Archiver la vidéo sur le web

Des documents ? Quels documents ?

En une année et demie, la vidéo en ligne a connu un essor tel que, désormais, la grande question en suspens est celle des modèles économiques (*business models*) appelés à s'imposer, entre gratuit et payant, distribution des films par les fournisseurs d'accès à Internet (FAI) ou par les diffuseurs classiques (distributeurs, chaînes de télévision), sortie simultanée ou différée par rapport aux autres écrans.

Alain Carou

Bibliothèque nationale de France
alain.carou@bnf.fr

Quant au fait de savoir si l'accès individuel à la demande a un avenir ou non, la question ne se pose plus. La télévision à la demande (*pay-per-view*) avait échoué, probablement du fait d'une offre restreinte pour des raisons techniques ; cette limite n'est plus avec Internet. En consultant les résultats du premier *crawl* (collecte automatisée du web) fait par la BnF sur les sites français à l'automne 2005, on mesure tout ce qui s'est passé depuis [1] : apparition d'une offre riche de vidéo à la demande, décollage spectaculaire de la plate-forme d'échange Dailymotion, intégration de reportages aux sites de presse, etc. L'intérêt d'organiser l'archivage des vidéos en ligne est multiple : création formelle (fondée notamment sur les principes de recyclage et de détournement et sur l'utilisation des caméras miniatures), mais aussi politique (satire), diffusion d'opinions, diffusion du savoir (très riche offre de conférences enregistrées)… Le département de l'Audiovisuel de la Bibliothèque nationale de France, en charge depuis 1975 du dépôt légal de l'édition vidéographique [2], participe donc activement aux débuts de l'entreprise d'archivage de ces réalités nouvelles, qui enrichissent considérablement les contenus présents sur le web et redéfinissent ses pratiques et ses usages.

L'état de l'offre

Si la vidéo a envahi le web, elle y est à l'évidence inégalement répartie. On pourrait cerner plusieurs grandes familles de sites qui en concentrent l'essentiel [3] :

– les sites de vente de fichiers pour un usage restreint dans le temps

1. Les collectes faites à titre expérimental par la BnF avant la promulgation de la loi Dadvsi sont encore pour le moment non accessibles au public. Rappelons que la loi Dadvsi du 1er août 2006 a chargé la BnF et l'Ina de la collecte du web au titre du dépôt légal.

2. Et par ailleurs depuis 1938 de l'édition phonographique, depuis 1992 de l'édition électronique sur support.

3. État nécessairement provisoire : l'évolution très rapide de l'offre (YouTube n'a été créé qu'en février 2005 !) va se poursuivre. À l'heure de la rédaction de cet article (janvier 2007), les fondateurs de Kazaa et de Skype lancent Joost, technologie de diffusion de contenus à la demande par le *peer-to-peer*, censé offrir une qualité de visionnement supérieure au classique client-serveur. Elle comprend notamment la rétribution des ayants droit, la gestion et la mise en commun des *playlists*. Par ailleurs, il faut noter le développement rapide du nombre d'abonnés à la télévision à la demande par IP (IPTV), qui permet d'acheter et de voir un film sur son téléviseur par l'intermédiaire d'une « box ».

Archiviste-paléographe de formation, conservateur des bibliothèques, ***Alain Carou*** *est chef du service Images au département de l'Audiovisuel à la BnF. Il est membre du comité technique de l'Association internationale des archives sonores et audiovisuelles (IASA).*

(VOD, Video on Demand, assimilable à de la location) ou pour un usage permanent (EST – Electronic Sell Thru –, assimilable à de la vente). Cette offre, apparue au printemps 2006, se développe en France à un rythme encore modéré[4]. La mise de fonds initiale étant nettement moins élevée que pour une édition sur support, ce mode de diffusion peut laisser espérer la valorisation des fonds de catalogues, mais également de modestes productions indépendantes, « de niche » ou de cinématographies nationales qui trouvent difficilement des débouchés en salle, en DVD et a fortiori à la télévision[5].

– les plates-formes d'échange communautaire, où les contenus sont envoyés par les internautes. Pour les ressources sous droits (actuellement surtout des clips musicaux), le modèle économique « YouTube 2006 », fondé sur le partage de contenus de manière libre et en quelque sorte autogérée par les usagers, a d'ores et déjà vécu : en atteignant très rapidement un volume d'activité énorme, ce nouveau modèle devait s'intégrer à l'ordre du copyright. La gratuité va être à présent liée à la publicité (bandeaux et spots diffusés en pré-générique) et à la répartition de ses gains entre les éditeurs des sites et les ayants droit des contenus. Cela devrait s'accompagner d'une surveillance des fichiers mis en ligne[6]. L'économie qui se met ainsi en place est comparable à celle des grands portails liés aux moteurs de recherche (ce qui explique en partie le rachat de Youtube par Google). Pour les ressources sans copyright (vidéos réalisées par les internautes en particulier), l'inflation se poursuit à un rythme effréné, mais divisée entre une majorité de documents qui ne sont là que pour le plaisir de leur auteur et une minorité qui suscite un écho sur le web et, parfois, un grand mouvement d'intérêt pendant quelque temps (*buzz*) : séquences « volées », détournement d'émissions télévisées, *clips* autoproduits, etc.

– enfin, les sites de publication, d'accès gratuit, ayant un axe éditorial défini par une thématique ou un genre précis (art vidéo, *machinima*, militantisme, conférences, etc.). De tels sites comptent de moins d'une dizaine à plusieurs milliers de vidéos. Les éditeurs peuvent être titulaires des droits de diffusion (notamment lorsqu'ils sont les auteurs des ressources) ou non[7].

Pour simplificateur qu'il soit, cet état de l'offre permet une première approche des questions qui se posent pour l'archivage de la vidéo en ligne. Et en premier lieu de la définition d'un périmètre de collecte prioritaire :

– dans la continuité du dépôt légal de l'édition sur support : sont concernés les sites de VOD proposant des titres absents des catalogues DVD, les sites de création jouissant d'une certaine notoriété, les sites de communication institutionnelle, etc.;

– dans la continuité des campagnes thématiques de collecte décidées par la BnF (par exemple les campagnes électorales) : sont concernés les documents vidéo générateurs d'un *buzz* sur le thème ciblé.

> La « documentation » des ressources vidéo sur le web apparaît en fait comme un enjeu aussi essentiel dans une logique de lecture publique que dans une logique patrimoniale

Constitution du document

Il faut souligner que la vidéo n'existe pas toute seule sur le web. À vrai dire, il en allait déjà ainsi avec la vidéo sur support : le support, le conditionnement, la présentation matérielle dans son ensemble, sont porteurs d'éléments d'information paratextuels. Mais on va un pas plus loin avec les ressources vidéo en ligne : elles n'ont pas d'existence sociale sur le réseau si elles ne sont pas appelées par des liens (par exemple au sein d'une page de présentation) ou par des termes d'indexation. Leur paratexte éditorial (résumé, présentation, etc.) et leur balisage définissent leurs modalités d'accès – c'est-à-dire la manière dont on peut les trouver, mais aussi la « grille » d'appréhension à travers lesquelles on va les regarder. Il faut archiver les composantes liées du document : les films et leur valorisation éditoriale sur les sites de VOD, les vidéos postées sur les plates-formes et les *tags* (indexation libre) choisis par les utilisateurs des plates-formes d'échange, etc.

Cela supposera fréquemment de déborder les limites d'un site. Ainsi, les blogueurs utilisent aujourd'hui couramment Dailymotion ou Blip.tv pour y héberger commodément des fichiers vidéo : c'est alors le blog qui donne sens à la ressource. Il importe

4. Touchant les œuvres cinématographiques, les discussions achoppent encore début 2007 sur la place à donner à la sortie VOD dans la chronologie des médias (sortie salle, sortie DVD, diffusions télévisées).
5. Voir par exemple www.universcine.com, www.cinezime.fr
6. Voir par exemple « YouTube supprime 300 000 vidéos de ses serveurs », *Le Monde*, 23 octobre 2006.
7. Voir par exemple le site américain Ubuweb (fondé en 1996) qui met à disposition quantité d'œuvres cinématographiques et sonores des avant-gardes, sous droits et introuvables dans le commerce. Peu d'ayants droit sont intervenus pour faire retirer leurs documents (selon les informations données sur www.ubu.com/resources/shame.html).

donc de bien déterminer ce qui constituera la ressource en document.

La question gagne en épaisseur si l'on considère que, dans ce qu'il est à présent convenu d'appeler le « web 2.0 », les usagers commentent et réindexent la ressource. Les messages associés à la ressource (annotations) s'enrichissent au fil du temps, formant une « queue de comète » potentiellement de plus en plus longue. Si l'on juge pertinent de les inclure dans la définition du document, celui-ci sera donc une entité en évolution, différente d'une collecte à l'autre.

Élément majeur d'information et de choix des internautes[8], la « documentation » des ressources vidéo sur le web (paratexte éditorial, indexation, annotation) apparaît en fait comme un enjeu aussi essentiel dans une logique de lecture publique que dans une logique patrimoniale.

Dans une perspective de lecture publique, proposer aux usagers des médiathèques des réservoirs de VOD « de qualité » à un tarif préférentiel répond à l'objectif de démocratisation de l'accès pour tous ; mais reste entier le problème de la construction d'une médiation alternative entre le public et la création audiovisuelle. Autrement dit, comment faire en sorte que les bibliothèques puissent pointer comme elles l'entendent vers les ressources, plutôt que de renvoyer seulement vers un portail commercial généralement construit selon des logiques de marketing (les dix titres les plus vendus, « ceux qui ont acheté ceci ont aussi acheté cela »)[9] ? Il faudrait que les médiathèques puissent profiter du passage au « en ligne » pour documenter à leur manière les ressources dont elles ont négocié les droits (du moins celles qu'elles veulent défendre), et ainsi prolonger avec une visibilité accrue le travail d'invitation à la découverte et à la curiosité qu'elles accomplissent avec la vidéo sur support.

Dans la perspective de constitution d'une archive historique, il est au contraire primordial que le traitement bibliothéconomique ne se substitue pas à celui qui existe sur le web. Si l'on est convaincu de l'importance de la « documentation » des ressources pour comprendre les modalités de consultation contemporaines de leur existence en ligne et de leur réception, il paraît essentiel d'en garder trace, et de ne pas chercher à y substituer des accès élaborés selon les règles du traitement documentaire professionnel.

Enfin, outre ce qu'elle dit de l'usage des ressources vidéo au moment de leur publication et de leur appropriation par les usagers, leur « documentation » sur le web (paratexte, indexation, voire annotation) doit être archivée aussi tout simplement parce qu'elle sera pour longtemps la seule voie d'accès à l'archive pour l'usager. Le traitement bibliographique traditionnel à l'unité est en effet exclu (ou au mieux réduit à la portion congrue) pour des raisons de masse évidentes[10], et l'indexation automatique des contenus audio et vidéo en est encore à ses prémices[11].

Le traitement bibliographique traditionnel à l'unité est en effet exclu pour des raisons de masse évidentes, et l'indexation automatique des contenus audio et vidéo en est encore à ses prémices

Les difficultés de mise en œuvre

Deux modes de collecte des sites sont en œuvre à la BnF. Le premier est l'archivage des ressources telles qu'elles sont délivrées aux utilisateurs du réseau : schématiquement, un « robot » logiciel (*crawler*) parcourt les sites de liens en liens et enregistre les résultats. Ce mode de collecte présente l'avantage de rendre compte des ressources telles qu'elles ont été effectivement mises à disposition du public à un moment donné.

Le second mode de collecte est le dépôt, c'est-à-dire la prise de contact directe avec l'éditeur du site pour définir des procédures de versement des fichiers lorsqu'ils ne peuvent pas faire l'objet d'une collecte automatique. Il s'applique en particulier au « web invisible », tels les réservoirs de données auxquelles on ne peut accéder que par l'intermédiaire d'un formulaire de recherche ou sur authentification de l'internaute (abonnement, mot de passe). Par principe, cette solution n'est considérée que comme un recours si la procédure de collecte ne peut être menée à bien. Seule la

8. Il en existe certes d'autres, par exemple dans la presse (voir la page hebdomadaire consacrée maintenant aux ressources vidéo en ligne dans *Télérama*).

9. La première étude commandée par le Centre national de la cinématographie sur la VOD montre la grande influence des choix de valorisation des éditeurs de sites sur les décisions d'achat des usagers : trois quarts des utilisateurs se laissent guider par le site, notamment la page d'accueil et les listes de meilleures ventes. *Pratiques de la VOD en France*, étude réalisée par le laboratoire Novatris à la demande du CNC, décembre 2006 : www.cnc.fr/CNC_GALLERY_CONTENT/DOCUMENTS/publications/etudes/PublicVOD061206_.pdf (consulté le 22 janvier 2007).

10. Dailymotion (site français, faut-il le rappeler) annonce 15 000 nouvelles vidéos chaque jour. Le département de l'Audiovisuel de la BnF assure chaque année le catalogage d'environ 9 000 vidéogrammes édités sur support.

11. L'indexation automatique de la parole et celle des textes affichés à l'écran atteignent aujourd'hui des performances relativement intéressantes, mais seulement avec des contenus fortement normalisés comme les journaux télévisés (voir par exemple le site www.Blinkx.tv, qui annonce 6 millions d'heures de télévision indexées via des fonctions de reconnaissance de langage et de décryptage de discours). Elle en reste au stade de la recherche pour la reconnaissance de formes. On est loin encore de la sémantisation des contenus. En France, le projet Quaero, lancé fin 2005, vise à faire progresser ces domaines.

collecte offre en effet la garantie d'archiver des ressources effectivement mises à disposition du public.

Ce cadre étant rappelé[12], le dépôt légal de la vidéo en ligne se retrouve face à des problèmes plus accentués que pour d'autres types de ressources.

– Les questions de territorialité : la loi (et son décret à venir) ne peut faire obligation de dépôt légal qu'aux opérateurs français. Or le caractère international des industries et de la diffusion de l'audiovisuel, déjà prononcé avec les médias classiques, se renforce encore sur le web. Quand le filtre de l'importation n'existe plus et que chacun peut accéder sans intermédiaire à toute ressource mise en ligne n'importe où dans le monde[13], la délimitation de la collecte doit évoluer par rapport à l'édition sur support. Il n'est plus envisageable que le dépôt légal continue à couvrir exhaustivement non seulement la production nationale mais aussi la production étrangère mise à disposition sur le territoire français. Mais la remise en cause va plus loin, car maintes productions « nationales » peuvent se trouver diffusées par des opérateurs étrangers (par exemple : quid des réalisations françaises sur Youtube ?).

– La diffusion selon des modalités particulières, qui ne sont pas toujours bien gérées par le robot de collecte. C'est le cas du *streaming* (ou lecture en continu) dynamique, utilisant par exemple le protocole MMS (Multimedia Messaging Service).

– Le caractère non pérenne d'une majorité de ressources vidéo sous leur forme disponible en ligne. Elles sont souvent encodées dans des formats propriétaires (.ram, .wmv, .mov pour ne citer que les plus répandus)[14], donc menacées de ne plus pouvoir être relues dans les environnements informatiques futurs. Ce problème peut être résolu par des opérations de migration de masse (réencodage du contenu dans un format ouvert). Probablement plus grave est l'apposition de DRM (Digital Rights Management) sur les fichiers vidéo vendus en ligne, afin de limiter le nombre de copies et/ou leur durée d'utilisation. Si l'usage des DRM s'impose et qu'ils continuent à se sophistiquer[15], leur contournement à des fins de conservation (autorisé par la nouvelle loi sur le dépôt légal) pourra se révéler compliqué et coûteux. La récupération des fichiers avant apposition des DRM serait la meilleure option.

> Il n'est pas question d'élaborer une politique documentaire au sens classique, mais de définir des priorités de collecte, suivant une logique mixte : valeur documentaire intrinsèque et représentativité

À l'inverse, si riches soient les réservoirs de vidéos en ligne, tout un éventail de stratégies éditoriales est souvent mis en œuvre pour rendre visible l'existence des contenus : feuilletage, nuages de *tags*, communautés… Un moissonnage approfondi des sites permettra la mise à jour d'une grande partie des ressources (le reste pouvant être considéré comme jouissant d'une notoriété très minime). Le problème est donc d'ici peu celui du « web invisible ».

La nécessité de passer provisoirement ou durablement par la procédure du dépôt dépend essentiellement des trois problèmes spécifiques pointés précédemment. La problématique de l'archivage d'un site riche en vidéos va se trouver paramétrée par la localisation géographique du site (le dépôt, opération relativement lourde, se trouvant pratiquement exclu avec un site étranger), le protocole technique de diffusion (susceptible d'exclure l'option de collecte automatique, le temps d'une évolution des techniques de capture, et conduisant donc à rechercher pendant un certain temps des dépôts) et la capacité à pérenniser les fichiers diffusés par le web (le dépôt étant susceptible de donner accès aux mêmes ressources sous des formes plus aisées à pérenniser que ceux moissonnés lors de la collecte automatique).

Organiser un ciblage

Compte tenu de ces difficultés d'ordres multiples, le ciblage et l'évaluation de la collecte automatique sont essentiels. Il n'est pas question d'élaborer une politique documentaire au sens classique, mais de définir des priorités de collecte, suivant une logique mixte : valeur documentaire intrinsèque et représentativité. Rien de nouveau en soi : la prospection du dépôt légal des documents édités sur support s'opère déjà ainsi dans la pratique (implicitement ou explicitement), entre recherche d'exhaustivité et logique d'échantillonnage, suivant les secteurs.

D'un point de vue pratique, au sein du service Images du département de l'Audiovisuel de la BnF, une correspondante dépôt légal web[16]

12. Sur les stratégies d'archivage du web à la BnF, cf. Illien, Gildas ; Game, Valérie, « Le dépôt légal d'Internet à la Bibliothèque nationale de France : Cadre juridique, modèle de collecte, évolutions des métiers », *BBF*, 2006, nº 3, p. 82-85. http://bbf.enssib.fr, et http://www.bnf.fr/pages/infopro/depotleg/dli_intro.htm

13. Exception faite de systèmes de VOD verrouillant, via un système de géolocalisation, la possibilité d'acheter à partir de pays pour lesquels les droits de diffusion n'ont pas été négociés.

14. A contrario, les formats de fichiers .mpg ou .ogg sont complètement documentés.

15. Cela n'est pas tout à fait certain. Ainsi, les éditeurs phonographiques, après avoir été de grands adeptes des DRM, paraissent actuellement faire machine arrière après s'être aperçus qu'ils rendent impopulaire le téléchargement légal.

16. Des correspondants du DL web ont été désignés dans presque tous les départements de collections de la BnF.

se charge à la fois de l'actualisation des « signets » Internet du site bnf.fr dans le domaine cinéma-audiovisuel et de l'enrichissement de la liste des ressources à cibler pour l'archivage. Les signets sont construits selon une logique de signalement de référence : la bibliothèque qualifie et « labellise » des sites qui constituent des points d'entrée jugés fiables et relativement stables dans un paysage documentaire mouvant. Le dépôt légal d'Internet répond, lui, à une logique de miroir : il s'agit de collecter de vastes échantillons de la production, bonne ou mauvaise, durable ou éphémère, pour témoigner de ce qu'elle a été à un moment donné. Certes, les critères à appliquer diffèrent grandement entre signets et dépôt légal et le repérage des ressources existantes sur le web réclame souvent beaucoup de temps : il est donc rationnel que les deux tâches soient assurées par la même personne, qui devra évaluer ses découvertes sous deux éclairages distincts.

Les sites ciblés et ne se prêtant pas à la collecte automatique des ressources vidéo feront l'objet d'une démarche de dépôt – démarche déjà mise en œuvre pour certains sites à la BnF, mais encore à ses débuts pour les ressources vidéo et son. La détermination du jeu de métadonnées à livrer avec la ressource sera essentielle pour que la ressource vidéo archivée forme un document. Ces métadonnées doivent être interopérables et pérennes, donc converties dans des formats ouverts. On s'orientera dans un premier temps vers la récupération d'un jeu de données Dublin Core, sans perdre de vue, à terme, l'intérêt d'une « documentation » de la ressource plus riche (par exemple avec des données témoignant de leur appropriation : nombre de consultations, annotations…)[17].

Quels accès ?

La consultation des archives collectées est pour l'heure cantonnée à un usage interne à la BnF, situation appelée à évoluer après qu'auront été définies les conditions d'accès du public chercheur. De par la loi, l'accès à l'archive du DL Internet sera restreint aux enceintes de la BnF – même pour des documents en accès libre sur Internet.

Les possibilités d'identification des ressources archivées vont grandement évoluer avec leur prochaine indexation (elles ne sont aujourd'hui accessibles que par leur URL). Mais pour les contenus vidéo, l'heure n'est pas encore venue d'une indexation automatique structurée, à l'instar de ce qui existe pour les contenus textuels. Les outils d'indexation plein-texte qui vont commencer à être utilisés à grande échelle sur les archives du web à la BnF ne porteront que sur ces derniers. Avoir archivé, en même temps que les ressources vidéo, le signalement contemporain de leur présence sur le web (leur « documentation ») sera donc très précieux pour orienter le chercheur futur. En complément, compte tenu du relatif sous-signalement des ressources vidéo sur le web, il pourrait se révéler utile de traduire les connaissances acquises lors de la prospection des ressources existantes en guides pour les investigations futures.

Janvier 2007

17. Le versement dans l'archive numérique conduira à associer dans un second temps d'autres métadonnées : métadonnées de pérennisation (format de données par exemple), métadonnées juridiques (droits d'accès), etc.