuxième édition de la journée d’étude CO.SHS – qui a marqué la mi-parcours du projet – s’est tenue le mercredi 10 octobre 2018 à l’Université de Montréal. Elle a permis aux différentes équipes qui participent aux développements technologiques de la **cyberinfrastructure ouverte pour les sciences humaines et sociales** de présenter leurs avancées au cours de la dernière année, les défis et problèmes rencontrés, ainsi que leurs réussites. Plus généralement, cette journée avait aussi pour objectif de poursuivre la réflexion entourant l’intégration éventuelle de certains projets à la plateforme Érudit.

→ Pour en savoir plus sur les projets de recherche et les outils développés, les équipes et partenaires participants, les corpus mis à disposition et les dernières nouvelles : **co-shs.ca**

AXE PRODUIRE

Nicolas Sauret et Marcello Vitali-Rosati, Chaire de recherche du Canada sur les écritures numériques – **Stylo : éditeur sémantique pour les sciences humaines**

→ Voir la présentation

→ Tester le prototype Stylo

→ Consulter la documentation de Stylo

→ Voir le code source de Stylo

Au cours de la dernière année, un prototype fonctionnel de l'outil Stylo a été mis en ligne. Stylo articule dans une même interface l'édition du corps de texte, des métadonnées et de la bibliographie et combine ainsi un éditeur de texte avec balisage interne, un éditeur de métadonnées et un outil de gestion de la bibliographie. Il comprend également un module de conversion en différents formats et un module de versionnage. Il permet donc de générer des publications multiples à partir d'un seul document.

Stylo est désormais pleinement utilisé par la revue *Sens public*, prenant le relais de la chaîne de publication déjà utilisée par les éditeurs. Les auteurs sont progressivement invités à utiliser l'outil en phase d'annotation (aperçu des épreuves) et en phase de rédaction. Les membres de la Chaire de recherche du Canada sur les écritures numériques poursuivront les expérimentations avec six revues pilotes dans le cadre du projet Revue 2.0 (revue20.org) et souhaitent convaincre davantage d'équipes éditoriales de tenter l'expérience pour adapter et améliorer leur prototype.

Robin Varenas et Sylvain Aubé, Laboratoire NT2 – **SARQC : structure d'accompagnement des revues québécoises et canadiennes**

→ Voir la présentation

→ Voir le code source du module Views OAI-PMH
drupal.org/project/views_oai_pmh

→ Lire l'entrevue réalisée en août dernier

Les efforts du Laboratoire NT2 ont en grande partie été consacrés cette année à l'interopérabilité avec la plateforme Érudit à partir du cas de la revue *Captures* – l'une des quatre revues savantes et culturelles qui font partie du projet pilote, aux côtés de *Voix et images*, *Lettres québécoises* et *Estuaire*. La revue *Captures* possède déjà son propre site web (revuecaptures.org) et le travail réalisé jusqu'ici visait l'exportation des données vers Érudit.

Pour ce faire, les membres de l'équipe ont œuvré à la structuration des contenus Drupal et au développement d'un module nommé « Views OAI-PMH », qui permet de convertir un document en format HTML vers du XML (notamment vers le schéma EruditArticle, propre à Érudit) et qui génère ensuite une sortie OAI-PMH pour la diffusion des données de la revue. Ils sont maintenant en phase de validation du résultat de la sortie OAI-PMH avec l'équipe d'Érudit et poursuivent le développement de la distribution Drupal interopérable avec la revue *Lettres québécoises*.

Davin Baragiotta, Érudit – **OJS sur Érudit : efforts dans le cadre de CO.SHS**

→ Voir la présentation

→ Voir les dépôts de code source d'Érudit github.com/erudit et gitlab.erudit.org

→ Voir la documentation du schéma EruditArticle

Plusieurs initiatives techniques liées à la production de revues XML ont été menées par Érudit au cours de la dernière année, en collaboration notamment avec la Chaire de recherche du Canada sur les écritures numériques, le Laboratoire NT2 et le Public Knowledge Project (PKP). L'objectif est d'améliorer les chaînes de production d'Érudit, en favorisant le traitement en amont et l'automatisation d'une partie du traitement, tout en assurant un strict contrôle de la qualité.

À ce jour, 7 revues OJS2 ont été ajoutées sur Érudit en traitement minimal, 15 revues OJS3 sont en attente de traitement minimal, et le traitement complet de revues OJS3 est en cours de développement. L'extraction des références bibliographiques, la structuration semi-automatique du corps du texte et la migration du schéma XML EruditArticle vers JATS (Erudit Publishing Schema) demeurent les principaux défis.

Juan Pablo Alperin, Public Knowledge Project – **PKP et XML**

→ Voir la présentation

→ En savoir plus sur le Consortium Substance

→ Télécharger Texture

L'objectif de PKP est d'amener l'ensemble des 10 000 revues utilisant OJS à produire un balisage XML complet, et ce, à travers le monde, dans toutes les disciplines. Cela suppose une double tâche : automatiser la création en XML, ce qui est très complexe, et intégrer la structure XML dans le processus éditorial des revues, ce qui est plus simple. Leurs efforts actuels se concentrent sur la réalisation de cette dernière tâche.

Le Consortium Substance, incluant PKP, CoKo, Érudit, SciELO et eLife, est une initiative communautaire en soutien au développement et à l'entretien continus d'outils libres pour l'intégration XML. Ils sont notamment en train de développer Texture, un éditeur de textes en ligne qui produit des documents JATS-XML. La première version a été lancée lors de la conférence SciELO 20 Years en septembre dernier. Les prochaines étapes seront de faire remonter les documents XML de plus en plus en amont dans le processus éditorial sur OJS3 : depuis la fin de la production vers le début.

Toutefois, le problème de l'automatisation de la création XML demeure. Open Typesetting Stack reste la meilleure option pour le balisage à ce jour, mais le service est difficile à maintenir et à développer. Le Semantic Extraction Working Group explorera aussi d'autres solutions, telles que Grobid.

AXE
DÉCOUVRIR

Fabrizio Gotti, Philippe Langlais et Vincent Letard, RALI (Recherche Appliquée en Linguistique Informatique) – **Extraction d'information ouverte**

→ Voir la présentation

→ En savoir plus sur l'extraction de triplets d'information : arXiv:1809.08962

→ Lire l'entrevue réalisée en septembre dernier

Le travail réalisé par le RALI au cours de la dernière année a surtout été axé sur l'amélioration du processus d'extraction des concepts et des relations verbales. L'un des plus grands défis de l'OIE (*open information extraction*) réside en effet dans l'extraction des concepts exprimés par des paraphrases ou des indices non verbaux, qui sont très ambigus pour une machine. Afin d'évaluer l'ampleur de ce problème, les membres de l'équipe ont réalisé un banc d'essai en annotant manuellement des triplets d'information contenus dans 57 phrases (Léchelle, Gotti et Langlais, 2018). Ils sont désormais en phase d'évaluation des paraphrases dans l'optique de les injecter dans leur extracteur Distylium, qui fonctionnera pour les langues anglaise et française.

L'équipe poursuit en outre le travail sur les liens internes – qui permettront de suggérer à un.e lect.eur.rice des articles connexes – et sur les liens externes – qui permettront d'ajouter des ancrages vers des bases externes. Un effort sera fait également pour étudier l'ajout d'annotations marginales.

Luis Meneses et Ray Siemens, ETCL (Electronic Textual Cultures Lab) – **Social Media Engine**

→ Voir la présentation

→ Tester le prototype du Social Media Engine
sme.dhinstitute.ca/solr/sme/browse/?q=cinema

→ Lire l'entrevue réalisée en mars dernier

Le Social Media Engine est basé sur l'idée selon laquelle les articles devraient être reliés au savoir, et vise à transformer en ce sens l'expérience de recherche dans les bibliothèques numériques ou les dépôts en libre accès. Il permet de réorganiser les résultats de recherche sur un sujet donné selon le nombre de mentions ou de partages des articles sur les réseaux sociaux. Il est basé sur une combinaison de *topic modeling* et de TF-IDF et emploie un transfert de données Altmetric.com (Digital Science). Au cours de la dernière année, les membres de l'équipe d'ETCL ont créé et amélioré le prototype et l'interface utilisateur. Ils ont également lancé un processus d'évaluation, à la fois qualitatif et quantitatif, du cycle d'utilisation. L'équipe a été très active dans la communauté de recherche en humanités numériques.

Le ETCL poursuivra son travail dans cette voie, plus précisément pour améliorer l'interface utilisateur et pour étendre son cercle d'interlocuteurs.

Juan Pablo Alperin, Public Knowledge Project – **Altmetrics libres : Paperbuzz**

→ Voir la présentation

→ Tester le service web Paperbuzz

→ Voir le code source de l'API et de PaperbuzzViz

→ En savoir plus sur le projet Facebook : arXiv:1809.01194

Depuis 2017, avec le lancement du nouveau service Crossref Event Data, PKP œuvre au développement d'un outil libre de mesures d'impact alternatives (*altmetrics*) appelé Paperbuzz. Il s'agit d'une solution de rechange communautaire aux entreprises commerciales telles que Altmetric.com (Digital Science). Crossref rassemble des événements Twitter, Wikipédia, Facebook, Reddit et WordPress, et donne accès à ces données à un niveau granulaire, mais ne calcule pas les indicateurs de mesure (*metrics*). C'est pourquoi PKP s'est associé à l'initiative non commerciale Impactstory pour concevoir un service qui agrège ces événements en indicateurs de mesure et les rend accessibles à travers une API. PKP a également développé une visualisation Javascript qui utilise l'API de Paperbuzz pour afficher les métriques par jour, par mois, par année et au total. PaperbuzzViz est maintenant prête pour l'utilisation et sera disponible pour toutes les revues OJS grâce à un plugin, et sur les pages des articles des revues hébergées par la Open Library of Humanities.

Toutefois, la collecte de données *altmetrics* demeure pleine de défis, d'autant plus que les grandes plateformes de réseaux sociaux sont des boîtes noires qui ne divulguent pas la manière dont ils assemblent et cartographient les données. PKP a donc lancé un autre projet indépendant pour colliger des données Facebook de meilleure qualité, avec l'objectif de les rediriger vers Crossref Event Data par la suite.

AXE EXPLORER

Christopher Collins, Vialab – **Vue d’ensemble des projets Vialab : des outils interactifs pour les chercheurs en sciences humaines et sociales**

→ Voir la présentation

→ Tester l’outil Textension

→ Voir le code source de divers projets de Vialab

L’équipe de Vialab développe une vaste gamme d’outils de recherche et de visualisation conçus pour les sciences humaines et sociales. Dans la dernière année, les membres de l’équipe ont complété la conception de l’outil Textension, qui sera bientôt entièrement libre et accessible. Un autre projet bien entamé et rebaptisé Slow Analytics (présenté l’année dernière sous le nom Document Analyzer), offre de nouvelles façons d’explorer un vaste corpus comme celui d’Érudit; il permet par exemple de téléverser un document d’intérêt et d’extraire des termes de ce document pour les réemployer dans le moteur de recherche.

Les trois autres chantiers sont des projets à part entière qui ont été entamés au cours de l’année. La carte des connaissances d’Érudit (Érudit Knowledge Map) vise à cartographier le transfert de connaissances qui se produit au sein du corpus d’Érudit, par exemple, les réseaux d’auteurs dans des revues précises, ou par institution, par date, par auteur ou par titre d’article. Document Matching est conçu pour aider les usagers individuels à trouver la revue qui les intéresse en fournissant en guise d’exemple un document pertinent, grâce à la fonction glisser-déplacer. Enfin, Citation Galaxies est un outil de visualisation pour la bibliométrie qui contribue à l’analyse du contexte des citations. Il permet également de construire manuellement des jeux de données pour l’analyse du ton d’un texte.

Maxime Sainte-Marie, Chaire de recherche du Canada sur les transformations de la communication savante – **Le nettoyage de la collection BAnQ : état des lieux**

→ Voir la présentation

L'approche proposée l'an dernier pour le nettoyage du corpus de revues et journaux numérisés de BAnQ s'est révélée une impasse. En effet, la qualité de la numérisation est extrêmement variable d'un document à l'autre et parfois très mauvaise. Les documents, qui contiennent une forte proportion de doublons, sont pour la plupart des images au format JPG, mais aussi des fichiers PDF et TXT. Parmi ces derniers, certains fichiers ne sont que des conversions vides de PDF sans balisage texte, tandis que d'autres sont complets et prêts à l'utilisation.

Il est donc nécessaire de mener une rétro-ingénierie du corpus afin d'en réduire la taille, puis de procéder à un traitement d'image à grande échelle – en gardant à l'esprit que le traitement d'image en soi constitue un défi de taille. Ce n'est qu'une fois ces étapes critiques complétées que la reconnaissance optique de caractères et le post-traitement linguistique pourront être réalisés.

Vincent Larivière, Observatoire des sciences et des technologies – **Érudit : Une plateforme d’édition devient une base de données bibliométrique**

→ Voir la présentation

→ En savoir plus sur les services et les bases de données de l’OST

Au cours de l’année 2017-2018, l’équipe de l’Observatoire des sciences et des technologies a officiellement démarré le projet d’index de citations basé sur le corpus de revues savantes d’Érudit. Pour rappel, un prototype d’index de citations a été réalisé en collaboration avec Érudit, l’OST et l’Agence universitaire de la francophonie entre 2015 et 2017, mais le travail réalisé dans le cadre de CO.SHS permettra de constituer une base de données relationnelle complète et exploitable. La structure des tables sera similaire à celle utilisée par l’OST pour ses autres bases de données.

La base est actuellement en cours de peuplement avec les données d’Érudit. L’un des points critiques du projet sera la désambiguïsation des noms d’auteur.e.s, des affiliations et des revues. Ce travail pourra en partie être automatisé grâce aux algorithmes que l’OST utilise chaque année pour constituer la base de données du Web of Science, mais un important nettoyage manuel devra aussi être effectué. Plusieurs types de recherche pourront être menés grâce à cet index de citations, notamment sur la place des femmes en sciences humaines et sociales au Québec.