# LA FRANCE ET L'EUROPE FACE À L'ENJEU DU BIG DATA

## L'exemple de la collecte et du traitement des données médiatiques

**Mathieu Gallet**

**Editions Choiseul |** *Géoéconomie*

**2014/2 - n° 69**
**pages 7 à 23**

**ISSN 1620-9869**

# Apartés

Mathieu GALLET

# La France et l'Europe face à l'enjeu du *Big Data*

## L'exemple de la collecte et du traitement des données médiatiques

*Mathieu Gallet est Président-directeur général de l'Institut national de l'audiovisuel (Ina). Il prendra la présidence de Radio France en mai prochain.*



Depuis plusieurs années, nous assistons à une croissance exponentielle de la production de données numériques de toutes sortes, s'accumulant en une masse gigantesque que l'on a baptisé *Big Data* : le développement massif de *data centers* aux capacités toujours plus grandes, ou l'essor rapide de pratiques comme le *cloud computing* sont autant de symptômes tangibles de cette nouvelle orientation de l'économie mondiale.

Prise isolément, chacune de ces informations n'a souvent qu'une valeur infime, voire nulle ; mais pour qui sait les collecter, les agréger, les interroger, les analyser, elles deviennent une clé de compréhension décisive des comportements des individus comme de la société dans son ensemble, faisant du *Big Data* un enjeu essentiel, tout à la fois économique, sécuritaire et, plus largement, sociétal, pour ne pas dire « civilisationnel ».

Trois grandes catégories d'acteurs se consacrent aujourd'hui dans le monde à la collecte et au traitement de données « en masse », suivant des logiques disparates : d'abord les acteurs du renseignement et de la surveillance d'État, comme l'a illustré encore récemment le

scandale autour de la *National Security Agency* (NSA) ; les géants de l'économie numérique, ensuite, qui, à l'instar de Google ou Facebook, se sont constitués un gigantesque patrimoine de données - notamment personnelles - qu'ils s'attachent à monétiser avec une efficacité toujours plus grande ; enfin, une troisième catégorie d'acteurs, autrement moins médiatisés : les institutions patrimoniales et scientifiques, qui se consacrent à l'archivage d'informations (le plus souvent publiquement accessibles) pour les générations futures.

Dans ces trois domaines, les chefs de file sont incontestablement américains. Et tandis que de nouveaux acteurs émergent progressivement dans la région Asie-Pacifique, l'Europe se maintient à une timide deuxième place, loin derrière les États-Unis : une position dont ne saurait se satisfaire un continent qui a fait de l'« économie de la connaissance » son cheval de bataille.

Dans le cadre de cet article, nous souhaiterions explorer un champ particulier de cette problématique du *Big Data*, où la France se trouve être particulièrement en pointe : celui de la collecte des données médiatiques. Nous entendons par là, au sens large, les informations circulant publiquement sur les réseaux numériques, qu'il s'agisse de textes, d'images, de vidéos ou d'enregistrements audio, générés tant par des médias traditionnels (chaînes de télévision ou de radio, journaux) que par des *pure players*, voire par de simples personnes privées (comme dans le cas de comptes Twitter ou de blogs, par exemple).

Comparées aux données personnelles – qui permettent de connaître les goûts d'un individu, de prévoir son comportement, de lui proposer une publicité adaptée ou de l'inciter à acheter un bien ou un service –, ces informations disponibles à tout un chacun pourraient sembler avoir une faible valeur économique. Leur traitement n'en soulève pas moins un enjeu décisif, car c'est en les archivant et en les interrogeant qu'une société démocratique peut progresser dans la connaissance qu'elle a d'elle-même et de son histoire.

D'où une série de questions essentielles : l'Europe est-elle capable d'exister dans ce domaine face aux États-Unis, de se forger ses

propres outils technologiques pour collecter, structurer et interroger les données numériques qu'elle génère ? En un mot : peut-elle rester maîtresse de sa propre mémoire ? De sa propre histoire ?

## Une mémoire d'Internet sous contrôle étatsunien ?

Inventeurs et pour une large part gestionnaires d'Internet, les États-Unis sont aussi les premiers à s'être préoccupés de son archivage. Non par le vote d'une grande loi fédérale sur le dépôt légal à l'ère numérique, mais à travers une initiative privée : Internet Archive.

Il s'agit en somme d'une histoire très américaine, à la frontière des mondes de la philanthropie et de l'entreprenariat. Elle commence sur les hauteurs de San Francisco en 1996, époque où le Web commence à atteindre une certaine masse critique. Le « père fondateur », Brewster Kahle, est un informaticien et entrepreneur visionnaire, qui a fait fortune en vendant ses parts des deux sociétés high-tech qu'il a cofondées : Wais, logiciel d'édition en ligne cédé à AOL en 1995, et Alexa, entreprise spécialisée dans l'analyse du trafic sur le Web mondial, acquise par Amazon en 1999. Il réinvestit cet argent dans la création d'Internet Archive, un organisme à but non lucratif (dit « 501 (c) non-profit »), statut qu'il a toujours conservé depuis. Cela ne signifie pas pour autant que son financement passe exclusivement par des donations : la *Digital Library*, comme elle se définit elle-même, tire une part de ses recettes – 15 millions de dollars en 2010 – de ses services payants d'archivage du Web[1].

Dès ses débuts, Internet Archive prend acte du caractère global du réseau : son champ de collecte sera le monde. L'objectif ? « Construire la bibliothèque d'Internet[2] ». Ou, pour reprendre le crédo mille fois répété de Brewster Kahle : « Donner un accès universel à l'ensemble du savoir ». Pionnier d'une aventure encore inédite, l'organisme va définir ses propres standards méthodologiques et techniques

1. Voir le site www.archive-it.org.

2. Voir le site https://archive.org/about/

en termes de collecte et de stockage. Avocat d'un Web ouvert – il s'est depuis illustré par sa critique de la mainmise de Google sur la numérisation des livres, au point de lancer une initiative concurrente – Brewster Kahle adopte une démarche souple et pragmatique, fondée sur une infrastructure informatique évolutive et l'utilisation chaque fois que possible de logiciels libres. Diffusés à travers l'*International Internet Preservation Consortium* (IIPC) – qui rassemble depuis 2003 la plupart des bibliothèques et centres d'archives actifs en la matière –, ces outils seront par la suite adoptés dans le monde entier, non sans susciter aujourd'hui une certaine dépendance à l'égard de logiciels dont la maintenance n'est pas toujours assurée de manière suffisante.

Près de deux décennies plus tard, Internet Archive revendique la sauvegarde de 388 milliards de pages Web (au 1er février 2014), pour un volume total de 10 pétaoctets. Des fonds d'une richesse unique, particulièrement pour les premières années, que l'organisme est le seul à avoir archivées. En France, tant la Bibliothèque nationale que l'Ina ont puisé dans ce fonds pour compléter leurs collections concernant la période antérieure à la mise en œuvre du dépôt légal dans l'hexagone, ce grâce à une extraction et une copie réalisées pour leur compte par Internet Archive.



Dès l'origine, Brewster Kahle s'est donné pour mission non seulement de collecter, mais aussi de rendre les résultats de ce travail accessibles au plus grand nombre. C'est le cas depuis la mise en ligne, en 2001, de la *Wayback Machine*[3] : une « machine à remonter le temps » permettant à tout un chacun de consulter les pages conservées par Internet Archive. On peut ainsi visiter le site de l'Élysée tel qu'il était le 12 février 1998 (jour de la première sauvegarde), ou encore se rendre sur celui du *New York Times* à la date du 11 septembre 2001 (quatre sauvegardes distinctes pour cette seule journée).

Comme chacun pourra s'en rendre compte en testant la *Wayback Machine*, l'expérience peut toutefois souvent se révéler décevante, car

3. Voir le site www.archive.org/web/.

l'affichage des pages se fait le plus souvent de manière imparfaite ou incomplète : des formats comme *Shockwave* ou *Flash* n'ont pas été archivés, des liens sont rompus, des pages dynamiques se figent, etc. Ce n'est pas faire injure à la prouesse technique réalisée par Internet Archive que de considérer ces insuffisances comme la conséquence d'une approche privilégiant une exhaustivité holistique à une démarche plus qualitative.

En d'autres termes : on aurait tort de penser que, puisqu'Internet Archive fait déjà le travail, il serait inutile de procéder à d'autres sauvegardes de la toile. Le Web est un objet si complexe et évolutif qu'un seul type de collecte ne saurait en épuiser la richesse. À cet égard, maintenir une certaine diversité dans les approches apparaît essentiel – notamment au profit de collectes aux « mailles » plus fines – sans même parler du danger qu'il y aurait à confier toute la mémoire de notre société numérique à un acteur unique, fût-il désintéressé.

## Les dépôts légaux du Web en Europe : une mosaïque d'initiatives

Il n'est pas aisé, dans le cadre de cet article, de donner une vision exhaustive des initiatives touchant à l'archivage d'Internet dans le monde. L'examen attentif de la carte des membres de l'IIPC[4] permet toutefois de donner un premier aperçu et d'identifier trois grands pôles.

L'Amérique du Nord, tout d'abord : aux États-Unis, plusieurs grandes bibliothèques et universités travaillent de concert avec Internet Archive, telles que la bibliothèque du Congrès à Washington, Harvard ou Columbia. De l'autre côté de la frontière, à Ottawa, Bibliothèque et archives du Canada (BAC) recueille depuis 2005 un échantillon représentatif de sites locaux, comme le fait également son homologue québécois. Dans la zone Asie-Pacifique, plusieurs

4. Consultable dans une version pour l'instant incomplète, à l'adresse http://netpreserve.org/timeline (cliquer sur « Map »). Voir aussi la liste complète des membres de l'IIPC sur http://netpreserve.org/about-us/members

initiatives se distinguent au Japon, en Corée, en Chine, à Singapour, en Australie et en Nouvelle-Zélande. Mais c'est en Europe que l'on constate le foisonnement le plus dense, avec environ vingt-cinq acteurs différents !

Si l'adhésion récente du Chili annonce peut-être une évolution, on peut remarquer au passage l'absence quasi-totale dans ce domaine, des pays dits « du Sud », et en particulier de l'Afrique, à l'exception notable de la bibliothèque d'Alexandrie, laquelle assure, pour des raisons éminemment symboliques, la conservation d'une copie de sécurité des données d'Internet Archive (« site miroir »).

Si la directive communautaire 2001/29/EC, transposée en droit français en 2006, recommande fortement la mise en place d'un dépôt légal du Web dans chaque pays à l'échelle nationale, il n'existe pas pour autant de cadre vraiment contraignant en la matière, d'où une mosaïque de situations particulières. Trois grands cas de figure peuvent toutefois être distingués : en premier lieu, les pays – somme toute minoritaires – où n'existe encore aucun dispositif de collecte significatif : c'est par exemple le cas de l'Italie, de l'Irlande, de la Grèce ou de la Hongrie. À l'opposé, on trouve les pays ayant introduit dans leur législation un dépôt légal du Web en bonne et due forme (dont les modalités peuvent bien sûr varier), le plus souvent assuré par la Bibliothèque nationale : on relève dans cette liste la présence de nations aussi diverses que la France, l'Estonie, ou plusieurs pays scandinaves, ces derniers faisant montre d'un dynamisme certain (la Suède a ainsi été le premier pays européen à initier une collecte, dès 1997).

Entre ces deux catégories, une zone « grise », celle des pays où n'existe pas forcément un dépôt légal du Web obligatoire et formalisé sur un plan législatif, mais où un ou plusieurs acteurs - qui peuvent aussi être des initiatives privées ou citoyennes - se consacrent néanmoins à l'archivage d'Internet. En Allemagne, par exemple, la structure politique fédérale se reflète dans la multiplicité des parties prenantes, dont la Bibliothèque nationale ; au Portugal, le *Portuguese Web Archive* dépend de la Fondation pour la science et la technologie du ministère de l'Éducation : autant de pays, autant de configurations possibles.

Quelques points communs, toutefois, à la majorité de ces dispositifs, qui ne se distinguent pas seulement d'Internet Archive par leur caractère principalement public. Tout d'abord, aucun ne vise à l'exhaustivité, chacun se définissant un périmètre de collecte, correspondant généralement aux frontière ô combien insaisissables du Web « national » : cela suppose nécessairement des « trous » dans le maillage (mais aussi, sans doute, un certain nombre de zones de recouvrement) à l'échelle de l'Europe. Par ailleurs, au contraire d'Internet Archive et de sa *Wayback Machine*, la consultation - quand elle est possible - n'est généralement prévue que dans un cadre restreint (dans la plupart des cas : réservée aux étudiants et chercheurs), les règles de la propriété intellectuelle, dans les législations tant européennes que nationales, empêchant la mise à disposition auprès du grand public des contenus ainsi collectés[5].

## Le dépôt légal du Web en France : un dispositif original

En France, le dépôt légal du Web se distingue par le fait qu'il est partagé entre deux acteurs : la Bibliothèque nationale de France (BnF) et l'Ina. Dès le tournant des années 2000, ces deux institutions patrimoniales, conscientes de cet enjeu de mémoire émergent, ont commencé à développer des dispositifs de collecte et de stockage, dans une démarche de dialogue et de concertation. Il faudra toutefois attendre la loi DADVSI du 1er août 2006 - complétée depuis par le décret d'application du 19 décembre 2011 - pour que le dépôt légal soit officiellement étendu aux « signes, signaux, écrits, images, sons ou messages de toute nature, faisant l'objet d'une communication au public par voie électronique ».

Déjà en charge depuis 1992, du dépôt légal de la radio et de la télévision, l'Ina est responsable de l'archivage d'un segment

5. Des cas d'accessibilité totale ou partielle en ligne existent cependant au Royaume-Uni, au Portugal, ainsi qu'en Catalogne, Croatie, Slovénie et Estonie.

particulier de l'Internet français : le Web lié au secteur audiovisuel. Un champ qui pourrait de prime abord sembler limité, puisqu'il concerne « seulement » 12 000 sites : certains services de médias audiovisuels à la demande, dits « Smads » (télévision de rattrapage, vidéo à la demande, web-TV et web-radio), mais aussi les sites officiels de diffuseurs (par exemple : le site d'actualité en temps réel francetvinfo.fr), de programme, ou encore les blogs et *fansites* consacrés aux émissions radio et télé. Compte tenu du « rafraîchissement » fréquent de ces pages, régulièrement alimentées en contenus nouveaux, et surtout de la place centrale qu'y tiennent les vidéos, particulièrement gourmandes en données, ces quelques milliers de domaines n'en représentent pas moins une part majeure du volume global de l'Internet français. Ce qui contribue à expliquer que l'archive constituée par l'Ina, forte de 217,9 téraoctets (2,3 pétaoctet avant compression et « déduplication ») et de 26,8 milliards de versions d'URL à la fin 2013, soit la deuxième au monde en nombre d'URL conservés, après celle d'Internet Archive.



Le dispositif mis en œuvre par l'Institut se caractérise par un effort constant de recherche et développement, afin notamment d'adapter les outils de collecte à l'évolution technique du Web dit « vivant », ainsi que la fréquence et la « profondeur » de la collecte selon les sites : celui d'une grande chaîne de télévision, mettant son contenu à jour en continu, ne saurait être traité de la même manière qu'un petit blog peu actif, accueillant un nouvel article tous les un à deux mois. Autre enjeu de taille, compte tenu des volumes concernés : limiter la redondance, afin, par exemple, de ne pas stocker x fois la même vidéo, diffusée par différents canaux. Le format de stockage DAFF (*Digital Archive File Format*) développé par l'Ina permet ainsi un facteur d'économie en stockage équivalent à 91 %. Grâce à ces outils développés en interne, en toute autonomie vis-à-vis d'Internet Archive, l'Ina est ainsi en mesure de s'adapter avec souplesse aux évolutions du Web, tout en assurant le maintien d'une comptabilité avec les autres formats d'archivage en vigueur dans le monde.

Lancés en 2009, des ateliers méthodologiques mensuels ont permis d'instaurer un dialogue constant avec la communauté des chercheurs et de construire ensemble un projet patrimonial qui demeure par essence un *work-in-progress*. Depuis 2011, les archives constituées sont accessibles au centre de consultation de l'Ina THEQUE, dans les six délégations régionales de l'Ina et dans plusieurs bibliothèques associées, afin d'assurer des accès démultipliés sur l'ensemble du territoire.

Quant à la BnF, elle est chargée – si l'on ose dire – du « reste », c'est-à-dire de l'archivage de la totalité du Web français, hormis les sites audiovisuels captés par l'Ina. À l'aide du robot Heritrix, logiciel libre initialement développé par Internet Archive, la bibliothèque procède ainsi une à deux fois par an à une collecte dite « large », portant sur des sites du domaine « .fr » dont la liste est fournie par l'AFNIC (Association française pour le nommage Internet en coopération). Cette opération est complétée par des collectes plus fréquentes et plus ciblées (par exemple, pour les sites de presse, mis à jour au moins quotidiennement).

Toute la difficulté reste de définir le périmètre du Web français, qui est évidemment loin de se limiter aux domaines en « .fr », mais concerne l'ensemble des sites édités par des acteurs hexagonaux. Lors d'une collecte réalisée en 2011 et concernant un peu plus d'un milliard d'URL, la BnF a ainsi collecté 63 % de « .fr », 27 % de « .com » et 10 % d'autres extensions (« .net », « .tv » etc.). Or, selon les estimations de l'AFNIC, l'Internet hexagonal était alors à 32 % en « .fr » et à 46 % en « .com ». Si les différentes méthodologies de décompte peuvent toujours être discutées, on voit néanmoins apparaître là des écarts significatifs, qui semblent suggérer qu'une part non négligeable du Web français (notamment des domaines en « .com ») échapperait encore à notre dépôt légal, obligeant le cas échéant le chercheur à se tourner vers d'autres sources comme… Internet Archive.

## Décrypter les données médiatiques : un enjeu économique et démocratique

L'exemple du dépôt légal du Web illustre la problématique de la collecte de masse de données, de leur indexation et de leur mise à disposition au public, dans un cadre plus ou moins restreint ou ouvert. Reste la question cruciale, celle d'élaborer des outils qui permettent de « faire parler » cette masse d'informations, de l'interroger et de l'interpréter. Par exemple : comment un sujet d'actualité est-il abordé ? Qui produit l'information ? La multiplication des supports garantit-elle la diversité ? Autant de questions qui revêtent un intérêt crucial pour les professionnels des médias, mais aussi pour la collectivité dans son ensemble, car c'est bien d'un véritable décryptage de la société en temps réel qu'il s'agit. À cet égard, les bénéfices à la fois économiques et démocratiques induits par une telle démarche justifient pleinement une action publique volontariste.



C'est précisément à ces interrogations que s'est attelé l'Observatoire transmédia (OTMedia), projet de recherche piloté par l'Ina avec six autres partenaires[6], sur la période mi-2011-2013. Cette plateforme de description, d'unification et d'analyse des actualités présente l'originalité de considérer l'écosystème médiatique dans son ensemble, sans se limiter à un seul type de contenu : une tâche complexe, qui a nécessité le concours tant de chercheurs en informatique et sciences humaines et sociales que de professionnels de l'information.

Riche de cinq millions de documents à la fin 2013, le corpus OTMedia agrège cinq grands types de sources, couvrant tout le spectre du paysage médiatique français : les programmes d'information de huit chaînes de télévision[7], ainsi que ceux de neuf

6. Ina (coordinateur), INRIA (Zenith), Syllabs, LIA, AFP, Université Paris 3 (CIM), LATTS

7. Les journaux télévisés de la mi-journée et du soir pour TF1, France 2 et France 3, ceux du soir pour Canal +, France 5, M6 et Arte, plus les chaînes d'information en continu, I-télé, BFM-TV.

chaînes de radio[8], quelque 1 800 sites Internet traitant de l'actualité capturés *via* leurs flux RSS (dont 70 sites ou portails de presse), les messages de 20 000 comptes Tweeter qualifiés, et enfin l'intégralité du flux des dépêches de l'Agence France presse (AFP), partenaire du projet.

Les enjeux technologiques du projet sont donc liés au volume, mais aussi à la diversité des sources d'information prises en considération. Il s'agit d'élaborer des référentiels de représentation homogènes des données, intégrant le traitement des modalités visuelles, sonores et textuelles. Dans un deuxième temps, la difficulté réside dans la mise en œuvre de différentes phases de fouille sur ces données automatiquement enrichies, potentiellement bruitées et incomplètes. Pour y parvenir, plusieurs outils ont été développés : recherches textuelle et visuelle (avec identification d'images similaires ou partiellement similaires), génération de courbes temporelles comparatives (portant sur des occurrences ou co-occurrences de termes ou de champs structurants), toutes ces interfaces communiquant entre elles de manière à enchaîner les tâches nécessaires à une analyse donnée.

Le champ d'application d'une telle plateforme est vaste. À titre d'exemple, une étude réalisée pour l'ONU-AOC a permis d'analyser le rôle de la thématique « immigration » lors de la campagne présidentielle de 2012 : l'augmentation sensible des occurrences autour de ce thème témoigne du fait qu'il s'est bien agi là d'un thème de campagne important, mais à la différence des élections de 2007 et 2002, on constate que les événements rapportés relevaient davantage de questions politiques que de sécurité : droit de vote des étrangers aux élections locales, immigration économique, gestion de l'immigration au niveau européen.

En rapprochant et comparant une masse de sujets sur un thème donné, un outil comme OTMedia permet également d'évaluer le taux

8. RTL, France Inter, Europe 1, RMC Info, France Culture, France Info, RFI, BFM et Radio Classique.

de reprise d'une source donnée – typiquement, une dépêche AFP, qui peut être livrée « telle quelle » ou légèrement modifiée – ou encore, inversement, la proportion de contenus originaux produits par un média donné. Une manière de réduire le foisonnement apparent des communications médiatiques à un nombre de messages plus réduits circulant de place en place, et ainsi de tester en grandeur nature la réalité de ce que Pierre Bourdieu appelait la « circulation circulaire » de l'information. Les analyses sont en cours d'évaluation.

Pionnier à bien des égards, OTMedia n'est toutefois pas le seul projet de recherche à s'être lancé dans de telles investigations. On ne sera sans doute pas surpris d'apprendre qu'il a un « concurrent » américain, baptisé News Rover. Porté par l'université de Columbia, ce consortium accueillant en son sein de nombreux journalistes vise à fournir aux rédactions un outil d'appréhension et de contextualisation globale de l'information à partir d'une analyse transmodale sur un choix de médias américains. Il propose notamment certaines fonctionnalités avancées, comme la reconnaissance de visage ou l'extraction de texte dans une image.



Et en Europe ? Le *Joint Research Centre* de la Commission mène depuis plusieurs années une initiative passionnante, le *Europe Media Monitor* (EMM). Accessible en ligne, cette plateforme collecte chaque jour 80 à 100 000 articles issus de 2 200 sources Web sélectionnées, qui alimentent ensuite plusieurs systèmes d'analyse automatisés permettant de dégager des tendances à court ou long terme. La grande force de ce projet est d'agréger en une même base des données en 50 langues différentes (dont les 23 officielles de l'Union), permettant ainsi une perspective vraiment internationale : une véritable performance, lorsque l'on sait que le seul nom de Barack Obama (ou Barack Hussein Obama) peut faire l'objet de plusieurs dizaines de variantes orthographiques dans différents idiomes ! À l'inverse des deux projets décrits plus hauts, l'EMM ne traite toutefois pas un corpus plurimédia : l'outil n'indexe que des données textuelles en ligne, ne donnant donc qu'un aperçu partiel du paysage médiatique européen et international. Nous manque donc encore l'outil qui permettrait

de réaliser, à l'échelle de l'Europe et au-delà, le travail initié pour la France par OTMedia.

Le développement de tels projets n'est toutefois pas sans susciter des interrogations quant au cadre juridique, tant communautaire que français, concernant la question du *data mining*. En effet, « fouiller » les données suppose de les extraire et de les copier : des opérations susceptibles d'entrer en conflit avec le droit des auteurs de ces documents ainsi qu'avec les droits *sui generis* des producteurs de base de données. La directive 2001/29/CE - dont il a été question plus haut - prévoit certes une exception dite de « copie provisoire », mais sa rédaction apparaît fort restrictive[9].

Aux États-Unis, à l'inverse, l'exception du *Fair Use* permet à des acteurs comme Google de faire du *data mining*, y compris à des fins commerciales, sans risque d'être inquiétés. Alors que la Commission européenne vient de lancer une consultation publique sur la révision des règles de l'Union européenne en matière de droit d'auteur, en vue de la rédaction d'un livre blanc sur la question, ne faudrait-il pas remédier à cette « asymétrie législative » en créant une exception spécifique aux droits de propriété intellectuelle pour permettre la fouille de données ?

## Conclusion

À travers l'exemple de l'archivage d'Internet et de l'analyse des données médiatiques, on perçoit un réel dynamisme européen : un foisonnement d'initiatives, souvent d'une haute qualité scientifique et technique, associé à une conception largement partagée du rôle de

9. Directive 2001/29/CE, relative à l'harmonisation du droit d'auteur : « Lorsque l'œuvre a été divulguée, l'auteur ne peut interdire : la reproduction provisoire présentant un caratère transitoire ou accessoire, lorsqu'elle est partie intégrante et essentielle d'un procédé technique et qu'elle a pour unique objet de permettre l'utilisation licite de l'œuvre ou sa transmission entre tiers par la voie d'un réseau faisant appel à un intermédiaire ; toutefois, cette reproduction provisoire qui ne peut porter que sur des œuvres autres que les logiciels et les bases de données ne doit pas avoir de valeur économique propre. »

l'action publique dans ces domaines, garante de neutralité. Mais face aux géants américains (et Internet Archive en est un, à sa manière), les acteurs européens se distinguent aussi par leur dispersion et leur manque de coordination.

En ce début d'année 2014, la Commission européenne vient justement de lancer son nouveau programme de recherche « Horizon 2020 », doté de 80 milliards d'euros sur sept ans. C'est dans ce contexte que doit être imaginé une politique de R&D européenne plus intégrée et plus internationale, porteuse de projets forts. Parmi les enjeux sociétaux prioritaires identifiés par « Horizon 2020 » : la construction de « sociétés réflexives » (*reflective societies*), en d'autres termes : capables de réfléchir sur elles-mêmes. Ne sommes-nous pas là au cœur de l'enjeu de la collecte et du traitement des données médiatiques ? Construire ensemble un observatoire transmédia européen serait assurément un projet emblématique de cette nouvelle voie à suivre.




## Résumé

*Alors que la production de données numériques connaît une croissance exponentielle, leur collecte et leur analyse est devenu un nouvel enjeu stratégique. Le présent article se propose d'explorer un versant souvent méconnu de cette problématique du* Big Data *: la collecte et le traitement des données médiatiques, à des fins patrimoniales et scientifiques. L'Europe est-elle capable d'affirmer son autonomie dans ce domaine à l'égard des États-Unis, de constituer et d'interroger sa propre mémoire numérique ? Face au « géant » Internet Archive, fondation américaine qui archive le Web mondial depuis 1996, l'Europe présente un foisonnement d'initiatives disparates, depuis l'instauration d'un dépôt légal d'Internet jusqu'à des initiatives privées ou citoyennes. En France, l'Ina, chargé par le législateur du dépôt légal du Web audiovisuel, est l'un des rares acteurs à avoir développé une recherche et développement autonome de celle d'Internet Archive. Analyser et interpréter la masse d'informations ainsi collectée s'affirme aujourd'hui comme un enjeu à la fois économique et démocratique, en cela qu'une telle démarche*

*permet de « lire » notre société en temps réel et de la décrypter. Plusieurs initiatives innovantes – dont l'Observatoire transmédia développé par l'Ina – ont récemment apporté des percées dans ce domaine, mais la politique de recherche au niveau européen reste encore insuffisamment coordonnée et structurée en la matière.*

## Abstract

*With digital data production growing exponentially, data collection and analysis have become a new strategic challenge. This article sets out to explore an often neglected aspect of the "Big Data" issue: the collection and processing of media data for heritage and scientific purposes. Is Europe capable of asserting its autonomy in this field in relation to the United States, and of constituting and retrieving its own digital memory? Compared with the giant that is the Internet Archive (the American foundation that has been archiving the worldwide web since 1996), Europe today has a profusion of disparate initiatives, ranging from the establishment of a legal deposit system for the Internet to private or citizens' initiatives. In France, Ina, placed in charge of the legal deposit for the audiovisual media web by the legislator, is one of the few players to have developed a research and development capability that is autonomous in relation to Internet Archive. Analysing and interpreting the mass of data collected in this way is today clearly a challenge that is both economic and democratic, as this kind of approach makes it possible to "read" and decipher our society in real time. Several innovative initiatives – including the OTMedia project developed by Ina – have recently triggered breakthroughs in this area, but research policy at European level is still insufficiently coordinated and structured in this field.*