RES
PA
DON

Réflexions méthodologiques

Le retour d'expérience du datasprint ResPaDon et les travaux préliminaires issus des quatre groupes permettent de dégager quelques réflexions d'ordre méthodologique pour étudier de manière complémentaire les archives du web et le web vivant à des fins de recherche.

1. Constitution du corpus de recherche

Les quatre comptes rendus d'expérimentation[1] montrent qu'il existe plusieurs manières d'envisager la constitution du corpus de recherche. Il existe des points de tension à interroger entre la volatilité et la stabilité des données du corpus, entre l'apport des approches synchroniques et diachroniques, entre une démarche cherchant à rendre compte de l'évolution entre web archivé et web vivant et une démarche fondée - a contrario - sur l'étude de la complémentarité de ces ressources.

Il est possible de choisir d'étudier une ressource particulière en ligne, fixe dans le temps, comme l'a fait le groupe qui a choisi d'étudier la communauté politique qui soutient Jean-Luc Mélenchon en construisant son corpus de départ sur un ensemble fixe de pages wikipedia. L'avantage de cette démarche est qu'elle facilite la comparaison des évolutions dans le temps du corpus car la structure des pages Wikipedia ne subit pas de changements majeurs et les pages elles-mêmes sont pérennes. Elle nécessite cependant d'avoir dès le début de la démarche de recherche une idée précise de la manière dont le corpus sera circonscrit, ce qui s'oppose en partie à une démarche plus exploratoire.

La volatilité des ressources issues du web et qui constituent le corpus est un enjeu majeur à considérer lorsque l'on envisage de travailler à la fois sur le web vivant et sur le web archivé. Une des manières d'approcher ce problème est de construire le corpus de recherche à partir de listes d'URLs précédemment répertoriées dans le cadre de collectes ciblées qui concernent des sites de référence sur un champ disciplinaire donné. Ce type de corpus permet - a minima - de décrire dans les grandes lignes les marqueurs structurants d'un sujet donné en adoptant une approche synchronique. Ainsi, le groupe qui a travaillé sur la critique en ligne des arts du spectacle a pu cartographier les différents types d'acteurs (critique, théâtre, institutionnel, festival, édition, …) mais également la nature de la critique (individu, professionnel, presse, académique, …) qui composaient les relations entre les différentes pages web de leur corpus, issues de la collecte ciblée de 2021 de la BnF.

L'intérêt pour les études diachroniques s'appuyant sur les ressources hautement éphémères du web soulève de nombreux problèmes méthodologiques et techniques. En effet, dans de nombreux cas, le·a chercheur·euse va vouloir observer une évolution des contenus des pages web, que ce soit au niveau micro par exemple les vocabulaires employés pour qualifier tel phénomène, de leurs structures, de leurs apparitions ou de leurs disparitions, ou macro en

[1] Un modèle de document (disponible en ligne), témoin de l'activité de recherche de chaque groupe a été réalisé et adapté pour le datasprint ResPaDon.

analysant les liens hypertextes qu'elles entretiennent les unes avec les autres et qui structurent leur écosystème.

Or, les modalités de collecte et d'archivage ne sont pas figées, elles évoluent au gré de la redéfinition des périmètres de collecte, de la diversification des stratégies de captation des données et des technologies employées. On voit bien à quel point le web archivé ne peut être étudié comme un miroir du passé mais bien comme une reconstruction à partir de traces collectées dont il convient de connaître le processus d'élaboration.

Afin d'améliorer ce processus, la constitution du corpus de recherche peut également être vue et utilisée comme un outil de contrôle qualité des collectes et une opportunité pour enrichir leurs métadonnées.
C'est un des points saillants qui ressort de l'expérimentation menée par le groupe qui a travaillé sur la représentation de la crise de la Covid-19 dans les archives de la BnF. Les cartographies réalisées à partir des sélections réalisées par les chargé·e·s de collection dans le cadre de la collecte sur la Covid-19 sont un très bon outil pour présenter la collection et ses contenus en première approche, d'en dégager les lignes de force et les faiblesses (sites centraux mais non archivés par exemple).
La démarche adoptée par le groupe qui s'est intéressé à la représentation de la génomique d'un point de vue politique a permis de mettre en lumière l'intérêt d'une approche complémentaire en identifiant quels étaient les sites manquants dans l'archive et en allant chercher dans le web vivant les ressources pour enrichir leur corpus d'étude.

2. Représentation et visualisation des données explorées

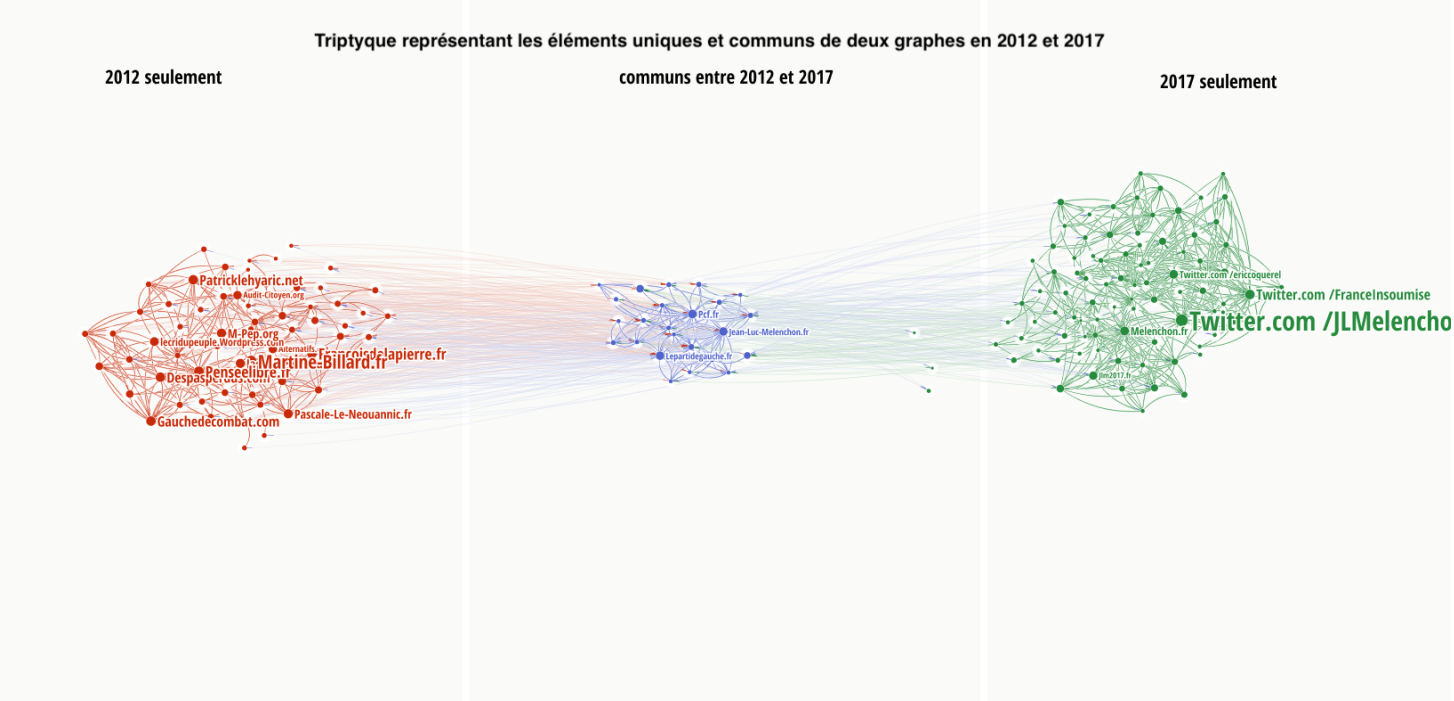
Les considérations méthodologiques sur la constitution des corpus de recherche amorcées ci-dessus se sont poursuivies, dans chacune des expérimentations citées, par des réflexions portant sur la manière de rendre compte des processus d'évolution et de comparaison entre corpus, entre web archivé et web vivant.
L'outil Hyphe, utilisé pour la curation de corpus, propose de représenter les entités qui le composent (par exemple des pages web) et les liens qu'elles entretiennent les unes (les liens hypertexte) avec les autres sous la forme d'un graphe.
Or, dans le cadre de recherches sur les archives du web, la dimension temporelle du ou des corpus est un point central à prendre en compte dans les analyses à mener. Si les graphes dynamiques permettent de voir se jouer les évolutions dans les représentations, ils sont bien moins satisfaisants lorsqu'il s'agit de faciliter l'analyse, notamment en raison du caractère mobile et volatile des données tant d'un point de vue temporel (apparition, disparition d'un noeud) que dans la spatialisation[2] du graphe (le même noeud ne sera pas représenté à la même place dans le graphe au temps t0 et t1). Qui plus est, ce que l'on cherche à rendre visible dans ce type de corpus sont les phénomènes d'apparition, de permanence, d'instabilité et de disparition de

[2] La spatialisation des graphes est généralement réalisée à partir d'algorithmes basés sur les forces. Ils suivent un principe simple : les noeuds reliés s'attirent et les noeuds non reliés se repoussent. Ainsi, l'évolution dans les données des corpus représentés a un impact direct sur la place des noeuds dans le graphe.

sites et de liens, marqueurs de restructuration de l'espace de la thématique étudiée sur le web à un temps t0, t1, t2.

Deux solutions émergent des travaux menés lors du datasprint ResPaDon pour outiller la comparaison entre graphes et rendre compte de l'évolution temporelle des éléments d'un même graphe.

Le groupe travaillant sur la structuration de communautés politiques propose la mise au point d'un script sur mesure développé pour visualiser des réseaux en "triptyque" permettant de comparer les données dans le temps. Un tel modèle visuel permet de faire ressortir la structure pérenne des entités web communes aux deux périodes mais aussi de visualiser les entités et communautés apparues et disparues entre les deux réseaux.



Le groupe ayant travaillé sur la cartographie de la crise de la Covid-19 dans les archives de la BnF s'est concentré, quant à lui, sur la comparaison de deux graphes entre eux, en fixant des nœuds pérennes d'une période temporelle à une autre comme repères dans la spatialisation des graphes successifs. Pouvoir retrouver des mêmes éléments au même endroit dans deux graphes différents facilite l'observation des processus d'évolution des données. Comparer des graphes en maintenant une cohérence dans la spatialisation des nœuds permet également de faire un contrôle qualité sur les données représentées, ce qui s'avère particulièrement utile quand on travaille sur de grands corpus de données et que le processus de collecte a été délégué.

Les retours d'expérience des quatre groupes montrent que ces problématiques de visualisation sont au cœur du travail d'analyse mais également nécessaires à l'effort de restitution des travaux tant il paraît pertinent de mobiliser les représentations cartographiques comme autant de points d'entrée dans les archives du web et ainsi faciliter la compréhension et l'utilisation des collections à des fins de recherche.

Pour aller plus loin

Site web de restitution des résultats du datasprint ResPaDon :
https://respadon.medialab.sciencespo.fr/
Carnet Hypothèses :
https://respadon.hypotheses.org/
Twitter :
@Respadon_Projet

Pour en savoir plus, contactez :
Audrey Baneyx
audrey.baneyx@sciencespo.fr
Eleonora Moiraghi
eleonora.moiraghi@sciencespo.fr