IFLA WLIC 2018
KUALA LUMPUR

Submitted on: 20/08/2018

Archiver les plateformes vidéo contributives à la BnF. Expérimentations, réalisations et projets

Alain Carou
Département de l'Audiovisuel, Bibliothèque nationale de France, Paris, France.
E-mail address: alain.carou@bnf.fr

Abstract:

Dès 2007-2008, la BnF a entrepris d'archiver des échantillons de la plate-forme vidéo contributive Dailymotion. Inscrit dans le cadre du dépôt légal du web, ce programme de collecte ciblée a permis au cours des années de réunir de précieux témoignages des premiers usages vernaculaires de l'image animée en ligne. Il connait aujourd'hui un nouveau rebond avec un programme de collecte de YouTube. L'archivage de corpus de vidéos vernaculaires présente un intérêt évident pour l'histoire des pratiques culturelles et des usages du web. Cela recoupe aussi à présent des questions historiques, politiques et sociales brûlantes. Ainsi le département de l'Audiovisuel de la BnF s'est-il engagé dans un programme de recherche comportant un volet de préservation de vidéos vernaculaires du conflit et de la révolte en Syrie.

Keywords: Vidéo, archivage du web, Youtube, images vernaculaires

Introduction : quel dépôt légal pour les médias sociaux ?

En 2006, une nouvelle loi sur le dépôt légal était adoptée en France : elle étendait le périmètre de la collecte aux sites internet et en confiait la responsabilité à la Bibliothèque nationale de France (BnF) et à l'Institut national de l'audiovisuel (INA). Exactement à la même époque, de nouvelles plates-formes de publication gratuites et faciles d'utilisation faisaient leur apparition et faisaient du même coup exploser la masse documentaire – déjà considérable – disponible sur le web. Cela posait dès lors des questions cruciales de politique de la collecte

(puisque le dépôt légal du web abandonne officiellement l'objectif d'exhaustivité et qu'il faut donc opérer des choix). Et cela soulevait notamment cette question : que devait-on faire avec les contenus amateurs ou, pour les nommer d'une manière moins ambiguë, avec les contenus vernaculaires, c'est-à-dire issus d'une pratique non professionnelle, et qui banalisée, de ce média par des particuliers ? Devait-on collecter uniquement des contenus de référence ou émanant d'éditeurs professionnels bien identifiés, ou bien aussi conserver un échantillon représentatif des pratiques ordinaires du web ?

Ces questions concernaient tous les chargés de collection de la BnF dans leur domaine de veille (que collecter de Facebook, de Twitter ?). Mais elle était adressée avec une acuité particulière au département de l'Audiovisuel de la BnF, du fait de l'essor pris par la vidéo en ligne dans les années 2005-2010. Jusqu'alors, les images animées (films puis vidéos) réalisées par les particuliers étaient presque toujours cantonnées dans le cadre privé. Cette situation a radicalement évolué avec les plates-formes contributives telles que YouTube et DailyMotion, plate-forme française née en 2005, qui fut pendant quelques années un véritable challenger de YouTube. Peu après avoir commencé à mettre en œuvre une politique d'archivage des sites internet français dans le cadre du dépôt légal, nous avons dû commencer à nous préoccuper de conserver des traces significatives des vidéos produites par les simples utilisateurs (*user-generated contents*). Comment choisir ce qui sera conservé dans une masse aussi gigantesque de contenus ? Comment leur donner sens et les constituer en corpus de recherche ? A quels enjeux mémoriels peut répondre l'archivage de portions de la production vidéo vernaculaire sur le web ? Ce sont nos expériences depuis dix ans, nos tâtonnements et nos perspectives actuelles qui vont être présentées ici.

Eléments techniques

Faisons au préalable un bref rappel des principes de l'archivage du web. La collecte est réalisée selon des processus qui sont automatisés : des listes d'URL sont fournies à l'outil logiciel appelé « robot ». Si l'on met à part les collectes larges qui consistent à collecter une fois par an un nombre limité d'URL sur tous les sites relevant de notre compétence, les listes d'URL sont constituées par les chargés de collections, qui déterminent les sites à collecter fréquemment ou de manière approfondie, et qui sélectionnent les ressources à retenir au sein des très grands sites. Le robot se connecte comme le ferait un internaute, archive une copie des pages et de tous leurs contenus, explore tous les liens contenus dans la page, rebondit sur ces autres URL, et ainsi de suite. L'archive qui en résulte est constituée d'instantanés de pages web (*snapshot*), reliés entre eux par les liens hypertexte et aussi proches que possible de l'apparence des pages originales. Conformément à la loi sur le dépôt légal, elle est accessible uniquement sur des postes de consultation dédiés, dans les emprises de la BnF et dans un réseau de bibliothèques en régions. La BnF fait partie d'IIPC (consortium international de préservation de l'Internet) destiné à assurer l'interopérabilité des archives et à faire progresser les outils et les méthodes de collecte.

1. Images d'« amateurs » : collecter une plate-forme vidéo à ses débuts

L'histoire de la collecte des vidéos a commencé au printemps 2007. La BnF a alors effectué une collecte ciblée de sites web relatifs à l'élection présidentielle en France. Le constat a été fait que le robot parvenait à capturer les vidéos de la plate-forme Dailymotion, qui était alors la plus prisée des internautes français. Dans la foulée, un nouveau projet a été lancé, pour collecter un échantillon de contenus représentatifs de Dailymotion, de quelque nature qu'ils soient. Dans un premier temps, il s'agissait principalement, pour le département de

l'Audiovisuel, de conserver des productions de communication des pouvoirs publics, des entreprises, des ONG, tous types de contenus institutionnels et militants que nous avions l'habitude de collecter au titre du dépôt légal. La grande différence avec la simple collecte des vidéogrammes est que dans le cas des plateformes vidéos du web, c'est aussi un contexte riche en informations qui est archivé : le nombre de vues de la vidéo au moment où est passé le robot, les commentaires d'internautes, le réseau de relations avec d'autres membres inscrits sur la plateforme, etc.

Néanmoins, nous nous sommes rapidement avisés du fait que cette politique de sélection orientée vers les comptes « institutionnels » risquait de laisser une vision tout à fait tronquée de cette plate-forme et de ses usages les plus populaires. Sur Dailymotion prédominaient des gags, des enregistrements de performance sportive ou chorégraphique, des clips enregistrés par des musiciens amateurs, des réemplois critiques d'images de télévision par des citoyens engagés, des captations d'événements d'intérêt politique ou social par de simples citoyens munis d'un téléphone portable, etc., en un mot des contenus situés hors de toute continuité avec le type de publication vidéo que nous avions l'habitude d'archiver. La facilité pour chaque internaute de republier les vidéos qui lui plaisent permettait d'accélérer les phénomènes de recommandation. Elle produit ce qu'on appelait dès lors une diffusion « virale », de proche en proche, par opposition avec l'habituelle diffusion « radiale » des contenus des industries médiatiques vers un large public.

Ainsi, une équipe de quatre chargés de collections a travaillé, à raison de deux à trois semaines par an, au repérage de comptes populaires et de pratiques très significatives afin qu'ils soient archivés. Le parti pris majeur de cette sélection a été de le construire autour de la notion de compte, donc de membre actif, plutôt que de faire de la sélection des vidéos à l'unité, travail chronophage et qui efface les acteurs et leurs trajectoires.

De plus, pour restituer de manière la plus neutre possible l'arrière-plan sur laquelle se détachaient les contenus que nous avions identifiés comme spécifiquement à collecter, nous avons alors choisi de capter une ou deux fois par an l'intégralité des contenus postés au cours d'une journée choisie aléatoirement. Au total, près de 3 000 comptes (membres actifs) et près de 200.000 vidéos ont été collectées sur Dailymotion entre 2008 et 2012. Brèves, spontanées, les premières vidéos relevant d'une pratique vernaculaire associée à une logique de partage nous apparaissent dix ans après comme de précieux jalons dans l'histoire des usages du web.

Cela dit, il est particulièrement malaisé pour un chercheur d'explorer le web du passé, dès lors que l'on sort des sites de référence, identifiés et bien organisés, et que l'on entre dans le maquis des plateformes où se déploient les pratiques vernaculaires. Les archives du web disposent pour le moment de moteurs de recherche très limités. Nous avons donc entrepris de les éditorialiser, c'est-à-dire de rendre compte de nos sélections et de fournir aux usagers quelques portes d'entrée dans les archives. Un « parcours guidé » leur a été consacré dans l'interface d'accès aux archives, pour mettre en lumière quelques exemples emblématiques des usages sociaux des images. Le parcours guidé souligne les dynamiques qui font des images un support de sociabilité numérique : partage de témoignages, partage d'exploits, citations directes ou détournées. Il rappelle aussi les tâtonnements dont s'accompagne ce mouvement d'émergence de l'usager comme producteur de contenu, par exemple le street-reporting, tentative sans lendemain pour créer un média d'information exclusivement à base de vidéos d'amateurs.

2. Vers une politique de collecte de Youtube

La sophistication croissante des protocoles de diffusion de la vidéo nous a malheureusement empêchés durant plusieurs années de poursuivre la collecte des vidéos, qui s'est interrompue entre 2013 et 2017. L'une des principales difficultés réside dans le fait que le robot de collecte Heritrix utilisé par la communauté IIPC ne peut collecter les vidéos comme il collecte les contenus qui obéissent aux standards ouverts du web. Les internautes disposent bien de solutions pour télécharger les vidéos, notamment l'outil sous licence libre Youtube-dl. Cependant, nous ne voulons pas collecter uniquement des vidéos, mais aussi la page dans lequel elles s'inscrivent, qui est leur contexte de publication, ainsi que leurs métadonnées. Et nous ne voulons pas collecter des vidéos à l'unité, mais collecter des chaînes entières, c'est-à-dire la production d'un membre actif dans sa totalité. La solution technique la plus efficace et cohérente avec les méthodes d'archivage du web à la BnF nous a été apportée par un script de l'University of North Texas. Celui-ci fait que l'outil Heritrix lance lui-même youtube-dl à chaque fois qu'il rencontre une vidéo à collecter. Testé avec succès au printemps dernier, cette méthode a été mise en production pour la première fois en juillet 2018. 42 comptes sélectionnés par nos soins ont été collectés intégralement, soit 28063 vidéos, pour un poids total de 1,8 To. A court terme, c'est 5 To (et donc approximativement 80 000 vidéos) qui pourront être collectés chaque année.

Entre la collecte de Dailymotion (2008-2012) et la collecte régulière de Youtube qui commence à présent, la situation a beaucoup évolué. L'amateur qui mettait en ligne gratuitement des contenus a cédé la place au youtubeur, nouvelle figure de professionnel de l'audiovisuel, rémunéré par la redistribution des rentrées publicitaires de la plate-forme. L'espace de partage sans contrôle qu'étaient les plateformes vidéo à leurs débuts est devenu un média *mainstream* à part entière. Les vidéos vernaculaires, quand elles remplissent une fonction de partage et de conversation au sein d'un groupe, se retrouvent désormais davantage sur Facebook, Twitter ou Snapchat. Ce tournant dans l'économie et les usages d'un des principaux médias sociaux valide rétrospectivement notre politique de collecte des premières vidéos vernaculaires à des fins de constitution de sources pour l'histoire. Car celles-ci appartiennent désormais à un passé largement révolu.

La politique documentaire de collecte de Youtube sera par conséquent adaptée à ce qu'est devenu ce média aujourd'hui. L'orientation que nous prenons actuellement, et qui évoluera peut-être au fur et à mesure que la réflexion s'enrichira, consiste à sélectionner les comptes les plus influents, dans tous les domaines : politique, diffusion des savoirs, culture et divertissement. Cette influence peut se mesurer de manière chiffrée : quel qu'en soit le sujet, un compte qui cumule plus d'un million de vues doit être collecté à nos yeux. Mais la question peut se poser de manière plus subtile. De nombreux contenus sont diffusés à travers plusieurs médias sociaux et pas seulement sur Youtube. Youtube n'est que l'observatoire et le point de collecte le plus commode pour nous. C'est donc sur une estimation de cette influence transverse et combinée qui doit entrer en ligne de compte dans notre réflexion, en nous fondant sur ce qui s'en dit dans les médias et sur notre propre expérience d'internautes.

3. Sauvegarder des témoignages historiques en danger : le projet Shakk

Youtube a maintenant une douzaine d'années d'histoire, et de nombreux contenus intéressants pour l'histoire en disparaissent quotidiennement. Cela soulève des questions historiques et mémorielles majeures, dont l'enjeu dépasse de beaucoup l'histoire des pratiques culturelles et des usages du web. Ainsi, pendant les « révolutions arabes » de 2011

et leurs suites, des vidéos tournées au téléphone portable et postées sur les médias sociaux ont été à la fois pensées comme des témoignages à opposer aux médias officiels et comme des facteurs de mobilisation. Dans le cas du conflit en Syrie, des vidéos ont été ainsi mises en ligne pendant plusieurs années, non seulement par des comptes clairement affiliés aux groupes combattants, mais aussi et chronologiquement d'abord par des individus engagés dans le mouvement. Des volumes considérables de ces vidéos ont disparu de la plateforme, car Youtube a procédé en 2017 à l'élimination automatique de milliers de contenus qualifiés indistinctement de violents par un algorithme de modération. La plateforme a admis avoir commis des erreurs et a rétabli les vidéos retirées lorsque une réclamation a été faite. Mais là intervient un deuxième facteur de disparition de ces vidéos : devant la tournure qu'a pris le conflit, de nombreuses personnes engagées dans la révolte voudraient tourner la page, oublier, voire éviter de se mettre en danger en laissant des traces en ligne. La victoire du régime en place a pour conséquence la disparition des images produites par les vaincus.

Des initiatives fragiles travaillent aujourd'hui à préserver les témoignages du conflit qui existent en ligne, tels les sites Creative Memory et Syrian Archive. Des centaines de vidéos n'existent plus que sur les disques durs des chercheurs-ses qui les ont enregistrées pour les besoins de leurs travaux. Le département de l'Audiovisuel de la BnF a répondu à la sollicitation d'une équipe de chercheurs-ses en sciences sociales pour constituer une archive pérenne et authentifiable. Un consortium constitué de l'Ecole des hautes études en sciences sociales, de l'Institut français du Proche-Orient de Beyrouth et du département de l'Audiovisuel de la BnF vient ainsi de se mettre en place. Financé par l'Agence nationale de la recherche sur trois ans (2017-2020), le programme de recherche « Shakk-Syrie : conflits, déplacements, incertitudes » vise à rassembler les sources présentes sur le web vivant aussi bien que celles qui ont été sauvées de la disparition mais sont dans des conditions précaires.

Les archives ainsi constituées seront donc de deux natures avec des traitements documentaires différents : archivage du web par capture en ligne des vidéos dans leur contexte d'une part ; archivage de fichiers vidéo à trier, identifier et reclasser d'autre part. L'accès restreint à ces ressources qu'offre la bibliothèque sera ici une opportunité beaucoup plus qu'une contrainte. Parce qu'elles ne seront accessibles que sur place, à des chercheurs accrédités, et sans possibilité de mettre en place des logiciels de fouille de données, leur constitution en archive ne contreviendra pas à la demande de retrait et d'oubli que portent une partie des acteurs et actrices de la révolte. Elle préserve la possibilité de l'histoire et de la remémoration fondées sur la plus grande pluralité possible de sources.

Nous avons ici un cas d'école d'images réalisées et diffusées en réaction à un moment historique crucial. L'enjeu mémoriel de leur sauvegarde est très différent de celui des collectes de comptes français sur Dailymotion ou sur Youtube, tout comme les outils, les méthodes et la relation avec les missions de la BnF diffèrent. Mais dans tous les cas, construire une politique d'archivage des plateformes vidéo contributives nous apparait comme une nécessité centrale, et ce pour deux raisons. D'abord (et c'est l'angle sous lequel on les prend le plus volontiers au sérieux aujourd'hui) elles sont comme l'ensemble des médias sociaux un foyer majeur de diffusion d'informations, vraies ou fausses, donc un lieu d'influence, de production de l'opinion publique. Mais elles sont également un foyer d'expressions personnelles originales, un laboratoire d'écritures, un lieu de construction du regard, et c'est aussi à ce titre-là qu'il nous apparait important de collecter d'une manière représentative la vidéo vernaculaire aussi bien que les grands comptes.

Aussi puissantes et solides paraissent-elles, ne prenons pas les plateformes vidéo pour des archives ou quasiment des archives. Elles sont soumises à quantité d'aléas : effets de mode, pressions extérieures, coûts, etc... C'est ce que nous décidons et parvenons à archiver qui constituera le socle documentaire d'une mémoire collective interrogeable partageable.

Acknowledgments

Merci aux collègues du service Images du département de l'Audiovisuel et du service du Dépôt légal numérique, ainsi qu'à Cécile Boëx (EHESS).