# Medialab de Sciences Po : cartographier le web pour les sciences sociales

Tommaso Venturini

ina EXPERT
Enseignement
Formation
Recherche
Etude et Conseil

# Médialab de Sciences Po : cartographier le web pour les sciences sociales

Par Tommaso Venturini, Sciences Po, médialab, Paris, France

(enseignant-chercheur, responsable recherche du médialab de Sciences Po)

Juin 2012

**Tommaso Venturini** est diplômé de l'Université de Bologne, cursus en Sciences de la Communication. Il a été visiting student à l'Université de Californie Los Angeles (Ucla) où il a travaillé avec Philip Bonacich sur la simulation des phénomènes sociaux par des modèles orientés agents. Il a suivi le programme doctoral international "Quality of Life in the Society of Information" de l'Université de Milan Bicocca, y développant des recherches sur les controverses avec pour thème "biopiraterie et modernisation des systèmes agroalimentaires traditionnels". En tant que chercheur post-doc de l'Université de Bologna et chercheur invité au Centre d'étude des techniques des connaissances et des pratiques de l'Université de Paris 1 (Cetcopra), il s'est occupé de sociologie des sciences et des techniques. Depuis 2008, il organise et enseigne les cours de Cartographie des Controverses de Sciences Po et depuis 2009, il assume la coordination des projets de recherche du médialab. Il a fondé et dirigé Studio Ideaedi une agence de développement Web et a participé à de nombreux projets innovants de communication en ligne.

---

**Le médialab de Sciences Po est un laboratoire de moyens numériques créé, en mai 2009, pour mettre la révolution numérique au cœur d'une des plus des plus anciennes et prestigieuses communautés des sciences sociales françaises. Un goulot d'étranglement, pourtant, empêche encore l'équipe du médialab de profiter pleinement de la source de données la plus important de l'âge numérique : le World Wide Web. À ce jour, aucun instrument n'existe pour permettre aux chercheurs en sciences sociales de sélectionner, extraire et archiver un corpus d'informations en ligne. L'objectif du projet HCI (Hypertext Corpus Initiative) au sein du médialab est de développer cet instrument pour le mettre au service non seulement de l'équipe du médialab et des chercheurs de Sciences Po, mais aussi de tout chercheur intéressé par l'étude du Web. En leur offrant des outils pour cartographier le Web, HCI ambitionne de renouveler l'usage des données numériques dans les sciences sociales.**

---

Depuis quelques années, les sciences sociales se retrouvent dans une situation tout à fait nouvelle. Encore relativement jeunes, ces sciences étaient encore il y a peu loin de pouvoir se doter des énormes machines capables d'engranger des données, comme celles dont disposent les sciences naturelles. Contrairement aux physiciens jonglant avec des milliards de particules dans leurs accélérateurs, ou aux biologistes cultivant des millions de microbes sous leurs microscopes, les sociologues ne pouvaient suivre que quelques centaines d'êtres humains et étaient condamnés à deviner la forme des phénomènes collectifs à travers ces aperçus partiels. Au cours des dernières années, toutefois, cette situation a été bousculée par l'arrivée des médias numériques.

Les médias numériques ont une caractéristique intéressante : toutes les interactions qui les traversent y laissent des « traces » et ces traces peuvent être facilement enregistrées, conservées et retransmises. Cette caractéristique a des conséquences capitales pour les sciences sociales [Lazer et al., 2009]. Au fur et à mesure que le numérique infiltre les sociétés modernes, la vie collective devient de plus en plus « traçable » [Mitchell, 2009]. Au fur et mesure que les archives publiques et privées sont avalées par la mémoire des ordinateurs, que les transactions économiques migrent en ligne, que les réseaux sociaux s'enracinent dans le Web, la quantité de traces accessibles aux chercheurs croît exponentiellement.

## LES SCIENCES SOCIALES FACE À UNE SOUDAINE ABONDANCE DE DONNÉES

Soudainement, les sciences sociales se retrouvent confrontées à autant de données que les sciences naturelles, mais avec une différence cruciale : les sciences sociales n'ont rien fait pour le mériter. Elles n'ont pas construit leurs radiotélescopes, leurs microscopes, leurs

séquenceurs. Les données numériques ont été recueillies pour des finalités autres que la recherche scientifique. Il s'agit d'informations récoltées pour des besoins de marketing (comme dans le cas des cartes de fidélité), de surveillance (comme dans le cas des déplacements aériens), de maintenance techniques (comme dans le cas des réseaux de télécommunication), de transparence (comme dans le cas de Wikipedia). Il s'agit, pour le dire autrement, de « données d'occasion », qu'investissent les sciences sociales sans que les chercheurs ne puissent maîtriser leur production et, surtout, sans que les chercheurs puissent s'y préparer. La situation des sciences sociales ressemble à celle de certains pays ruraux poussés à une brusque industrialisation par les pressions de l'économie internationale. Nées dans une époque de pénurie, les sciences sociales accèdent à un âge d'abondance trop vite et sans préparation.

Cela est particulièrement vrai pour les chercheurs qui s'intéressent au World Wide Web comme source de données pour les sciences sociales. Si à ses débuts, le Web pouvait être considéré comme un domaine à part [la cyberculture de Negroponte, 1996 ; les communautés virtuelles de Rheingold, 2000 ; les identités en ligne de Turkle, 1995], sa diffusion par capillarité transforme ce média en une sorte de papier-carbone d'une part grandissante de la vie collective [Rogers, 2009]. Dans les blogs, les wikis, les médias sociaux , les phénomènes collectifs laissent une quantité croissante de traces qui n'attendent que d'être exploitées par les chercheurs.

Les promesses des méthodes numériques risquent pourtant d'être vaines sans les outils conceptuels et techniques nécessaires pour transformer les traces Web en données de recherche. La distinction est cruciale, puisque l'abondance des traces Web ne se traduit pas automatiquement par un accroissement des données à disposition des sciences sociales. Les données ne sont jamais simplement « données », elles sont toujours construites par le travail de chercheurs [Latour, 1993]. Pour mettre la traçabilité du Web au service de la recherche, il ne suffit pas de recueillir les informations disponibles sur la Toile, il faut savoir les extraire, les nettoyer, les indexer, les préparer à l'analyse. Il faut, en d'autres termes, les constituer en corpus et cela reste, à ce jour, un goulot d'étranglement majeur des sciences sociales numériques.

Deux solutions existent aujourd'hui pour la constitution de corpora  hypertextuels, mais aucune ne correspond aux conditions de la recherche académique. D'un côté, les bibliothèques publiques réunies dans l'International Internet Preservation Consortium (IIPC [1] ) ont lancé de vastes initiatives de préservation du Web. Et pourtant, bien que méritoires, ces initiatives ne peuvent conserver qu'une partie infinitésimale du Web. L'Internet Archive [2] , à présent la plus large collection publique de documents Web, ne contient que 150 billions de pages tandis que, d'après les estimations des ingénieurs Google, la Toile dépasse le trillion de pages [Alpert, Hajaj, Google, 2008]. Le Web est tout simplement trop vaste et trop dynamique pour qu'on puisse espérer l'archiver exhaustivement. Toutes les initiatives d'archivage centralisé se concentrent, par conséquent, sur des régions bien délimitées de la Toile. Indispensables à fin de conservation historique, les collections publiques sont donc incapables de répondre aux

besoins de la recherche, puisque leurs échantillons ne correspondent que rarement aux intérêts des chercheurs.

De l'autre côté, les traces Web font l'objet d'une attention croissante de la part des entreprises de marketing commercial et politique. Ces entreprises disposent d'archives bien plus limitées que les collections publiques, mais ont l'avantage d'une souplesse qui leur permet de s'adapter aux besoins de leurs clients (voir par exemple l'approche innovante de Linkfluence [3] , entreprise pionnière dans l'archivage et l'analyse du « Web social » et partenaire recherche du médialab de Sciences Po). Cette solution, bien qu'assez satisfaisante, présente le désavantage majeur de subordonner la recherche académique aux stratégies de la Recherche et Développement privée.

## HYPERTEXT CORPUS INITIATIVE : DONNÉES HYPERTEXTUELLES POUR LES CHERCHEURS

Pour dépasser le goulot d'étranglement de l'archivage du Web, le médialab de Sciences Po [4] a lancé en 2010 l'Hypertext Corpus Initiative (HCI) [5] . L'objectif de cette initiative est d'offrir une alternative aux grands projets d'archivage public ou privé. HCI vise à encourager la constitution d'une multitude de micro-corpora hautement spécialisés, en donnant à chaque chercheur la possibilité de recueillir et conserver les échantillons du Web qui l'intéressent.

Cette stratégie se base sur l'observation que la Toile, loin d'être un espace uniforme, est au contraire un paysage varié, fait de discontinuités, de limites, de lacunes. Le Web est un tissu d'hyperliens, mais pas un tissu sans couture. Loin d'être des défauts, les nœuds et les déchirures de la Toile sont les raisons de son succès. C'est précisément parce qu'il n'est pas homogène que le Web a pu devenir l'habitat d'une multiplicité de groupes qui y ont trouvé leur niche. Admettre la nature intrinsèquement discontinue de la communication en ligne implique de reconnaître que l'exhaustivité de l'archivage est probablement moins importante que la capacité d'identifier les frontières qui séparent les différentes régions de la Toile. Pour que le Web puisse être mis au service des sciences sociales, il faut remplacer la stratégie du stockage massif et centralisé par une approche visant à donner aux chercheurs les moyens de définir les limites des territoires qu'ils/elles souhaitent investiguer. La constitution d'un corpus Web n'est pas seulement une question d'archivage, mais aussi et prioritairement, une question de cartographie.

Sous le nom de « cartographie du Web », on rassemble une série de notions, méthodes et outils permettant l'exploration des territoires numériques [Boullier, 2009]. La base de cette cartographie repose sur l'observation que, malgré le coût minime de la création des hyperliens, les acteurs du Web se montrent relativement prudents dans la création de connexions entre documents et sites. De ce fait, le Web n'est pas un espace chaotique. En choisissant à quels discours lier leur propre discours en ligne, les utilisateurs du Web établissent des hiérarchies et des « clusters » (groupes d'intérêt) [Gibson, Kleinberg and Raghavan, 1998]. Exactement comme Internet, le Web n'est pas un réseau classique, mais un « réseau de réseaux » : un

graphe où de zones densément connectées  côtoient des zones caractérisées par une raréfaction des liens. La cartographie du Web permet de délinéer ces zones et de caractériser les phénomènes collectifs qu'elles désignent. Développées pour représenter la topographie des territoires en ligne, les méthodes de la cartographie du Web offrent une base idéale pour la définition des limites d'un corpus numérique. L'objectif de l'Hypertexte Corpus Initiative est de développer ces méthodes et les implémenter dans un outil permettant aux chercheurs en sciences sociales de tracer, constituer et archiver leur corpora Web.

Plusieurs résultats remarquables ont été obtenus par le médialab de Sciences Po, qui représente déjà un centre d'innovation reconnu dans le domaine de la cartographie du Web. Avant même la fondation du laboratoire, les membres de l'équipe du médialab étaient déjà actifs dans l'exploration de territoires numériques, à la fois dans le cadre du « web-mining », la fouille de données (WebAtlas [6] ), et de la cartographie des controverses (Mapping Controversies du médialab de Sciences Po [7] , MappingControversies.net). Depuis sa fondation, le médialab est donc au centre de la recherche sur la cartographie du Web et participe activement au développement de ses outils (en particulier Navicrawler et Gephi).

## LES ÉTAPES DE LA CARTOGRAPHIE DU WEB

Les bases conceptuelles et techniques du projet HCI sont donc déjà posées. Sur ces bases reste à construire une méthodologie de création de corpus Web à la fois simple et solide, capable de rivaliser avec les méthodes quantitatives ou qualitatives classiques. Autour de cette idée, le médialab a réuni les principaux acteurs actifs dans le domaine de l'archivage et de l'analyse du Web. L'Hypertext Corpus Initiative rassemble plusieurs acteurs académiques (Sciences Po, Institut de Système Complexe de Paris), institutionnels (BnF, Ina) et privés (Linkfluence, WebAtlas, le Consortium Gephi), non seulement en France mais aussi à l'étranger (Digital Méthodes Initiative d'Amsterdam, Density Design Lab de Milan). L'accomplissement des objectifs du projet HCI passe par deux étapes, l'une plutôt méthodologique, l'autre plutôt technique.

- **Perfectionnement des méthodes de la cartographie du Web**

La cartographie est une discipline très jeune, qui a produit des résultats très encourageants, mais qui mérite d'être développée davantage. Deux questions méthodologiques en particulier seront abordées par ce projet.

En premier lieu, la cartographie du Web s'est jusqu'à maintenant concentrée exclusivement sur les hyperliens. Si cette approche a l'avantage de simplifier les activités d'analyse, elle néglige complètement les contenus des pages Web. L'approche topographique doit donc être complétée par une approche lexicométrique capable de prendre en compte au moins le texte des pages analysées (les contenus multimédias restant hors de la portée de ce projet). Les difficultés impliquées par cette expansion de la méthode cartographique sont méthodologiques

avant d'être techniques : comment rendre compte des flux d'idées, rumeurs, opinions qui circulent sur le Web [Leskovec, 2009] ? Comment conceptualiser une citation non formalisée par l'ajout d'un lien hypertextuel ?

En deuxième lieu, les méthodes dont nous disposons aujourd'hui sont capables de photographier l'état courant du Web, mais impuissantes quant à la question de la dynamique de la communication en ligne. Cela est, plus généralement, un problème reconnu de la « network analysis » (analyse des réseaux) encore relativement incapable de décrire la transformation des réseaux et la circulation de flux sur leurs connexions. Il s'agit ici d'une difficulté majeure qui n'a encore pas trouvé une solution standard. Un des objectifs de ce projet sera donc de tester les différentes solutions proposées et d'identifier celles qui semblent plus efficaces pour rendre compte de la dynamique du Web.

- **Développement d'un outil pour la constitution de corpora Web**

La recherche numérique en sciences sociales est impossible sans des outils à la hauteur des quantités des données qu'elle souhaite traiter. Cela est particulièrement vrai pour la cartographie du Web qui implique l'exploration, la sauvegarde et l'analyse d'une multiplicité de sites. Pour spécialisé qu'il soit, un corpus Web (surtout s'il est archivé et mis à jour instantanément) peut facilement attendre plusieurs milliers (voire plusieurs dizaines de milliers) de pages. La méthodologie de constitution des corpora Web nécessite donc d'être « implémentée » dans un outil capable d'automatiser les opérations répétitives et chronophages, tout en laissant aux chercheurs le maximum de contrôle sur la définition des données.

Hyphen, l'outil développé par HCI, vise à équiper les chercheurs avec un instrument capable de :

- accompagner les chercheurs dans l'exploration du Web, en gardant la trace des sites qu'ils visitent (sites-visités) et des sites cités par se sites (sites- voisins);
- assister la sélection des ressources, en offrant aux chercheurs une série d'indicateurs topographiques et lexicographiques sur les sites-visités et sur les sites-voisins ;
- faciliter l'extension du corpus par des techniques d'exploration semi-automatique (crawling) ;
- permettre la qualification des ressources du corpus selon un « codebook » (manuel de codage ) individuel ou partagé ;
- extraire, stocker, indexer, mettre à disposition les contenus des sites du corpus ;
- favoriser la mise à jour du corpus par la possibilité de planifier des sessions d'indexation automatique ;
- permettre l'exportation des informations récoltées dans une pluralité de formats afin d'en faciliter le traitement par les logiciels d'analyse statistique, les Caqdas (Computer Assisted Qualitative Data Analysis Software), les logiciels de manipulation de réseaux.

## UN POTENTIEL DE RENOUVELLEMENT POUR LES SCIENCES SOCIALES

Le projet HCI vise deux résultats principaux. D'une part, HCI entend équiper les chercheurs en sciences sociales qui souhaitent utiliser le Web comme source de données. Très concrètement, il s'agit mettre à point les techniques de sélection, d'extraction et d'archivage de traces numériques et de les implémenter dans un outil diffusé avec une licence open source. Le choix de renoncer à l'exploitation commerciale des techniques développées par le projet et d'inscrire HCI dans le cadre juridique et idéal du logiciel libre n'est pas anodin. Ce choix vise non seulement à faciliter la diffusion des résultats du projet, mais aussi à assurer la transparence des méthodes inscrites dans l'outil (tout monde pourra vérifier jusqu'à la dernière ligne du code d'Hyphen) et, surtout, à encourager la création d'une communauté d'utilisateurs-partenaires.

Le développement d'un outil pour la constitution de corpus Web n'est que l'objectif à court terme de ce projet. Les traces numériques ne sont pas intéressantes en soi, mais pour le potentiel de renouvellement qu'elles offrent aux sciences sociales. Une fois levé le goulot d'étranglement de la constitution de données numériques, c'est ce potentiel que le projet vise à explorer. Le deuxième objectif d'HCI est l'établissement de l'équipe du médialab comme « hub » de recherche sur les méthodes numériques au cœur de Paris. Ce réseau, mobilisé autour du développement d'un outil, vise à se stabiliser pour les années à venir. Conçu comme un projet d'expérimentation, HCI ambitionne de renouveler à long terme l'usage des données numériques dans les sciences sociales.

**Tommaso Venturini**, enseignant-chercheur, responsable recherche du médialab de Sciences Po

Mise en ligne (juin 2012)

---

1. Voir le site de l'International Internet Preservation Consortium,Netpreserve.org.

2. Voir le site de l'Internet Archive, Archive.org.

3. Voir le site de Linkfluence. Social Web Insight .

4. Voir le site du médialab de Sciences Po.

5. Voir le site de médialab Sciences Po — Hypertext Corpus Initiative.

6. Voir le site WebAtlas.fr.

7. Voir le site des Mapping Controversies du médialab de Sciences Po, medialab.sciences-po.fr/controversies.

_____

## RÉFÉRENCES BIBLIOGRAPHIQUES

ALPERT (Jesse), HAJAJ (Nissan), “We Knew the Web was Big…”, Google Official Blog, 25 July 2008.

BOULLIER (Dominique), « Au-delà des territoires numériques en dix thèses », in Frantz ROWE (ed.), **Sociétés de la connaissance et prospective. Hommes, organisations et territoires**, Nantes, Lemna, 2009.

GIBSON (David), KLEINBERG (Jon), RAGHAVAN (Prabhakar), « Inferring Web communities from link topology », in **Proceedings of the ninth ACM conference on Hypertext and hypermedia. HyperText ’98**, pp. 225-234, New York, USA, ACM Press, 1998.

LATOUR (Bruno), « Le topofil de Boa Vista ou la référence scientifique–montage photo-philosophique », in **Raisons Pratiques**, numéro 4, pp. 187–216, Paris, Éditions de l’Ehess, 1993.

LAZER (David) et al., “Computational Social Science” in **Science**, Volume 323, n° 5915, pp. 721-723, Washington, AAAS (American Association for the Advancement of Science), février 2009.

LESKOVEC (Jure), BACKSTROM (Lars), KLEINBERG (Jon), “Meme-tracking and the dynamics of the news cycle”, in **Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining**, pp. 497–506, New York, ACM (Association for Computing Machinery), 2009.

MITCHELL (Tom M.), “Mining our reality” in **Science**, Volume 326, n° 5960, pp. 1644-1645, Washington, AAAS (American Association for the Advancement of Science), décembre 2009.

NEGROPONTE (Nicholas), **Being digital**, New York,Vintage Books, 1996.

RHEINGOLD (Howard), **The Virtual Community. Homesteading on the Electronic Frontier**, Cambridge (Mass.), Londres, MIT Press, 2000.

ROGERS (Richard), **The End of the Virtual. Digital Methods**, (p. 36), Amsterdam, Amsterdam University Press, 2009.

TURKLE (Sherry), **Life on the Screen. Identity in the Age of the Internet**, New York, Simon & Schuster, 1995.