Constituer des corpus pour la recherche contemporaine :
de l'archivage du web à son analyse

Conférence du consortium international pour la préservation de l'internet (IIPC)
Paris, Bibliothèque nationale de France, 19 mai 2014

Vers un Web temporel ?

Sous *l'effet-diligence* qui pousse à penser chaque nouveau médium à l'aune des précédents, l'archivage du Web a d'abord été pensé sur le modèle de la bibliothèque, comme une collecte de documents. C'est alors l'ampleur et la ramification des « rayonnages » qu'on imaginait devoir conserver qui paraissaient constituer la principale difficulté. Dans un deuxième temps, l'expérimentation de nouvelles techniques de captation a déplacé l'enjeu sur la question de l'instabilité des contenus. Bien connue des internautes exposés aux « liens cassés », celle-ci a commencé à être perçue comme un problème épistémologique majeur à mesure que la production scientifique intégrait elle-même de plus en plus de références « en ligne ». Ignorant souvent les différents projets de préservation de l'Internet, les chercheurs étaient condamnés à bricoler leur archivage, en déplorant l'impossibilité de travailler sur des corpus homogènes et stables. Le dépôt légal du Web trouvait dès lors sa principale légitimation dans la garantie du recours à la source comme fondement de toute transmission. Ce faisant, les instituions patrimoniales pouvaient légitimement prétendre se maintenir dans leur mission, en assurant à la collectivité la maîtrise de sa mémoire, contre l'obsolescence technologique et les logiques de rentabilité économique qui prévalent à la pérennité des accès. À mesure que l'architecture du Web et ses usages évoluaient, cependant, force était de reconnaître que l'instabilité relevait moins d'un dysfonctionnement procédural contingent, que d'une dynamique indispensable à la circulation des informations et aux interactions en ligne. Avec la multiplication des sites dynamiques et l'importance croissante des « effets de réseau » (viralité, conversation, remix, recommandations…), il devenait essentiel d'adapter l'archive à cette plasticité. C'est ce qu'a notamment fait l'Ina, en mettant au point un outil de collecte tenant compte de la taille des sites et de la fréquence des mises à jour pour suivre au plus près l'élasticité du Web.

Paradoxalement, en épousant plus étroitement les variations du Web vivant, on crée cependant un Web archivé de plus en plus éloigné de l'idée de restitution. Contaminée par l'incohérence temporelle, l'archive fabrique de fait des sites qui n'ont jamais existé, en assemblant ou reliant des versions se rapportant à des dates différentes. Le recentrage sur la notion de ressource permet toutefois d'envisager cette incohérence sous un autre jour. Si on considère que le Web n'est pas constitué de documents instables (dont certains changeraient plus vite ou plus souvent que les autres), mais de représentations momentanément négociées et projetées au gré des accès d'une ressource elle-même inaccessible, on admettra plus facilement que l'archivage participe du même processus.

Mais que devient, dans cette perspective, la visée mémorielle de l'archive ? À la lumière des perceptions successives de l'instabilité numérique, on comprend que la « mémoire du Web » a longtemps été pensée avec des références spatiales. Procédant en surface (tout le .fr ou le .dk à un instant *t*) ou en carottage (tous les sites se rapportant à un événement), les stratégies de collecte ont d'abord reconduit une conception classique du corpus comme stabilisation et circonscription d'un *terrain*. Dès lors que les contenus en ligne sont envisagés comme des redocumentarisations

incessantes, en revanche, le Web comme son archive deviennent *temporels*. Procédant par recouvrement continu de situations, à l'instar des sédimentations du souvenir, ils se modifient à mesure qu'ils sont appelés, là où le stock n'est qu'une mémoire morte. À l'extraction cartographique des graphes de liens, peut alors se substituer la recherche d'une mise en contexte plus « narrative », où le lien marque un mouvement plutôt qu'une simple connexion. Dans cette perspective, le corpus lui-même n'est plus une liste de sites, mais un *environnement* : un réglage temporaire des distances informationnelles qui donne à voir un moment.

Reste à savoir si les méthodologies académiques sauront se réformer pour parvenir à naviguer ainsi dans le temps. On voudrait en tout cas suggérer que le recours croissant aux ressources et archives du Web ne pourra se généraliser qu'en développant de nouvelles intelligibilités du flux.

Bibliographie indicative

Brügger Niels, « L'historiographie de sites Web : quelques enjeux fondamentaux »,
Le Temps des médias, 2012/1 n° 18, p. 159-169. DOI : 10.3917/tdm.018.0159

Delaforge Nicolas, Gandon Fabien et Monnin Alexandre, « L'avenir du web au prisme de la ressource », in Séminaire IST Inria, *Le document numérique à l'heure du web de données* (2012) p. 229-252.

Frey Valentine, Treleani Matteo (dir.), *Vers un nouvel archiviste numérique*, Editions L'Harmattan 2013, 224 p.

Mussou Claude, « Et le Web devint archive : enjeux et défis »,
Le Temps des médias, 2012/2 n°19, p. 259-266.

MOTS-CLÉS

Instabilité, Web temporel, Mémoire, Environnement

*

Towards an temporal web ?

According to the "stagecoach-effect" which encourages us to think of each new medium in the light of the previous one, web archiving was first thought on the model of the library, as a collection of documents. At the time, the main difficulty seemed to be the scale and branching of "shelves" that needed to be preserved. Later, experimentation with new techniques of harvesting has moved the problem to the question of the instability of contents. Well-known by users exposed to "broken links", this began to be seen as a major epistemological problem as scientific production incorporated more and more "online" references. Often ignoring the various Internet conservation projects, researchers were forced to do their own archiving, lamenting the inability to work on homogeneous and stable corpora. Digital legal deposit found in this situation its main legitimation by ensuring the use of sources as the basis of any transmission.

In this way, heritage institutions could legitimately claim to remain in their mission, ensuring the community control of its memory, against technological obsolescence and the logic of economic efficiency that ignore the sustainability of access. As the architecture of the Web and its uses evolved, however, it was recognized that the instability was less a procedural failure, than a prerequisite for the flow of information and online interaction dynamics. With the proliferation of dynamic sites and the growing importance of network effects (virality, conversation, remix,

recommendations...), it became essential to adapt the archive to this plasticity. This was especially done by Ina, who developed a collecting tools taking into account the size of sites and the frequency of updates in accordance with the elasticity of the Web.

Following changes of living Web more closely, however, paradoxically creates a web archived further away from the idea of restitution. Contaminated by the time inconsistency, the archive invents sites that have never existed, assembling or linking versions pertaining to different dates. However, the focus on the concept of resource allows us to consider this inconsistency in a different light. If we consider that the Web is not made up of unstable documents (some of which would change faster or more frequently than others), but of momentarily negotiated and projected representations of a *resource* which is itself unreachable, it becomes easier to assume that archiving participates in the same process.

But what, in this perspective, does the memorial purpose of archiving become ? In the light of successive perceptions of digital instability, we understand that the memory of the web has long been thought with spatial references. Conducting in surface (all .fr or .dk at time t) or by core drilling (all sites related to an event), collecting strategies have first continue a classic idea of corpora as stabilization and demarcation of *grounds*. If online content is considered as incessant redocumentarization, however, the web and its archive become *temporal*. Proceeding by the continuous recovery of situations, like the sedimentation of memory itself, they change as they are called, contrary to the stock which is a ROM. Mapping extraction of links graphs can then be replaced by searching for a more “narrative” contextualization, where the link marks a movement rather than a simple connection. In this perspective, corpora themselves are no longer a list of sites, but an environmen : a temporary setting of informational distances which allows us to capture a moment.

KEYWORDS

Instability, Temporal Web, Memory, Environment