Indexer les corpus numériques

Emmanuelle Perrin

Indexer les corpus numériques

PAR EMMANUELLE PERRIN · PUBLIÉ 19 AVRIL 2019 · MIS À JOUR 27 JUIN 2019

Un séminaire sur l'indexation des corpus numériques a été organisé les 1er février et 29 mars 2019 à l'université Jean Monnet de Saint-Étienne, par l'Institut d'histoire des représentations et des idées dans les modernités (IHRIM, UMR 5317), avec le soutien du consortium Cahier et de la Maison des sciences de l'homme Lyon-Saint-Étienne. L'édition critique est l'un des axes fondateurs de l'IHRIM et le site de Saint-Étienne rassemble plusieurs projets d'édition numérique en XML TEI : la correspondance du libraire Marc-Michel Rey, celle du jésuite Matteo Ricci, le *Parallèle des Anciens et des Modernes* de Charles Perrault et les œuvres complètes de Charles Fontaine. L'organisation de ce séminaire se trouve directement liée aux questionnements que soulève la structuration des données de ces corpus de recherche, et en particulier le traitement de l'information biographique, bibliographique et géographique. Coordonné par Fabienne Vial-Bonacci (CNRS, IHRIM) et Emmanuelle Perrin (Université de Lyon, Archéorient), il a réuni sur les deux séances une cinquantaine de participants – enseignants-chercheurs, ingénieurs, documentalistes, doctorants.

Volontairement ancré dans une perspective pluridisciplinaire – littérature, géographie, histoire, économie, sociologie – ce séminaire avait pour objectif de montrer comment la structuration et l'indexation des données permettent d'enrichir, d'analyser et de diffuser les corpus numériques. Au croisement de l'informatique et des sciences de l'information, entre modélisation et interprétation, la constitution d'index apparaît comme un véritable enjeu technique et intellectuel sur deux plans, celui du corpus et celui, plus large, de l'ouverture des données de la recherche dans le contexte des humanités numériques. L'indexation suppose l'analyse d'un contenu et sa description à l'aide d'un langage documentaire, d'un vocabulaire contrôlé que fournissent les référentiels et les notices d'autorités. Au niveau du corpus, les index relèvent, comme l'annotation, de l'exploitation scientifique des textes édités : ils sont des points d'entrée dans un corpus, des outils de navigation et ils participent à la compréhension et à l'étude d'un texte en localisant dans son flux et en identifiant de manière univoque les personnes, les lieux et les œuvres mentionnés, ainsi que les sujets traités. La structuration des données et l'utilisation de formes normalisées apportent ainsi de la cohérence aux données et préparent à leur analyse et à leur exploitation. Au second niveau, dans le contexte de l'ouverture des données de la recherche, la structuration et l'indexation favorisent, selon les principes FAIR, la découverte, l'accès, l'interopérabilité et la réutilisation des données scientifiques. Les référentiels et les notices d'autorité constituent des données de confiance pour le développement de l'*open data* et du web sémantique.

Programme de la première séance

Programme de la seconde séance

Encodage de textes ou encodage des données ?

C'est la question posée par Lou Burnard pour introduire les propositions de la *Text Encoding Initiative* au sujet de l'indexation. Le co-fondateur de la TEI distingue trois manières d'aborder un texte à l'aide d'un ordinateur. On peut le considérer comme un objet physique et travailler son aspect visuel au niveau de la mise en page. Si l'on se focalise sur son contenu linguistique, la dimension formelle du texte disparaît. On peut enfin le traiter indépendamment de sa forme et de son lexique, pour les descriptions de lieux, d'événements, d'objets ou de personnes qu'il contient. On franchit alors la frontière entre texte et données. Les données apparaissent ainsi comme un type particulier de texte.

Explorer et formaliser les catégories

La question de l'indexation invite à adopter un recul critique sur les catégories explorées dans un corpus. Iona Galleron (Sorbonne nouvelle – Paris 3, LaTTiCe UMR 8094) et Cécile Meynard (Université d'Angers, CIRPaLL), ont exposé les travaux du groupe de travail R2 Cahier qu'elles mènent, avec Fatiha Idmhand (Université de Poitiers, CRLA-Archivos – ITEM UMR8132), sur les études littéraires assistées par ordinateur et la définition du personnage littéraire. À partir d'une expérience de « *crowdreading* », elles s'intéressent au repérage des personnages. En considérant les chaînes de coréférence comme des chaînes de caractérisation, elles encodent en TEI les différents éléments qui contribuent à définir un personnage (tempérament, situation sociale, âge, etc.), dans le but de comparer ces personnages entre eux — Madame de Rênal et Mathilde de la Mole apparaissent ainsi

physiquement et moralement proches — ou les styles de caractérisation de différents auteurs — on constate par exemple que les personnages de Borges sont fort peu caractérisés.

C'est une longue fréquentation des sources qui a permis la construction, dans le cadre du programme Europange, de l'outil d'analyse prosopographique concernant les officiers des principautés angevines du XIII^{e} au XV^{e} siècle, qu'a présenté Thierry Pécout (UJM, LEM-Cercor UMR 8584). Il s'agit de comprendre la formation d'un corps social par l'indexation fine d'éléments biographiques issus du dépouillement de sources nombreuses, grâce à une nomenclature précise des charges, des statuts et des lieux. Systèmes anthropologiques (filiation et clan) et contextes différents ont conduit les participants au projet à réfléchir sur la manière de caractériser au mieux les liens entre les individus : comment qualifier par exemple ce qui unit ceux qui ont eu un même professeur, ou bien l'amitié d'un roi pour ses subalternes ? Thierry Pécout souligne également que l'on utilise bien souvent des catégories historiques postérieures aux corpus étudiés : c'est le cas des catégories *religieux* et *laïc*, construites au cours du XVIII^{e} siècle, avec lesquelles on considère le Moyen-Âge, une période rétive à la formalisation. Michel Bellet (UJM, GATE LES UMR 5824) emploie même le terme d'*écrasement* pour évoquer les interprétations dont il est nécessaire de se dégager pour revisiter le saint-simonisme. Seul un travail rigoureux sur un corpus permet d'étudier très précisément le contexte d'emploi d'une notion chargée d'interprétations polysémiques (ouvrier, république, propriété, etc.).

Indexation et analyse des données textuelles

Plusieurs projets ont recours aux outils d'analyse textuelle. Il en est ainsi du projet présenté par Thierry Joliveau (UJM, EVS UMR 5600), qui combine l'annotation automatique d'informations spatiales avec la plateforme Perdido et une approche textométrique (TXM), pour extraire les entités spatiales nommées (odonymes) dans un corpus de romans parisiens du XIX^{e} siècle. La cartographie de l'espace romanesque met en évidence la structure spatiale d'un roman, permet de comparer l'empreinte géographique de différentes œuvres et offre la possibilité de travailler sur le corpus d'un auteur, d'une période ou d'une école littéraire. Ancrée dans un espace réel, la cartographie romanesque ouvre des perspectives en analyse littéraire, en histoire urbaine et dans le domaine du tourisme culturel.

Première séance du séminaire, le 1er février 2019

Pascal Vallet (UJM, Centre Max Weber UMR 5283), dans le cadre du projet Émergence de l'art contemporain en Auvergne Rhône-Alpes, a également utilisé l'analyse lexicométrique (Iramuteq) pour étudier des transcriptions

d'entretiens avec des acteurs de l'art contemporain, dans un triple objectif : circuler dans ces données, fouiller dans ces données afin d'en retirer des informations inattendues, et rendre ces données accessibles au plus grand nombre. L'analyse textométrique a permis de mettre au jour de manière inductive des univers discursifs significatifs pour la sociologie de l'art.

Gérald Chatagnon (UJM, DSI) et Ahmad Fliti (CNRS, LEM-Cercor) ont détaillé le travail d'indexation du corpus sur le saint-simonisme qu'ils ont réalisé avec Solr et Drupal. Le paramétrage de la recherche en texte intégral qu'autorise Solr est ici combinée avec la description, sur le modèle Dublin Core et avec des formes d'autorité, des différents textes réunis dans le corpus. Les catégories du Dublin Core ont notamment servi à concevoir la recherche à facettes et des parcours de lecture.

Les notices d'autorité, pivots pour l'interopérabilité des données

La question des formes d'autorité a été abordée par François Mistral, responsable d'IdRef, qui a présenté ce référentiel de l'Agence bibliographique de l'enseignement supérieur et son offre de service à destination de la recherche pour enrichir les données et accroître leur interopérabilité. En normalisant les points d'accès aux notices bibliographiques, les notices d'autorité permettent d'identifier de manière univoque et documentée les personnes, les collectivités, les noms géographiques et les œuvres. On peut ainsi retrouver toutes les œuvres d'un auteur, toutes les éditions d'une œuvre, tous les documents d'un même sujet. Un signalement adossé à des notices d'autorité favorise l'interopérabilité des données. En cela IdRef fournit des identifiants pivots pour exposer les données de la recherche dans le web sémantique. IdRef a également ouvert la production de ses notices à des projets de recherche. Prelib (projet de recherche en littérature de langue bretonne), et Siprojuris(Système d'information des professeurs de droit), par exemple, alignent leurs données sur les notices d'autorité d'IdRef et contribuent à enrichir ces mêmes notices grâce à un service de connexion d'inter-application.

Dans une communication sur l'utilisation des référentiels publics pour l'encodage sémantique des textes historiques, Francesco Beretta (CNRS, LAHRA) a présenté les travaux du Pôle histoire numérique du LAHRA sur la question de la modélisation et de l'interconnexion des données produites par les historiens (plateforme Symogih, Data for History). Il s'est notamment félicité que des initiatives comme IdRef mettent en valeur le savoir-faire des chercheurs dans l'élaboration et la mutualisation de données de qualité. L'alignement sur les identifiants IdRef et l'interrogation de différents points d'accès SPARQL permet de rassembler une documentation très importante, qu'il appartient ensuite au chercheur de critiquer et d'interpréter. Il faut également souligner la valeur heuristique de l'agrégation et de la visualisation des données, qui permettent de percevoir des phénomènes sur lesquels on applique la méthode critique. Les conditions de l'interopérabilité s'appuient selon lui sur trois piliers : l'alignement des objets dont on parle sur les identifiants d'un référentiel tel qu'IdRef, l'utilisation d'un vocabulaire contrôlé pour désigner les concepts et la modélisation de l'information selon un modèle standard comme CIDOC-CRM. En ce qui concerne l'annotation sémantique, le pôle histoire numérique du LAHRA préconise une méthode associant un modèle de données sémantique avec l'encodage en XML-TEI et l'analyse textométrique avec le logiciel TXM.

Seconde séance du séminaire, le 29 mars 2019

L'enrichissement des données avec des référentiels a également été abordé par Laurent Capelli (CNRS, Huma-Num) qui a décrit la nouvelle interface et les nouvelles fonctionnalités d'Isidore, la plateforme d'accès aux ressources numériques des SHS, dont il est le chef de projet. Les mots du titre, les mots clés et les résumés sont enrichis à l'aide de différents thésaurus comme Rameau, le Répertoire d'autorité-matière utilisé par la Bibliothèque nationale de France et ceux de la Library of Congress et de la Bibliothèque nationale d'Espagne. La classification par disciplines utilise les référentiels des plateformes HAL et OpenEdition. Enfin les métadonnées sont normalisées. Les noms des auteurs sont par exemple alignés sur les Orcid (*Open Researcher and Contributor ID*) et IdRef.

Maud Ingarao (ENS, IHRIM) et Elysabeth Hue-Gay (Lyon 2, HISOMA) ont présenté la combinaison de l'éditeur XMLMind XML Editor(XXE) avec le gestionnaire de thésaurus OpenTheso, qui permet de construire un vocabulaire propre à un projet de recherche et de l'aligner sur des référentiels extérieurs. Avec le plugin Pluco, mis au point par le pôle Document numérique de la MRSH de Caen, le thésaurus se trouve directement intégré à l'environnement numérique et le travail d'indexation s'effectue en même temps que le travail d'édition.

La perspective pluridisciplinaire de ce séminaire, tout comme la dimension expérimentale de certains des projets présentés, ont largement contribué à la richesse des débats. La méthodologie des outils numériques réserve en fin de compte une large place à *l'interaction et à l'interprétation humaines*. L'automatisation de la reconnaissance des entités nommées (Stanford Named Entity Recognizer, Recogito, Perdido) nécessite toujours une validation humaine. Les représentations graphiques obtenues avec le calcul sur les données apparaissent comme des matériaux à traiter comme des sources. La TEI en tant que « manière de s'exprimer sur un texte » pour en révéler

le sens, doit se faire *à la main*. L'encodage, comme l'indexation, reflète un *point du vue* sur le texte. Et la structuration des données oblige à dire quelque chose sur ces données.

Bibliographie

Beretta F. 2016. Pour une annotation sémantique des textes: le projet symogih.org et la Text Encoding Initiative, *Bruniana e Campanelliana* XXII (2). [en ligne] https://hal.archives-ouvertes.fr/halshs-01505635v1

Frontini F. 2018. Données liées et annotation de corpus, Atelier de formation annuel du consortium CAHIER, Montpellier, juin 2018. [en ligne] http://cahier.hypotheses.org/files/2018/08/LOD-annotation-Cahier-Frontini.pdf.

Galleron I., Idmhand F., Meynard C. 2018. Que mille lectures s'épanouissent… Modélisation du personnage et expérience de "crowdreading", *Digital Humanities Quaterly* 12/1. [en ligne] https://hal.archives-ouvertes.fr/halshs-01815606.

Les conférences de la première séance du séminaire "Indexer les corpus numériques" sont en ligne sur le site de la MSH Lyon-Saint-Etienne à cette adresse : http://25images.msh-lse.fr/Portails/indexer_corpus_numeriques-ihrim-fev19/fr

Les conférences de la deuxième séance du séminaire "Indexer les corpus numériques" sont en ligne sur le site de la MSH Lyon-Saint-Etienne à cette adresse : http://25images.msh-lse.fr/Portails/indexer_corpus_numeriques-ihrim-mars19/fr

L'auteur

Formée aux humanités numériques, Emmanuelle Perrin participe au projet HyperThesau (labex IMU) et est rattachée au laboratoire Archéorient.

Pour citer ce billet : Emmanuelle Perrin. Indexer les corpus numériques, *ArchéOrient - Le Blog*, 19 avril 2019, **[En ligne] https://archeorient.hypotheses.org/11944**