Dépasser la liste : quand la bibliothèque entre dans la danse des corpus web

Cynthia Pedroja, Anne L'Hôte, Elise Chapoy, Elisabeth Levain

Dépasser la liste : quand la bibliothèque entre dans la danse des corpus web

DH2016 – 14/07/2016
Auteurs : Pedroja Cynthia, L'Hôte Anne, Elise Chapoy, Elisabeth Levain
Direction des ressources et de l'information scientifique, Sciences Po

Depuis 2014, la bibliothèque de Sciences Po développe une offre de service à destination des laboratoires de recherche de l'Institution. Traditionnellement assez éloignée de ces publics, elle essaye d'adapter ses pratiques et de trouver de nouvelles méthodologies de travail pour accompagner les chercheurs dans leurs travaux.
La révolution numérique a introduit un bouleversement dans la constitution des fonds des bibliothèques : comment rendre compte de l'état d'un sujet d'actualité pour les générations à venir lorsque le discours ne se bâtit plus uniquement dans des médias coutumiers, mais dans le flux des réseaux ? La bibliothèque peut-elle opérer des carottages thématiques du web et concevoir les dispositifs qui en conserveraient la trace ?
Les corpus web tentent de répondre à cette double problématique documentaire et de recherche en s'appuyant sur des techniques d'exploration et de visualisation de réseaux web. S'inscrivant dans la lignée des corpus d'étude outillés proposés par Corinne Welger-Barboza, ils permettent de structurer et d'identifier les discours sous-jacents, tout en apportant des moyens d'appropriation adaptés.
De la construction des données aux premiers constats scientifiques, cette communication dresse un bilan de ce travail, expérimental pour la bibliothèque tant au niveau des outils mobilisés et développés que des processus documentaires mis en œuvre, à travers l'exemple d'un corpus des acteurs de la question des changements climatiques.

Construction

En 2015 se tient à Paris la Conférence des Nations Unies sur les changements climatiques, la COP21. Sciences Po a retenu cette année ce sujet d'actualité comme thème fort de l'institution ; le contenu des enseignements, des manifestations scientifiques et de vulgarisation, des expositions en est irrigué. La bibliothèque et le médialab[1], partenaires de ce projet, l'ont choisi comme problématique de leur premier corpus. Quels sont les acteurs de la discussion autour des changements climatiques ? Quelle est leur position quant à la responsabilité de l'homme ? Qui parle avec qui ? Faut-il réduire les émissions de gaz à effet de serre ? Voici la liste des questions que nous avons posées. Pour y répondre, nous avons développé un processus en 3 temps : construction, exploration et exposition des données.
Le crawler Hyphe est l'outil qui nous permet de constituer les données. Pour cette opération, itérative, deux types d'actions sont mises en œuvre : identifier, sélectionner. La première consiste à trouver des sites (entités) pertinents pour la thématique. C'est le cœur du corpus[2]. Les crawlers parcourent les liens hypertextes de ce cœur afin d'identifier de nouvelles entités. La seconde à distinguer ceux qui sont pertinents et qui seront crawlés, de ceux qui ne le sont

[1] « Laboratoire de recherche centré sur les méthodologies numériques qui présente la particularité de s'appuyer sur une approche théorique forte et originale en sciences sociales, la théorie de l'acteur-réseau » [http://www.medialab.sciences-po.fr/fr/about/]
[2] L'équipe a retenu comme point de départ les 12 thèmes identifiés dans [Venturini, 2014]

pas. De proche en proche, nous prospectons le web et de nouveaux sites apparaissent. De 60 entités de départ, 40.000 ont été identifiées et 600 ont été conservées. Ce sont ces dernières qui forment le corpus [Jacomy, 2015].
Cette étape repose sur des pratiques à la fois connues et nouvelles pour la bibliothèque : l'analogie entre la sélection d'ouvrages à partir de sources fiables (ici pertinentes) est assez aisée à comprendre. En revanche, crawler, et ainsi automatiser l'opération d'identification reste une action assez éloignée des processus traditionnellement mobilisés en bibliothèque. D'autant plus que le crawler utilisé, Hyphe, conserve l'empreinte des liens qui unissent les entités constitutives du corpus : le réseau. Au-delà de la liste des sites de référence sur les changements climatiques, Hyphe nous permet donc de garder la trace des discussions en ligne. La compétence méthodologique sur les réseaux est apportée ici par le médialab.

Exploration

L'exploitation envisagée nécessite que les entités web soient enrichies. Les métadonnées des sites ne sont pas standardisées, à l'inverse des objets traditionnellement traités en bibliothèque (ouvrages, revues, documents audiovisuels) qui sont, eux, décrits avec des langages contrôlés. C'est pourquoi une série de catégories a été créée : type d'acteur, responsabilité de l'homme dans le changement climatique, nature des actions soutenues (atténuation, adaptation).
Ce travail est la combinaison de compétences documentaires et scientifiques. Le chercheur spécialiste de la thématique propose, et valide les catégories choisies, mais il aide également à circonscrire le périmètre du corpus, en optant par exemple pour la conservation de sites anglophones uniquement. Les catégories sont construites sur des index élaborés par la bibliothèque qui les éprouve lors des phases de test. De plus, pour chaque site, un résumé est rédigé afin de garder une trace de son contenu et ainsi pallier la nature éphémère du web [Corey, 2010].
En l'absence d'outil de catégorisation pour ce type de corpus, le développement d'une solution ad hoc est nécessaire. Elle doit répondre à des contraintes claires, dont le travail collaboratif synchrone et le maintien de l'intégrité des données entre Hyphe et le site web.
La phase exploratoire ultime est celle de la visualisation : « faire parler » le réseau en s'appuyant sur la catégorisation[3].

Exposition

Le troisième temps est celui de l'exposition : comment dépasser la liste, par laquelle une bibliothèque présente traditionnellement les ressources web sélectionnées ? En proposant un outil d'exploration en ligne, qui permet à l'utilisateur non seulement de consulter les sites du corpus, mais également de se les approprier. En combinant les catégories au graphe, l'usager peut jouer la partition écrite par la bibliothèque : si la liste, rassurante, car connue, est toujours là, il est en plus possible de visualiser, par exemple, les acteurs institutionnels qui soutiennent les mesures d'atténuation des émissions de gaz à effet de serre et de les confronter à celles des blogueurs climato-sceptiques en manipulant la visualisation. Le graphe ajoute de la profondeur à la liste en faisant apparaître la dimension de connexion entre les entités : c'est un atout du corpus.

[3] Quelques visualisation sont disponibles sur le site http://medialab.github.io/double-dating-data/#/

Les interfaces web existantes exposent rarement cette dualité liste/visualisation. C'est pourquoi, une nouvelle fois, une solution adaptée aux besoins des utilisateurs a été développée[4]. Ce projet entre dans un cadre de libération du code et d'ouverture des données.

Conclusion

Ce travail est le fruit d'une collaboration étroite entre les différents corps de métiers inhérents aux Digital Humanities. 11 personnes ont participé à cette expérimentation : 1 chef de projet, 4 bibliothécaires dont le référent « environnement », 1 développeur, 2 informaticiens spécialistes des réseaux web, 3 chercheurs du domaine. Il aura fallu une année pour mener à terme ce projet qui avait valeur de test : la bibliothèque a-t-elle les moyens de proposer un service d'accompagnement des chercheurs pour l'étude des réseaux web ? Répond-elle également à la question de l'archivage du web ?

Le premier constat, positif, est celui de la faisabilité : le collectif a été en mesure de finaliser ce corpus. Le deuxième, tout aussi engageant est celui des premiers résultats scientifiques basés sur la visualisation du réseau. L'équipe de recherche a été surprise de voir que les climato-sceptiques sont encore très actifs sur le web, alors même qu'ils semblent minoritaires dans les discussions physiques. Qui sont-ils ? Comment interagissent-ils avec les institutions, les entreprises et la société civile ? Pourquoi ne forment-ils pas un groupe à part, mais apparaissent-ils comme parties prenantes de la conversation en ligne ? Ces questions doivent maintenant être analysées en profondeur. D'objet documentaire, ce corpus sur le changement climatique est devenu objet d'étude outillé. Chaque carte peut entériner ou infirmer les problématiques initiales et parfois générer des intuitions, interrogations qui doivent être vérifiées et prouvées scientifiquement.

Ce projet a permis à la bibliothèque de collaborer avec un laboratoire de recherche et ainsi d'aller au-delà de ses logiques traditionnelles de documentation. De nouvelles méthodes de travail ont été identifiées. Il s'agit maintenant de les fixer et de voir s'ils sont opérants sur d'autres thématiques en achevant la conception de la chaîne d'outils et en faisant monter en compétence les équipes sur leur utilisation et la connaissance des réseaux web. La phase suivante est également celle de l'archivage : la documentation du processus de construction et d'exploration de chaque corpus est une étape nécessaire pour garder une trace pérenne des ressources web et assurer les conditions de réutilisation des corpus.

S'appuyant sur la décomposition du travail autour de la construction, l'exploration et l'exposition des données, l'équipe a su répondre à la double problématique à laquelle l'institution est confrontée. Et ainsi, comme une valse, la bibliothèque est entrée en trois temps dans la danse des corpus web.

[4] http://corpusweb.sciencespo.fr/app/#/

Annexes

Bibliographie

COREY Davis., 2014. “Archiving the Web: A Case Study from the University of Victoria”, *Code4Lib*. 2014, n° 26. URL : http://journal.code4lib.org/articles/10015 (consulté le 28/10/2015

GIRARD Paul. HyperText Corpus Initiative : how to help researchers sieving the web?. In : *Out of the Box conference : Using Web Archives*, 2011, Velika dvorana, Slovénie. URL : http://spire.sciencespo.fr/hdl:/2441/5coittpe7h8g695h172cg34d3e (consulté le 28/10/2015)

JACOMY Mathieu, « L’analyse visuelle de réseaux. Explorer le social grâce au numérique », *I2D – Information, données & documents* [en ligne]. 2015, vol. 52, n° 2 , p. 60-61. URL : www.cairn.info/revue-i2d-information-donnees-et-documents-2015-2-page-60.htm (consulté le 28/10/2015)

NIU Jinfang, “Functionalities of Web Archives”, *D-Lib Magazine* [en ligne]. 2012, vol. 18, n° 3-4, URL : http://dx.doi.org/10.1045/march2012-niu2 (consulté le 28/10/2015)

VENTURINI Tommaso, BAYA LAFFITE Nicolas, COINTET Jean-Philippe et al., “Three maps and three misunderstandings: A digital mapping of climate diplomacy”, *Big Data & Society* [en ligne]. 2014, vol. 1, n° 2, p. 1-19. URL : http://spire.sciencespo.fr/hdl:/2441/1lpf2c85nl8a5bvji6lcpcc4bp (consulté le 28/10/2015)

WELGER-BARBOZA Corinne, Corpus d’étude outillés [en ligne]. URL : http://observatoire-critique.hypotheses.org/category/corpus-detude-outilles (consulté le 28/10/2015)

WELGER-BARBOZA Corinne, Quelques réflexions sur l’effet propédeutique des catalogues des collections des musées en ligne. *DH2010*, 10 juillet 2010, Londres, Royaume-Uni. URL : https://docs.google.com/viewer?url=http%3A%2F%2Fdh2010.cch.kcl.ac.uk%2Facademic-programme%2Fabstracts%2Fpapers%2Fpdf%2Fbook-final.pdf (consulté le 28/10/2015)