Da Sylva Lyne, Cuxac Pascal (dir.). Analyse et exploitation des bibliothèques numériques

Florence Thiault

THIAULT, Florence. Lyne Da Sylva et Pascal Cuxac (dir.). Analyse et exploitation des bibliothèques numériques. Études de communication, 2019, vol. 52, p. 211-215.

Note de lecture

DA SYLVA Lyne, CUXAC Pascal (dir.). Analyse et exploitation des bibliothèques numériques. Document numérique - RSTI série DN - Volume 20, N° 2-3, Mai-Décembre 2017.

Ce numéro spécial de la Revue des sciences et technologies de l'information (RSTI) coordonné par Lyne Da Sylva enseignante à l'EBSI (École de bibliothéconomie et des sciences de l'information, Montréal Canada) et Pascal Cuxac ingénieur de recherche à l'INIST-CNRS fait suite au colloque « Analyser la science : les bibliothèques numériques comme objet de recherche » qui a eu lieu à Montréal dans le cadre du 85e congrès de l'ACFAS (Association francophone pour le savoir). À la croisée des enjeux techniques, sociaux et culturels, les questions traitées donnent l'occasion aux chercheurs d'exposer dans une perspective pluridisciplinaire (principalement issus des sciences de l'information, de l'informatique et des sciences du langage) les derniers résultats de leurs travaux et de leurs réflexions. Les sept contributions sélectionnées sont centrées autour de la problématique de l'analyse et du traitement du contenu de bibliothèques, archives et musées numériques, d'un point de vue théorique ou pratique. Une courte introduction rappelle les axes du colloque et présente successivement l'ensemble des textes sans hiérarchie ni organisation thématique. Quelques textes se détachent nettement, par leurs spécificités techniques, amplifiées notamment par l'ajout de documents d'accompagnement (tableaux, figures, copies d'écran). Les études empiriques développées dans ce numéro montrent bien les changements en cours dans la façon de mener des recherches. En effet, le passage d'une recherche individuelle à des projets de taille importante nécessite la contribution d'équipes qui mettent en œuvre des approches innovantes via des méthodes automatiques de traitement spécialisées.

Ce numéro débute et se termine par deux articles présentant des études de cas qui exploitent l'extraction de métadonnées dans des corpus numériques scientifiques. Ces deux exemples permettent d'aborder des questions liées à la gestion de corpus multithématiques comportant un gros volume de données de nature hétérogène, pas toujours structurées. Le premier article de ce numéro (Kergosien et al., p. 11-30) illustre exactement le sujet du colloque. Les auteurs (au nombre de 11) présentent une démarche méthodologique d'analyse de corpus par traitement automatique des langues et fouille de textes pour l'extraction du vocabulaire du domaine. Dans le cadre du projet interdisciplinaire TERRE-ISTEX[1], les expérimentations sont menées sur un corpus hétérogène constitué de thèses électroniques (données de l'Atelier National de Reproduction des Thèses et du portail theses.fr) et d'articles scientifiques provenant des bibliothèques d'ISTEX (bibliothèque d'archives numériques scientifiques) et de l'archive ouverte Agritrop[2] du centre de recherche CIRAD (Centre de coopération internationale en recherche agronomique). L'objectif est de développer un démonstrateur web de recherche d'information géographique (RIG).

[1] https://www.istex.fr/terre-istex/
[2] https://agritrop.cirad.fr/

En croisant trois dimensions (spatiale, thématique et temporelle), il est ainsi possible de comprendre quelles recherches ont été menées sur quels territoires et à quel moment. La question de la représentation de l'espace géographique est au centre de l'étude menée par François Dominic Laramé (p. 159-177) qui clôt ce recueil. Dans cette étude de cas, l'auteur s'appuie sur deux corpus d'articles tirés de l'Encyclopédie de Diderot (bibliothèque numérique du projet ARTFL[3]) dans le but de caractériser les représentations géographiques existantes dans ces documents numérisés. Cette étude permet de cerner la représentation des lieux produite par l'Encyclopédie, en particulier en mettant en exergue l'évolution liée au changement d'imaginaire de l'espace entre la version initiale de Diderot et celle de l'auteur de la majorité des articles, le chevalier Louis de Jaucourt. Un second corpus sur l'Amérique a permis de caractériser la production de cet espace en tant qu'objet de curiosité et de convoitise tel qu'imaginé par les Encyclopédistes. Chaque corpus a été soumis à des mesures de statistique textuelle par une approche par « sacs de mots » (identification du vocabulaire) courante en linguistique de corpus et en analyse de discours.

Dans la perspective d'une science ouverte, deux articles relatent la mise en œuvre d'une démarche de pratiques communicationnelles de la recherche académique. Le premier article, particulièrement dense, porte sur l'analyse bibliométrique de recherches en traitement automatique de la parole et du langage naturel (Mariani, Francopoulo, Paroubek, p. 31-78). Le corpus NLP4NLP[4] contient les articles publiés dans les conférences et revues principales du domaine, sur une période de 50 ans (1965-2015). Différentes études ont été menées sur ces données : évolution au fil du temps du nombre d'articles et d'auteurs, collaborations entre auteurs, citations entre papiers et entre auteurs, évolution des thèmes de recherche, détection des innovations et des ruptures épistémologiques, réutilisation des articles et plagiat, tout ceci dans le cadre d'une analyse globale ou comparative entre sources. Dans le contexte de l'Open Acess et de la communication scientifique, Annaïg Mahé et Camille Prime-Claverie (p.79-96) analysent dans leur contribution la présence numérique des chercheurs en sciences humaines et sociales (SHS) à partir de l'archive ouverte HAL-SHS[5] et la plateforme de blogs scientifiques Hypotheses.org[6]. Sur la base d'une analyse quantitative, elles étudient la présence des chercheurs sur HAL-SHS par discipline, leur participation via une typologie d'usagers de HAL-SHS et leur présence conjointe sur les deux dispositifs. Les auteurs soulignent la croissance régulière du nombre de contributeurs ainsi que la fluctuation des périodes de dépôts sur HAL-SHS, ainsi que la prépondérance de certaines disciplines sur les deux dispositifs. Cette étude exploratoire met en évidence quelques caractéristiques de nouvelles formes de communication scientifique.

Les usages et les parcours d'usagers de bibliothèques numériques institutionnelles (Gallica[7] et Inathèque[8]) sont examinés dans deux contributions. L'article sur les usages de la bibliothèque numérique de la Bibliothèque Nationale de France expose une méthodologie originale d'étude fondée sur la vidéo-ethnographie (Rollet, Beaudouin, Garron, p. 97-114). Afin d'observer les parcours des utilisateurs et les spécificités de leurs pratiques sur la bibliothèque numérique Gallica, les auteurs ont effectué des captations audiovisuelles des activités à l'écran, complétées par des entretiens d'autoconfrontation afin de donner sens aux observations.

[3] ARTFL : American and French Research on the Treasury of the French Language
[4] http://www.nlp4nlp.org/
[5] https://halshs.archives-ouvertes.fr/
[6] https://cleo.openedition.org/openedition/hypotheses
[7] https://gallica.bnf.fr
[8] http://inatheque.ina.fr/

Ils relèvent ainsi le rôle de la sérendipité pour enrichir les phénomènes de catégorisation et d'évaluation d'une liste de résultats. Ces vidéos explorent la dimension écologique de l'usage de Gallica et montrent que l'utilisateur est engagé dans de nombreuses opérations associant environnement numérique, matériel et social. Les professionnels de la documentation de l'Institut national audiovisuel (INA) dans leur article sur « l'automatisation dans les outils de consultation et de documentation de l'Institut national de l'audiovisuel » (Alquier, Carrive, Lalande, p. 115-136) réfléchissent à la conception d'interfaces de consultation adaptées aux différents besoins des professionnels de l'audiovisuel (journalistes, producteurs) et chercheurs du monde académique. En effet, la refonte du modèle de données documentaires de l'INA interroge aujourd'hui l'évolution des pratiques documentaires des usagers. Cet article étudie l'évolution des usagers et de leurs pratiques en les confrontant à deux cas pratiques d'automatisation : la segmentation automatique d'enregistrements vidéos et l'enrichissement des métadonnées par le biais du liage avec le Linked Open Data.

L'apport du Web sémantique et de données liées est questionné précisément dans le contexte de collections muséales numériques. L'analyse de la Collection of Historical Scientific Instruments de l'université Harvard[9] (Sainte-Marie, Gauvin, Larivière, p. 137-158) présente un exemple de fouille de données muséales à des fins de recherche, en dressant un portrait historique, géographique et matériel des objets contenus dans la base de données en ligne Waywiser[10] de l'université Harvard. La perspective de ce projet est de sensibiliser les autres musées des sciences à la mise en commun de leurs bases de données, collaboration qui permettrait d'élargir les perspectives d'analyse et de recherche en matière d'histoire des sciences et des technologies.

Tout en proposant une grande diversité d'objets d'étude et d'approches, ce numéro nous offre un panorama des questions actuelles liées à l'analyse et l'exploitation des bibliothèques numériques. Sa lecture permet d'ouvrir des pistes de réflexion qui soulignent la richesse des problématiques que les bibliothèques numériques soulèvent encore aujourd'hui. Cet ouvrage s'adresse aux chercheurs en sciences de l'information et de la communication, en informatique et en traitement automatique des langues travaillant sur le traitement de données et de documents numériques. Il intéressera également plus largement les chercheurs en SHS souhaitant s'interroger sur les usages et pratiques autour des bibliothèques numériques. En conclusion, la lecture de ce numéro de la revue RSTI permet au lecteur, de disposer d'un bon panorama des recherches menées dans le domaine de l'analyse de corpus d'articles scientifiques.

Florence Thiault
Laboratoire PREFICS
Université Rennes 2

[9] https://chsi.harvard.edu/
[10] https://chsi.harvard.edu/waywiser