# *Enjeux numériques*

## Usages et archivages des contenus du Web

UNE SÉRIE DES
ANNALES
DES MINES
FONDÉES EN 1794

N° 10 - JUIN 2020

*Publié avec le soutien de l'Institut MinesTélécom*

# Usages et archivages des contenus du Web



***Ce numéro est coordonné par Michel SCHMITT***

# Introduction

Par **Michel SCHMITT**
Conseil général de l'Economie (CGE)

Il n'est pas si loin, le temps des disquettes informatiques souples qui avaient une capacité de 128 kilo-octets… Ceci fait sourire les anciens, car aujourd'hui, Internet et le *cloud* ont balayé ces dispositifs, en nous donnant la sensation que nous disposons d'un espace infini pour y déposer nos contenus. De surcroît, une infinité de contenus sont à notre disposition : des sites nous proposent des bibliothèques de films, de musiques ou de livres en nombre vertigineux.

Cette profusion de ressources sur la toile a modifié nos comportements. Recherche-t-on une information, une définition ? Wikipédia sur son *smartphone* que nous avons à portée de (ou dans) la main. La réponse est instantanée, alors que se déplacer jusqu'à sa bibliothèque, chercher son dictionnaire préféré et tourner les pages prend bien plus de temps. Et puis, l'édition de ce dictionnaire date de quelques années, rendant son information parfois un peu datée, et, le comble, le texte n'est pas indexé. Cette anecdote montre que la quantité [(1)], la mise en réseau et l'immédiateté ont changé notre rapport aux connaissances et ont forgé de nouveaux usages.

Par ailleurs, nous participons tous à accumuler les connaissances sur la toile, au travers des réseaux sociaux, des encyclopédies participatives, par exemple. L'une des qualités paradoxales de cet infini est sa taille humaine puisque quelques clics permettent d'accéder à tout document. Comme toujours, les choses ne sont pas si simples, et l'objet de ce numéro des *Annales des Mines* consacré aux « Usages et archivages des contenus du Web » est d'examiner comment notre relation aux contenus se modifie progressivement, comment cet infini s'organise, se structure et crée une nouvelle économie, faisant émerger de nouveaux concepts et de nouveaux métiers. Se pose également la question de garder une trace de ces contenus.

## Les usages, ou notre relation aux contenus

La révolution numérique a profondément modifié notre accès à la connaissance et de nombreux usages en ont été transformés ou ont émergé. En cette période de confinement, tous les parents pensent à assurer la continuité de l'éducation de leurs enfants. Les MOOC permettent l'accès aux enseignements de manière simple, sans se déplacer, et la classe, géographiquement éclatée, se reforme grâce à un réseau social dont le professeur devient modérateur. Le livre numérique, qui a cependant quelques difficultés à s'imposer, rend obsolètes les mètres linéaires d'étagères ou la lourde caisse pour partir en vacances. De même, les jeux en réseau ont transformé le quotidien de nos jeunes… et des moins jeunes également.

Côté producteur de contenus, la course à la visibilité est engagée. De nouveaux métiers – dans le sens d'une source de revenus – sont apparus, tels les *youtubeurs*, les *influenceurs*. Des sociétés de service proposent de faire mieux référencer les sites Web. Les *cookies* permettent d'adapter les propositions de contenus aux utilisateurs. Ainsi, deux personnes faisant la même recherche n'obtiendront plus le même résultat, résultat qui sera influencé par le comportement passé sur la toile. Ce qui frappe,

(1) Il ne faudrait pas en conclure que l'on trouve tout sur le Web ! J'avais en projet de faire en 2018 une leçon inaugurale de probabilités à mes étudiants aux Mines de Paris. On trouve, certes, des introductions au sujet, mais pas une leçon digne de ce nom sur les sites Web des grandes universités. J'ai donc dû prendre mon bâton de pèlerin pour la construire… et elle se retrouvera sur la toile prochainement.

c'est que le monde se rétrécit grâce au Web, rend proches les unes des autres les personnes, et ce rapprochement est tel que, dans cet univers virtuel, nous ne voyons plus notre voisin !

## Structurer, ou comment s'y retrouver

Côté utilisateur, comment s'orienter dans cette jungle de contenus ? C'est comme si l'on devait faire le point dans l'océan des données. Autrefois, les hauturiers, afin de naviguer en haute mer (par opposition aux caboteurs qui restaient toujours en visibilité des côtes), ont dû inventer de nouveaux outils, et c'est ainsi que sont apparus les sextants, les éphémérides et les horloges précises qui ont permis de faire le point indépendamment de repères terrestres. Les outils pour se diriger sur le Web sont les moteurs de recherche qui nous facilitent bien la tâche. Les premiers moteurs cherchaient des occurrences de mots dans les pages Web. Ils emploient maintenant des formules et des algorithmes plus complexes, certains issus de l'intelligence artificielle, pour parcourir, indexer, sélectionner et trier les résultats qu'ils présentent. Nous assistons en parallèle à des évolutions de la toile dans de multiples directions : le Web social (Web 2.0) qui nous connecte toujours plus, le Web sémantique (3.0) qui rend le Web plus accessible aux machines et applications, le Web des objets qui inclut jusque dans nos applications les objets connectés aux réseaux, etc. Et dans cette expansion galopante, nous examinerons plus particulièrement le Web sémantique, qui a pour rôle de garder l'ensemble des ressources que nous versons au Web universellement compréhensibles et utilisables. Il est cependant à noter que ces évolutions ne disent rien sur la qualité des contenus eux-mêmes [(2)].

## Aspects économiques et émergence des GAFAM

Derrière tout cela, se jouent des enjeux économiques colossaux. Les plus grandes capitalisations boursières sont le fait des géants de l'Internet, les GAFAM… Nous regarderons dans ce numéro les stratégies des diffuseurs de contenus, en tête desquels Netflix, YouTube et la manière dont le droit d'auteur évolue, puisque dupliquer un contenu est techniquement le fait d'un clic… Son coût marginal est quasi nul. Sans oublier que plus de la moitié du flux de données vers l'utilisateur est constituée de vidéo…

## Archiver le Web ou le Web comme archive ?

Qui n'a pas été agacé par la disparition d'une page Web ? Si l'on regarde le Web comme une immense bibliothèque de contenus, se pose la question de son instabilité et, par suite, de son archivage. Plusieurs projets ont vu le jour, mais ils concernent avant tout le patrimoine – les livres, les objets, les monuments, etc. – et l'on est plus dans une démarche d'archiviste, dont le métier se transforme. Nous sommes toujours dans l'angoisse d'une perte irrémédiable, comme autrefois face à l'incendie d'une bibliothèque ou d'un monument. Là encore se posent de nouvelles questions : que signifie archiver un livre ou un écrit ? Comment le rendre accessible ? Que devient le dépôt légal lorsque l'écrit n'est plus que numérique ?

De plus, de nouveaux types de contenus ont vu le jour. Par exemple, que vont devenir les contributions dans les réseaux sociaux ? Chaque minute, 50 000 nouvelles photos sont postées sur Instagram, 500 000 nouveaux tweets sont diffusés. *Quid* des logiciels qui ont été développés ? Voici quelques angles d'attaque que nous examinerons dans ce numéro.

(2) L'immédiateté de l'accès est l'une des caractéristiques de ces contenus numériques. L'édition scientifique est un bel exemple du bouleversement des usages. Ecrire un article scientifique, le soumettre à une revue, le faire *reviewer* (relire), le corriger et enfin l'insérer dans la revue… : ce processus peut prendre plusieurs années. Il est censé garantir la qualité de la recherche, et c'est sur lui que repose la bibliométrie. Cependant, ceci est en contradiction avec l'immédiateté d'Internet, et puisque chacun peut déposer ce qu'il veut – neutralité du Net oblige –, les archives ouvertes ont fleuri. La contrepartie est que c'est l'utilisateur qui doit juger de la qualité.

# Mutation des écosystèmes de médias

Par **Olivier BOMSEL**
MINES ParisTech, PSL Research University

Fin mars 2020, la France est confinée. Les règles de cette urgence inédite sont fixées au jour le jour par le gouvernement. Les médias numériques, dont certains sont redevenus gratuits, relaient ces instructions et les nouvelles du monde. Pour trois milliards de reclus sur la planète, ce sont les seuls moyens de socialisation et de délassement. Mais que sont les médias ? Quelles fonctions remplissent-ils ? En quoi la numérisation les a-t-elle transmués ? Comment celle-ci a-t-elle, par exemple, changé le cadre économique et institutionnel de la musique, des *news*, du cinéma ?

## Que sont les médias ?

Les médias ont d'abord été perçus comme des systèmes techniques de communication. Paul Starr, qui en décrit l'histoire (Starr, 2003), la fait commencer à l'imprimerie et cible sa vocation comme le partage de l'information, voire de la connaissance. Pourtant, avec le numérique, cette définition s'écroule car, non seulement, toute l'information converge sur un système unique, Internet, mais avec le codage binaire, c'est la notion même d'information qui change de nature. L'*information* ne désigne plus la mise en forme signifiante, mais toute chaîne de *bits* pouvant circuler sur un réseau (Varian et Shapiro, 1999).

Dès lors, il faut reconstruire l'opération symbolique qui constituait la *mise en forme signifiante* évoquée dans le mot *information*. Cette opération s'effectuait à travers deux protocoles différents : celui de correspondance, où le sens se construit dans l'échange entre interlocuteurs identifiés, et celui de publication, dans lequel un émetteur s'adresse à des récepteurs indistincts (Bomsel, 2010).

Dans la correspondance, les agents, qu'ils soient hommes ou machines, s'identifient et acceptent (ou non) d'échanger pour un bénéfice mutuel. Le sens s'élabore progressivement à cette fin. L'échange s'arrête quand son utilité – la perspective de gain – s'annule. De nombreux systèmes techniques ont concouru à ce protocole : la voix, la tablette et le calame, le courrier, le télégraphe, le télex, le téléphone, le fax, etc. La correspondance a le plus souvent un usage transactionnel, intime, diplomatique, commercial. Elle a vocation à rester secrète, à ne pas être dévoilée.

A l'inverse, la publication organise un dévoilement. Un émetteur adresse un message à des récepteurs indistincts. Le message est préparé à l'avance, généralement dans le secret. Il est ensuite dévoilé et diffusé dans le public. La publication produit un « effet de sens », une mythification qui s'ajoute au récit du message publié. Cet effet traduit l'événementialité du dévoilement, le fait que le monde change d'état entre « avant » et « après ». Il dépend d'une part de l'identité de l'émetteur – de celui qui parle, et d'autre part du *protocole éditorial*, du processus de dévoilement et de diffusion choisi. Les médias sont des opérateurs de publication (Bomsel *et al.*, 2013). L'effet de sens qu'ils produisent est à ce point puissant que Barthes (1957) appellera « sémiologie » l'exégèse des *mythologies* de la médiatisation. Et fera dire à McLuhan (1964), autre pionnier des médias, que le message, c'est le médium lui-même.

Les médias sont les outils de protocoles éditoriaux dont la vocation est de construire l'image, la représentation sociale d'une institution, d'un bien, d'un service, d'une expression, d'un récit… Ces protocoles sont d'une immense variété symbolique et technique. Les Sumériens édifient des monuments, des outils de monstration, pour signifier le hiératisme et la pérennité de leurs

Détail de la stèle du Code de Hammurabi, v. 1750 av. J.-C, Musée du Louvre.

règles sociales – le code de Hammurabi (Glassner, 2013). Alfred Sloan, patron charismatique de General Motors, institue le *design* pour éditer l'innovation technique de ses voitures (Sloan, 1963). Steve Jobs invente le *keynote* pour mettre en scène le génie de ses inventions (Isaacson, 2012). Les industries créatives sont plus que d'autres des industries de monstration : *show business*. Elles recourent aux premières, aux vernissages, aux défilés, aux *showcases*, aux festivals, à la circulation des épreuves, pour ritualiser l'apparition, le statut et la mise en marché des œuvres. Quant à Donald Trump, il utilise le tweet et les lettres capitales pour impulser l'emphase de sa médiatisation.

Les médias combinent ainsi une dimension sémantique, institutionnelle et économique. Sémantique en raison de l'effet de sens et de sa portée. Institutionnelle en ce qu'ils publient des règles sociales, en régulent la légitimité et le rayonnement international. Économique car ils sous-tendent le fonctionnement des marchés et des firmes, tout en étant eux-mêmes structurés en industrie et consommés sur des marchés. Pour cette raison, leur analyse convoque la plupart des disciplines de sciences humaines.

## Organisation industrielle des médias

Les médias fabriquent la médiatisation, autrement dit la représentation sociale d'institutions, d'organisations, de personnes, de faits sociaux, de marchandises, de récits… Cette opération dépend du protocole éditorial retenu. Elle engage plusieurs organisations qui concourent à construire et à diffuser la représentation sociale, le complément médiatique de la chose publiée. La médiatisation de l'automobile, par exemple, passe par la presse spécialisée, les salons, les sports mécaniques, la publicité sur les médias classiques et en ligne, l'exposition en concession, mais aussi le *design* des véhicules en circulation, etc. Celle de la musique et de ses artistes combine le concert, la radio, la télévision, le disque, et, désormais, les médias sociaux, les plateformes de *streaming*. Ces médias sont complémentaires et interagissent pour créer et diffuser l'image de chaque produit, voire, dans le cas de la musique, le produit lui-même. Ces interactions, que les économistes appellent des « externalités croisées », interviennent au sein d'écosystèmes, de grappes d'organisations adjacentes liées par des arrangements formels et informels.

Les écosystèmes de médias sont associés à des protocoles éditoriaux spécifiques. Les biens informationnels, notamment ceux fondés sur le copyright, issus des industries créatives ou culturelles, suivent des protocoles éditoriaux originaux. La publication d'expressions artistiques ou littéraires obéit à des codes qui identifient les auteurs, les cautionnent à travers des marques éditoriales et des outils de monstration. Ces protocoles étaient jadis associés à des supports matériels dont l'économie encadrait la publication : le livre, le film, le disque, le journal, la radio, etc. Ils ont été les plus affectés par la numérisation qui a touché aussi bien la production et la distribution des biens, que les écosystèmes sous-tendant leur médiatisation. En outre, l'utilité des biens sous copyright repose sur le sens, le récit et l'expérience de la perception. La publication modifie le sens des expressions dévoilées par des effets d'apparition, de statut, de contexte. En conséquence, la numérisation bouleverse les industries du copyright, non seulement à travers la production et la distribution, mais aussi à travers l'utilité sémantique de leurs produits. C'est sur elles que cet article va désormais se concentrer.

Pour simplifier, on peut considérer que les industries culturelles (y compris la presse ou les *news*) ont pour vocation de produire des récits. La musique ou les jeux vidéo entrent aussi dans ce cadre dès lors qu'on considère que la séquence des émotions qu'ils procurent a une dimension narrative. C'est ce qu'ont bien compris les médias sociaux en introduisant des fils, des séquences de récit qui fidélisent les utilisateurs à une histoire dont ils sont les héros. Ces récits, comme le suggère le théoricien du roman Mikhaïl Bakhtine (1978), sont caractérisés par des *chronotopes*, des espaces-temps qui identifient un genre, un contexte structurant la perception et l'utilité du consommateur.

Les *news*, le sport, la fiction audiovisuelle ou littéraire, la musique, classique ou de variété, le jeu vidéo sont autant de *chronotopes* identifiés.

Leurs écosystèmes historiques étaient fondés sur des systèmes techniques de production et de distribution ainsi que sur des médias de résonance facilitant la promotion des produits. Ces écosystèmes étaient structurés par les caractéristiques économiques originales des biens culturels. Résumons-les rapidement :

- biens informationnels, non rivaux (sans rareté), aux coûts de reproduction, de radiodiffusion ou de distribution numérique quasi nuls ;
- biens d'expérience, qu'on ne connaît qu'après les avoir consommés, exigeant une forte signalisation pour structurer l'espérance d'utilité du consommateur ;
- générateurs d'effets de réseau (effets de mode, d'appartenance culturelle) et d'addiction, où la consommation de récits fractionnés (*breaking news*, feuilletons, séries) accroît l'utilité des épisodes suivants... ;
- s'adressant à des préférences très diverses et, pour cela, devant être tarifés non pas au coût marginal, mais à l'utilité marginale des consommateurs, impliquant des schémas de discrimination tarifaire (vente groupée, *versionnage*) propres à chaque écosystème.

Les technologies numériques et l'essor des médias sociaux ont profondément bouleversé les structures exploitant ces caractéristiques économiques. Les écosystèmes médiatiques sont alors entrés en mutation, avec, en conséquence, une transformation des incitations à la création d'œuvres et de récits.

## Mutation des écosystèmes industriels

### La musique

Dans le vieil écosystème du disque, la radio diffusait gratuitement les nouveautés pour créer des effets de mode et inciter le consommateur à acheter l'album. Lequel était à la fois un bouquet de titres et un contexte identifiant l'univers de l'artiste. Il donnait lieu à des tournées, des concerts étendant la notoriété et la perception de son registre. C'est cet écosystème qu'a changé d'abord le piratage lié au déploiement d'Internet et des médias sociaux, puis l'adoption du *streaming* redonnant aux ayants droit des moyens de recettes.

Avec l'écosystème du *streaming,* le modèle tarifaire de la musique a radicalement changé. Le consommateur n'achète plus les albums des artistes qu'il préfère, souvent un achat d'impulsion, mais paie un abonnement pour des millions de titres. Les artistes ne sont plus rémunérés à la vente d'albums, mais au prorata du nombre de clics enregistrés sur les plateformes, autrement dit à la part de marché de chaque titre en *streaming.* Ainsi, l'artiste qui pouvait vendre son disque à l'utilité marginale du consommateur reçoit désormais un montant fixe pour l'écoute de chaque titre, indépendamment des goûts de l'auditeur. Le prix du clic [1] est fixé par le chiffre d'affaires de la plateforme divisé par le nombre total de clics. Plus encore que l'ancien système, ce modèle tarifaire encourage les gros hits et le matraquage promotionnel visant les grands cliqueurs, autrement dit la jeunesse. Or, celle-ci a un pouvoir d'achat limité mais une part d'audience très forte. La rémunération se fait alors au détriment de la diversité prisée par les mélomanes aux goûts plus éclectiques et au consentement à payer supérieur.

(1) Pour autant que l'écoute dépasse trente secondes.

## Les *news*

Les *news* désignent ici ce qu'on appelait jadis les nouvelles ou l'information (voir plus haut). L'écosystème des *news* était initialement constitué de la presse écrite, quotidienne et hebdomadaire, à quoi se sont successivement ajoutés la radio, la télévision, les chaînes d'information en continu, la presse en ligne, les médias sociaux... Ce faisant, le *chronotope* des *news* a changé. Du temps de la presse écrite, le journal rapportait les nouvelles de la veille. La radio et la télévision ont élargi l'espace et raccourci la périodicité des récits. Avec les chaînes d'info en continu, la presse en ligne et les médias sociaux, le *chronotope* des *news* est le monde au présent. En outre, leur protocole éditorial qui s'appuyait sur des acteurs garantissant la fiabilité des sources a muté vers un processus plus complexe laissant un très grand nombre d'émetteurs propager des rumeurs et des fausses nouvelles. En cause, le développement des médias sociaux capables d'inscrire la *news* dans le récit personnel de chaque membre d'un réseau.

Ainsi, la numérisation des *news* ne se limite pas à la dématérialisation de la presse ou à l'obsolescence croissante de la radio, mais à une mutation profonde de tout l'écosystème qui construisait et diffusait les récits d'actualité. Cette mutation fait surgir un écosystème dans lequel les éditeurs traditionnels ne renoncent pas à leur support, mais se redéploient en ligne en intégrant de nouveaux récits (vidéos, *podcasts)* et en usant des médias sociaux pour accroître leur audience. Là encore, cette mutation bouleverse les modèles tarifaires, tant dans l'accès aux *news* en ligne que dans le partage des externalités croisées avec les médias sociaux. Le risque des grands titres de presse est de voir leur pouvoir contextuel – issu de la qualité journalistique de leurs articles – s'éroder au profit des médias sociaux. De là l'introduction de nouvelles règles de copyright devant permettre aux éditeurs de *news* de licencier leur création.

## Le cinéma

Le cinéma est le medium historique de la fiction audiovisuelle. Son organisation industrielle s'est structurée en un écosystème – studios, presse spécialisée, *star system,* festivals, académies – valorisant l'exploitation du film en salle. L'essor de la télévision a fait surgir d'autres formats, d'autres récits, parmi lesquels le cinéma a trouvé à s'insérer. Ce second média a initié une exploitation *versionnée* des films, une discrimination tarifaire fondée sur l'exploitation successive des films en salle, puis à la télévision. L'apparition de nouveaux médias tels que la télévision payante, la VHS et le DVD a élargi progressivement ce modèle. En France, cette dynamique a sous-tendu l'organisation institutionnelle et industrielle de la télévision et déployé autour d'elle un vaste écosystème de production, de distribution et de médiatisation. Dans cet écosystème, le cinéma a bénéficié d'une place privilégiée visant à rehausser l'offre de télévision. Mais avec la multiplication des chaînes, cette offre s'est banalisée au point d'atteindre aujourd'hui 6 000 films programmés par an.

En outre, la télévision est un medium fondé sur la synchronisation, autrement dit sur l'organisation du temps du spectateur (Ortoleva, 2013). Tant qu'elle a été le mode exclusif de l'accès domestique à l'image animée, elle a pu agréger dans ses programmes des récits aux *chronotopes* très divers : *news*, météo, jeux, sports, variétés, documentaires, spectacles vivants, films, fictions, etc. Les grilles de programme se fondaient sur des carrefours d'audience offrant au spectateur des choix de soirées composites. L'essor d'Internet a, au contraire, favorisé une consommation asynchrone permettant le visionnage individué des récits au passé, notamment des films et des fictions. La vidéo à la demande (VOD) a brisé la logique de rendez-vous de la télévision pour proposer une consommation à volonté de fictions sérielles et addictives. Le format narratif de la série a pris le pas sur celui, plus éclaté, du cinéma, lequel est devenu un produit de complément. Enfin, la diffusion de la VOD sur l'Internet ouvert a contourné l'accès contrôlé des chaînes de télévision et les règles qui s'y appliquaient. Au final, c'est l'ensemble de l'écosystème du cinéma et de la télévision qui se

trouve ainsi contesté. Le phénomène s'amplifie d'autant que le coronavirus paralyse l'exploitation en salle. Aux États-Unis, les studios s'adaptent bon an, mal an à ce nouveau paradigme en déployant des plateformes de VOD. En France, la modification de règles institutionnelles rigides se heurte aux intérêts catégoriels des agents, rendant toute réforme impopulaire. Au final, la destruction créatrice issue du nouveau paradigme recomposera le vieil écosystème (Bomsel, 2017).

## Bibliographie

BAKHTINE M. (1978), *Esthétique et théorie du roman*, Paris, Gallimard.

BARTHES R. (1957), *Mythologies*, Le Seuil.

BOMSEL O. (2010), *L'Économie immatérielle*, NRF Essais, Gallimard.

BOMSEL O. *et al.* (2013), *Protocoles éditoriaux*, Armand Colin.

BOMSEL O. (2017), *La Nouvelle Économie politique*, Folio, Gallimard.

GLASSNER J.-J. (2013), « Premières institutions de l'écrit », *Protocoles éditoriaux*, *op. cit.*

ISAACSON W. (2012), *Steve Jobs*, J.-C. Lattès, Paris, pour la traduction française.

McLUHAN M. (1964), *Pour comprendre les médias*, Toronto, Points essais, 1968 pour la traduction française.

ORTOLEVA P. (2013), « Télévision : l'anti-édition ? », *Protocoles éditoriaux*, *op. cit.*

SLOAN A. P. (1963), *My Years with General Motors*, Doubleday.

STARR P. (2003), *The Creation of the Media*, Basic Books, New York.

VARIAN H. & SHAPIRO C. (1999), *Information rules*, Harvard Business School Press, Boston, Ma.

# Nouvelles plateformes entre télévision et cinéma : quelles mutations en cours ? Quels impacts sur les contenus ?

Par **Valery MICHAUX**
Neoma Business School

Partout dans le monde, le secteur de la SVOD (*Subscription Video On Demand*) est en train d'exploser avec une croissance qui a plus que triplé en cinq ans (entre 2014 et 2019 [1]). Cette croissance, qui va encore s'accentuer avec la pandémie du coronavirus, transforme profondément les secteurs du cinéma et de la télévision.

## La SVOD a réinventé les modèles économiques d'accès à la télévision

Quand les plateformes de SVOD ont commencé à se développer aux Etats-Unis dans les années 2010, et en France à partir de 2014, on parlait de services OTT (*Over The Top*). L'expression symbolisait l'idée que des acteurs, tels que Netflix, passaient au-dessus des opérateurs de réseau traditionnels (câble, satellite, ADSL) et de leurs programmes ou bouquets de chaînes qu'ils proposaient par défaut. Ce phénomène de désintermédiation permis par Internet, c'est-à-dire de contournement des intermédiaires en contact avec les clients finaux, émerge dès la fin des années 1990 (Leebaert, 1999) et touche progressivement depuis cette date, tour à tour, tous les secteurs de l'économie en instaurant un mode de distribution *DTC* (*Direct to Consumer*). Netflix a donc ouvert la porte à une désintermédiation du monde de la télévision et a eu pour conséquence un déclin rapide et régulier jusqu'à nos jours de ce qu'on appelle la télévision payante. Aujourd'hui, cette distinction entre *OTT* et télévision payante devient d'ailleurs caduque à mesure que les anciennes télévisions payantes (AMC aux Etats-Unis ou Canal+ en France, par exemple) sont en train elles-mêmes de devenir des plateformes en mode *DTC* accessibles *via* Internet. Mais, au-delà de la seule télévision payante, l'impact des plateformes de *streaming* sur les habitudes de consommation a favorisé l'émergence de comportements nouveaux (liberté de regarder ce que l'on veut quand on veut, *binge-watching* ou consommation compulsive de séries, etc.) qui ont détourné les spectateurs de la télévision linéaire entrecoupée de publicités (Plothe et Lang, 2020). Les chaînes de télévision gratuites vont alors progressivement basculer aussi dans le modèle des plateformes (par exemple, TF1 en France) et la distinction entre télévision linéaire (*live*) et télévision de rattrapage (*replay*) va émerger.

Aujourd'hui, les frontières se brouillent encore entre plateformes et télévision avec l'émergence de trois modèles économiques complémentaires accessibles *via* Internet. Le modèle dominant reste les plateformes sur abonnement (SVOD). C'est ce modèle qui a initié la mutation du secteur. Mais en parallèle se développe aujourd'hui le modèle appelé TVOD (*Transactionnal Video On Demand)* qui correspond à l'accès à des contenus à la carte, payables au cas par cas, et qui tend à remplacer la traditionnelle vente de vidéos à la demande. Enfin, émerge également le modèle AVOD (*Advertising-based Video On Demand*) qui correspond à des plateformes

(1) $12.5 billion to $38.2 billion, 2019-2023, PWC.
https://www.pwc.com/gx/en/industries/tmt/media/outlook.html

accessibles *via* Internet dont l'accès aux contenus est gratuit grâce à de la publicité. Qu'y a-t-il de vraiment nouveau ? Parmi les plateformes qui sont en train d'émerger, certaines restent des *pure players* de la SVOD, comme Netflix, pendant que d'autres s'en éloignent. Par exemple, aux Etats-Unis, Hulu, créée un peu avant 2010, est une plateforme commune aux groupes Disney, NBCUniversal et WarnerMedia, qui est accessible en mode SVOD mais également en mode TVOD et AVOD. Elle tirait déjà, en 2019, 70 % de son audience du modèle AVOD (O'Donnell, 2020). Autre caractéristique : Hulu se positionne de plus en plus comme une méta-plateforme qui donne accès à une cinquantaine de chaînes nationales américaines, gratuites ou payantes, et donc, à des programmes linéaires « *Live* ». Elle constitue donc aujourd'hui une méta-plateforme hybride, entre un Netflix, une télévision gratuite et une télévision payante, diffusant des contenus exclusifs dont la qualité et la diffusion mondiale n'ont rien à envier à Netflix. N'oublions pas que *The Handmaid's Tale / La Servante écarlate* (Hulu, 2017, distribué en France par OCS) est la première création originale à avoir reçu la récompense ultime aux Emmy Awards (le prix du meilleur *drama* 2017), avant même Netflix.

## Trois catégories de nouvelles plateformes issues d'acteurs aux racines et identités différentes

Cette grille de lecture SVOD, AVOD et TVOD permet de comprendre l'incroyable consolidation du secteur audiovisuel aux Etats-Unis avec quelques méga-fusions spectaculaires autour de trois types de secteurs très différents.

1) Les entreprises du secteur des télécoms cherchent depuis quelques années déjà à se diversifier pour trouver de nouvelles sources de profit et pour ne pas se retrouver cantonnées à gérer des infrastructures qui coûtent de plus en plus cher à installer (4G, 5G, etc.). AT&T, l'opérateur télécom, a acquis en 2018 Time Warner (HBO, CNN et Warner Bros) et a créé WarnerMédia [(2)]. L'entreprise va lancer courant 2020 sa plateforme HBO+, initialement sur un modèle SVOD, mais intégrera rapidement un modèle AVOD pour mieux potentialiser sa base de clients d'AT&T. HBO, positionné haut de gamme, est le plus ancien service de télévision payante accessible *via* le câble. Très tôt positionné sur des contenus originaux et exclusifs aux Etats-Unis, c'est le plus ancien concurrent de Netflix. Son catalogue est très complet et compte de très beaux succès planétaires (*Friends, Sex and the City, Game of Thrones*) récompensés par de nombreux prix. L'entreprise annonce déjà trente et une nouvelles séries originales pour 2020. Comcast, le premier câblo-opérateur américain (également fournisseur d'Internet) a acquis Sky, le puissant groupe de télévision payante britannique, pour s'établir en Europe et développer le maximum de synergie en matière de contenus exclusifs. Perdant toujours plus d'abonnés au câble, Comcast se prépare à lancer en 2020 son propre service de streaming, « Peacock », qui sera proposé *via* NBCUniversal, sa filiale. Peacock suivra le modèle hybride du AVOD pour élargir sa base de clients, TVOD sur des contenus au catalogue et SVOD pour accéder à des contenus *premium* exclusifs.

2) Les entreprises du secteur de la télévision et du cinéma cherchent à éviter la disruption, à profiter de l'explosion du secteur de la VOD et à concentrer leurs moyens financiers pour être en mesure de développer des contenus originaux exclusifs. L'exemple le plus emblématique est l'acquisition de la 21st Century Fox par Disney et le lancement de sa plateforme Disney+ fin 2019. A la grande surprise de la direction de l'entreprise, Disney+ a séduit 10 millions d'abonnés aux Etats-Unis en seulement vingt-quatre heures alors que l'objectif avait été établi autour de 8 millions de souscripteurs au 31 décembre 2019 [(3)]. L'entreprise communique sur la synergie de

(2) https://www.warnermediagroup.com/

(3) https://www.capital.fr/entreprises-marches/netflix-disney-et-amazon-prime-declarent-la-guerre-de-plateformes-de-svod-1362637

ses différents studios : Disney+, Marvel+, Star Wars+, Pixar+, National Geographic, en illimité. Fin 2019, Disney avait soixante et un projets de séries ou de films originaux prévus à court terme. Parallèlement, les groupes américains Viacom et CBS ont fusionné en 2019, créant un des tout premiers groupes mondiaux du cinéma et de la télévision qui va bien entendu lancer sa propre plateforme de *streaming* regroupant toutes les plateformes qui existaient déjà au niveau du groupe sous l'ombrelle de ViacomCBS. Parallèlement, on voit un peu partout dans le monde apparaître des alliances locales entre acteurs de la télévision publique et privée pour créer des plateformes communes, comme Salto en France.

3) Les « géants de l'Internet » comme Apple et Amazon cherchent à diversifier leurs services pour fidéliser encore davantage leurs consommateurs. Amazon propose déjà un abonnement par défaut à Amazon Prime Vidéo depuis 2011 aux Etats-Unis (2016 en France) à tous ceux qui s'abonnent à la formule Amazon Prime. La plateforme a adopté la même stratégie que Netflix en misant sur des productions originales, accessibles de façon exclusive, qui commencent à remporter des prix très prestigieux. Mais contrairement à son concurrent, Amazon Prime Vidéo s'éloigne du modèle pur de la SVOD pour adopter au fur et à mesure un modèle hydride de méta-plateforme SVOD-TVOD qui permet d'accéder aujourd'hui à des chaînes de télévision payantes ou gratuites locales et internationales. Apple, qui avait déjà lancé Apple TV, est allé plus loin en lançant sa plateforme en 2019, AppleTV+, positionnée également comme une méta-plateforme qui donnerait accès à de nombreuses chaînes gratuites et payantes. Contrairement à ses concurrents, Apple n'a pas de fond de catalogue mais possède de gros moyens financiers et commence à investir dans des productions très haut de gamme comme *The Morning Show* avec Jennifer Aniston, d'ailleurs primé.

En face de tous ces nouveaux entrants, Netflix est le précurseur, le leader (71 % du marché mondial en 2018 [4]), celui qui a initié les transformations du secteur mais qui pourrait potentiellement être rattrapé. Il reste aujourd'hui un *pure player* de la SVOD. Dès 2013, Netflix avait basculé vers le rôle de producteur de séries et, dès 2015, de films, car il était trop dépendant des contenus produits par d'autres. Depuis 2018, il renforce chaque année toujours plus ses investissements dans les contenus exclusifs.

Parallèlement à ces grosses plateformes, de nombreuses plateformes SVOD locales de niche sont apparues partout dans le monde depuis quelques années. Il y avait en tout en 2019 un peu moins de quarante plateformes en France (Thuillas et Wiart, 2019 a,b ; CNC [5]).

Ainsi, le monde des plateformes est aujourd'hui un paysage hétérogène en pleine mutation avec des acteurs aux identités et aux racines très différentes. Les facteurs-clés de succès de ces plateformes sont les contenus, l'expérience client, le système de recommandation ou la qualité éditoriale et l'usage intelligent des *data*. En d'autres termes, d'un côté le financement de contenus exclusifs et de l'autre des investissements technologiques considérables qui donnent peut-être une avance aux grands acteurs de la tech et à Netflix qui possèdent une redoutable expertise en la matière. Parallèlement, beaucoup d'experts considèrent que le temps d'attention du public est arrivé à saturation. Chaque minute se gagne sur les concurrents. Dans un tel contexte, la force des images de marque devient un avantage concurrentiel. En témoigne la guerre actuelle des talents que se livrent les grandes plateformes et qui est en train d'alimenter une bulle financière autour des stars et des grands réalisateurs. Les coûts des productions auraient été multipliés par 3 ou 4 ces dernières années [6]. On estime que Netflix a mis entre 150 et 160 millions sur la table pour financer la dernière réalisation du célèbre Martin Scorsese, *The Irishman* (Joe Pesci, Robert De Niro,

(4) Parrot Analytics 2018 Global Television Demand Report.
(5) https://www.cnc.fr/documents/36995/961345/201905-+Barom%C3%A8tre+de+la+V%C3%A0DV%C3%A0DA_mai2019.pdf/47d91293-5bf0-024d-01dd-bef4b857ec34
(6) « La guerre du *streaming* fait flamber le prix des séries télé » (*Les Echos*, 24 septembre 2019).

Al Pacino). Enfin, cette croissance des plateformes pose la question de l'hyperchoix. A combien de services de SVOD un spectateur est-il prêt à s'abonner ? On comprend pourquoi on observe d'une part, une croissance du modèle AVOD capable de capter les spectateurs plus largement que le modèle SVOD et, d'autre part, un développement du modèle de la méta-plateforme capable d'intégrer de nombreuses chaînes et autres plateformes partenaires, à la manière de Molotov [7] en France. Ainsi, aujourd'hui, on assiste à une très grosse recomposition du paysage. Certains acteurs retirent leurs contenus des plateformes concurrentes (Disney de Netflix, par exemple), tandis qu'elles nouent des partenariats nouveaux (Disney avec Amazon, par exemple) en fonction des positionnements et repositionnements de chacun. En France, cette recomposition a suscité pendant trois années consécutives (2017, 2018, 2019) des tensions entre les chaînes et les opérateurs télécoms pour une meilleure répartition des revenus.

## Quels impacts des mutations en cours sur les contenus ?

Les plateformes et le *binge-watching* ont-ils tué le cinéma ? Les avis sont partagés (Baldacchino, 2019 ; Campion, 2019). Il y a ceux qui estiment que les séries ont adopté aujourd'hui les pratiques du cinéma (plans larges, plans séquences, etc.), qu'elles utilisent les mêmes moyens techniques et que leur grammaire visuelle se rapproche beaucoup de celle des films. Ceux-là estiment que les plateformes financent des films d'auteur qui n'auraient jamais trouvé d'autres moyens de financement. Pour eux, « plateforme » est synonyme de créativité renforcée et d'une plus grande liberté de création. Les moyens donnés à certains types de films n'ont jamais été aussi importants qu'aujourd'hui. En effet, les plateformes ont une base internationale de spectateurs tellement importante qu'il leur serait plus facile de sortir du cadre avec une masse critique suffisante d'audience. Cette tendance est d'ailleurs en train de s'accentuer. Des sujets comme la transidentité qui étaient plutôt initialement l'apanage de Netflix se sont étendus à d'autres plateformes avec une série comme *Bisexual* créée pour Channel 4 (chaîne publique anglaise) et diffusée sur Canal+ ou Hulu, ce qui n'aurait peut-être pas été possible auparavant. L'effet Netflix tient aussi à la façon dont la plateforme a rendu possible la diffusion à très grande échelle de contenus très locaux aux côtés de contenus globaux susceptibles de résonner au-delà des frontières, des cultures et des langues (Michaux, 2018).

Parallèlement, il y a ceux qui estiment qu'il y a bien trois mondes différents qui correspondent à des façons distinctes de raconter des histoires. Ce que l'on appelle le « film de cinéma » correspond à une scénarisation adaptée à la salle de cinéma avec une durée et une grammaire spécifique (Quinton, 2020). La télévision a inventé ses propres contenus et sa propre grammaire basée sur des rendez-vous quotidiens : reportages, émissions de variétés, débats, feuilletons à épisodes, etc. Les plateformes ont inventé un nouveau modèle de séries qui emprunte les codes du cinéma. Ces trois modèles de contenus cohabitent aujourd'hui. Ceux-là sont plus réservés quant à l'impact des plateformes sur la créativité car souvent elles sont obligées de trouver un équilibre entre plaire au plus grand nombre et être originales et finissent par arrêter les séries qui restent trop confidentielles en termes d'audience. Ceux-là sont également plus critiques sur l'impact du modèle AVOD qui repose la question des audiences, là où le modèle SVOD l'avait supprimé.

Enfin, il y a ceux qui considèrent que le cinéma est en danger et peut y perdre son identité. Auparavant, il pouvait y avoir plusieurs années entre l'idée, le tournage et la diffusion d'un film ou d'une série. Les plateformes tendent à raccourcir et à industrialiser ce processus, avec des budgets cinquante à soixante fois supérieurs à celui des télévisions payantes traditionnelles (Baldacchino, 2019). Ceux-là reconnaissent que les plateformes prennent des risques créatifs,

(7) https://www.molotov.tv/

comme Netflix en finançant des films d'auteur comme *Roma* d'Alfonso Cuarón, *Okja* de Bong Joon-ho ou *The Irishman* de Scorsese. Néanmoins, ils estiment que ces exemples traduisent une fuite des auteurs vers les plateformes et qu'il y a danger à mettre la création sous dépendance des seules plateformes pendant qu'en salles on ne trouve plus que des *blockbusters* mondiaux. Ceux-là s'insurgent contre la disparition de la salle de cinéma dans le parcours d'un film, réduit à quelques passages seulement pour avoir le droit de se présenter aux Oscars, par exemple. Ceux-là défendent bien évidemment la chronologie française des médias qui, d'ailleurs, dans ce contexte évolutif, va être à nouveau rediscutée au parlement français dès la fin de la crise du Coronavirus. Ils s'inquiètent, comme Scorsese lui-même, de la vie des œuvres cinématographiques visionnées uniquement sur les smartphones. 70 % des spectateurs ayant commencé à visionner *The Irishman* sur Netflix (13,2 millions de spectateurs en cinq jours aux Etats-Unis) se seraient arrêtés avant la fin (Quinton, 2020).

## Les contenus du futur

La 4G, le wifi, le *cloud* et le mobile ont permis l'avènement puis l'explosion des nouvelles plateformes. La 5G va permettre une autre mutation : la réalité virtuelle et l'interaction avec le spectateur. Néanmoins, la 5G est une révolution lente qui ne fait que commencer et qui va s'étaler sur les dix prochaines années, entraînant une diffusion très progressive dans le temps de ces nouveaux contenus. Par contre, devant l'hyperchoix offert aujourd'hui, les spectateurs deviendront de plus en plus exigeants en termes de conseils. L'impact de l'intelligence artificielle, tant sur l'ergonomie des plateformes que sur leur système de recommandation, devrait être assez rapide. De plus, l'avenir du modèle AVOD est la publicité personnalisée où les progrès de l'intelligence artificielle devraient permettre des expérimentations de plus en plus poussées également. C'est d'ailleurs un des axes de la nouvelle loi sur l'audiovisuel en France qui va commencer à autoriser la publicité ciblée. Enfin, l'expérience client, qui dépend exclusivement des investissements technologiques, restera centrale dans la réussite des plateformes, au côté des contenus de plus en plus originaux et exclusifs.

## Références

BALDACCHINO J. (2019), « Netflix, l'usine à séries que le cinéma déteste adorer », *Nectart*, Vol. 9, n°2, 2019, pp. 116-125.

CAMPION B. (2019), « Plateformes de SVOD : les nouveaux *networks* de la télévision américaine ? », *Télévision*, Vol. 10, n°1, pp. 53-69.

CNC (2019), « Baromètre de la vidéo à la demande », mai.

LEEBAERT. D. (1998), *The Future of the Electronic Marketplace*, MIT Press, Cambridge.

MICHAUX V. (2018), « Netflix, talentueux colosse aux pieds d'argile », *The Conversation*, mai.

O'DONNELL V. (2020), *Television Criticism,* 3e édition, Sage, USA.

PLOTHE T. & LANG P. (2020), *Netflix at the Nexus: Content, Practice and Production in the Age of Streaming Television,* Peter Lang Edition, New York.

QUINTON J.M. (2020), « Ce que le XXe siècle nous enseigne sur Netflix », Slate.fr, janvier.

THUILLAS O. & WIART L. (2019a), « Les plateformes de VOD cinéphiliques : des stratégies de niche en question, *Les Enjeux de l'information et de la communication*, Vol. 1, n°20, pp. 39 à 55.

THUILLAS O. & WIART L. (2019b), « Plates-formes culturelles : quelles alternatives aux géants du numérique ? », *Nectart*, Vol. 8, n°1, pp. 78-87.

# La dynamique stratégique des plateformes digitales : analyse du marché de la formation en ligne

Par **François ACQUATELLA**
IAE – Ecole universitaire de Management
CREOP – EA 4332, Université de Limoges

## Introduction

Dans sa définition consensuelle, la plateforme offre simultanément un intérêt d'usage pour deux (ou plusieurs) catégories d'agents (faces), représentant autant de versants d'un marché biface ou multiface. Les différentes faces agrégées forment l'écosystème de la plateforme (Hagiu et Wright, 2015). Ainsi, dans le modèle de plateforme, la valeur du service est proportionnelle au nombre d'acteurs présents sur chaque face du marché, en ce que l'écosystème se construit d'externalités d'effets de réseau. Cette conception majoritairement répandue au sein de la littérature dédiée procède d'une approche holistique nécessaire au cadrage et à la compréhension de la dynamique stratégique des plateformes dans leur caractère générique. Pour autant, elle n'envisage pas les spécificités techno-stratégiques à l'aune de la diversité des modèles d'affaires observés dans la sphère digitale. Les méthodes et leviers de création et de développement de réseaux de valeur, et de propositions de valeur (Chanal et Craron-Faisan, 2008), y sont en effet présentés de façon générique.

Notre article vise à mieux appréhender les dynamiques stratégiques de plateformes sur le marché de la formation, notamment en analysant le poids grandissant des technologies les plus récentes dont, au premier plan, celui de l'intelligence artificielle (IA).

Nous portons donc une attention particulière aux plateformes de formation utilisant les données collectées pour développer des calculs dans le but de déployer des modèles d'algorithmes, dans une optique de classification et/ou de prédiction de l'usage et du comportement des utilisateurs. Ces plateformes bâtissent en effet des stratégies fondées sur la donnée et sur son traitement algorithmique pour développer de nouveaux *business models*.

Nous présentons dans un premier temps une revue de la littérature émergente relative à l'impact du *big data* et des technologies d'IA sur la « plateformisation » de l'économie. Nous analysons, dans les parties qui suivent, les différentes trajectoires stratégiques dont les plateformes de formation se saisissent pour accroître leur création de valeur, en pointant notamment le rôle et les caractéristiques de l'IA pour améliorer l'efficacité de leurs choix stratégiques.

## Des technologies d'IA au cœur des processus stratégiques des plateformes

Le développement des technologies de l'IA, au travers de ces nouvelles formes d'organisation que sont les plateformes, génère des architectures de marché reposant sur le traitement de l'information (Benavent, 2016). En effet, les modèles d'affaires des plateformes « multifaces » reposent en très grande partie sur des mécanismes techno-stratégiques de création de valeur basés sur l'exploitation

des volumes de données massives, et ce, dans le but de générer une dynamique d'accroissement et de fidélisation des utilisateurs. De fait, en recommandant des produits et des services, en mettant en relation des personnes et des objets, en évaluant les potentialités de futurs marchés, en définissant des prix ou bien en qualifiant des offres, les plateformes étendent et affinent leurs activités grâce au traitement automatisé d'un ensemble de données et de métadonnées adossées (Hartmann, 2014).

Les technologies de l'IA les plus visibles, telles que les agents virtuels de conversation *Chatbot,* ou bien l'aide à la prise de décisions, fondent des mécanismes de motivation et d'action par une amélioration de l'expérience utilisateur (Benavent, 2016 ; Schneider *et al,* 2017), de sorte qu'elles participent à redéfinir les périmètres et les enjeux stratégiques des plateformes. Ainsi, la mise en œuvre de ces technologies permet le développement d'une connaissance toujours plus intimiste de l'utilisateur et permet d'affiner la segmentation client pour, à terme, verser vers l'ultra-segmentation.

Les technologies de l'IA telles que le *Machine Learning, Deep Learning* contribuent quant à elles à perfectionner le design des interfaces des plateformes, en déterminant l'évolution du cadre d'usage des interfaces sur la base d'un jeu d'incitations et de contraintes pour leurs usagers.

En résumé, les technologies de l'IA portent la dynamique stratégique des plateformes et constituent de plus en plus une composante structurelle de l'architecture technologique des plateformes, d'autant plus qu'elles procurent aux plateformes une agilité dans leur capacité à gérer et développer des modèles organisationnels écosystémiques hybrides, notamment en développant la connectivité des interfaces, généralisant de la sorte les stratégies d'interopérabilité des plateformes à d'autres plateformes ainsi qu'aux objets connectés (Acquatella *et al.,* 2019).

## Principales dynamiques stratégiques observées

Trois dynamiques stratégiques se dégagent vis-à-vis des leviers technologiques qui fondent leur caractère disruptif. Les distinctions proposées visent à mieux appréhender les dynamiques stratégiques à la lumière du rôle, des incidences et des enjeux des technologies de l'IA pour améliorer l'efficacité des choix stratégiques des plateformes sur le marché de la formation. Ainsi, ces principales trajectoires stratégiques identifiées ne sont pas mutuellement exclusives, mais complémentaires. Il s'agit de la désintermédiation des marchés, de l'appariement d'acteurs et de l'architecture technologique.

### La désintermédiation-réintermédiation

Cette approche stratégique se caractérise par une forme de désintermédiation-réintermédiation des marchés ; les plateformes s'appuient sur l'accaparement d'une partie de la chaîne de valeur dans le but de disrupter les intermédiaires traditionnels. Cette dynamique stratégique se caractérise par la capacité à développer et à reconfigurer de façon agile des partenariats, en permettant de nouvelles modalités de distribution de l'information. Les plateformes réintermédialisent des segments de marché de biens et services, sans en acquérir les actifs. La puissance de l'offre de la plateforme crée alors une modification massive des usages des internautes.

En offrant un design algorithmique, les plateformes proposent une concentration et une présentation plus fluide des informations et deviennent alors les intermédiaires privilégiés des utilisateurs. Elles revalorisent une offre de formation existante en proposant une alternative en termes de choix ainsi qu'une mise en perspective de leurs attributs différenciateurs leur permettant d'acquérir un statut de prescripteur auprès du consommateur.

Le développement constant des capacités analytiques en IA permet aux plateformes de recommander aux utilisateurs l'offre susceptible de correspondre avec précision aux

attentes explicites ou implicites des utilisateurs. Ces plateformes bénéficient ainsi d'un statut d'intermédiaire de confiance, de nature à accroître la fidélité des usagers pour ainsi conserver une position concurrentielle avantageuse sur leur marché. Maintenir ce rôle d'intermédiaire de confiance impose aux plateformes une maîtrise fine des flux d'informations générés par leurs différents partenaires pour coaliser leur écosystème autour de stratégies fédératives et collectives. Par exemple, les plateformes MOOCs telles que Class cental ou Quick Code maîtrisent l'ensemble des informations inhérentes aux formations qu'elles exploitent (par exemple : nombre d'apprenants actifs, taux de complétion, taux de visualisation, taux de rétention, taux de satisfaction…) et ajustent ainsi leur offre de formation pour recommander les formations sur un marché de l'emploi ou certaines compétences spécifiques connaissent des tensions élevées sur le recrutement. Dès lors, le développement par la performance des outils de recommandation traduit l'exigence d'un contenu de plus en plus algorithmique, pour continuellement stimuler la fluidité des informations transmises (actualités, enseignements proposés, sujets populaires...) et améliorer l'expérience utilisateur pour intensifier l'engagement des apprenants. La combinaison des informations personnelles issues des interactions des apprenants sur et avec la plateforme affine la compréhension des attentes et des besoins des utilisateurs. Connaître les besoins de formation, le comportement d'apprentissage, les attentes exprimées ou tacites des utilisateurs à partir des informations et des métadonnées d'usage et de navigation permet à la fois de faire évoluer l'expérience utilisateur par l'amélioration du design de la plateforme mais également de recommander l'offre de formation correspondant à leurs *desiderata*.

## L'appariement d'acteurs

Cette trajectoire stratégique supporte la mise en œuvre d'un nouveau réseau de valeur lui permettant de diffuser massivement l'innovation. Cette dynamique stratégique réside précisément dans la capacité de la plateforme à créer de nouvelles interrelations de valeur entre différents acteurs du marché afin d'amplifier et de faire évoluer le périmètre de son activité pour créer de nouveaux marchés. Par exemple, la plateforme Coursera propose une nouvelle forme d'intermédiation et d'interaction entre différents agents économiques (faces), se définissant par les institutions académiques sur l'un de ses versants et les individus souhaitant se former sur l'autre de ses versants. Ainsi, cette nouvelle forme d'intermédiation permet l'émergence et la création d'une nouvelle offre de prestations de formation en mobilisant des actifs (cours, ressources pédagogiques…) sous-exploités qui, en étant valorisés d'une manière nouvelle, créent et coordonnent un marché en construisant une nouvelle demande de formation et de diplomation. Le développement de ce mode de création de valeur se produit en permettant aux détenteurs d'actifs, à savoir les institutions académiques, de tirer avantage de leurs ressources organisationnelles (enseignements, cours, enseignants…) d'une manière nouvelle. La plateforme est alors vectrice de nouveaux modes de consommation de ces actifs au travers d'une forme inédite d'intermédiation entre les internautes et les établissements pourvoyeurs de contenus de formation. Dès lors, la capacité des plateformes à créer de nouveaux réseaux de valeur, centrée autour d'une offre distinctive à son marché, repose sur la création de nouveaux actifs (Moocs) et supporte la dynamique de développement de ce type de plateforme. Créer et coordonner de nouvelles formes d'intermédiation repose sur des itérations permanentes de ces plateformes avec le marché, pour tester continuellement différentes propositions de valeur dans une posture effectuale. Ainsi la plateforme génère, assemble et redistribue de façon semi-automatique des propositions de valeur grâce aux données générées par leurs différents acteurs partenaires pour réfléchir aux possibles stratégiques. Le poids des algorithmes dans leurs analyses stratégiques tend à automatiser la recherche de nouveaux gisements de valeur exploitables. La plateforme Coursera, en faisant évoluer la manière dont les utilisateurs consomment des produits de formation et le format des formations qu'ils consomment, a ainsi ouvert un nouveau marché de la formation en ligne.

### L'architecture technique

Dans son caractère générique, la conception de l'architecture de la plateforme est élaborée comme une base technique supportant un développement continuel d'innovations (Baldwin et Woodard, 2009). La maîtrise technologique de l'architecture de la plateforme a pour but de maîtriser et d'orienter les standards des innovations techno-pédagogiques sur différents marchés. Dans son caractère spécifique, le modèle de conception de l'architecture permet aux plateformes d'opter pour une stratégie « propriétaire » ou « standardisée ». La stratégie « propriétaire » consiste à créer un avantage compétitif en se portant sur le marché avec une offre technologique totalement intégrée. La stratégie « standardisée » consiste à drainer les enjeux de la compétition technologique du marché par la capacité d'une plateforme technologique à percoler un large ensemble de systèmes techniques par un flux continu d'innovations incrémentales effectuées par la communauté du Web.

Certaines plateformes telles que Coursera définissent seules leurs interfaces avec un outil de développement très intégré. Le design d'interaction, ainsi qu'une importante majorité des composants techniques associés à l'architecture de la plateforme, sont propriétaires. Cette *internal platform* au sens de Gawer (2011) développe et implémente une famille de produits en déployant un ensemble de composants et de nouvelles fonctionnalités qu'elle produit et qu'elle fonde sur ses propres ressources. L'architecture de la plateforme se présente comme un ensemble de systèmes, sous-systèmes représentant des interfaces qui forment une structure propriétaire gérée en interne.

*A contrario,* certaines plateformes (notamment la plateforme EDX) ont opté pour une solution de développement de leurs interfaces en *open source.* Elles proposent une architecture dite « standardisée », leur permettant une forme d'externalisation des processus de développement technique de la plateforme. Les plateformes de formation dont l'architecture est dite « standardisée » se positionnent sur le marché avec une infrastructure modulaire et englobante [(1)]. La principale différenciation entre ces deux modèles de développement technique fonde avant tout des ambitions stratégiques différenciées de conquête de marché. Les leviers d'une stratégie propriétaire reposent sur la création d'une nouvelle offre de services en construisant une nouvelle demande. Les leviers d'une stratégie standardisée reposent *a contrario* sur une volonté de la plateforme de préempter les systèmes techniques du marché.

## Conclusion

Si ces technologies de l'IA renvoient à une pluralité de modèles et d'ambitions stratégiques, elles portent intrinsèquement des caractéristiques qui imposent peu à peu un standard stratégique participant à architecturer des marchés. De fait, il semblerait que toutes les organisations en modèle de « plateforme » aient vocation, par nature, à intégrer une composante algorithmique pour continuellement développer et/ou revitaliser leur modèle d'affaires. C'est le cas des plateformes de formation qui élargissent leurs capacités d'apprentissage automatique au travers de la collecte de données auprès des faces agrégées en écosystème.

(1) Cette solution de développement n'exclut pas la possibilité de contrôler le degré d'ouverture sur un certain nombre de dimensions (techniques ou non), tel que le niveau d'accès aux informations, ou bien le niveau de support attendu par les tiers souhaitant s'impliquer.

## Références

ACQUATELLA F. (2019), "Platforms as technical infrastructures, architects of a dominant design", R&D Management, International Conference 2019, innovation challenge - Polytechnique-HEC, Paris, France.

BALDWIN C. Y. & WOODARD C. J. (2009), "The architecture of platforms: A unified view", *Platforms, markets and innovation,* 32.

BENAVENT C. (2016), *Plateformes. Sites collaboratifs, marketplaces, réseaux Sociaux, Comment ils influencent nos choix,* FYP éditions.

CARON-FASAN M. L. & CHANAL V. (2008), « Des scénarios pour explorer les modèles d'affaires », *L'Expansion Management Review*, (1), 108-119.

GAWER A. (2011), *Platforms, Markets and Innovation*, Edward Elgar Publishing Ltd.

HAGIU A. & WRIGHT J. (2015), "Multi-sided platforms", *International Journal of Industrial Organization*, 43, 162-174.

HARTMANN Ph., ZAKI M., FELDMANN N. & NEELY A. (2014), *Big Data for Big Business? A Taxonomy of Data-driven Business Models used by Start-up Firms*, Cambridge Service Alliance. eng.cam.ac.uk.

SCHNEIDER D., LINS S., GRUPP T., BENLIAN A. & SUNYAEV A. (2017), *Nudging Users into Online Verification: The Case of Carsharing Platforms*., ICIS. 2017.

# Les stratégies de visibilité, le rôle des plateformes

Par **Philippe BOUQUILLION**
Laboratoire des sciences de l'information et de la communication (Labsic), Université Sorbonne Paris Nord
Labex ICCA : Industries culturelles et création artistique

## Introduction

L'essor des plateformes numériques, à la fin des années 2000 et dans les années 2010, a facilité la diffusion d'une masse considérable de contenus créés et diffusés par des créateurs auto-entrepreneurs ; par des « amateurs » à l'origine des *User Generated Content* (UGC) ; par des marques cherchant à promouvoir leurs offres, mais aussi par des acteurs historiques des industries culturelles ou par de nouveaux entrants, tel Netflix par exemple. Aux contenus de format long ou moyen se sont ajoutés de très nombreux contenus de format court.

Du fait de l'abondance des contenus pouvant être offerts par une même plateforme, la question de la visibilité des contenus s'est posée en des termes particuliers et différents par rapport aux stratégies de visibilité mises en œuvre dans les industries culturelles et médiatiques plus anciennes, dont la radiodiffusion ou l'édition de livres et de disques. Au sein des plateformes, la mise en visibilité s'exerce principalement grâce à des systèmes de recommandation qui sont des dispositifs algorithmiques permettant d'organiser la collecte et le traitement des données massives. Ils sont destinés à proposer de façon personnalisée des contenus convenant, en principe, aux attentes et aux goûts exprimés, ou même non exprimés, des utilisateurs des plateformes.

Les plateformes offrant des contenus culturels forment un ensemble disparate. A défaut de pouvoir en produire ici une catégorisation complète, l'accent sera mis sur une distinction centrale. Certaines plateformes négocient en amont des droits de diffusion, voire produisent elles-mêmes des contenus. D'autres plateformes, en revanche, mettent en relation les utilisateurs avec les offreurs de contenus, sans intervenir dans la construction des offres. Elles se contentent d'imposer aux offreurs et aux utilisateurs un contrat d'adhésion pour régler les relations entre les divers agents recourant à la plateforme, qu'ils offrent ou consomment des contenus. Des économistes, dont Hagiu et Wright (2015), à la suite d'une longue suite de travaux, dont ceux de Rochet et Tirole (2003), distinguent ainsi les « magasins » des « plateformes » dites multi-versants. La théorie des industries culturelles, quant à elle, distingue le modèle du club de celui du courtage (Mœglin, 2007). Cette distinction est importante car selon les rapports que les opérateurs de plateformes entretiennent avec les offres de contenus, les dispositifs de recommandation sont différents.

Il est d'autant plus important de replacer les algorithmes de recommandation dans leur cadre socio-économique qu'ils ne visent pas seulement à maximiser la satisfaction des consommateurs mais qu'ils poursuivent aussi d'autres objectifs. Dans le cadre imparti à cette contribution, l'accent sera mis sur les enjeux des stratégies de visibilité développées par les plateformes de *streaming* vidéo et musical à ambition transnationale, productrices de contenus ou ayant signé des contrats préalables avec des producteurs tiers.

L'argumentation s'appuie sur les enseignements de diverses recherches antérieures, mais aussi sur les résultats d'enquêtes originales sur les plateformes de *streaming*, en particulier celles valorisées par l'abonnement comme Netflix (Bouquillion, 2019).

Notre proposition est que les plateformes développent des stratégies destinées à forger la croyance dans la qualité et l'efficacité de leurs algorithmes de recommandation (ce qui sera envisagé dans le premier volet), afin de renforcer leur position face aux utilisateurs, à d'autres acteurs industriels comme face aux autorités publiques et de régulation (ce qui sera examiné dans le second volet).

## Les dispositifs de recommandation, entre dimension matérielle et dimension idéelle

Les dispositifs de recommandation, comme plus largement l'ensemble des *Big Data* et des algorithmes, sont à la fois des « réalités matérielles » et des « réalités idéelles » selon la distinction de l'anthropologue Maurice Godelier (1984). D'un côté, il s'agit d'un ensemble de systèmes techniques de mesure et de calcul, fruits de stratégies de différents acteurs sociaux. D'un autre côté, ces dispositifs sociotechniques font l'objet d'importants discours. Ce faisant, ils sont porteurs de valeurs et ils reposent sur des visions différentes de la culture, de la société, de ce que peuvent et de ce que doivent être (dimension normative) les industries culturelles à l'« ère des plateformes ». Parallèlement aux discours promotionnels des acteurs industriels, les experts et certains chercheurs, tout spécialement les chercheurs non critiques, par leurs travaux et leurs écrits, contribuent non pas seulement à éclaircir les modes de fonctionnement des *Big Data*, mais aussi à construire leur dimension idéologique et la croyance en la puissance de leurs « effets » sociaux. A la décharge des chercheurs, il faut rappeler que les acteurs industriels, et singulièrement les opérateurs des plateformes, pratiquent très largement le secret industriel, ne communiquant que les informations qui s'insèrent dans leurs stratégies de construction de l'image de puissance et d'efficacité de leurs algorithmes. Les chercheurs doivent donc s'appuyer sur les informations communiquées par les industriels comme sur les réflexions de consultants ou d'experts, eux-mêmes fréquemment reliés aux intérêts industriels.

La perspective idéologique est bien présente lorsque certains opérateurs de plateformes communiquent autour de leur système de recommandation, qu'il s'agisse d'acculturer les utilisateurs au système de recommandation et de les inciter à dévoiler plus de données, comme c'est le cas de Netflix (Drumond, Coutant et Millerand, 2018), ou pour vanter la capacité de leur système de recommandation à favoriser la diversité culturelle. Celle-ci est entendue non pas seulement par la quantité ou la variété des contenus offerts mais aussi par la capacité des algorithmes de recommandation à adresser des contenus totalement inconnus aux utilisateurs. Ce faisant, ils favoriseraient la « diversité consommée ». Tel serait, par exemple, le rôle de Discover Weekly de Spotify, plateforme de *streaming* musical.

Discover Weekly associe, à l'instar de la plupart des systèmes de recommandation, deux logiques : d'une part, celle dite de *collaborative filtering* et, d'autre part, celle dite de *content-based filtering*.

Dans le cadre du *collaborative filtering*, les recommandations sont fondées sur des probabilités de goûts similaires entre des individus. Dans l'exemple de Spotify, les catégories d'individus ne sont fondées que sur des pratiques musicales communes et non pas sur des similitudes sociologiques.

Le *content-based filtering* conduit à analyser des données textuelles et des données sonores. L'étude des premières permet de construire des catégories de contenus à partir des commentaires dont ils font l'objet sur tous les supports analysables (sites, réseaux sociaux, *etc.*). Sont tout spécialement étudiés les termes considérés comme des *top terms*, qui qualifient de manière déterminante le

contenu faisant l'objet du commentaire. Chaque *top term* se voit associé un certain poids « *selon la probabilité que quelqu'un décrive l'artiste ou la chanson en employant ce terme* [1] ». Lorsque des contenus différents font l'objet des mêmes commentaires, des mêmes *top terms*, ils sont considérés comme relevant de la même catégorie. Des individus ayant consommé un ou plusieurs contenus de cette catégorie se voient alors proposer des contenus relevant de la même catégorie.

L'analyse des données sonores, quant à elle, conduit à l'élaboration de catégories de contenus selon les proximités musicales que présentent les différents contenus. En fonction des contenus qu'ils ont préalablement consommés, les utilisateurs se voient proposer des œuvres qu'ils ne connaissent pas, qui peuvent avoir été créées par des artistes tout aussi inconnus d'eux, mais qui présentent des similitudes sonores avec les contenus qu'ils ont déjà consommés. Ces dispositifs sont donc supposés à la fois favoriser la personnalisation par les jeux de probabilités et la découverte de contenus.

Il peut y avoir des variantes. Par exemple, Netflix construit des catégories d'utilisateurs en tenant compte des informations que ceux-ci ont communiquées dans leur « profil ». De même, l'avis et les notations des contenus par les utilisateurs sont pris en compte.

Si différents acteurs socio-économiques du e-marketing présentent les *Big Data* comme constituant des facteurs de diversité culturelle évoquant « *un futur dans lequel l'innovation dans la recommandation servira la cause de la diversité semble se dessiner avec de plus en plus de force* [2] », d'autres voix, plus critiques, se font entendre en soulignant que d'autres problématiques sont en jeu.

## Dispositifs de recommandation et affirmation des opérateurs des plateformes

Les systèmes de recommandation sont au cœur de la dynamique des plateformes et de leur valorisation. Ce faisant, ils soulèvent trois ensembles d'enjeux, vis-à-vis des utilisateurs, des fournisseurs de contenus et des politiques publiques.

Un premier enjeu a trait à la captation des utilisateurs. L'efficacité de la recommandation a été remise en cause (tant l'importance de cette dernière dans les choix des utilisateurs que sa capacité à effectivement promouvoir une plus grande diversité consommée). Dans l'exemple des plateformes musicales en *streaming,* des chercheurs notent :

« Les usages des recommandations par les usagers, […] restent mesurés, inégalement distribués, et les usages de découverte musicale, s'ils orientent les usagers vers des musiques de la mid-tail et de la long-tail, demeurent marginaux. […] Si la diversité des consommations sur la plateforme est plus grande que sur d'autres dispositifs, les écoutes demeurent fortement concentrées sur les artistes les plus populaires. » (Beuscart, Coavoux, Maillard, 2019)

En revanche, les algorithmes de recommandation peuvent être des éléments de promotion de l'offre. Des enquêtes ont montré qu'au-delà de l'efficacité directe de la recommandation, celle-ci constitue « un moyen d'améliorer la qualité globale du dispositif tel qu'il est perçu par les consommateurs, même s'ils n'y ont que faiblement recours » (Farchy, Meadel, Anciaux, 2017). La qualité de la recommandation est donc un facteur de distinction par rapport à la concurrence.

De surcroît, ainsi que le note Philip Napoli (2019), « l'intégration verticale de la production et de la diffusion du contenu […] encourage l'orientation de l'attention de l'auditoire vers le contenu

(1) Source : Delight,
https://medium.com/delightblog/parlons-recommandation-partie-2-le-mod%C3%A8le-spotify-5173958a41b5
(2) *Ibid.*

produit à l'interne, plutôt que vers une gamme diversifiée d'offres de contenu provenant d'une gamme diversifiée de sources ». La mise en avant des contenus « propriétaires » correspond à divers enjeux pour les opérateurs de plateformes dont, dans un objectif promotionnel, la mise en avant des succès de « leurs » contenus.

Un deuxième enjeu pour les plateformes est d'attirer les fournisseurs de contenus, voire de générer un rapport de force favorable vis-à-vis d'eux. Grâce à l'efficacité des algorithmes de recommandation, il est suggéré que les fournisseurs de contenus obtiendront plus d'avantages que sur d'autres plateformes (meilleur ciblage des consommateurs, plus de revenus, etc.) (Bullich, 2019). Lorsque les plateformes sont en mesure d'obtenir l'exclusivité des droits, des cercles vertueux entre qualité des contenus, nombre d'utilisateurs et ressources peuvent alors s'enclencher au bénéfice des opérateurs de plateformes les plus puissants mais au détriment de la concurrence, celle entre plateformes comme celle entre plateformes et fournisseurs. On rejoint alors le constat formulé par trois chercheurs : « La menace sur la diversité culturelle proviendra peut-être moins de la toute-puissance des algorithmes que de l'éventuel pouvoir oligopolistique qu'un petit nombre d'acteurs maîtrisant les données de comportements des consommateurs pourraient construire » (Farchy, Meadel, Anciaux, 2017). Cette dissymétrie n'est pas seulement liée aux dispositifs de recommandation, elle renvoie aussi à la puissance financière et à la forte dimension transnationale des opérateurs de plateformes les plus puissants face à certains fournisseurs de contenus et opérateurs de plateformes, d'ampleur seulement nationale.

Un troisième enjeu est lié aux politiques publiques. Les opérateurs de plateformes, en mettant en avant l'efficacité de la recommandation, dans un contexte d'abondance de contenus, font apparaître les politiques publiques comme obsolètes – et singulièrement, les soutiens publics à la production. Par ailleurs, certains acteurs, dont Netflix, insistent sur l'importance de leurs contributions à la diversité dans les différents sens donnés à ce terme. Ils soulignent qu'ils produisent les films de réalisateurs issus de l'immigration ou qu'ils favorisent la participation des femmes ou des membres des minorités aux productions qu'ils soutiennent. Dans le cas de Netflix, les productions non américaines, et dans des langues autres que l'anglais, sont mises en exergue dans les discours de légitimation, au prétexte qu'une audience mondiale est offerte à ces œuvres. Le territoire qui est dessiné par la communication de Netflix n'est pas un assemblage de nations ou de régions mais bien un territoire transnationalisé construit par Netflix, et selon les principes et visions du monde que promeut l'entreprise.

Les réactions des institutions en charge des politiques publiques et de la régulation ainsi que celles des chercheurs oscillent entre deux positions, qui ne sont pas exclusives l'une de l'autre. La première est de considérer que les *Big Data* et l'approfondissement de la transnationalisation des plateformes et des contenus obligent à reconsidérer radicalement l'action publique et les régulations. Par exemple, Philip Napoli (2019) met l'accent sur la nécessaire régulation des algorithmes. Il considère que les algorithmes devraient être « modifiés de façon à encourager plus agressivement la découvrabilité et la consommation du contenu diversifié ». La seconde position est celle qui consiste à étendre aux nouvelles réalités industrielles les règles anciennes ou, du moins, à maintenir leurs principaux objectifs. En France, l'exemple qui peut être cité est celui du projet de loi sur l'audiovisuel, dont certaines dispositions visent à étendre aux opérateurs des plateformes les obligations de financement de la création audiovisuelle et cinématographique nationale et européenne.

## Conclusion

Pour produire une critique des stratégies des plateformes dans les algorithmes de recommandation, sans doute faut-il envisager l'économie de la culture et des industries culturelles non pas seulement comme une économie de l'attention (Simon, 1971) – perspective au cœur de la dimension idéelle

de la recommandation –, mais aussi telle une économie de l'incertitude, comme le font différents auteurs de la théorie des industries culturelles (Miège 1984 ; Mœglin 2007) ou de l'économie de la culture (Caves, 2000).

Les pratiques culturelles sont alors vues comme étant radicalement imprévisibles – ce que Caves a résumé par le slogan *Nobody Knows* – tandis que l'accent est mis sur les modalités grâce auxquelles les filières des industries culturelles sont organisées pour mettre sur le marché plus de contenus que les consommateurs ne pourront jamais en absorber. Le regard sur l'abondance de contenus est alors bien différent.

Dans cette perspective, la dimension stratégique de la recommandation peut être mise en tension avec sa fonction affichée, celle de l'appariement de l'offre et de la demande. Le rôle de ces dispositifs dans les divers rapports de force qui se nouent, y compris dans la position de force prise par les plateformes dans la valorisation des contenus « surabondants », peut alors être étudié, tandis que les plateformes sont replacées dans la longue histoire des relations des industries culturelles avec des activités industrielles dont le cœur de métier a trait pour tout ou partie aux technologies.

## Bibliographie

BEUSCART J.-S., COAVOUX S. & MAILLARD S. (2019), « Les algorithmes de recommandation musicale et l'autonomie de l'auditeur. Analyse des écoutes d'un panel d'utilisateurs de *streaming* », *Réseaux*, La Découverte, 213, pp. 17-47. https://www.cairn.info/revue-reseaux-2019-1-page-17.htm

BOUQUILLION P. (2019), « Les plateformes numériques audiovisuelles, entre flux transnationaux et cadre national », *in* GEORGE Eric (dir.), *Numérisation de la société et enjeux sociopolitiques 1. Numérique, communication et culture*, ISTE Editions, pp. 127-136. / BOUQUILLION P. (2019), "Digital Audiovisual Platforms, Between Transnational Flows and National Frameworks", *in* GEORGE Eric, *Digitalization of Society and Socio-political Issues 1. Digital, Communication and Culture*, ISTE Edition, pp. 107-116.

BULLICH V. (2019), « Les agences de valorisation de vidéos et la structuration d'une économie des UGC », *Communication*, vol. 36/1 | 2019 : Vol. 36/1.

CAVES R. E. (2000), *Creative Industries. Contracts between Art and Commerce*, Cambridge, Massachusetts, London, Harvard University Press.

DRUMOND G., COUTANT A. & MILLERAND F. (2018), « La production de l'usager par les algorithmes de Netflix », *Les Enjeux de l'information et de la communication*, 2018/2 (N°19/2), pp. 29-44. DOI : 10.3917/enic.025.0029. URL : https://www.cairn.info/revue-les-enjeux-de-l-information-et-de-la-communication-2018-2-page-29.htm

GODELIER M. (1984), *L'Idéel et le Matériel*, Paris, Fayard.

FARCHY J., MEADEL C. & ANCIAUX A. (2017), « Une question de comportement. Recommandation des contenus audiovisuels et transformations numériques », *tic&société*, Vol. 10, N°2-3 | -1, pp. 168-198.

HAGIU A. & WRIGHT J. (2015), "Multi-Sided Platforms", Harvard Business School, Working Paper 15-037, March 16.

MIEGE B. (1984), « Postface » *in* HUET A., ION, J., LEFEBVRE A., MIEGE B. & PERON R. (1978), *Capitalisme et industries culturelles,* Grenoble, Presses Universitaires de Grenoble, pp. 199-213.

MOEGLIN P. (2007), « Des modèles socio-économiques en mutation », *in* BOUQUILLION P. & COMBÈS Y. (éd.), *Diversité et industries culturelles*, Paris, L'Harmattan, pp. 152-162.

NAPOLI P. (2019), « Diversité de contenus à l'ère numérique : découvrabilité de contenus diversifiés aux échelons local, régional et national », Commission canadienne pour l'UNESCO, 7-8 février.

ROCHET J.-C. & TIROLE J. (2003), "Platform Competition in Two-Sided Markets", *Journal of the European Economic Association*, vol. 1, n°4, pp. 990-1029.

SIMON H. (1971), "Designing Organizations for an Information-Rich World", *in* M. GREENBERGER, *Computers, Communication, and the Public Interest*, Baltimore, The Johns Hopkins Press, pp. 40-41.

# Gallica, mine d'or et source de culture

Par **Arnaud BEAUFORT**
Directeur des services et des réseaux et Directeur général adjoint,
Bibliothèque nationale de France (BnF)

## Introduction

Mise en ligne en 1997, Gallica, la bibliothèque numérique de la BnF et de ses partenaires, est passée au milieu des années 2000 d'une bibliothèque de l'honnête homme rassemblant les œuvres les plus significatives depuis l'Antiquité à une vaste collection universelle, encyclopédique et multiforme (livres mais aussi estampes, photographies, partitions, vidéos, marionnettes, médailles, etc.), promise à une croissance continue : elle franchissait le cap du million de documents en 2010, en comptait 5 millions en 2019 et 6 millions début 2020. Cette masse de documents est indexée par les grands moteurs de recherche et rencontre les intérêts tant des chercheurs que du grand public. Une étude d'usages approfondie menée en 2016 révèle le poids des recherches personnelles aux côtés des recherches plus académiques. Le site reçoit quelque 50 000 visites quotidiennes (ce chiffre passe à 70 000 en période de confinement). 50 % de ses contenus ont été consultés au moins une fois en 2019 (soit près de 3 millions de documents).

Les contenus librement accessibles de cette bibliothèque forment, en eux-mêmes, une collection organisée dont les contours sont sans cesse interrogés par l'évolution des modalités de diffusion, d'exploration et d'appropriation au sein d'un Web dans lequel les accès sont sans doute moins intuitifs, moins simples et limpides qu'il n'y paraît : la recherche d'information et la capacité à identifier des sources pertinentes relèvent davantage de compétences réelles que de l'intuition, et les choix faits en matière de transmission sont aussi importants que les contenus transmis eux-mêmes.

Son enrichissement est le fruit de nombreuses années de numérisation du patrimoine documentaire de la BnF auquel se sont adjoints d'importants gisements documentaires provenant d'autres bibliothèques : Gallica est une bibliothèque numérique collective qui entraîne avec elle à ce jour plus de 400 institutions.

Dans ce contexte, Gallica se définit davantage comme un dispositif évolutif multidimensionnel : le site gallica.bnf.fr est complété par des applications (pour Androïd et iOS), par une version intra-muros enrichie par des documents sous droits, par des marques blanches… La BnF développe ce dispositif selon trois pistes : l'amélioration constante du référencement, un positionnement en tant qu'acteur de confiance, et la capitalisation au profit d'un patrimoine numérique du XXI$^{e}$ siècle en cours de constitution.

## Référencer six millions de documents et rendre possible la trouvaille

### Gallica, un Web dans le Web, ou la multiplication des portes d'entrée

Tout comme le Web en général, Gallica requiert des contenus disponibles et directement accessibles, des serveurs capables de gérer des connexions massives, un moteur apte à traiter des requêtes complexes en temps réel, et des interfaces familières et continues.

### *Parler aux moteurs*

Moins d'un utilisateur sur cinq passe par la page d'accueil de la bibliothèque numérique, et 39 % des visites proviennent d'un moteur de recherche : Gallica est un réservoir dont le centre est potentiellement partout, et la circonférence, nulle part… Si la longue traîne, définie par un principe de rareté, est naturellement référencée par ces moteurs, il en va autrement des contenus les plus communs, les plus étudiés, les plus sujets aux homonymies. Excluant tout achat de liens sponsorisés, la BnF mène depuis dix ans une entreprise de traduction et de structuration de ses données dans des formats connus des moteurs, adoptant les principes du Web sémantique : cette traduction est au fondement de data.bnf.fr (2012), qui surplombe la variété des contenus numériques de la Bibliothèque, et dont 84 % de la fréquentation vient aujourd'hui des moteurs. Cette colonne vertébrale agit comme un pivot vers les autres sites de la BnF, elle hisse les contenus patrimoniaux des institutions aux premiers rangs dans les résultats des requêtes.

### *L'enjeu des API*

Chacun des atomes de Gallica peut être trouvé non seulement sous une forme classique (inséré dans une page du site) mais également directement *via* une ligne de code (URL). De ce point de vue, le développement de l'API IIIF a grandement favorisé l'interopérabilité des contenus et leur dissémination : elle permet en effet d'appeler et de manipuler des contenus iconographiques [(1)] depuis un site pour les diffuser sur un autre sans obliger l'internaute à changer d'interface. Parmi les utilisateurs de Gallica qui ont choisi cette solution, on peut citer par exemple un site de généalogie équine [(2)], les Archives polaires françaises [(3)], ou encore un site commercial de vente de coques personnalisables pour smartphones [(4)].

### *S'appuyer sur des relais*

Au-delà des progrès technologiques, ces usages relèvent globalement d'une politique de relais qui inclut aussi une présence active sur les réseaux sociaux et l'appui sur une communauté – les Gallicanautes – dont la configuration et l'implication évoquent les communautés des débuts du Web. Ils interagissent avec la BnF, partagent leurs trouvailles.

Les liens vers Gallica sur le site Wikipédia représentent également une source croissante de visites : les visites de Gallica en provenance de ce site ont progressé de 30 % entre 2018 et 2019.

Enfin, les marques blanches, qui permettent aux institutions de profiter de l'infrastructure de Gallica pour placer leurs collections numériques sur le chemin de leurs publics tout en enrichissant la collection collective de nouveaux documents, s'avèrent un véhicule de plus en plus plébiscité. De cette mutualisation de l'infrastructure de Gallica résulte un dispositif de coopération vertueux, qui répond aux enjeux de sobriété numérique et d'intelligence informationnelle. Dix marques blanches sont en ligne, dix sont en préparation [(5)].

## Outiller pour donner à voir l'inédit

La communauté professionnelle mobilisée pour augmenter la diffusion des contenus sur le Web se rassemble également autour de l'enjeu de leur exploration. Elle s'appuie pour cela d'une part sur un dialogue constant avec les chercheurs, d'autre part sur les perspectives technologiques, en particulier sur l'évolution du moteur de recherche de Gallica (Cloud view, de Dassault systèmes).

(1) Voir https://iiif.io/
(2) https://www.pedigreequery.com/
(3) https://www.archives-polaires.fr/
(4) https://cover.boutique
(5) https://www.bnf.fr/fr/cooperation-autour-de-gallica#bnf-gallica-en-marque-blanche

Si les outils de fouille au cœur de la totalité des images de Gallica sont encore en cours de développement – ils progressent activement à la faveur de l'intelligence artificielle, et des expérimentations sont déjà disponibles sur certains fonds – les outils de fouille de texte permettent d'accompagner le travail d'exploration et de dépouillement des collections (c'est notamment le rôle du rapport de recherche [6]) et d'interroger les documents selon des modalités inédites (ainsi en est-il de la recherche par proximité [7] ou de l'analyse de vastes corpus [8]).

Gallica contribue à la création de nouveaux métiers, mais elle révolutionne aussi les activités et métiers déjà existants par la richesse des contenus auxquels elle donne accès et le gain de temps qu'elle permet : les doctorants, les chercheurs, les journalistes, les auteurs, ou les dessinateurs s'en font régulièrement l'écho, tels Pierre Lemaître, Daniel Schneidermann, Alain Rey, Benoît Peeters, Maylis de Kerangal…

Cependant, la multiplication des portes d'entrée, des sites et des outils, pas plus que le renouvellement du dialogue avec les utilisateurs, ne suffisent à une navigation fluide et tranquille au sein de ces contenus : il faut y placer des repères.

## Gallica, source de culture et acteur de confiance sur le Web

### Editorialiser, conseiller

Comment trouver tout de suite la bonne édition des *Pensées* de Pascal ? Où figurent les premiers textes de référence sur l'intelligence artificielle ou bien sur les problématiques climatiques ? Gallica n'apporte pas tant l'information en tant que telle que la source de l'information. Elle seconde l'esprit qui explore. Elle se décline à travers des dispositifs visant certains publics en particulier : les étudiants du BTP (site *Passerelle(s)* [9]), le jeune public (Gallicadabra [10]), etc. Elle met en valeur le travail de lecture et d'analyse des contenus réalisé par des Gallicanautes… La capillarité entre data.bnf.fr, les marques blanches et les réseaux sociaux fait naître un niveau affiné de conseils (« Gallica vous conseille » en tête des listes de résultats du site), de sélections, de billets de blog, etc., à travers lesquels la bibliothèque joue son rôle de référence, et qui complètent les traitements algorithmiques par de l'intelligence humaine.

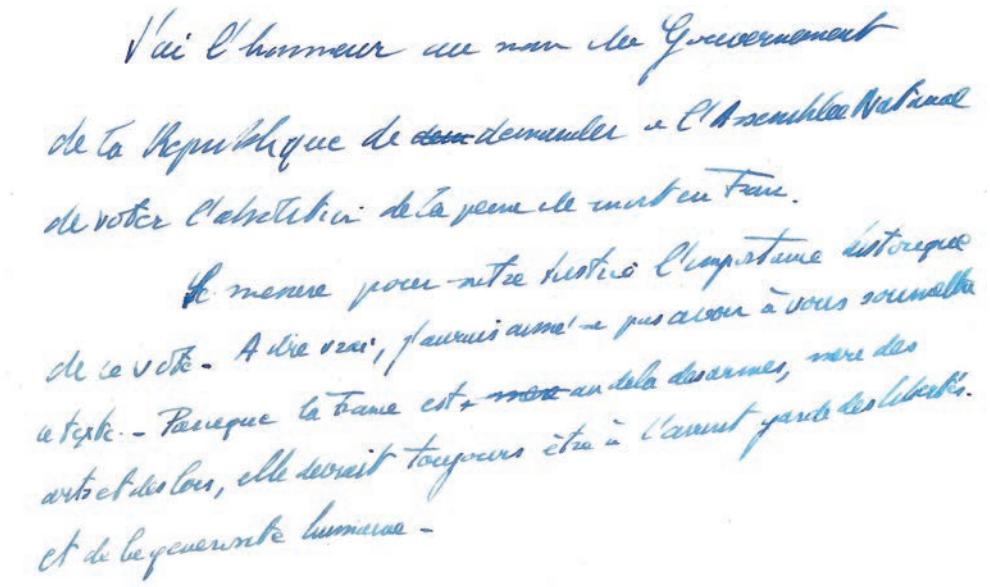
Ce travail de médiation s'intéresse en particulier aux trésors (du manuscrit du discours prononcé en 1981 par Robert Badinter contre la peine de mort [11] aux anciens numéros de la présente revue [12]…), aux essentiels (les éditions de référence des classiques de la littérature, du droit, de la politique…), et aux documents plus ou moins anciens qui entrent en résonance avec l'actualité.

### Ethique du moteur interne et du site tout entier

En tant que source, en tant que service public, la galaxie Gallica respecte des principes de sécurité, de neutralité, d'ouverture, de légalité, et de stabilité.

Les contenus consultés lors de sessions précédentes, les parcours des internautes, ne sont pas utilisés pour modifier l'ordre des résultats dans la liste, ni pour enfermer un utilisateur dans un environnement déterminé par ses préférences. Le travail d'un chercheur qui, après une patiente recherche, aurait rapproché plusieurs titres, n'est pas divulgué *via* des suggestions de consultation aux autres internautes.

(6) https://c.bnf.fr/G97
(7) https://c.bnf.fr/Haa
(8) Voir en particulier Pierre-Carl LANGLAIS, *Reconstituer les genres romanesque sur Gallica : essai de classification automatisée de 1500 romans* (1815-1850), https://scoms.hypotheses.org/986
(9) http://passerelles.bnf.fr/
(10) https://c.bnf.fr/Hav
(11) https://c.bnf.fr/Hap
(12) Par exemple, cette série : https://c.bnf.fr/Has



Robert Badinter, manuscrit autographe du discours sur l'abolition de la peine de mort prononcé à l'Assemblée nationale le 17 septembre 1981 (https://c.bnf.fr/IaD)
© Bibliothèque nationale de France

Impliquée dans les débats nationaux et internationaux sur les questions juridiques, la BnF est parvenue en outre à une position d'équilibre en matière d'accès à ses ressources, *via* une formule articulant les enjeux de l'*open data*, ceux de l'*open science*, et les règles de l'univers marchand. Elle opère une distinction entre les données – entièrement libres – et les contenus numérisés, soumis à différents régimes. La politique de l'*open data* mise en œuvre dès 2014 s'inscrit dans une politique de l'État visant à faire émerger de nouveaux usages citoyens, voire de nouvelles opportunités économiques. En ce qui concerne la réutilisation des contenus numérisés librement accessibles en ligne, la BnF promeut la gratuité dans l'univers gratuit, éducatif et académique, et soumet à des redevances les acteurs de l'univers marchand.

Enfin, quiconque cite un document de Gallica, peut être sûr non seulement que le lien restera valable (la BnF va même jusqu'à proposer son propre outil pour raccourcir les URL, c.bnf.fr, qui garantit la pérennité des liens courts) mais que le contenu cité restera le même. Ainsi, l'étude d'usages de 2016 a fait apparaître une très forte progression des pratiques de consultation attentive en ligne [13] entre 2011 et 2016, qui rend inutile le téléchargement de sauvegarde.

## Gallica, un terreau pour le patrimoine numérique du XXIe siècle

Le terrain exploré par la BnF dans le cadre des évolutions de Gallica concerne d'une certaine manière le patrimoine à venir : en tant que source, Gallica conduit à de nouvelles créations, et les dispositifs mis en place permettent d'étendre au-delà du Web les fonctionnalités et les usages qu'il a lui-même suscités.

### L'ouverture des possibles et la co-construction

Outre son mérite de déplacer l'évaluation des résultats vers un champ davantage qualitatif que quantitatif, la focalisation sur les usages à laquelle nous invite le dossier de ce numéro permet de

(13) 66 % des répondants disent le faire « souvent » ou « à chaque fois », contre 31 % en 2011. https://multimedia-ext.bnf.fr/pdf/mettre_en_ligne_patrimoine_enquete.pdf

montrer combien une approche globale gagne à considérer non seulement les usages individuels, mais aussi ce qu'ils peuvent avoir de réticulaire : il s'agit moins d'une somme d'usages particuliers juxtaposés que d'interactions constantes et inspirantes. Le besoin de personnalisation – que la BnF assume à travers des dispositifs comme « Gallica vous conseille », le rapport de recherche, la numérisation à la demande, Adoptez un livre, etc. – ne s'envisage pas sans une dimension collaborative, sans un transport collectif.

L'univers du Web, qui associe étroitement la lecture et l'écriture, favorise cette conception des usages. A la BnF, par exemple, le hackathon de 2016 a donné naissance à Gallicarte, à présent implanté dans Gallica : un mois après le lancement de cette fonctionnalité de géolocalisation, 5000 points supplémentaires en bénéficiaient grâce à l'implication des internautes (14). L'usage peut donc aussi consister en de l'annotation, de la correction d'OCR, de la transcription de manuscrits… Comme l'expliquent Henri Verdier et Nicolas Colin, dans *L'âge de la multitude*, « il y aura presque toujours plus d'intelligence, plus de données, plus d'imagination et de créativité à l'extérieur qu'à l'intérieur d'une organisation » (15).

Pour cette raison, la BnF doit offrir, maintenir et encourager une multiplicité d'usages possibles, en commençant par les plus basiques : imprimer, télécharger, etc., sans négliger, en regard, les usages physiques. « Souvent qualifiés d'usagers "distants", les Gallicanautes nous rappellent que les offres physiques et numériques se nourrissent l'une de l'autre : 38 % disent avoir déjà fréquenté les espaces physiques de la BnF » en 2016 (16). La force de la bibliothèque est de ne pas dissocier ces formes de matérialité, et elle est sans doute l'un des lieux les plus indiqués pour garantir leur coexistence.

### L'exploration du dépôt légal numérique

Enfin, la puissance de l'outil Gallica profite aux documents numériques protégés par la propriété intellectuelle et à la recherche documentaire en général : *Gallica intra-muros* donne aujourd'hui accès à un million de documents supplémentaires. À la faveur du dépôt légal numérique, elle sera demain bien plus importante que Gallica, et les chercheurs pourront aussi faire de la fouille sur place, dans le DataLab de la BnF. Le cercle commencé avec Gallica est vertueux : les outils seront d'autant plus puissants que la BnF recevra ce dépôt légal protégé. C'est un enjeu de supériorité informationnelle.

## Conclusion

Si la dimension révolutionnaire de la bibliothèque numérique est à nuancer – elle est le reflet de pratiques transdisciplinaires et documentaires qui lui préexistent – la visibilité spécifique du patrimoine sur le Web et l'extension de ses usages recouvrent un triple enjeu: l'entraînement des intelligences artificielles sur des documents non contemporains (en particulier sur le patrimoine iconographique), l'intégration massive de documents francophones dans les corpus numériques, et l'éclairage des débats d'actualité.

Pour honorer ces enjeux, la BnF s'appuie à la fois sur l'engagement professionnel de ses personnels, sur une attention fine à ses publics, sur la technologie – de la fouille perfectionnée à l'interrogation des contenus en langage naturel – et sur la coopération, afin que le patrimoine en ligne soit véritablement partagé.

(14) https://gallica.bnf.fr/blog/21032018/gallicarte-arrive-dans-gallica?mode=desktop
(15) Armand Colin, 2015, p. 58
(16) https://gallica.bnf.fr/blog/10052017/resultats-de-lenquete-2016-aupres-des-usagers-de-gallica?mode=desktop

# Artificial Intelligence: Challenges for the future

By **Michalis VAZIRGIANNIS**
Professor at DaSciM, LIX [(1)], Ecole Polytechnique

*"The means of production are only decommissioned if a more productive replacement is found. In other words, we can only talk about the abolition of slavery if only the goods can be produced without the help of slaves: Daedalus or the tripods of Hephaestus, about whom the poet says that they automatically appeared in the assembly of the gods, and if the shuttles weaved on their own and the guitar strings played on their own, then the architects would not need servants neither the free people slaves."*

*Politics, Aristotle.*

## Introduction

Artificial Intelligence (AI) is a term coined by Prof. J. McCarthy, one of the "founding fathers" of the area in 1955. AI may assist, augment or eventually automate processes. Colloquially AI can be used to describe machines (or computers) that mimic "cognitive" functions associated with the human mind, such as "learning" and "problem solving" [6] and can be used to tackle many of the problems usually solved by humans such as classification, summarisation, translation, diagnosis, grouping, finding outliers. AI algorithms have already found applications in most domains of human activity including finance, national security, health care, justice, transportation, smart cities, industrial processes education, etc.

AI algorithms function effectively they need sufficient data in order to be trained but as well, especially recently with the advent of deep learning, significant computing power. Hopefully in the last decades data have become abundant – the term *big data* has been coined to describe the volume, speed of production and heterogeneity of the the data produced by socioeconomic processes [4].

As for computing resources there are different aspects. The *storage capacity* that has been increasing almost exponentially in the last decades [5] enables storage of unprecedented amount of data that correspond to practically any activity that can produce data – i.e. user behaviour at large, image and video, sensors' readings from different processes, loads of financial transactions, etc. Here, the role of cloud storage (i.e. distributed remote storage of information with availability guarantees) and software-as-service are significant as they alleviate small businesses and governments from the cost of purchase and maintenance of complex specialized infrastructures.

As regards the computational power the CPU capacity increases continuously in the last decades, converging though to a plateau. On the other hand, the advent of DL required the development and wide deployment of GPUs. This is a critical infrastructure which is currently the heart of resources needed for the large-scale computations.

(1) http://www.lix.polytechnique.fr/dascim/

The main stream AI methods are emanating from the field of *machine learning* where the scenario is that there is a training set (i.e. a set of emails labelled as "spam") where the data points (i.e. emails) are embedded in a feature space (i.e. the union of the words in the email set). The aim it to fit a mathematical function (i.e. regression) or other structure (i.e. decision tree) that given data of unknown classification produces a prediction (i.e. "spam/no spam"). This type of learning is known as "supervised" since there is some prior knowledge in the training set (i.e. which emails are "spam").

On the other hand, we have cases where supervision (i.e. prior knowledge on the data) is not available. In this case we have only the positions of the data in the feature space and we are seeking the organization of the data into cohesive groups indicating correlation – this is the case of *unsupervised learning*. The prevalent methods in this area belong to the family of algorithms under the term "clustering" [2]. Other approaches are seeking for repeating patterns and associations in the data. A famous approach includes the "association rules" methods that seeks, usually in retail data, highly correlated product purchases trying to detect a kind of causality based on conditional probabilities [3]. Numerous applications include retail, biomedical applications, recommendations, etc.

As mentioned above, in the last decade due to the presence of abundant data and computing power as well as the vast industrial investment (i.e. GAFA, etc.), we testified the rise and dominance of Deep Learning (DL) – as a reincarnation of Artificial Neural Networks [19] – long standing methods idling for decades. DL is based on the perceptron base learner [18] and consists of potentially complex architectures with (say $k$) consecutive layers of perceptrons? with or without feedback. Deep learning essentially – in the presence of a large enough amount of data, say X - is able to learn an internal representation Y of the input data X via the successive mappings through the layers aforementioned. This representation Y can be called embedding or even be represented by a potentially very complex non-linear function $Y = f_1(f_2 (..f_k(X)))$, where $k$ the number of layers from the input to the output. The coefficients of this function are essentially the weights of the connections among the perceptrons of the different layers and are learned via an error function that is used and the famous back-propagation process [7]. The representation Y can be used in supervised learning tasks (i.e. image or document classification) or end to end tasks (i.e. machine translation, document summarization, etc.). An interesting variance of DL is the auto encoder [8] approach where the task is to train an architecture whose output (decoder) should be the data themselves. Recently the area of graph neural networks (GNNs), considering graph as input to the DL architectures, is arising as a powerful alternative to traditional matrix-based data representations. The importance of DL is already cornerstone as in the presence of lots of data tasks of unprecedented complexity can be efficiently tackled by them.

Unsupervised learning is the most challenging and promising part of AI since the pace of data production is overwhelming and supervision is less and less possible.

## AI impact

Several years after IBM's Deep Blue defeated chess grandmaster G. Kasparov, human-computer cooperation bloomed. However, in recent years AI systems are so so good at playing games (i.e. chess) that their human counterparts seem to be of less value and this can be regarded as a precursor to what might happen at a more pervasive level. Another recent approach, DeepMind's Alpha Zero [13], went from utter ignorance to creative mastery in under four hours without the help of any human guide, to dominate the world's best AlphaGo players and programs.

AI is beyond a technology wave as it penetrates almost all aspects of human life - work, privacy, government – even questioning the human nature with advanced robots (cyborgs). Therefore, a

large-scale discussion among all social and economic parties [11] should decide the limits of AI with regards to ethics, in terms of how much decision power should be left to AI, its impact to democracy. Another important aspect is *explainability* - whether we can ask for an explanation for AI decisions coming from very complex algorithms.

The *impact of AI* in economy and society is already very visible and will be increasingly significant in decision making, business models, risk mitigation, and system performance in many different application domains. The adoptions of AI algorithms in economy and production is expected to affect the world's GDP that will be increased by $15.7 trillion, a full 14%, by 2030. A significant part of this will be in China [9].

## AI challenges for the future

The *future of AI* is a challenging debate. There are theories that AI will converge to *singularity* meaning that our technological creations exceed the computing power of human brains. There are predictions, based on Moore's Law and the general trend of exponential growth in technology, that singularity will come before the mid-21st century according to Ray Kurzweil, Google's director of engineering [10]. This means that AI will supersede human intelligence and that AI algorithms will potentially realise their existence and start to behave selfishly and cooperatively.

On the other hand, significant figures of Science and Technology such as Stephen Hawking, Elon Musk, warn [12] towards AI and the singularity. Potential risks include that AI will relieve us or deprive us from our work and a lot of jobs will be replaced by automated processes, even non repetitive ones such as chatbots responding intelligently to human spoken requests.

Also, the aspect of national and European sovereignty for AI is significant and discussed by the Villani report [11] aiming at maintaining local resources and talent in this strategic area. In this context since AI resources and capabilities are of cornerstone importance. We need to reflect of the following aspects:

- *Control and access to the data*. Data are like fuel to the AI engine. Currently the vast amounts of data produced by user's behaviour and interactions are mainly owned by private entities that capitalise on knowledge extracted from these data. This enables them to develop further their competences and products on AI. Therefore, it becomes essential that data are available to the research community and the governments for policy and decision making. A key political question of our era is "how do you regulate the ownership of data?" [1]. Of course, then the issue or privacy becomes important.
- *Computing resources/GPUS*. As mentioned above the capability to apply DL algorithms on bigdata depends totally on the availability of GPUs - the specialized processors for fast matrix computations necessary in deep learning. In the western world, GPUs are produced practically by a handful of industries mostly based in one country. On the other hand, large scale installations of GPUs – enabling the necessary computations are can only be afforded by a few large industries (mainly GAFAs) or governments. This oligopoly apparently poses threats for the future capacity of other governments, SMEs and academia to continue doing competitive research and advancing technology.
- *Access to AI algorithms*/methods. AI algorithms form the heart of AI, so the capacity to design and develop algorithms gives a very high added value and competitive advantage to the owners of the algorithms. For example, searching in the web is a ubiquitous activity giving access to information matching users' queries and is dominated in the western world by Google. It's ranking algorithm though – responsible for prioritising the best results – is a well-kept secret. On the other hand, a large number of popular machine learning algorithms is available as opens source code within the *scikit-learn* library [15] that has made possible the dissemination

of machine learning throughout the academic and industrial community with significant added value accumulated. Similarly for DL Tensorflow [16] and Pytorch [17] have made possible the breath-taking development of deep learning methods. The value of having algorithms as open source code guarantees transparency and increases productivity and progress in this area. A very interesting such case at governmental level is the Translago portal [20] of the French state providing access to all the algorithms used by the state administration to take decisions.

- *Deepening inequality gap among states and social classes.* According to a recent report of McKinsey, half of the world's jobs could be automated by 2055 [14]. This wave of automation will affect classic middle-income white-collar jobs, such as bank tellers, insurance underwriters, loan officers and case-file workers, essentially each job that includes following rules and making few decisions. Thus, the rich-get-richer dynamics of the digital economy is leading to monopolising sectors, where companies like Facebook, Google, Apple and Amazon are eliminating competition.

Here we only present challenging aspects of AI for the future. Tackling these problems is a highly complex task that touches upon the synergy of socio-economic and political players worldwide. The only certainty we have is that AI is already happening and transforms the world very fast. We can only hope that the decisions of societies and governments will lead humanity to enjoying the added value of AI rather than concentrating it to a handful of powerful players.

## References

[1] Harari, Yuval Noah; Spiegel & Grau (2018). *21 Lessons for the 21st Century*. ISBN 9780525512172.

[2] Rui Xu and D. Wunsch, "Survey of clustering algorithms," in IEEE Transactions on Neural Networks, vol. 16, no. 3, pp. 645-678, May 2005.

[3] Agrawal, R.; Imieliński, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207. CiteSeerX 10.1.1.40.6984. doi:10.1145/170035.170072. ISBN 978-0897915922.

[4] X. Wu, X. Zhu, G. Wu and W. Ding, "Data mining with big data," in IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, Jan. 2014.

[5] The Evolution of Hard Disk Drives and Storage Capacity, https://www.securedatarecovery.com/blog/hdd-storage-evolution

[6] Russell, Stuart J.; Norvig, Peter (2009). *Artificial Intelligence: A Modern Approach* (3rd ed.). Upper Saddle River, New Jersey: Prentice Hall. ISBN 978-0-13-604259-4.

[7] David E. Rumelhart, Geoffrey E. Hinton & Ronald J. Williams, "Learning representations by back-propagating errors", *Nature* volume 323, pages 533–536 (1986)

[8] Hinton, G. E., & Zemel, R. S. (1994). "Autoencoders, minimum description length and Helmholtz free energy". In *Advances in neural information processing systems* 6 (pp. 3-10).

[9] Sizing the prize, PwC's Global Artificial Intelligence Study: Exploiting the AI Revolution, What's the real value of AI for your business and how can you capitalise? https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.htm

[10] https://www.sciencealert.com/google-s-director-of-engineering-claims-that-the-singularity-will-happen-by-2029

[11] Villani, Cédric & Schoenauer, Marc & Bonnet, Yann & Berthet, Charly & Cornut, Anne-Charlotte & Levin, François & Rondepierre, Bertrand. (2018). *Donner un sens à l'intelligence artificielle : Pour une stratégie nationale et européenne*

[12] https://www.bbc.com/news/technology-30290540

[13] Silver, David; Hubert, Thomas; Schrittwieser, Julian; Antonoglou, Ioannis; Lai, Matthew; Guez, Arthur; Lanctot, Marc; Sifre, Laurent; Kumaran, Dharshan; Graepel, Thore; Lillicrap, Timothy; Simonyan, Karen; Hassabis, Demis (December 5, 2017). "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm". arXiv:1712.01815 [cs.AI].

[14] https://www.mckinsey.com/featured-insights/digital-disruption/harnessing-automation-for-a-future-that-works

[15] https://scikit-learn.org/stable/

[16] https://www.tensorflow.org/

[17] https://pytorch.org/

[18] Rosenblatt, F. (1958). "The Perceptron: A Probabilistic Model For Information Storage And Organization In The Brain". Psychological Review. 65 (6): 386–408. doi:10.1037/h0042519.

[19] Kleene, S.C. (1956). "Representation of Events in Nerve Nets and Finite Automata". Annals of Mathematics Studies (34). Princeton University Press. pp. 3–41.

[20] https://www.transalgo.org/le-projet/

# Une toile de fond pour le Web : lier les données et lier leurs vocabulaires sur la toile, pour un Web plus accessible aux machines

Par **Fabien GANDON**
Directeur de Recherche
Inria

Attention…Top ! Je parle toutes les langues. J'ai plus de trois milliards d'utilisateurs directs. Je peux être privé ou public. Je m'étends sur tous les réseaux du monde. Je suis derrière votre réservation de vacances, dans votre téléphone, vos échanges avec une assurance ou votre livre électronique… je suis … je suis… je suis… Le Web, la toile ou l'ouaibe, pour nos cousins d'outre-Atlantique. Nous avons toujours cherché dans l'Histoire des moyens efficaces de collecter et d'accéder aux masses d'informations que nous créons, c'est notamment la raison d'être des bibliothèques. Ce fut aussi la motivation de Tim Berners-Lee lorsqu'il proposa en 1989 de tisser un système d'hypertexte global [F] au CERN pour mieux partager les informations dans un campus où plusieurs milliers de personnes se croisent avec de multiples spécialités et instruments.

Aujourd'hui encore, il est frappant de voir à quel point le Web est à la fois très connu et mal connu, comme en témoigne la confusion tenace entre les termes Web et Internet que l'on rencontre encore bien trop souvent. Malgré le fait que leurs inventeurs respectifs aient reçu deux prix Turing bien distincts, respectivement en 2004 et en 2016 pour deux inventions bien différentes, Internet et Web sont encore trop souvent utilisés de façon interchangeable.

Redisons-le : Internet permet l'interconnexion des réseaux d'ordinateurs et objets connectés en général. Il fournit une infrastructure de communication qui supporte au-dessus d'elle de nombreuses applications comme : la messagerie électronique (mail), la téléphonie et la vidéophonie… et le Web, cet hypermédia distribué qui devient l'architecture logicielle majoritaire des applications sur Internet. Une autre information encore moins connue à propos du Web est que depuis la fin des années 1990, il n'est plus uniquement consulté et utilisé par nous, les humains, mais aussi par les machines, notamment dans ses versions que l'on appelle Web de données et Web sémantique.

## Un réseau hypermédia de ressources

Dès 1996, Tim Berners-Lee fait une relecture de l'architecture du Web [E] en insistant sur trois concepts-clés : l'adressage (les adresses Web ou URL et URI de la forme http://inria.fr), le protocole de transfert de données pour le Web (HTTP) et le mécanisme de négociation de contenu. Ce dernier est un mécanisme du protocole HTTP qui permet à un serveur Web de fournir pour une même adresse URL différentes représentations d'une même ressource en fonction de ce qu'il sait sur celui qui l'interroge. Par exemple, s'il connaît les langues parlées par celui qui accède à une adresse, il pourra préférer lui servir une version de la réponse dans une de ces langues. Cette négociation de contenu a lieu à chaque fois que nous accédons au Web et sans même que nous le voyions.

Les possibilités de la négociation de contenu vont plus loin que cet exemple et, d'une certaine façon, déclassent la célèbre « page Web » en nous donnant la possibilité de négocier auprès d'un serveur Web différents types de formats de réponse. Ainsi le Web n'est pas limité à une toile de documents

mais offre la possibilité de servir et lier tout et n'importe quoi. En effet, comme les URI permettent d'identifier tout type de ressources (une page, une image, une personne, un produit, une molécule, etc.) et pas uniquement les contenus du Web, on peut dès lors utiliser la toile et ses langages pour décrire et lier tout ce que l'on sait identifier dans le monde. Le Web affirme ainsi son indépendance à un modèle ou une structure de données et le langage HTML des pages Web redevient donc juste un prérequis pour un navigateur [T]. Il aura permis dans un premier temps de fournir un format uniforme de documents hypertextuels et de documentariser le réseau de ressources que devient le Web. La toile est prête à échanger beaucoup d'autres choses que des pages.

## Un Web plus accessible aux machines

En 1996, le langage PICS permet de standardiser le filtrage des contenus inappropriés, notamment pour les enfants. Incidemment, PICS ouvre aussi l'idée d'étiqueter les contenus avec des données pour les machines. Le Web s'ouvre à la notion de métadonnées en général ; commence alors une évolution vers un Web de documents et de données structurées.

Dans cette évolution, le langage de feuilles de styles CSS est une étape importante qui marque le début de la séparation du fond et de la forme sur le Web. La notion de feuille de style permet de sortir et séparer la mise en forme de la structure du document et d'utiliser une même mise en forme pour plusieurs documents, ou inversement de faire varier la mise en forme d'un même document. Peu après, le Web connaît une nouvelle évolution avec le standard XML permettant de créer et gérer ses propres structures de documents et données. Prolongeant cette évolution, Tim Berners-Lee publie en 1998 une feuille de route pour ce qu'il appelle le Web sémantique [AK]. Elle est dans la continuité de sa présentation de 1994 [S] et aussi de son article de la même année [T]. On peut lire dans cet article qu'il souhaite une évolution des objets du Web, qui sont à l'époque essentiellement des documents destinés aux humains, vers des ressources avec une sémantique plus orientée vers les machines pour permettre des traitements plus automatisés. La feuille de route de 1998 [AK] ouvrira la voie à tous les travaux sur le Web de Données et le Web sémantique et aux standards qui en découlent (RDF, RDFS, SPARQL, OWL, etc.).

## Web sémantique : quand le lien fait sens

Si la notion de Web sémantique date de 1998 et que l'article grand public le plus connu date de 2001 [BZ], elle peut, vingt ans après, s'expliquer en deux grandes étapes : premièrement, la notion de données liées et le Web de données et, deuxièmement, la notion de schémas liés et le Web sémantique.

### Données liées et Web de données

Le principe des données liées est de créer des liens entre les données, comme on crée des liens entre des pages et, par extension, de créer des liens entre les bases de données, comme on crée des liens entre les sites.

La première étape est d'utiliser les identifiants standardisés par le Web pour identifier les sujets et relations de ces données. Ainsi, je peux forger une adresse Web (URI) pour identifier une berline du parc automobile de mon entreprise : http://www.mon-entreprise.fr/voiture/berline-n3
On imagine bien que la voiture concernée n'est pas accessible par le Web à cette adresse mais que cette adresse peut la représenter dans des données qui la décrivent. Ce même identifiant pouvant être réutilisé dans plusieurs sources de données, il permet alors de faire des liens entre ces données et entre ces sources : on parle de données liées.

De même, je peux forger une adresse Web (URI) pour identifier une relation entre une personne et une voiture : http://www.mon-entreprise.fr/voiture/estConducteurDe

A nouveau, ceci fournit un identifiant pour représenter l'ensemble des occurrences d'une relation de conducteur entre une personne et une voiture. Cet identifiant peut être réutilisé par autant de jeux de données qui le souhaitent et ainsi favoriser une interopérabilité entre les applications consommant ou produisant ces données.

Dans ces deux exemples, les URI utilisent le protocole HTTP, *i.e.* elles commencent par http://... On parle d'alors d'URI HTTP. On ne les qualifie pas d'URL car contrairement à une adresse (par exemple http://fabien.info) elles ne correspondent pas à une ressource (par exemple une page Web) disponible sur la toile. Par contre, le fait d'utiliser le protocole HTTP permet d'avoir un mécanisme par défaut pour découvrir ce que ces URI identifient en accédant à l'adresse qu'ils donnent pour obtenir des données à leur propos. C'est ce que l'on appelle la déréférenciation.

Le Web de données met donc en relation des sources de données plus ou moins grandes en reposant sur l'architecture classique du Web et en l'étendant. Comme ce ne sont donc plus uniquement des pages qui sont liées sur le Web, mais des identifiants de ressources arbitraires, se pose la question de ce que l'on doit obtenir lorsque l'on accède à de tels identifiants. Lorsqu'un identifiant est consulté, les serveurs répondent en fournissant des données décrivant la ressource sans que celle-ci soit nécessairement sur le Web (par exemple, une voiture, une espèce animale, une protéine, un auteur...) et en s'adaptant à celui qui interroge les données. Ainsi, pour un même identifiant et grâce au mécanisme de négociation de contenu, un utilisateur devant son navigateur recevra une page Web en HTML pour sa lecture, là où un agent logiciel recevra des données à intégrer à sa base de données.

On identifie classiquement cinq étapes et critères incrémentaux de qualité pour la publication de données ouvertes liées cinq étoiles sur le Web :

| | |
|---|---|
| ★ | les données sont sur le Web sous licence libre |
| ★★ | *idem* + les données sont explicites et structurées |
| ★★★ | *idem* + les données sont dans un format non propriétaire |
| ★★★★ | *idem* + des URI HTTP sont utilisés pour identifier sujets, objets et types de relations |
| ★★★★★ | *idem* + les données sont liées à d'autres données |

L'évolution du Web documentaire vers le « Web des données » repose sur ces principes et standards, permettant de tout identifier et de tout décrire sur la toile, et de tisser ainsi un graphe de données mondial. En appliquant ces principes, là où on avait avant un Web de documents (les pages Web) essentiellement à consommation humaine, à partir des années 2006-2007, on ajoute un Web reliant des bases de données de toutes tailles et sur tous les sujets (Figure 1), essentiellement à consommation des machines, qui peuvent les parcourir, suivre les liens pour trouver de nouvelles sources et naviguer et chercher ce Web de données comme nous naviguons et cherchons sur les pages du Web. L'appellation « Web de données » insiste donc sur la possibilité d'ouvrir nos silos de données de toutes tailles, depuis notre agenda jusqu'aux immenses bases géographiques, et de les échanger, de les relier, de les composer selon nos besoins.

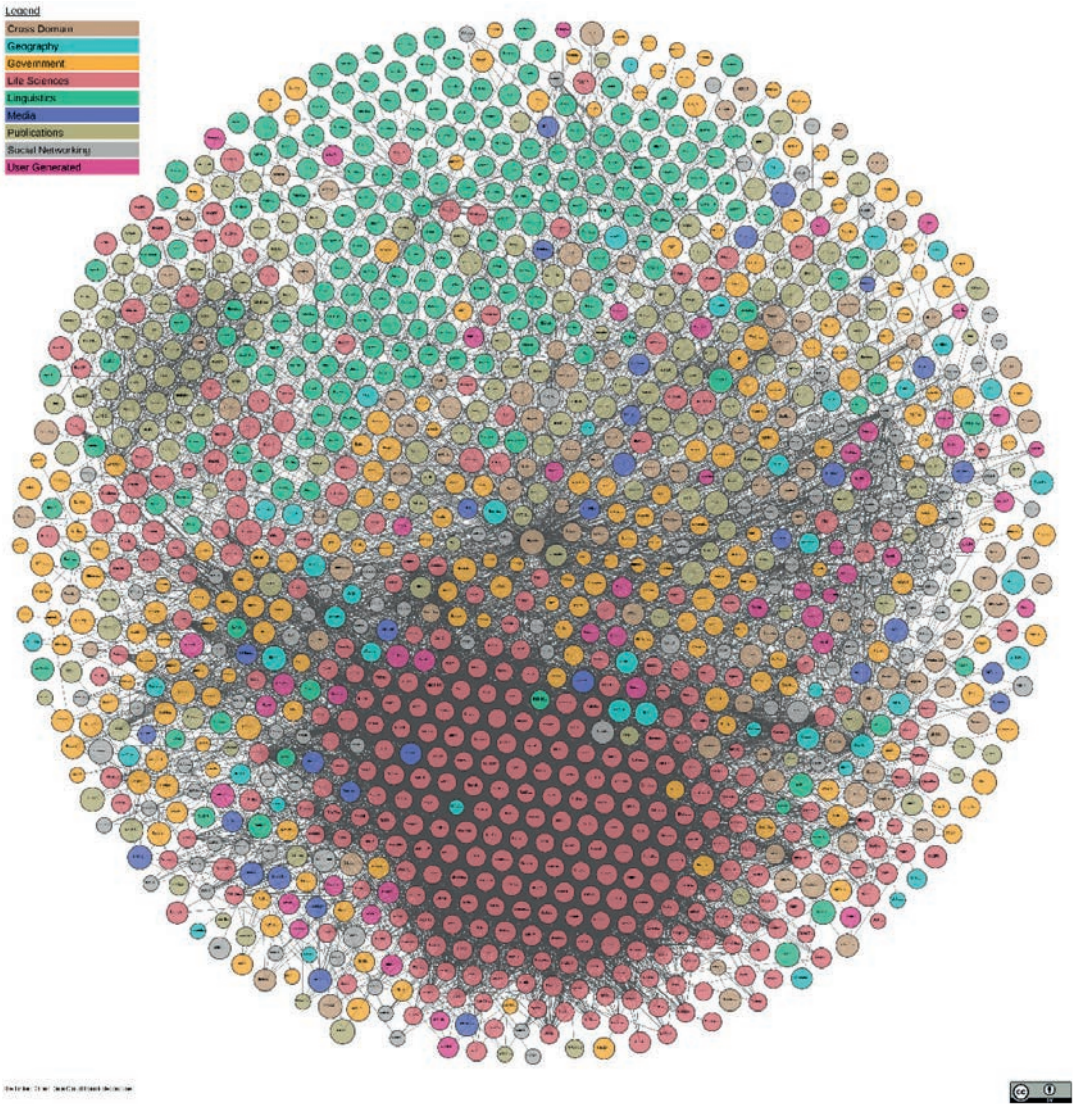



Figure 1 : Le nuage d'une partie des bases de données ouvertes liées du Web le 29/03/2019 par le site lod-cloud.net (The Linked Open Data Cloud, https://lod-cloud.net/).

Pour se représenter l'impact de ce changement et les volumes de données qui deviennent disponibles, nous allons prendre quelques exemples. L'une des bulles de la figure 1 représente DBpedia qui publie sur le Web les données liées que ce site extrait de l'encyclopédie Wikipédia. Au moment d'écrire cet article, cette seule bulle dans le nuage des bases de données ouvertes représente déjà 38 millions de sujets décrits par 3 milliards de données élémentaires (attributs et relations de ces sujets) issus de 125 langues différentes dans Wikipédia et mises ainsi à disposition comme des données structurées ouvertes. Une autre bulle représente le projet Wikidata qui permet de directement saisir et publier des données structurées actuellement à hauteur de 75 millions de sujets décrits par 23 000 utilisateurs. Dans des domaines spécialisés comme la biologie, une des bulles représente Uniprot fournissant 179 millions de données élémentaires, notamment à propos des protéines. Et ce nuage ne rend pas compte de beaucoup d'autres sources de données. Par exemple, les grands moteurs de recherche derrière l'initiative Schema.org nous permettent de

mettre des données structurées dans nos pages Web et des millions de sites le font pour être indexés plus précisément. Autre exemple, mais même technique, tous les sites que vous voyez arborer un bouton « Like » de Facebook incluent dans leurs pages des données structurées descriptives afin de nourrir ce bouton lorsque vous cliquez dessus. Les volumes de données structurées sur le Web ont donc explosé dans les quinze dernières années, sans que l'utilisateur lambda ne s'en aperçoive nécessairement.

Pour représenter et échanger ces données à l'échelle du Web, nous avons besoin de définir des standards pour leurs modèles, structures, formats et langages. RDF (pour *Resource Description Framework*) est au Web de données ce que HTML est au Web documentaire : le langage RDF permet de représenter et de relier des données à propos de ressources. RDF s'insère parfaitement dans l'architecture du Web, notamment en utilisant les URI pour identifier les ressources et les types de relations décrits par les graphes qu'il permet de représenter.

Par ailleurs, RDF fournit également un modèle de données servant de fondation à d'autres standards. Ainsi, au-dessus de RDF, le standard SPARQL fournit un langage d'interrogation et de modification des graphes RDF et un protocole pour soumettre de telles requêtes à un serveur distant. Par exemple, sur le site DBpedia (données RDF extraites de Wikipédia), qui est l'une des bases dans le nuage de la figure 1, on peut demander en SPARQL tous les URI des ressources nommées « Paris » en français. A partir des identifiants reçus, on peut à nouveau interroger le site pour avoir des données supplémentaires et ainsi passer de données liées en données liées comme on passerait de page en page. Un autre standard nommé SHACL permet quant à lui de valider des données en capturant des règles auxquelles la structure des graphes RDF doit se conformer pour être valide (par exemple tous les livres doivent avoir un titre). Cette validation est utile pour vérifier les données échangées entre applications, par exemple.

De telles descriptions peuvent provenir de n'importe quelle source sur le Web et être fusionnées avec d'autres. Le terme de « gigantesque graphe global » (*Global Giant Graph*) désigne parfois cette toile de données d'envergure mondiale tissée par des milliers de descriptions distribuées sur le Web déclarant des liens entre des nœuds identifiés par des URI.

## Schémas liés et Web sémantique

Dans cette deuxième étape nous publions en plus sur le Web les schémas de ces données, c'est-à-dire les vocabulaires et les règles qui régissent leurs valeurs, leurs structures, leur utilisation, leur interprétation... bref leur sens, leur sémantique. Ces schémas et leurs termes utilisent eux aussi des identifiants du Web (par exemple une URI identifiant la catégorie « femme ») et des liens pour déclarer des relations entre les notions qu'ils définissent (par exemple, une femme est une personne) leur donnant ainsi un sens et tissant un Web sémantique.

Le « Web sémantique » permet donc la formalisation, la publication et le liage des vocabulaires utilisés dans les descriptions RDF. Ces vocabulaires permettent aux applications d'utiliser plus efficacement les données du Web en reconnaissant les différents types de ressources et de liens qu'elles rencontrent, et en exploitant le sens et les raisonnements qui leur sont attachés. Une application peut ainsi faire la différence entre des ressources nommées « Charles de Gaulle » mais de types différents (l'homme, une rue, une résidence, le poète, l'aéroport, le porte-avions…).

Différents types de modèles sont conçus pour fournir des vocabulaires permettant de décrire notre monde sur le Web, on parle notamment d'ontologies informatiques et de thésaurus. En interrogeant et en raisonnant sur ces modèles informatiques, il est possible d'améliorer des fonctionnalités existantes et d'en proposer de nouvelles. Au-dessus de RDF se dresse ainsi la pile des langages de schémas, ayant une expressivité et un coût de calcul croissants : plus l'on monte dans la pile et plus les définitions logiques du vocabulaire permettent de capturer précisément les

structures et le sens des données, mais aussi plus les raisonnements qu'ils permettent sont coûteux en termes de complexité et donc de temps de calcul. Le premier niveau dit « des schémas légers » est celui de RDFS (*RDF Schema*) permettant de déclarer et de nommer les classes de ressources (comme les livres, les films, les personnes…) et leurs propriétés (comme l'auteur, l'acteur, le titre…) et d'organiser ces types dans des hiérarchies. On appelle aussi ces schémas des ontologies légères. Au-dessus de RDFS, la recommandation OWL (*Ontology Web Language*) permet de représenter formellement les définitions d'ontologies plus lourdes et s'organise en plusieurs fragments d'expressivité plus ou moins étendue, qui permettent des déductions supplémentaires en contrepartie de temps de calculs plus longs.

Dans la continuité du Web de données, le « Web sémantique » met donc l'accent sur la possibilité d'échanger les schémas de nos données et la sémantique associée. Formalisés et publiés selon des standards, ces modèles permettent d'enrichir la gamme des traitements automatiques qui peuvent être appliqués aux données. En ouvrant les données et leurs modèles, le Web de données et le Web sémantique ouvrent l'ensemble des utilisations qu'il est possible d'en faire.

## Un projet littéralement infini

Le Web de données et le Web sémantique sont déjà adoptés et déployés dans beaucoup d'applications. Ils n'en sont pas moins toujours sujets à de nombreux travaux de R&D sur des questions, par exemple, de recherche de plus d'efficacité dans les passages à l'échelle, d'intelligence dans les traitements, ou de robustesse face à l'hétérogénéité, la qualité ou l'incertitude dans les données. Mais de toute façon, pour le Web, il s'agit là d'une direction d'évolution parmi de nombreuses autres.

Si dans les années 1990, le problème de Tim Berners-Lee était de faire imaginer un monde disposant du Web avant que celui-ci n'advienne, nous sommes maintenant dans le cas inverse où les gens oublient ou n'imaginent plus ce que serait un monde sans le Web [M]. Cependant, la défense du Web et de son expansion ouverte reste un enjeu. Il est universellement utile et utilisé mais il reste fragile et son idéal de départ pourrait n'être qu'une parenthèse historique si l'on ne veille pas en permanence à sa préservation, notamment en évitant toute forme de recentralisation, par exemple la centralisation de certaines données par certaines firmes. Il faut toujours garder à l'esprit que le Web n'est pas une réalisation acquise mais, par conception, un interminable projet.

## Pour en savoir plus

Fabien GANDON, « Pour tout le monde : Tim Berners-Lee, lauréat du prix Turing 2016 pour avoir inventé... le Web », *Bulletin de la Société informatique de France,* 1024, Société informatique de France, 2017. https://hal.inria.fr/hal-01623368

Fabien GANDON, "A Survey of the First 20 Years of Research on Semantic Web and Linked Data", *Revue des Sciences et Technologies de l'Information - Série ISI: Ingénierie des Systèmes d'information*, Lavoisier, 2018, https://hal.inria.fr/hal-01935898

## Bibliographie

[E] BERNERS-LEE T. (1996), *The World Wide Web: Past, Present and Future*, August. https://www.w3.org/People/Berners-Lee/1996/ppf.html

[F] "Information Management: A Proposal" (March 1989), the original proposal for the software project at CERN that became the World Wide Web. https://www.w3.org/History/1989/proposal.html

[M] SAVAGE N., "Weaving the Web", Communications of the ACM, Vol. 60 No. 6, Pages 20-22, 10.1145/3077334 https://cacm.acm.org/magazines/2017/6/217732-weaving-the-web/fulltext

[S] BERNERS-LEE T., "Plenary Talk extracted slides", First WWW Conference, Geneva 94, https://www.w3.org/Talks/WWW94Tim/

[T] BERNERS-LEE T., CAILLIAU R., LUOTONEN Ari, NIELSEN H. F. and SECRET A. (1994), "The World-Wide Web, Commun", ACM, August, Vol. 37, n°8, 0001-0782, pp. 76-82, 10.1145/179606.179671, ACM, New York, NY, USA.

[AK] BERNERS-LEE T. (1998), "Semantic web road map", https://www.w3.org/DesignIssues/Semantic.html

[BZ] BERNERS-LEE T., HENDLER J. and LASSILA Ora (2001), "The semantic web", *Scientific American*, 284.5 (2001): 28-37.

# Métopes, édition et diffusion multisupports.
## Un exemple de déploiement à l'EHESS

Par **Emmanuel VINCENT** (1)
Responsable éditorial multisupport aux Éditions de l'EHESS

Depuis le début du XXIe siècle, l'édition est entrée dans une « convergence numérique » marquée par l'orientation vers le « tout dématérialisé ». La loi pour une République numérique (2016) (2) prend en compte toutes les possibilités offertes par cette évolution des conditions de production et de diffusion pour favoriser la circulation des savoirs. Dans ce contexte, le principal enjeu pour l'édition scientifique publique est de garantir une stratégie de diffusion et de modèle économique cohérents en déployant des dispositifs adaptés de valorisation (Roux, 2015), qu'il s'agisse de la mise en place de formes matérielles – livres et revues sur support papier – dans le réseau des librairies traditionnelles et dans les librairies présentes sur la toile, ou de la distribution sur Internet de formes dématérialisées, publications en ligne sur des plateformes et/ou mise à disposition de supports détachables, payantes ou en libre accès. C'est à ces problématiques de production multisupport, imprimée et numérique (sous toutes ses formes d'expression digitale), que Métopes entend apporter une solution en investissant les champs technique, organisationnel, économique et juridique de l'édition scientifique (IR Métopes, 2017).

## Le projet Métopes

Née aux Presses universitaires de Caen, puis développée avec le soutien de l'AEDRES (3) et de BSN (4) au sein du pôle Document numérique (PDN) de la MRSH de Caen, Métopes (Méthodes et outils pour l'édition structurée) est une chaîne d'édition multisupport labellisée infrastructure nationale de recherche en 2015 (5), dont l'utilisation est recommandée par l'INSHS (6) pour la production et la diffusion d'ouvrages et de revues. Dans ses grandes lignes, le projet Métopes vise à développer, dans un périmètre initialement limité aux établissements publics d'Enseignement supérieur et de recherche, un environnement de travail normé et standardisé qui permette la création de contenus éditoriaux structurés en flux offrant un très fort potentiel d'interopérabilité (7), autrement dit la possibilité d'une édition ouverte construite sur la mise en réseau de contenus, et à terme de données. Ces flux structurés sont ensuite transformés pour une diffusion sur le Web ou en d'autres formes éditoriales dérivées, livres et revues, qu'elles soient imprimées ou mises en ligne.

(1) Je remercie vivement Victor Deroin, éditeur multisupport aux Éditions de l'EHESS, Étienne Anheim, directeur des Éditions de l'EHESS, et Dominique Roux, directeur de l'infrastructure de recherche Métopes, pour leurs amicales relectures et suggestions.
(2) L'article 30 de la loi pour une République numérique (Légifrance, 2016) s'inscrit dans le mouvement international pour la libération des résultats et données de la recherche. Sur les dispositions de ladite loi et son champ d'application, on se reportera utilement au *Guide d'application de la loi pour une République numérique* – Art. 30 (CoSO, 2018).
(3) Association des éditeurs de la recherche et de l'Enseignement supérieur.
(4) Bibliothèque scientifique numérique.
(5) En 2015, l'infrastructure de recherche Numédif (NUMérique pour l'ÉDItion et la DIFFusion de la production scientifique) combinait dans une logique de complémentarité le projet Métopes et FMSH Diffusion, service et chaîne logistique de diffusion-distribution qui étaient destinés aux presses universitaires et éditeurs institutionnels. Depuis 2018, seule Métopes a été reconduite comme infrastructure de recherche à part entière.
(6) Institut des sciences humaines et sociales du CNRS.
(7) Dans le domaine du numérique, l'interopérabilité est la capacité que possède un système ou un produit à communiquer transversalement avec d'autres, existants ou à venir.

L'inclusion stricte des métadonnées requises pour leur référencement et la génération automatisée de fiches produits garantissent l'intégration dans les circuits de diffusion, que l'ouvrage soit rendu disponible à la vente ou mis à disposition en *Open Access* sur des plateformes ou archives ouvertes (IR Métopes, 2017 et PDN, 2020).

Au sein d'un même service ou d'une structure éditoriale à une autre, le passage à Métopes permet l'organisation de l'activité éditoriale suivant des méthodes communes, dans une volonté d'harmonisation et d'échange de bonnes pratiques. Le partage de mêmes normes et de standards opérationnels en phase avec les réalités du métier simplifie en effet les collaborations internes ou externes, qu'il soit question de la mise en œuvre d'un workflow éditorial ou de la mise en place de coéditions. De plus, si le processus de Métopes est davantage destiné à la production des nouveautés, livres, revues et contenus Internet, un traitement rétrospectif du fonds reste envisageable dès lors que l'éditeur dispose d'archives numériques exploitables. Cette rétronumérisation donne une seconde vie numérique à certains titres et apporte une réponse, notamment grâce à l'impression à la demande, à leur éventuelle indisponibilité sur support papier ainsi qu'à la possible perte des droits afférents, puisque contractuellement l'éditeur est engagé envers l'auteur à ne jamais rompre l'exploitation du titre.

Par ailleurs, la redéfinition claire des rôles joués par chaque partenaire du projet de publication (auteur, éditeur, diffuseur et distributeur) amène à une nette délimitation de ce qu'on peut nommer la « plus-value éditoriale », ce qui dans le contexte actuel s'avère primordial pour définir les modalités de réalisation du libre accès en regard des états du texte, des équilibres économiques, des objectifs de diffusion de la science et des impératifs d'application de la loi pour une République numérique [8]. L'éditeur prépare la copie en liaison étroite avec l'auteur, et à l'issue de cette phase rédactionnelle, apporte une réponse technique par Métopes qui se traduira par une structure enrichie d'où seront dérivées des formes adaptées à chacun des modes et canaux de diffusion (IR Métopes, 2017).

Dans le cadre d'un des volets de leur mission, les porteurs de l'IR Métopes interviennent directement dans les structures éditoriales pour déployer les outils et former les futurs utilisateurs à leurs usages. En France comme à l'étranger où la chaîne est aussi implantée, près de 100 structures d'édition publique universitaire et de recherche, soit plus de 600 personnes au total, ont ainsi été initiées aux méthodes d'édition structurée. La transmission de Métopes à l'étranger est assurée dans le cadre d'accords entre associations nationales d'éditeurs, mais aussi en activant les réseaux des Écoles françaises à l'étranger, des UMIFRE [9] et des collaborateurs étrangers d'OpenEdition… L'appropriation de la technologie par l'éditeur est d'autant facilitée que l'ensemble des briques logicielles et des compétences métier mobilisées sont en adéquation avec son profil de poste : s'il évolue dans ses tâches en produisant des contenus structurés et doit en conséquence revoir son organisation et ses méthodes de travail, pour autant, l'éditeur ne change pas fondamentalement de métier.

## Description fonctionnelle de Métopes

La chaîne Métopes est basée sur le modèle du *single-source publishing,* logique de flux unique à partir duquel sont dérivés papier et numérique, pour lequel Darnton (1999) avait envisagé des applications pratiques dans le champ de l'édition en histoire et qui est décrit dans le *Chicago*

(8) Pour rappel, en application de la loi, il est permis au chercheur de mettre à disposition gratuitement, dans un format ouvert et numérique, la *version finale de son manuscrit acceptée pour publication*, à l'expiration d'un délai de 6 mois maximum pour une publication dans le domaine des sciences, de la technique et de la médecine, et de 12 mois pour les sciences humaines et sociales (Légifrance, 2016).
(9) Unités mixtes des instituts français de recherche à l'étranger.

*Manual of Style* (2010) [10] avec ses incidences organisationnelles et « métier ». Du point de vue fonctionnel, Métopes se résume à un *process* de fabrication multisupport qui, dans un seul effort de production, à partir d'un fichier XML pivot, non-propriétaire et pérenne pour l'archivage, génère en bout de chaîne des formes éditorialisées grâce à des automates de composition : PDF certifié conforme pour l'impression concernant le *print,* et pour le numérique, PDF interactif, ePub ou flux XHMTL afin d'alimenter des sites ou les plateformes SHS (Cairn.info, OpenEdition Journals, OpenEdition Books). Un environnement de travail unique assure la transformation en XML et la validation du flux, permet de l'éditer (saisie et correction), de l'annoter (à l'aide de métadonnées, de référentiels et d'index), et enfin de produire les formes de diffusion. Les flux structurés au format XML-TEI [11] embarquent les métadonnées susceptibles d'être produites par transformation au format ONIX [12], qui selon la sortie envisagée sont disponibles sous forme d'une fiche produit, ou sélectionnées suivant l'exigence de chaque forme de diffusion et directement enchâssées dans le fichier ou le flux numérique. Le schéma suivant décompose les étapes qui, à partir d'un document produit dans un traitement de texte, amènent aux différents types de formes et de valorisations éditoriales :

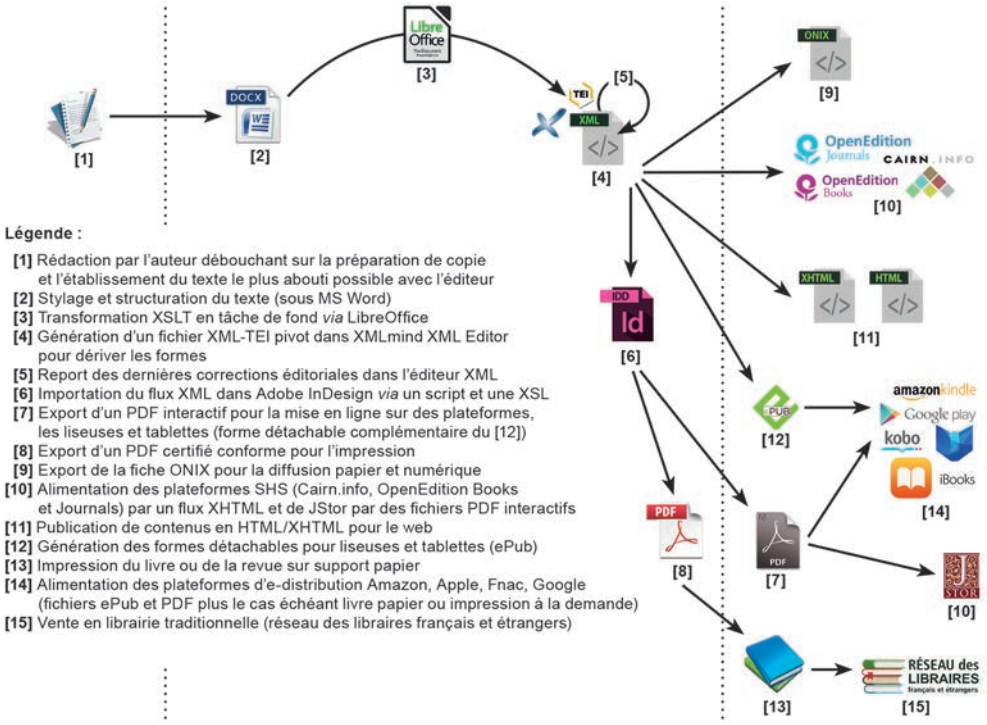



Schéma fonctionnel de Métopes
© Cellule Métopes EHESS.

(10) Voir notamment la section "Appendix A: Production and Digital Technology" et la figure A.5.
(11) Le XML (*Extensible Markup Language*) est un langage informatique de balisage utilisé pour structurer des contenus. Il repose sur un schéma de données, à savoir la TEI pour Métopes. Communauté internationale s'inscrivant dans le champ des humanités numériques, la TEI (*Text Encoding Initiative*) a défini des recommandations pour l'encodage de documents textuels, adaptant son modèle théorique aux technologies, d'abord sous la forme d'une DTD SGML, puis XML. Pour aller plus loin, voir Burnard (2015).
(12) ONIX est un format de métadonnées exprimé en XML et développé par EDItEUR, pour une utilisation dans le commerce du livre. À travers un jeu de balises, cette norme permet la description complète d'un ouvrage : données bibliographiques (titre, auteur, collaborateurs, lieu et date de publication, éditeur, ISBN, ISSN, etc.), données matérielles (type de support, format, nombre de pages, poids, etc.), données de commercialisation et de diffusion (prix, marché, choix de diffusion, etc.).

La transformation d'un format de fichier en un autre, de MS Word en XML et du XML en HMTL/XHTML ou en fichier ePub [13] est assurée par des XSLT [14] mises en forme par des CSS [15] pour les besoins de l'habillage graphique. L'import du flux XML dans InDesign, logiciel de mise en page à partir duquel sont produits les fichiers au format PDF, est quant à lui réalisé au moyen d'un script et d'un jeu de XSL [16]. Si Métopes repose sur des normes et standards, comme XML, TEI ou ONIX, l'utilisateur n'a en pratique qu'à appliquer une feuille de style sous MS Word pour s'acquitter des opérations de balisage visant à structurer les contenus. Il ne lui reste ensuite qu'à les enrichir de métadonnées dans l'éditeur XML, beaucoup étant déduites du flux structuré, et à finaliser les formes de diffusion. Il faut toutefois concéder que la tendance s'oriente vers une consolidation, donc une saisie accrue, des métadonnées (balisage des données bibliographiques pour leur repérage, par exemple), mais aussi vers un enrichissement des contenus et une sémantisation qui passe par leur annotation, l'objectif étant de les inscrire dans des systèmes d'information de plus en plus organisés, exigeants et interopérables. Métopes offre effectivement comme autre fonctionnalité la possibilité d'ajouter un jeu de données pour valider les autorités, autrement dit pour lever les homonymies d'auteurs, idRef [17], à partir duquel on peut aussi récupérer d'autres identifiants parmi Orcid [18], HAL, BNF, ISNI… Les contenus s'en trouvent donc enrichis sémantiquement par un ensemble de référentiels, des développements récents proposant l'interconnexion avec des bases de données comme GeoNames (pour les lieux géographiques), ou l'indexation à travers des vocabulaires contrôlés (thésaurus, taxonomies, etc.) [19]. Au final, le système d'information que véhiculent ces contenus scientifiques structurés, normalisés et enrichis crée un point de convergence avec d'autres usages et utilisateurs investis dans le numérique : chercheurs – ceux en SHS utilisent fréquemment la TEI dans leurs travaux –, documentalistes, archivistes, bibliothécaires, etc.

(13) Fondé sur le XML, ePub (acronyme de « *electronic publication* ») est un format ouvert standardisé pour les livres numériques. Ce type d'e-book a été conçu pour ajuster la mise en pages d'un contenu en redistribuant le texte en fonction du dispositif d'affichage et des choix ergonomiques du lecteur. Il est en effet composé de fichiers HTML et CSS pour l'ePub 2, de fichiers HTML 5 et CSS 3, de métadonnées et de JavaScript permettant une interactivité accrue avec des contenus enrichis (graphisme, typographie et multimédia) pour l'ePub 3.

(14) XSLT (*eXtensible Stylesheet Language Transformations*) est un langage de transformation XML de type fonctionnel qui permet de convertir un document XML dans un autre format.

(15) Les CSS (*Cascading Style Sheets*) forment un langage informatique décrivant la présentation des documents HTML et XML en vue de leur conférer un habillage graphique.

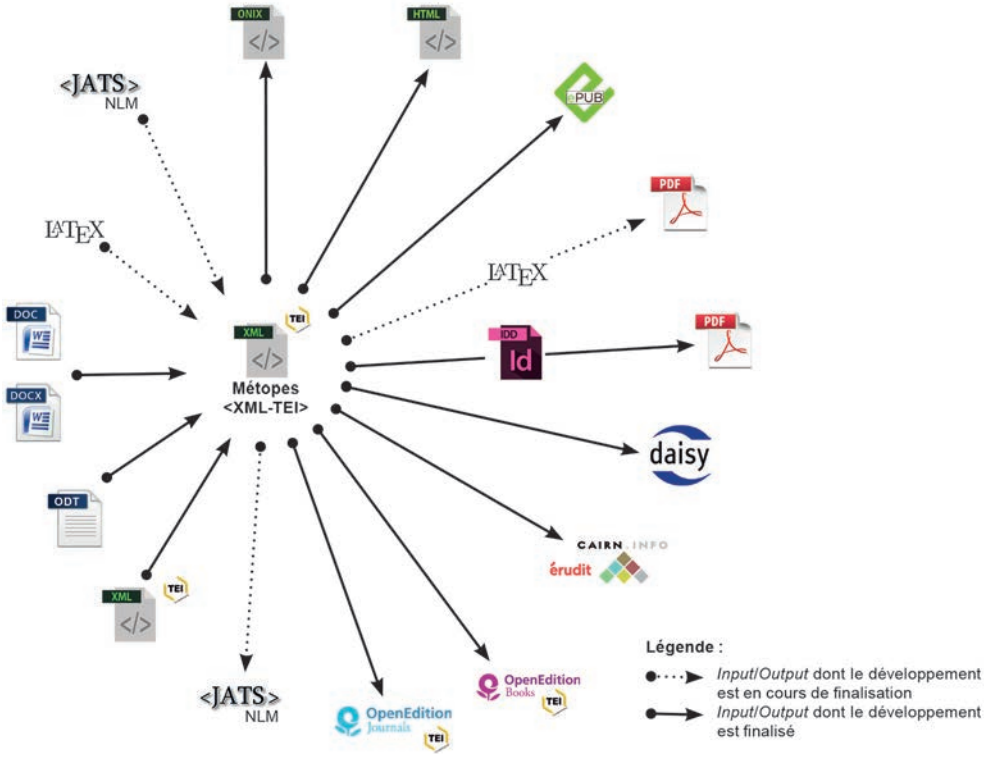
(16) XSL (eXtensible Stylesheet Language) est le langage de description de feuilles de style associé à XML. Une feuille de style XSL est un fichier indiquant comment doivent être transformés les documents XML basés sur une même DTD ou un même schéma.

(17) IdRef (Identifiants et référentiels) est une application Web développée et maintenue par l'ABES (Agence bibliographique de l'Enseignement supérieur). Tout auteur d'un texte catalogué dans le Sudoc a un idRef qui peut être retrouvé en interrogeant la plateforme www.idref.fr.

(18) ORCID (*Open Researcher and Contributor ID*), est un identifiant se présentant sous la forme d'un code alphanumérique non propriétaire, qui permet d'identifier de manière unique les chercheurs et auteurs de contributions académiques et scientifiques.

(19) Une partie de ces développements et l'adaptation de l'environnement de Métopes ont été réalisées dans le cadre du projet « Savoirs » piloté par Christian Jacob (EHESS). Il s'agit d'une plateforme en construction qui proposera une bibliothèque intelligente suggérant des parcours de lecture heuristiques, d'où le recours au Web sémantique.

Le schéma ci-après liste les *inputs* et *outputs* de Métopes, en précisant lesquels sont déjà pleinement opérationnels :




Formats d'entrée et de sortie pour Métopes [20]
© Cellule Métopes EHESS, adaptation du schéma publié en ligne par les porteurs du projet (PDN, 2020).

Lors du traitement de flux par Métopes, l'unité descriptive de base se situe au niveau de l'article ou du chapitre, dont l'assemblage complété de métadonnées spécifiques produit le numéro de revue ou la monographie, si bien que le jeu de métadonnées s'adapte à l'échelle de son implémentation, décrivant soit les unités éditoriales, soit le volume qui les collecte. En conséquence, le bassin de fichiers XML engrangés par l'éditeur, à la mesure de l'unité éditoriale, peut servir à puiser parmi des textes couvrant les mêmes thématiques et issus de volumes divers, afin de constituer de nouveaux ouvrages à moindre investissement car déjà relus et structurés, quitte à les agrémenter d'un appareillage inédit (préface ou postface) qui induise de nouveaux parcours de lecture. On touche ici à l'un des avantages certains du numérique, qui peut infléchir les politiques éditoriales en renouvelant les stratégies de valorisation. Ce type d'exploitation sous forme de compilation peut aussi apparaître comme mimétique du phénomène, fréquemment observé et étroitement lié à dématérialisation, de dislocation des unités physiques que sont les livres ou revues, au profit d'une recherche par mots-clés privilégiant chapitres, articles ou encore séries de textes, notamment

(20) OpenDocument (.odt) est un format ouvert sur lequel repose des logiciels de traitement de texte comme OpenOffice et LibreOffice. JATS (*Journal Article Tag Suite*) est un type de XML servant à la description et la publication d'articles scientifiques, notamment utilisé par Cambridge University Press ou le portail de p*eer-reviewing* Open Journal Systems. LaTex est un langage et système de composition de documents dont l'une des fonctionnalités est son mode mathématique qui facilite la mise en forme des formules complexes. Il est surtout employé dans les domaines scientifiques et techniques pour la production de mise en pages. Daisy (*Digital Accessible Information System*) est une norme et un format pour livres audio accessibles, conçus pour les personnes déficientes et empêchées de lire des documents imprimés.

lorsque le lecteur constitue ses propres recueils *ad hoc,* indépendamment d'un projet éditorial de départ, sinon de la cohérence du volume pensé dans sa globalité (Anheim, 2015).

En définitive, par son apport fonctionnel, Métopes rationalise les coûts d'édition multisupport : elle aide l'éditeur à s'extraire des logiques de reprise de contenus, très chronophages et nécessitant des moyens humains, auxquelles il est assujetti dès lors que l'ensemble des sorties sont produites indépendamment et en dehors d'un flux unique. Aussi, moyennant un coût humain relatif, l'éditeur peut-il dégager de réelles économies en internalisant totalement les processus de fabrication (hors frais d'impression), là où le prépresse était parfois confié à des prestataires et où s'ajoutaient des frais de rétroconversion des fichiers pour alimenter les plateformes. Sans compter que des bénéfices supplémentaires sont réalisables en exploitant de nouvelles formes de diffusion. Par son indépendance dans la mise en œuvre de sa production, et les formats ouverts sur lesquels Métopes s'appuie, l'éditeur devient pleinement propriétaire (dans le respect des droits d'auteur) de ses fichiers et maquettes. Vecteur de formes plurielles et dépositaire de la version définitive du texte, le XML issu de Métopes n'est pas propriétaire (tout comme l'ePub dérivé), c'est un format ouvert, stable et pérenne, susceptible d'être archivé en interne ou par le CiNEs [21], car produit dans un écosystème utilisant des standards et des normes documentées (XML et TEI), ainsi qu'Unicode pour l'encodage des caractères.

## L'exemple de Métopes à l'EHESS

Dès 2014, la chaîne Métopes a été mise en place pour traiter les monographies dans le service des publications des Éditions de l'EHESS, qui a fonctionné comme un terrain expérimental pour éprouver la solution, puis tel un incubateur à partir duquel le déploiement a ensuite été étendu par périmètres successifs. Depuis 2016, le projet a en effet pris une autre ampleur, avec la création d'une cellule Métopes dédiée à l'EHESS [22], l'intégration des revues des Éditions de l'EHESS au projet, et la couverture progressive de l'ensemble des collections publiées sous cette enseigne. À ce jour, trois-quarts des périodiques du portefeuille des Éditions ont basculé sous Métopes. Sauf exception, le prépresse a été internalisé, permettant aux rédactions de reprendre la main sur les fichiers de chaque numéro passant par la chaîne, et sur les maquettes de leurs séries. D'importantes économies ont ainsi été réalisées en se passant de la sous-traitance pour la mise en pages du *print,* la facture de fabrication se limitant dorénavant aux seuls coûts d'impression. Par ailleurs, certaines revues qui n'avaient plus d'existence sur OpenEdition Journals depuis plusieurs années sont à nouveau présentes en ligne, les lacunes ayant été comblées. Grâce à Métopes et aux accords passés avec les plateformes SHS, les Éditions de l'EHESS ne supportent plus que des coûts résiduels et aucun frais lié à la rétroconversion pour être mis en ligne sur Cairn.info, OpenEdition Journals et Books. Enfin, toujours sur le plan financier mais aussi pratique, la cellule Métopes a pris en charge l'informatique éditoriale de l'EHESS en collaboration avec la DSI, centralisant les investissements et les rationalisant, ce qui a permis des achats en volume de matériels et de licences logicielles [23]

(21) Établissement public situé à Montpellier, le CiNEs (Centre informatique national de l'Enseignement supérieur) est un service informatique utilisé pour la recherche publique en France dont l'une des missions consiste à archiver de façon pérenne des documents électroniques. Son archivage des fichiers produits par Métopes est envisageable sous réserve de validation par celui-ci du schéma de données et de la définition des modalités du service.

(22) La cellule Métopes de l'EHESS est composée de deux membres, un responsable éditorial multisupport et un éditeur multisupport. La feuille de route est définie en étroite collaboration entre les Éditions et les instances de l'EHESS. Depuis 2019, Métopes a été retenue comme projet d'établissement. Une convention est en cours de rédaction entre l'université de Caen, l'IR Métopes et l'EHESS pour définir les périmètres et champs d'intervention de cette cellule.

(23) Achat de licences pour des applications métier, à savoir Antidote et ProLexis pour l'assistance à la rédaction et la préparation de copie, suites Office (MS Word) et Adobe CC (InDesign, Photoshop, Illustrator et Acrobat Pro) pour le traitement de texte et la PAO, l'ensemble de ces briques logicielles étant en dehors de l'outillage proposé par Métopes, qui fournit l'éditeur XMLmind XML Editor avec un environnement de travail spécifique, les routines XSLT, le script pour l'import du flux XML dans InDesign.

afin d'équiper les structures pour accueillir au mieux la solution. La majeure partie des formes de diffusion des Éditions de l'EHESS sont mises au point *via* Métopes : outre les publications *online* sur les plateformes des revues ou de certaines collections et le support papier, plusieurs fichiers ePub et PDF interactifs ont été élaborés et diffusés sur Eden Livres. Pour l'habillage graphique des fichiers ePub, une CSS générique et adaptable d'une revue ou collection à une autre a été créée, qui respecte l'identité des séries et de la maison d'édition. L'élargissement graduel du déploiement à d'autres entités éditoriales de l'EHESS, revues ou collections hébergées dans les centres mais ne faisant pas partie du giron propre aux Éditions de l'EHESS, a débuté en 2019, de même que l'intégration de Métopes à des projets de recherche comme Savoirs [(24)], étendant la couverture fonctionnelle de la chaîne éditoriale aux besoins des chercheurs, et requérant de ses concepteurs des développements spécifiques. Fin 2016, en accord avec l'IR Métopes, la cellule de l'EHESS s'est positionnée comme référente sur la chaîne d'édition multisupport au sein de la ComUE PSL [(25)], partageant son expérience avec les structures éditoriales du Collège de France, de l'École nationale des chartes, de l'École pratique des hautes études et de l'École nationale supérieure des mines de Paris (Mines ParisTech). Depuis 2019, de nouveaux liens se sont créés sur les mêmes bases avec des services de publication du Campus Condorcet [(26)], comme Ined éditions.

Dans le périmètre qui lui incombe, la mission de la cellule Métopes est de faciliter au quotidien le déploiement. Servant de relais technique pour l'installation des outils ou leurs mises à jour, elle dispense soit un complément adapté (sur InDesign, Métopes ou la chaîne graphique) aux formations initiales délivrées par l'IR Métopes, soit une formation intégrale pour les nouveaux utilisateurs [(27)]. Elle documente la solution à l'aide d'un guide pas à pas actualisé, tout en assurant une veille sur les outils, les formats et les vocabulaires inhérents à l'édition électronique. Elle propose une assistance fonctionnelle (relevé de bogues pour transmission aux développeurs de l'IR Métopes et soutien opérationnel de proximité) à la fois rassurante et très bien perçue, qui, d'une aide temporaire et en présentiel pour les bouclages des numéros, y compris pour la mise en ligne, va jusqu'à l'autonomisation des structures. Enfin, sur le plan humain, les agents sont accompagnés dans le changement métier et la réorganisation de leur travail, se sentant moins isolés dans leur profession que par le passé car appartenant à une communauté d'utilisateurs. Un réseau s'est constitué autour de Métopes, de plus en plus dense, avec des échanges construits aussi bien autour de la solution que sur des projets transverses, plus particulièrement ciblés sur des aspects métier : modèle économique, diffusion et valorisation des savoirs, échange de compétences, processus d'évaluation des textes, etc.

En conclusion, si Métopes peut apparaître relativement rigide dans sa mise en œuvre pour un public non initié, la solution est très souple, robuste et productive pour la valorisation des contenus scientifiques. Elle rend réalisables toutes les possibilités d'une politique éditoriale : diffusion multicanal et multiforme, imprimée et numérique, commercialisation et *Open Access*, qui peuvent s'instancier différemment d'une structure éditoriale à une autre au sein d'une même institution comme l'EHESS.

(24) Voir note 19.
(25) Communauté d'établissements Paris sciences et lettres (université PSL), dont l'EHESS a été l'un des membres associés dès 2014.
(26) L'EHESS est l'un des onze établissements d'Enseignement supérieur ou organismes spécialisés dans les sciences humaines et sociales qui sont membres du Campus Condorcet.
(27) La cellule Métopes de l'EHESS a ainsi formé ou renouvelé les compétences sur la chaîne d'environ 90 collègues (EHESS, PSL, Condorcet). Chaque apprentissage s'effectue sur site, au plus près des préoccupations éditoriales des structures concernées, jusqu'à la pleine acquisition.

## Bibliographie

ANHEIM É. (2015), « Le numérique et l'économie éditoriale des revues scientifiques », *Revue d'histoire moderne & contemporaine*, 2015/5 (nº 62-4 bis), p. 22-32, en ligne sur www.cairn.info/revue-d-histoire-moderne-et-contemporaine-2015-5-page-22.htm.

BUARD P.-Y. (2015), « Modélisation des sources anciennes et édition numérique », thèse de doctorat, université de Caen Basse-Normandie, dactyl., 250 p., téléchargeable en version PDF sur https://hal.archives-ouvertes.fr/tel-01279385/document.

BURNARD L. (2015), *What is the Text Encoding Initiative?,* trad. de Marjorie Burghart, Marseille, OpenEdition Press, coll. « Encyclopédie numérique », nº 5, téléchargeable en version PDF sur https://books.openedition.org/oep/1237.

*The Chicago Manual Of Style* (2010), "Appendix A: Production and Digital Technology", 16th Edition.

CoSO (2018), *Guide d'application de la loi pour une République numérique – Art. 30,* téléchargeable en version PDF sur www.ouvrirlascience.fr/guide-application-loi-republique-numerique-article-30.

DARNTON R. (1999), « Le nouvel âge du livre », *Le Débat,* 105, p. 176-184.

IR MÉTOPES (2017), « Métopes : méthodes, outils, formations », contenu en ligne sur www.metopes.fr/index.html.

LÉGIFRANCE (2016), « Loi pour une République numérique », nº **2016-1321**, art. **30,** 7 octobre, version consolidée du 13 avril 2020, texte intégral de la loi en ligne sur www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000033202746.

PDN [Pôle Document numérique] (2020), « Métopes », contenu en ligne sur www.unicaen.fr/recherche/mrsh/document_numerique/projets/metopes.

ROUX D. (2015), « Les tournants numériques de l'édition scientifique. De la transition à la convergence : une perspective subjective, technique et... éditoriale », *in HENNY J.-M. (éd.), avec la collaboration de PIERROT D. & ROUX D., L'édition scientifique institutionnelle en France. État des lieux, matière à réflexions, recommandations,* rapport (dactyl.) établi pour l'AEDRES et remis au MENSR, p. 121-142, téléchargeable en version PDF sur www.enssib.fr/bibliotheque-numerique/documents/65757-l-edition-scientifique-institutionnelle-en-france.pdf.

# Software Heritage : l'archive universelle des codes sources du logiciel

Par **Roberto DI COSMO**
Software Heritage

Le logiciel est devenu en quelques décennies le moteur de notre industrie, le carburant de l'innovation, l'instrument essentiel que nous utilisons pour communiquer, nous entretenir, réaliser tout type de transaction et opération, nous organiser en société et former nos opinions politiques. Il contrôle le système embarqué dans nos moyens de transport ou de communication, les échanges commerciaux et financiers. Il est au cœur des équipements et des dispositifs médicaux ; il assure le bon fonctionnement des réseaux de transport et de communication, des banques et des établissements financiers. Le logiciel est crucial dans le fonctionnement des organisations économiques, sociales et politiques, qu'elles soient publiques ou privées, que ce soit sur des terminaux mobiles ou sur le *cloud.* Il est aussi le médiateur indispensable qui permet d'accéder à toute l'information numérique, et il constitue, avec les articles et les données, un des piliers de la recherche moderne (Noorden *et al.,* 2014).

Le logiciel est donc en train d'incorporer *une partie importante* de notre patrimoine scientifique, technique et industriel, et il porte avec lui des enjeux stratégiques majeurs.

Si on y regarde de près, il est aisé de se rendre compte que la vraie connaissance qui est contenue dans les logiciels se trouve non pas dans les programmes exécutables, mais dans le « code source » qui, selon la définition utilisée dans la licence GPL, est « la forme préférée pour un développeur pour apporter une modification à un programme [(1)] ». Le code source est une forme de connaissance spéciale : il est fait pour être *compris par un être humain*, le développeur, et peut être mécaniquement traduit dans une forme pour être *exécuté* directement sur une machine. La terminologie même utilisée par la communauté informatique est parlante : on utilise des « langages de programmation » pour « écrire » des logiciels. Comme Harold Habelson l'écrivait déjà en 1985, « les programmes doivent être écrits d'abord pour que d'autres êtres humains puissent les lire » (Abelson et Sussman, 1985).

Le code source des logiciels est donc bien une *création humaine* au même titre que d'autres documents écrits, et les développeurs des logiciels méritent le même respect que d'autres créateurs.

Enfin, le code source des logiciels est de plus en plus complexe, et est modifié régulièrement par des groupes de développeurs qui collaborent pour le faire évoluer : il est devenu essentiel, pour le comprendre, d'avoir accès aussi à son historique de développement.

Le code source des logiciels constitue donc un patrimoine de grande valeur, comme déjà soutenu par Len Shustek dans un bel article de 2006 (Shustek, 2006) et par Donald Knuth (Knuth, 1984), et il est essentiel de s'atteler à sa préservation.

C'est bien là une des missions que s'est donné Software Heritage, une initiative lancée en 2015 avec le soutien de l'Inria [(2)], pour *récolter, organiser, préserver et rendre facilement accessible* l'ensemble

(1) GNU91, Gnu general public license, version 2, 1991, Retrieved September 2015.
(2) Créé en 1967, l'Inria est un établissement public à caractère scientifique et technologique spécialisé en mathématiques et informatique, placé sous la double tutelle du ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation et du ministère de l'Économie et des Finances.

du *code source* disponible publiquement sur la planète, indépendamment d'où et comment il a été développé ou distribué. Le but est de construire une infrastructure commune qui permettra une multiplicité d'applications : bien sûr, préserver sur le long terme le code source contre les risques de destruction, mais aussi permettre des études à grande échelle sur le code et les processus de développement actuels, afin de les améliorer, et préparer ainsi un futur meilleur.

## Une tâche complexe

Archiver tous les codes sources disponibles est une tâche complexe : comme détaillé dans l'article d'Abramatic (Abramatic *et al.,* 2018), on doit déployer des stratégies différentes selon qu'on cherche à collecter du code source ouvert ou propriétaire, et on ne traite pas de la même façon le code source facilement disponible en ligne que celui qui se trouve sur des supports physiques anciens. Pour le code source des logiciels anciens, on doit mettre en place un véritable processus d'archéologie informatique, et nous avons déjà commencé ce travail dans le cadre d'une collaboration avec l'Université de Pise et l'UNESCO, qui a abouti au processus SWHAP utilisé pour retrouver, documenter et archiver des logiciels marquants de l'histoire de l'informatique en Italie (voir https://www.softwareheritage.org/swhap). Pour le code source ouvert et facilement disponible en ligne, l'approche la plus appropriée est de construire un moissonneur qui collecte automatiquement les contenus d'une grande variété de plateformes de développement collaboratif, comme GitHub, GitLab.com ou BitBucket, ou de plateformes de diffusion de paquetages logiciel, comme Debian, NPM. CRAN ou Pypi.

Si, au premier abord, cette approche peut paraître similaire à celle mise en place pour l'archivage du Web, en y regardant de plus près, on s'aperçoit rapidement que la tâche est ici bien plus difficile. Tout d'abord, comme il n'y a pas de protocole standard, il faut construire un adaptateur spécifique pour chacune des plateformes de développement et de diffusion des logiciels, afin d'extraire la liste des projets logiciels qu'elles hébergent (on appelle justement *listers* ces adaptateurs).

Ensuite, on se trouve confronté à pléthore de formats et de modèles de données différents, utilisés pour garder trace de l'historique de développement, ce qui pose un gros problème si on veut s'assurer que cet historique sera lisible dans le futur, même quand les outils utilisés pour le construire – comme git, darcs, subversion ou mercurial – seront devenus obsolètes. Pour cela, nous construisons une deuxième famille d'adaptateurs, que nous appelons *loaders,* pour convertir dans une structure de données commune, simple et maintenable, les informations contenues dans les différents systèmes de contrôle de version et formats de paquetages.

Cette structure de données est une généralisation des arbres de Merkle, inventés il y a plus de quarante ans, et dont les principes sont aujourd'hui largement utilisés dans des applications aussi variées que les systèmes de contrôle distribués, les *blockchains* ou les systèmes de fichiers distribués (Merkle, 1987). Elle présente de nombreux avantages : chaque artefact archivé se voit attribuer un identifiant intrinsèque, qu'on peut vérifier indépendamment (Di Cosmo *et al.,* 2020), les contenus sont dédupliqués, ce qui permet de réduire considérablement la taille de l'archive, et le graphe résultant permet de suivre la façon dont les mêmes codes sources sont réutilisés entre différents projets. L'architecture résultante est décrite dans la Figure 1 page suivante.

## Une mission universelle

Au-delà de la complexité technique, la mission universelle de Software Heritage pose des défis stratégiques considérables : comment s'assurer de sa viabilité à long terme ? Comment s'assurer qu'elle reste bien au service de tous, et ne soit pas privatisée ou mise sous contrôle par des intérêts particuliers ? Comment retrouver tous les codes sources écrits dans les dernières décennies ? Comment maximiser les chances que le précieux patrimoine ainsi collecté soit préservé sur le long terme ?

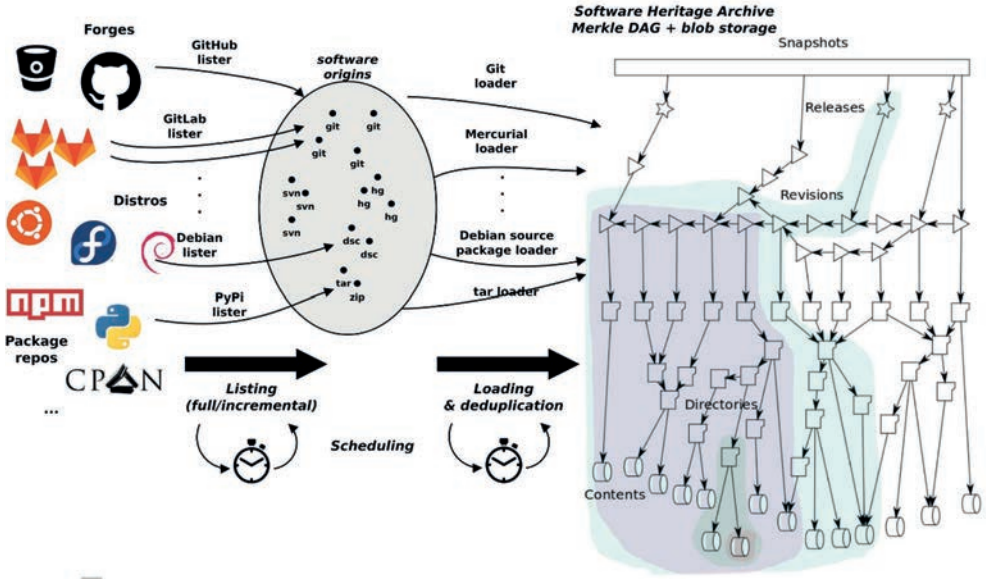



Figure 1 : Architecture du moissonneur de Software Heritage.

Ces questions ont été au cœur des réflexions qui ont mené à établir quelques principes fondateurs de Software Heritage (Abramatic *et al.*, 2018 ; Di Cosmo & Zacchiroli, 2017) : usage systématique de logiciels libres pour construire l'infrastructure de Software Heritage, afin de permettre de comprendre son fonctionnement, et de la répliquer si nécessaire ; construction d'un réseau mondial de miroirs indépendants de l'archive, parce qu'un grand nombre de copies est la meilleure protection contre les pertes et les attaques ; choix d'une structure sans but lucratif, internationale et multipartenaires, pour minimiser les risques d'avoir des points uniques de défaillance et pour s'assurer que Software Heritage sera bien au service de tous.

L'engagement dans cette initiative de personnalités qui ont une longue trajectoire au service du bien commun, comme Jean-François Abramatic et Stefano Zacchiroli, est un élément très important. Dans une telle mission, il faut bien sûr aussi une légitimité institutionnelle, et une véritable capacité d'ouverture pour fédérer un consensus large. L'accord-cadre signé entre Inria et UNESCO, le 3 avril 2017, est dans ce sens à la fois une reconnaissance de l'importance de la mission, et une grande opportunité d'établir des collaborations au niveau mondial pour l'accomplir.

Une importante étape dans cette direction a été la réunion du groupe d'experts internationaux organisée à l'UNESCO au mois de novembre 2018, qui a abouti à l'appel de Paris pour le code source des logiciels, disponible en ligne sur https://en.unesco.org/foss/paris-call-software-source-code. On y trouve une analyse détaillée des raisons pour lesquelles le code source des logiciels est devenu un enjeu majeur, et des recommandations pour des actions concrètes à mener afin de répondre aux défis qui sont posés. Parmi ces recommandations, figure le soutien à l'effort commencé avec Software Heritage pour construire une infrastructure internationale de préservation des codes sources des logiciels.

## Passé, présent, futur : bien plus qu'une archive !

Software Heritage est aujourd'hui une infrastructure qui grandit jour après jour, et si le plus gros du contenu de l'archive résulte du moissonnage automatique, des perles commencent à y être introduites par un patient travail de récupération de logiciels historiques marquants, suivant un processus d'acquisition qui a été mis au point en collaboration avec l'Université de Pise et l'UNESCO [3].

(3) Voir SWHAP sur https://www.softwareheritage.org/swhap


Figure 2 : Nombre de projets, fichiers sources et versions archivées dans Sofware Heritage au mois de janvier 2020 (voir https://www.softwareheritate.org/archive).

L'exhaustivité est encore loin d'être atteinte, mais l'archive contient déjà le plus grand corpus de codes sources disponible sur la planète, avec plus de 90 millions d'origines archivées, pour plus de 6 milliards de fichiers sources uniques, chacun équipé d'un identifiant intrinsèque basé sur des *hash* cryptographiques (Di Cosmo *et al.*, 2018).

Cette infrastructure unique a de multiples missions : bien sûr, il s'agit de préserver pour les futures générations les codes sources du *passé* qui ont fait l'histoire de l'informatique et de la société de l'information, mais aussi, et surtout, on cherche ici à construire le *très grand télescope* qui permette d'explorer l'évolution *présente* de la *galaxie du développement logiciel*, afin de mieux la comprendre, et de l'améliorer, pour construire un *futur* technologique meilleur.

## Un enjeu stratégique

L'archive de Software Heritage constitue déjà la collection de codes sources la plus importante de la planète, mais le chemin à parcourir est encore long : il est nécessaire de continuer à réunir les compétences scientifiques et techniques, ainsi que les ressources financières et humaines, afin de pouvoir construire la mémoire d'une partie importante de la technologie et de la science qui est au cœur de la transition numérique, à un moment où l'on peut encore espérer avoir accès à tout ce qui a été mis en œuvre dès le début de l'histoire, encore courte, de l'informatique.

Mais il y a bien plus que cela : à un moment où l'on voit clairement que le logiciel est devenu un composant essentiel de toute activité humaine, l'accès sans restriction aux codes sources des logiciels publiquement disponibles, ainsi qu'à l'information qualifiée sur leur évolution, devient un enjeu de souveraineté numérique pour toutes les nations.

L'infrastructure unique que construit Software Heritage, et son approche universelle sont un élément essentiel pour répondre à cet enjeu de souveraineté numérique, tout en préservant la dimension de bien commun propre à l'archive.

Il est donc de la plus haute importance que des acteurs institutionnels, industriels, académiques et de la société civile saisissent l'importance de ces enjeux, et que la France et l'Europe se positionnent rapidement, en fournissant les ressources nécessaires à faire grandir et pérenniser Software Heritage, en prenant leur place à côté des autres acteurs internationaux qui se sont déjà engagés, et en soutenant la création d'une institution internationale sans but lucratif qui porte cette mission sur le long terme.

## Pour en savoir plus

On peut en savoir plus sur le projet en visitant www.softwareheritage.org, annex.softwareheritage.org et wiki.softwareheritage.org
Il est possible d'explorer aisément les codes sources contenus dans Software Heritage sur archive.softwareheritage.org

### Licence



## Bibliographie

ABELSON H. & SUSSMAN G. J. S. with J. (1985), *Structure and Interpretation of Computer Programs*, The MIT Press.

ABRAMATIC J.-F., DI COSMO R. & ZACCHIROLI S. (2018), "Building the Universal Archive of Source Code", *Commun. ACM*, *61*(10), 29-31. https://doi.org/10.1145/3183558

DI COSMO R., GRUENPETER M. & ZACCHIROLI S. (septembre 2018), "Identifiers for Digital Objects: The Case of Software Source Code Preservation", *Proceedings of the 15th International Conference on Digital Preservation*, iPRES 2018, Boston, USA. https://doi.org/10.17605/OSF.IO/KDE56

DI COSMO R., GRUENPETER M. & ZACCHIROLI S. (2020), "Referencing Source Code Artifacts: A Separate Concern in Software Citation", *Computing in Science & Engineering*. https://doi.org/10.1109/MCSE.2019.2963148

DI COSMO R. & ZACCHIROLI S. (septembre 2017), "Software Heritage: Why and How to Preserve Software Source Code", *Proceedings of the 14th International Conference on Digital Preservation*, iPRES 2017. https://hal.archives-ouvertes.fr/hal-01590958/

KNUTH D. E. (1984), "Literate Programming", *Comput. J.*, 27(2), 97-111. https://doi.org/10.1093/comjnl/27.2.97

MERKLE R. C. (1987), "A Digital Signature Based on a Conventional Encryption Function", *Advances in Cryptology - CRYPTO '87, A Conference on the Theory and Applications of Cryptographic Techniques,* Santa Barbara, California, USA, August 16-20, 1987, *Proceedings*, 369-378. https://doi.org/10.1007/3-540-48184-2_32

NOORDEN R. V., MAHER B. & NUZZO R. (2014), "The top 100 papers", *Nature*, 514(7524), 550-553. https://doi.org/10.1038/514550a

SHUSTEK L. J. (2006), "What Should We Collect to Preserve the History of Software?", *IEEE Annals of the History of Computing*, 28(4), 110-112. https://doi.org/10.1109/MAHC.2006.78

# L'archivage du Web ou le Web comme mémoire des sociétés contemporaines

Par **Alexandre CHAUTEMPS**
Chef du service Dépôt légal numérique,
Bibliothèque nationale de France

## Bref historique de l'archivage du Web

Le Web a été inventé en 1989 par Tim Berners-Lee puis a été rendu librement utilisable en 1993 par son employeur de l'époque, le CERN. A peine plus de dix ans plus tard, le Web est devenu un outil de communication très largement employé. En 2004, 50 % des Français disposent d'un accès à Internet (1). Les individus, la société et les institutions ont déjà produit un contenu nativement numérique, et sans équivalent imprimé, d'un volume non négligeable. Pourtant, l'idée de collecter le Web pour créer des archives pérennes paraît encore étrange.

Seuls quelques visionnaires s'y consacrent. Dès 1997, Brewster Kahle, créateur de la fondation Internet Archive, avait attiré l'attention dans un article paru dans une grande revue de vulgarisation scientifique (Kahle, 1997) qui posait déjà les grands principes de ce qui allait devenir l'archivage du Web. Les premières initiatives d'archivage du Web avaient vu le jour en 1996, portées par Internet Archive et par quelques institutions : Bibliothèque royale de Suède (projet KulturarW3) et Bibliothèque nationale d'Australie (projet PANDORA). En France, la BnF avait, dès 1999, mené des collectes expérimentales du Web (cf. Masanès, 2002), la première collecte d'envergure étant réalisée à l'occasion de la campagne électorale de 2002. A la suite du travail d'information réalisé par la Bibliothèque nationale de France et le ministère de la Culture, le parlement a décidé d'inscrire dans la loi, en août 2006, une mission de dépôt légal (cf. Stirling *et al.,* 2011) portant sur « les signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique » (Code du Patrimoine, art. L131-2). Cette mission est confiée conjointement à trois institutions : l'Institut national de l'audiovisuel (INA) pour les programmes de télévision et de radio diffusés *via* l'internet, le Centre national de la cinématographie (CNC) pour les films diffusés en salle et la BnF pour tout le reste du Web français.

Le positionnement des activités d'archivage du Web dans le cadre du dépôt légal est un élément particulièrement important en ceci qu'il donne une assise juridique à ces activités, mais aussi par le processus de patrimonialisation qu'il implique (cf. Bermès, 2020, p. 68 *et sq.*). Les collections nativement numériques ainsi constituées s'inscrivent en effet dans la continuité des collections de documents imprimés entrés par dépôt légal, dont les plus anciens remontent au XVI$^{e}$ siècle.

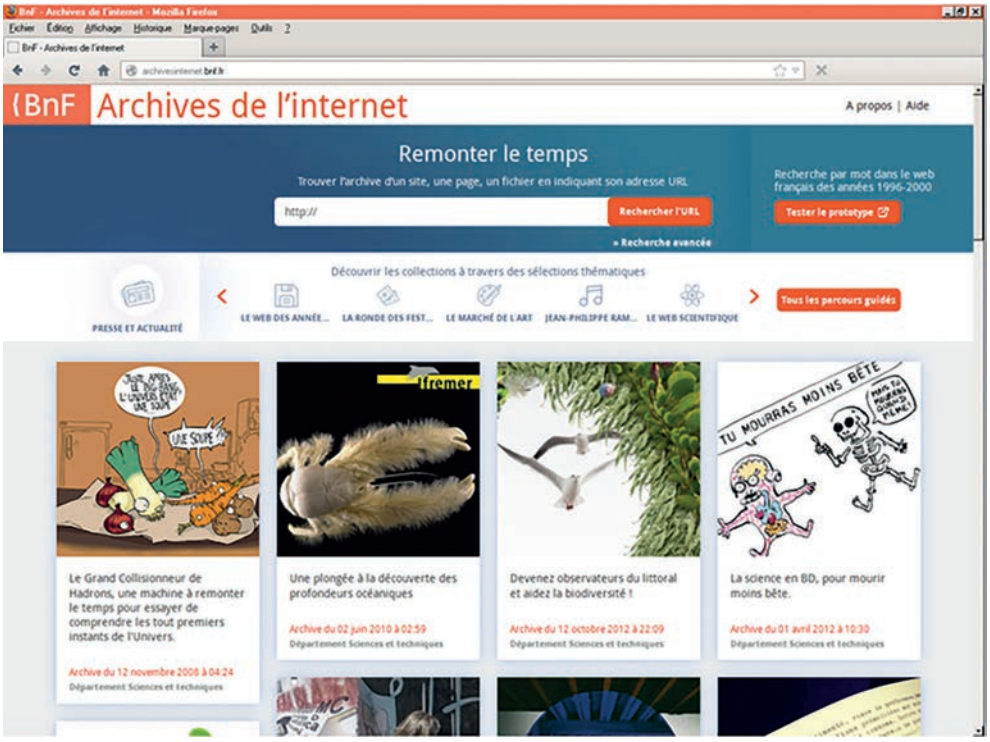
Chaque année, la BnF effectue une collecte large qui moissonne l'ensemble des domaines identifiés comme faisant partie du Web français. Il s'agit de tous les domaines en .fr ainsi que des autres domaines dont le producteur est localisé en France, quelle que soit leur extension (.com, .org, .net mais aussi les domaines de haut niveau liés aux territoires d'outre-mer et les « nouvelles »

(1) Cf. BIGOT R. (2004), *La Diffusion des technologies de l'information dans la société française*, Paris, Credoc, p. 83 et sq., https://bit.ly/2HAKt5X (consulté le 11 mars 2020).

extensions telles que .paris, .bzh ou .immo, .tools, etc. [2]). Cette collecte large est enrichie par des collectes ciblées, plus fréquentes et/ou plus profondes, donc plus complètes, qui portent sur des sites sélectionnés par les bibliothécaires de la BnF et d'établissements partenaires en régions, en fonction de critères documentaires. Certaines collectes ciblées sont réalisées à l'occasion d'événements : campagnes électorales, événements sportifs, mais aussi attentats, épidémies, célébrations, etc., de manière à garder une trace numérique desdits événements.

Depuis 2010, la BnF réalise par ses propres moyens l'ensemble de ses collectes (pour la description de l'architecture technique voir Le Follic *et al.,* 2012) et garantit la préservation des collections patrimoniales numériques ainsi constituées (cf. Derrot *et al.,* 2012).

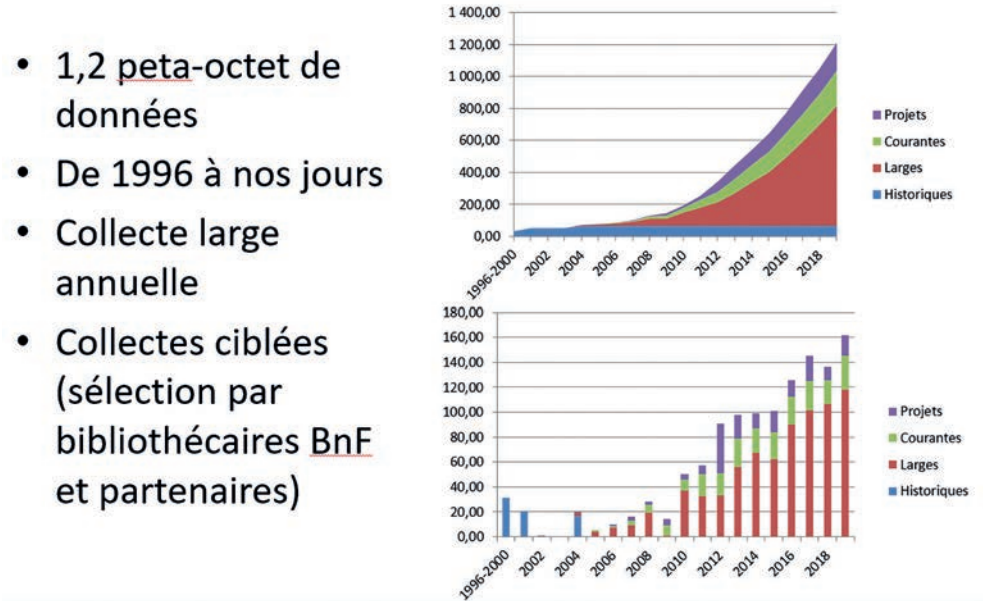
A la fin de l'année 2019, la BnF avait constitué une collection d'archives Web d'un volume d'environ 1 200 téraoctets, regroupant près de 35 milliards de fichiers, dont les éléments les plus anciens remontent à 1996 et dont l'accroissement se poursuit quotidiennement.




Page d'accueil de l'application de consultation des archives
© Bibliothèque nationale de France

L'accès aux Archives de l'Internet de la BnF est possible dans les locaux de l'établissement, en bibliothèque de recherche [3] ainsi que dans vingt bibliothèques partenaires en régions [4].

(2) Pour clarifier les notions de nom de domaine et d'extension, liée à un domaine de haut niveau, voir Bortzmeyer, 2018, chapitre « Noms de domaine et DNS », pp. 89 *et sq.*
(3) Voir description des Archives de l'Internet : https://www.bnf.fr/fr/archives-de-linternet et les conditions d'accès à la bibliothèque de recherche : https://www.bnf.fr/fr/bibliotheque-de-recherche (liens consultés le 11 mars 2020).
(4) Voir la localisation des différents points d'accès, en Métropole et en outre-mer : https://bit.ly/2PjOpfF (consulté le 11 mars 2020).




Archives de l'Internet BnF : volumétrie des collections début 2020
© Bibliothèque nationale de France

La collecte du Web par robot permet aussi de moissonner de nombreux livres et périodiques numériques, principalement aux formats EPUB et PDF. Néanmoins, les documents numériques distribués *via* des boutiques en ligne ne sont pas collectés par ce moyen. Des conventions sont conclues avec les éditeurs et distributeurs afin de procéder par la voie du dépôt, techniquement plus appropriée. Des dépôts de documents numériques en ligne sont ainsi réalisés à titre expérimental [5].

De nombreuses autres institutions de par le monde remplissent une mission d'archivage du Web analogue à celle de la BnF. Il s'agit de bibliothèques nationales mais aussi de centres d'archives, voire de fondations indépendantes. La plupart de ces institutions sont regroupées au sein de l'Internet International Preservation Consortium (IIPC) [6], créé en 2003, dont la BnF est membre fondateur (cf. Illien, 2011). IIPC regroupe à l'heure actuelle 57 institutions appartenant à 35 pays, parmi lesquels l'Europe, l'Amérique du Nord et dans une moindre mesure l'Asie sont les régions du monde les plus représentées. L'archivage du Web reste un chantier à ouvrir en Afrique, en Amérique latine et même en Inde et il s'agira d'un enjeu majeur des décennies à venir.

Si les institutions patrimoniales se sont donné pour mission de préserver le Web pour l'avenir, la collecte du Web est également utilisée pour servir les besoins immédiats de la recherche. Elle est alors pratiquée par un éventail plus large d'institutions, dont la plupart ne font pas partie du consortium IIPC : bibliothèques universitaires et laboratoires de recherche principalement. Dans la partie suivante de cet article, nous allons nous intéresser plus particulièrement aux utilisations possibles des archives Web dans le cadre de projets de recherche.

(5) Dans ce cadre, des conventions de dépôt ont été conclues par la Bibliothèque nationale de France avec l'éditeur de livres numériques publie.net et le distributeur de musique en ligne Idol.
(6) Voir http://netpreserve.org (consulté le 11 mars 2020).

## Usage(s) des archives du Web

Depuis le milieu de la dernière décennie, les sources nativement numériques, et singulièrement les archives Web, deviennent des sources légitimes pour la recherche dans de nombreuses disciplines des sciences humaines et sociales, et sont même déjà devenues des sources nécessaires et incontournables pour les travaux se focalisant sur la période récente (après l'an 2000).

Le savoir-faire du chercheur se modifie dans le contexte de l'essor des humanités numériques [7]. L'historien, le sociologue, l'anthropologue ou le linguiste, sans renoncer aux compétences propres à leur discipline, deviennent aussi des praticiens, et parfois des spécialistes des données numérisées et nativement numériques, de leur identification, de leur sélection, de leur manipulation, tout en étant conduits à s'intéresser aux circonstances de leur production et de leur collecte (voir Milligan, 2019).

Dans ce cadre, les archives du Web constituent un terrain d'exploration des groupes sociaux et des comportements.

Un premier exemple est constitué par le projet mené par Sophie Gebeil (Université Aix-Marseille) sur la mémoire de l'immigration maghrébine en France de 1999 à 2014 (cf. Gebeil, 2015). L'objet du travail de Sophie Gebeil n'était pas centré sur les événements eux-mêmes mais sur le processus mémoriel dont ils faisaient l'objet. Son approche s'inscrivait dans le cadre de l'histoire mémorielle, dans la continuité des travaux de Pierre Nora (voir Nora, 2008). Il s'agit de retracer la manière dont les Français issus de l'immigration nord-africaine produisent différents récits mémoriels en lien avec leur propre histoire individuelle ou familiale. Les archives Web de la BnF ont constitué le matériau principal de ces travaux. Le Web permet de mettre en rapport, sur un même plan, différents discours : celui des institutions (notamment le Musée de l'histoire de l'immigration), celui des médias traditionnels, celui des acteurs associatifs et/ou militants, celui des personnes s'exprimant à titre individuel (témoignages). L'essor du Web a en effet facilité l'expression des deux dernières catégories, et surtout, la diffusion de leur point de vue. Par ailleurs, le travail de Sophie Gebeil met en évidence une démarche de patrimonialisation de la mémoire des populations immigrées : Chibanis, mémoire des quartiers, mémoire des bidonvilles constituent autant d'objets mémoriels souvent laissés de côté ou traités de manière stéréotypée par les médias traditionnels. Sophie Gebeil a choisi de travailler sur le Web archivé et non sur le Web vivant car son besoin méthodologique était de disposer d'un corpus stable : les comptages d'occurrences, l'analyse des liens entre sites, la constitution de réseaux internes à la communauté étudiée, rendaient nécessaire le fait de disposer d'un matériau textuel figé dans le temps et protégé de la volatilité intrinsèque au Web vivant [8].

Autre exemple : le projet *La Grande Guerre sur le Web*, dirigé par Valérie Beaudouin (Télécom Paris-Tech) et soutenu par le Labex *Les passés dans le présent* (cf. Beaudouin *et al.,* 2018 et Bermès, 2020, p. 80 [9]). Ce projet s'est déroulé de 2013 à 2018. L'objectif initial consistait à étudier la manière dont les ressources patrimoniales numérisées mises en ligne par les bibliothèques étaient diffusées et réutilisées sur le Web, notamment dans le cadre de la commémoration. La première étape a été de sélectionner les sites en rapport avec le sujet et de les décrire par un jeu de métadonnées propres au projet. Sur cette base, les sites sélectionnés ont été collectés par la BnF à un rythme

(7) Lire à ce sujet le *Manifeste des Digital humanities*, https://journals.openedition.org/jda/3652 (consulté le 11 mars 2020).
(8) On trouvera davantage d'informations méthodologiques dans GEBEIL (2017). Voir également le parcours guidé *Les mémoires de l'immigration maghrébine* (1999-2014) sur le site de la BnF : https://www.bnf.fr/fr/centre-d-aide/parcours-guides-dans-les-archives-de-linternet#step03
(9) Voir le parcours guidé « Mémoire et histoire de la Grande Guerre sur le web » : https://www.bnf.fr/fr/centre-d-aide/parcours-guides-dans-les-archives-de-linternet#step01 ainsi que le billet de Philippe Chevallier, « La Grande Guerre sur le web : un éclairage inédit », publié en 2017 sur le carnet de recherche Hypothèses de la BnF : https://bnf.hypotheses.org/1588

semestriel, constituant ainsi le corpus de travail (volumétrie totale : 6,7 téraoctets). L'équipe projet s'est attachée à cartographier, à l'aide du logiciel Gephi, les relations que les sites Web consacrés à l'histoire de la Grande Guerre tissaient entre eux. Les sites dédiés à la commémoration s'inscrivaient dans un ensemble plus vaste, où les sites institutionnels côtoient les sites d'amateurs, d'associations, qu'ils soient centrés sur la généalogie, l'histoire des régiments, la mémoire familiale. L'approche cartographique et statistique mise en œuvre par l'équipe projet permet de mettre en évidence une polarisation marquée en deux ensembles, sites institutionnels et sites d'amateurs, correspondant à deux réseaux humains bien distincts, mais montre aussi l'existence d'une zone frontière où les deux réseaux communiquent ponctuellement entre eux.

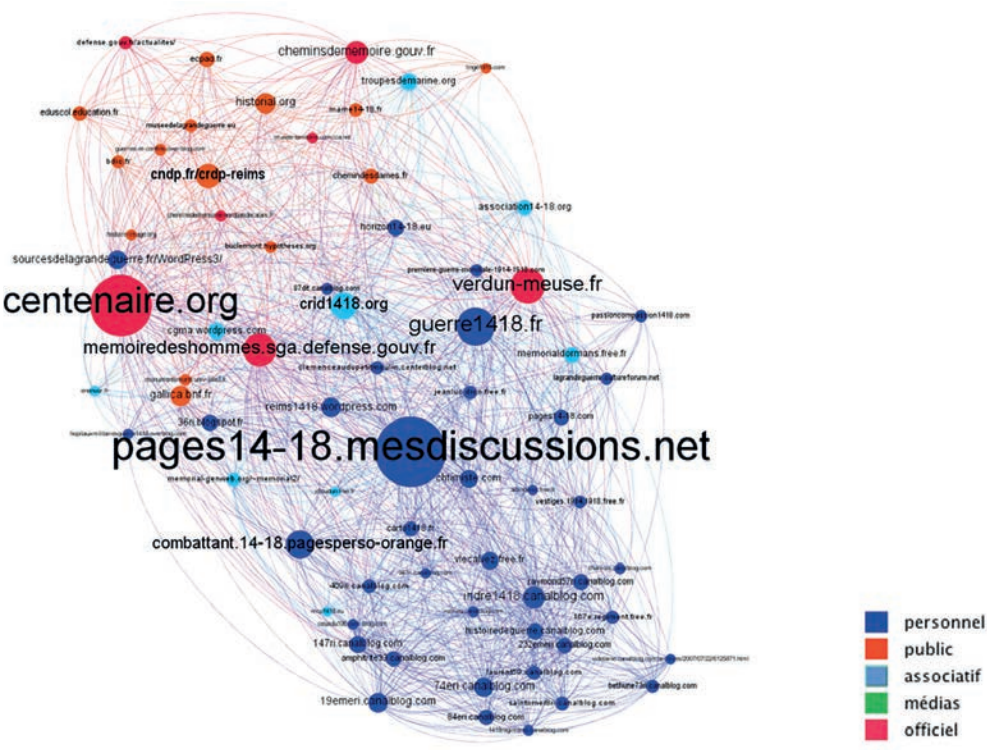


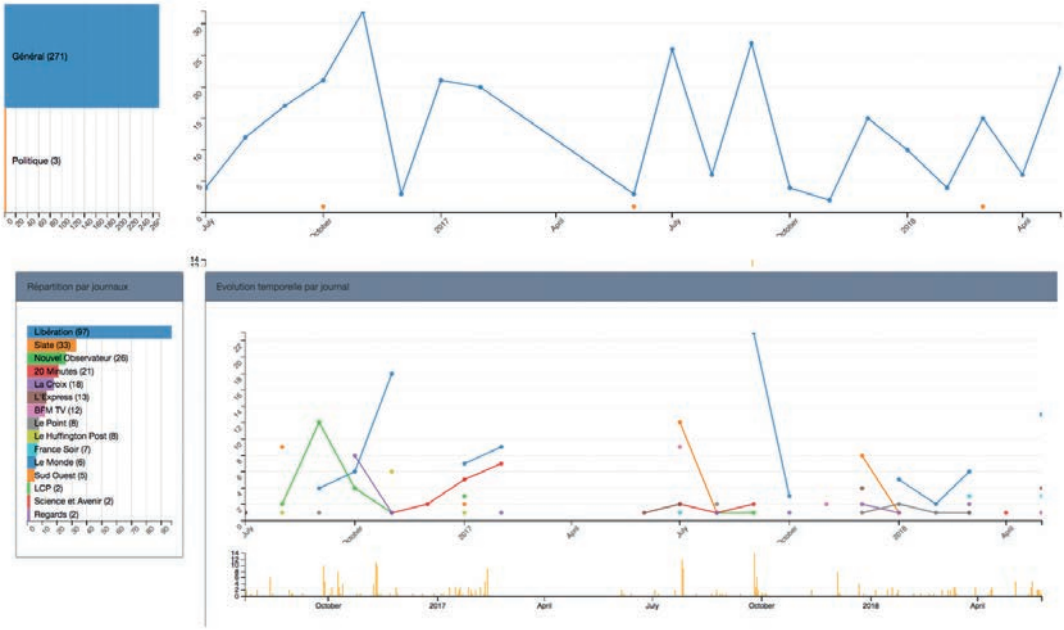

Figure produite par Valérie Beaudouin, Télécom Paris, à partir des données de la « Collecte Grande Guerre » de la BnF (archives du Web).
© Télécom Paris Tech

Le forum 14-18 [10], qui était l'un des lieux privilégiés de ces échanges, a d'ailleurs fait l'objet d'une vaste étude par fouille de données dans le cadre du deuxième volet du projet.

Le Web est également un terrain privilégié pour étudier l'évolution de la langue. Le projet Néonaute (cf. Aubry, 2018), mené de 2016 à 2018 sous la direction d'Emmanuel Cartier (Université Paris 13), avait pour objet l'étude du cycle de vie de trois ensembles de mots : 30 000 néologismes, un ensemble de mots français dont l'emploi est recommandé de préférence à des anglicismes, et un ensemble de termes féminisés. Le corpus choisi était celui des collectes Actualités réalisées par la BnF entre 2010 et 2017 (volumétrie totale : 13 téraoctets). Il est constitué des sites de presse français, dans leur partie accessible gratuitement, collectés à une fréquence quotidienne. Le

(10) Voir https://forum.pages14-18.com/ (ressource consultée le 29 mars 2020).

corpus a été indexé en texte intégral avec le moteur d'indexation Apache Solr puis a fait l'objet de différents traitements destinés à « nettoyer » les pages Web en en retirant tout ce qui n'est pas le texte des articles (boutons et liens de navigation, en-têtes, pieds de page, messages publicitaires, etc.). Ensuite, l'équipe projet a procédé à différents traitements d'analyse morpho-syntaxique, de lemmatisation et d'extraction de fragments de texte comportant chaque occurrence des termes analysés avec un environnement textuel suffisant pour permettre l'analyse de leur rôle dans la phrase. L'évolution de l'emploi des termes est étudiée en tenant compte de leur environnement d'apparition, en distinguant notamment presse d'information générale, presse magazine, presse technique, presse professionnelle, etc.




Projet Néonaute : visualisation multidimensionnelle interactive des occurrences.
© Université Paris 13

Le Web est aussi un gisement très riche d'articles scientifiques, dont une part importante est diffusée en accès ouvert [11]. De nombreux documents PDF sont collectés dans les différentes archives Web, mais il n'est pas possible de savoir de manière simple lesquels correspondent à des publications scientifiques. Un projet récent, mené par Internet Archive, a consisté à utiliser des technologies d'intelligence artificielle pour identifier des articles scientifiques au sein de la masse des documents PDF archivés. La première étape a été de définir, avec l'aide de bibliothécaires humains, un corpus d'entraînement constitué de sources importantes de publications scientifiques en accès ouvert sur le Web vivant. L'algorithme de *machine learning* a ensuite été entraîné à reconnaître un article scientifique. Après cela, il a été lancé sur les archives afin d'identifier les articles scientifiques parmi les PDF archivés, d'évaluer la qualité de la version archivée, de procéder à un archivage dans le cas où le document était inexistant ou de mauvaise qualité dans les archives, et enfin d'effectuer le signalement des versions archivées dans un catalogue en ligne de publications scientifiques : Fatcat [12]. Ce projet montre de manière convaincante les possibilités offertes par l'intelligence

(11) Pour une définition des différentes formes d'accès ouvert, voir le site web de l'Ecole des Ponts Paris-tech : https://espacechercheurs.enpc.fr/fr/open-access (lien consulté le 14 mars 2020).
(12) Voir https://fatcat.wiki (lien consulté le 14 mars 2020).

artificielle pour explorer les vastes gisements d'informations constitués par les archives Web, tâche qui, du fait de l'importance de la volumétrie, demeure irréalisable par des humains [13].

Enfin, les archives du Web constituent une source incontournable pour une discipline émergente : l'histoire du Web. Celle-ci, en effet, est d'abord racontée par le Web lui-même. C'est par les archives du Web que l'on peut reconstituer la succession des technologies (langages de programmation, protocoles, formats…) qu'il a utilisées. C'est également par les archives du Web que l'on peut appréhender les événements fondateurs de son évolution. Les différentes périodes et les moments-clés de l'histoire du Web y ont laissé de nombreuses traces : passage du Web 1.0 au Web 2.0, essor des blogs, des réseaux sociaux, des plateformes de mise en relation, mais aussi mouvements récurrents comme le Web activisme ou le *net art* (voir Brügger, 2017 et Schafer, 2018a et b [14]).

## En conclusion : quelques perspectives

Nous l'avons vu, les archives du Web offrent un nouveau terrain d'investigation aux chercheurs et mobilisent tout un arsenal de techniques spécifiques, dont l'enjeu est de faire face à une masse de données d'une ampleur sans précédent.

Comment connaître ces techniques et apprendre à les utiliser ? De nombreuses ressources existent, bien entendu, sur le Web [15]. Ces connaissances et leur mise en pratique font également l'objet d'une transmission par des humains, dans le cadre d'ateliers dédiés à la prise en main et à l'exploitation des outils de manipulation de données, de fouille de texte et de *data visualization* : les datathons. On suivra avec intérêt les activités du groupe d'historiens canadiens Archives Unleashed, particulièrement investi dans cette transmission, tout à la fois par la création d'outils logiciels [16] et l'organisation régulière de datathons.

Les bibliothèques développent également de nombreux services dédiés aux chercheurs travaillant sur des données numérisées et nativement numériques, parmi lesquelles les archives du Web tiennent une place importante. La Bibliothèque nationale de France travaille actuellement à la mise en œuvre d'une telle offre de service, BnF Data Lab, qui devrait voir le jour à l'automne 2020 (cf. Eloi *et al.,* 2019).

Si la mise en œuvre d'outils techniques est une étape indispensable, c'est cependant avant tout sur l'accompagnement humain que reposera la large diffusion des humanités numériques et l'utilisation de ces vastes collections dans de nombreux travaux de recherche.

## Bibliographie

AUBRY S., CARTIER E. & STIRLING P. (2018), "Néonaute: mining Web archives for linguistic analysis", communication à l'IIPC Web archiving conference, Wellington (Nouvelle-Zélande), http://netpreserve.org/ga2018/wp-content/uploads/2018/11/IIPC_WAC2018-Sara_Aubry_Emmanuel_Cartier_Peter_Stirling-Neonaute-mining_Web_archives_for_linguistic_analysis.pdf

BEAUDOUIN V. & PEHLIVAN Z. (2017), « Cartographie de la Grande Guerre sur le Web » : Rapport final de la phase 2 du projet *Le devenir en ligne du patrimoine numérisé : l'exemple de*

(13) Pour davantage d'informations sur ce projet, voir Praetzellis (2019).
(14) Voir également le site Web90 – Patrimoine, Mémoires et Histoire du Web dans les années 1990 : https://web90.hypotheses.org (consulté le 11 mars 2020).
(15) Un exemple parmi beaucoup d'autres : le site The Programming historian, https://programminghistorian.org (consulté le 11 mars 2020).
(16) Voir la page du site d'Archives Unleashed consacrée au toolkit : https://archivesunleashed.org/aut/ (consultée le 11 mars 2020), ainsi que Milligan (2016).

*la Grande Guerre*, Paris, Bibliothèque nationale de France; Bibliothèque de documentation internationale contemporaine ; Télécom ParisTech, 2017, https://hal.archives-ouvertes.fr/hal-01425600

BEAUDOUIN V., CHEVALLIER Ph. & MAUREL L. (dir.) (2018), *Le Web français de la Grande Guerre : réseaux amateurs et institutionnels*, Nanterre, Presses universitaires de Paris Nanterre, https://catalogue.bnf.fr/ark:/12148/cb45699767f

BERMES E. (2020), « Le numérique en bibliothèque : naissance d'un patrimoine : l'exemple de la Bibliothèque nationale de France (1997-2019) », thèse de doctorat, Paris, Ecole nationale des Chartes, 107 p., https://tel.archives-ouvertes.fr/tel-02475991

BORTZMEYER S. (2018), *Cyberstructure : l'internet, un espace politique,* Caen, C&F éditions, « Société numérique », https://catalogue.bnf.fr/ark:/12148/cb45637569v

BRÜGGER N. (dir.) (2017), *Web 25: histories from the first 25 years of the World Wide Web,* New York, Peter Lang publishing, "Digital formations", https://catalogue.bnf.fr/ark:/12148/cb453621987

BRÜGGER N. & LAURSEN D. (dir.) (2019), *The historical Web and digital humanities: the case of national Web domains*, London, Routledge, "Digital research in the arts and humanities", https://catalogue.bnf.fr/ark:/12148/cb457626551

DERROT S., FAUDUET L., OURY C. & PEYRARD S. (2012), "Preservation Is Knowledge: A community-driven preservation approach", communication à la 9e International Conference on Preservation of Digital Objects (iPRES), Canada, https://hal.archives-ouvertes.fr/hal-00925315

ELOI C., MOIRAGHI E. & ROSE V. (2019), « Un espace pour les humanités numériques à la BnF », *Bulletin des bibliothèques de France (BBF)*, n°17, p. 90-95. http://bbf.enssib.fr/consulter/bbf-2019-17-0090-009

GEBEIL S. (2015), « La fabrique numérique des mémoires de l'immigration maghrébine sur le Web français (1999-2014) », thèse de doctorat, Université Aix-Marseille, 2 vol. (1 109 p.), http://www.theses.fr/fr/2015AIXM3119#

GEBEIL S. (2017), « Quand l'historien rencontre les Archives du Web », billet publié sur le blog Web Corpora, 10 novembre 2017, https://Webcorpora.hypotheses.org/380)

ILLIEN G. (2011), « Une histoire politique de l'archivage du Web », *Bulletin des bibliothèques de France,* n°2, pp. 60-68, http://bbf.enssib.fr/consulter/bbf-2011-02-0060-012

KAHLE B. (1997), "Preserving the Internet", *Scientific American,* Vol. 276, N°3, March, pp. 82-83, https://www.jstor.org/stable/24993660?seq=1 (accès payant)

LE FOLLIC A., STIRLING P. & WENDLAND B. (2012), "Putting it all together: creating a unified Web harvesting workflow at the Bibliothèque nationale de France", communication au workshop IIPC How to fit in? Integrating a Web archiving program in your organisation, Paris, Bibliothèque nationale de France, novembre, https://hal-bnf.archives-ouvertes.fr/hal-00873759

MASANES J. *et al.* (2002), "A first experience in archiving the French Web", Actes de la 6e European conference on Research and advanced technology for digital libraries (ECDL), Rome (Italie), https://bit.ly/39JLpRv

MILLIGAN I. (2016), "Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives", UWSpace, http://hdl.handle.net/10012/10322

MILLIGAN I. (2019), "Historians' archival research looks quite different in the digital age", The Conversation, https://theconversation.com/historians-archival-research-looks-quite-different-in-the-digital-age-121096

NORA P. (2008), « Entre mémoire et histoire » : [article rédigé en 1984], *in* Les Lieux de mémoire, vol. I, Paris, Gallimard, « Quarto », pp. 23-43, https://catalogue.bnf.fr/ark:/12148/cb36695433c

PRAETZELLIS M. (2019), "From Open Access to Perpetual Access: Archiving Web-Published Scholarship", communication à l'IIPC Web archiving conference, Zagreb, http://netpreserve.org/ga2019/wp-content/uploads/2019/07/IIPCWAC2019-JEFFERSON_BAILEY__MARIA_PRAETZELLIS-From_open_access_to_perpetual_access-archiving_Web-published_scholarship.pdf

SCHAFER V. (2018a), *En construction : la fabrique française d'internet et du Web dans les années 1990,* Bry-sur-Marne, INA, « Études et controverses », https://catalogue.bnf.fr/ark:/12148/cb45632257t

SCHAFER V. (dir.) (2018b), *Temps et temporalités du Web,* Nanterre, Presses universitaires de Nanterre, « Intelligences numériques », https://catalogue.bnf.fr/ark:/12148/cb45466310p

STIRLING P., ILLIEN G., SANZ P. & SEPETJAN S (2011), « La situation du dépôt légal de l'Internet en France : retour sur cette nouvelle législation, sur sa mise en pratique depuis cinq ans, et perspectives pour le futur », communication au 77e congrès de l'International federation of library associations (IFLA), Porto Rico, https://www.ifla.org/past-wlic/2011/193-stirling-fr.pdf

# Le Programme interministériel d'archivage VITAM

Par **Jean-Séverin LAIR**
Chargé du programme Tech.Gouv
Direction interministérielle du numérique (DINUM)

## Le lancement d'un programme sur l'archivage numérique public

### L'archivage, un besoin essentiel... même dans le numérique

Pour la grande majorité des lecteurs, et encore plus sans doute pour ceux férus de numérique, à la seule évocation du mot *archivage,* arrivent instantanément à l'esprit des images de parchemins, de rayonnages poussiéreux et de gestionnaires grisonnants. Un peu de modernité, que diable !, avec le numérique on n'a plus besoin de cela, la mémoire est à l'échelle de la planète, diront certains. Et pourtant…

Je ne jette la pierre à personne car je n'étais pas loin d'entériner cette vision caricaturale des archives (et des archivistes !) avant d'avoir été « éduqué » [(1)]. Revenons aux fondements des archives publiques, retranscrites par le béotien que je suis [(2)] :

> La raison d'être des archives publiques est de garder la trace des travaux et décisions administratives ayant un impact sur les droits et devoirs des personnes physiques et morales, pour les faire valoir durant leur durée d'effet, pouvoir les invoquer avec force probante en justice, assurer leur transparence et la capacité à rendre des comptes des décideurs et finalement, pour certaines archives, en garder la mémoire historique.

Je tiens à souligner que :

- n'en déplaise aux plus intéressés par l'histoire et contrairement à la croyance populaire, la première utilité de l'archivage n'est pas l'étude scientifique, mais bien l'activité administrative, au profit des organisation et des personnes, et ce, depuis l'aube même des archives (la conservation des titres nobiliaires et de propriétés au niveau royal...). L'archivage est un élément-clé de la continuité de l'État et de la validité des décisions prises en son nom.
- nulle part dans cette définition je n'ai évoqué de papier et il est évident que de plus en plus de travaux et décisions sont maintenant sous forme numérique. Alors qu'il y a encore dix ans on dématérialisait mais avec des étapes qui restaient encore sous forme papier, on atteint enfin maintenant le tout numérique dans les procédures administratives. Ceci veut dire qu'en l'absence d'archivage numérique, nous perdons des traces essentielles ou du moins qu'elles perdent de leur force. *Quid* du cadastre avec ses tracés et ses servitudes, des décisions de justice, des naturalisations... ? Sans archivage numérique nous risquons un « Alzheimer d'État » – sans parler de la perte historique que l'on pourra constater dans quelques décennies.

(1) « Sorti de l'état sauvage » pour les étymologistes avertis.
(2) Je prie les archivistes qui pourront lire ces lignes d'excuser cette vision simpliste, mais il s'agit de vulgariser...

L'archivage est important, mettons ! Mais dans le numérique on sait conserver les traces sur nos disques durs, et même plus, puisque dans le *cloud* on a des capacités énormes. Alors pourquoi ce besoin d'archivage ?

L'archivage consiste à garder de manière sûre et suffisamment bien ordonnée une information pour pouvoir la retrouver plusieurs dizaines d'années plus tard, le tout garanti par des processus éprouvés et des personnes mandatées pour garantir toute la valeur probante et historique nécessaire. Empiler les informations sur des disques, comme on a l'a fait depuis vingt ans, ou les entasser dans des gisements de fichiers avec un moteur de recherche ne répond que très partiellement au besoin. Le numérique standard est parfait pour le temps court et la performance, mais les archives ont besoin de temps long et de garantie. Faire de l'archivage numérique nécessite de réconcilier la nanoseconde avec le siècle, le flux et la preuve. Il s'agit de projeter une pratique professionnelle ancestrale sur d'autres médias avec d'autres outils, en maintenant l'exigence du temps long et de la preuve qui manque tant à l'informatique.

### Le constat au niveau de l'Etat français en 2011 : nécessité d'agir

Le ministère de la Culture, qui est garant des aspects interministériels de l'archivage, a été le seul dès 1982 à se doter d'un procédé d'archivage du numérique, dit « Constance ». Reposant essentiellement sur des procédures manuelles et des moyens jugés maintenant artisanaux (des bandes, des bases documentaires…), il a permis d'archiver des contenus importants, comme certaines grandes enquêtes INSEE (recensement de 1962, par exemple) ou des données statistiques de différents ministères. Mais les volumes d'informations qui pouvaient être versés ont grandi de manière exponentielle. En 2011, les Archives Nationales ont constaté que les contenus prêts à être versés par les ministères étaient plus volumineux que l'ensemble des archives collectées depuis 1982. Il fallait agir !

### Le cadrage du programme VITAM : un long chemin

En France, trois ministères couvrent globalement l'essentiel de l'archivage historique de l'État, à grosse maille les ministères des Armées et des Affaires étrangères pour leurs besoins propres, le ministère de la Culture pour les autres (sauf Bercy...). En 2011, ils réfléchissaient tous à l'archivage numérique et faute de moyens suffisants, la seule voie était la coopération. Fait important, ce sont non seulement les grandes entités gestionnaires d'archives des trois ministères qui ont collaboré et élaboré ce qui n'était au début qu'un projet, mais aussi leurs DSI.

Il a fallu attendre près de quatre années pour réunir les moyens nécessaires. Un acteur essentiel a permis finalement le lancement, le Directeur Interministériel des Systèmes d'Information et de Communication de l'époque, Jacques Marzin. Fort sponsor, il a su mobiliser le Programme d'Investissement d'Avenir du CGI [3] pour compléter l'apport des ministères.

Début 2015, le programme Valeurs Immatérielles Transmises aux Archives pour Mémoire (VITAM) était né.

## Le programme VITAM

### L'essentiel du programme VITAM en quelques zooms

Les objectifs du programme :

La réalisation de la solution logicielle libre VITAM d'archivage numérique permettant la prise en charge, la conservation et la consultation sécurisée de très grandes volumétries d'archives numériques, pouvant même être classifiées ;

(3) Commissariat Général à l'Investissement

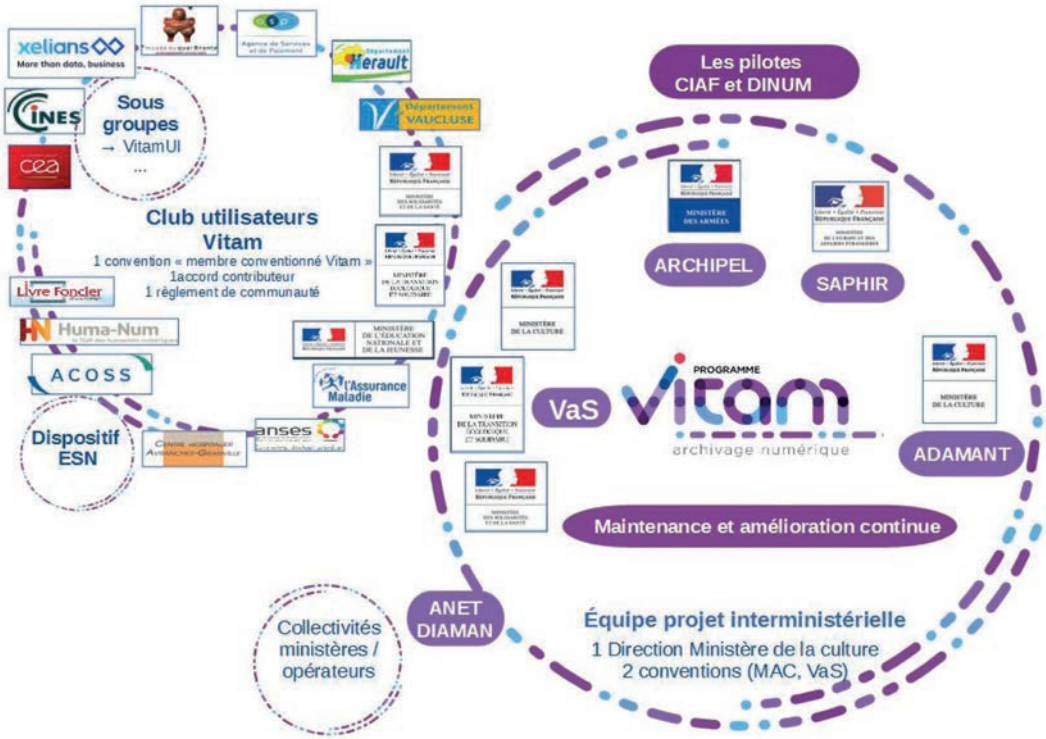
La mise en place de plateformes d'archivage utilisant la solution logicielle VITAM, dans chacun des trois ministères, *via* les projets ministériels : Saphir (MEAE), Adamant (MC) et Archipel (MinArm) ;

La diffusion et la réutilisation la plus large possible de la solution logicielle VITAM pour des usages d'archivage historique et intermédiaire.

Le « manifeste » [4] de la solution logicielle :

*Pour les services de l'Administration,*
*Afin de satisfaire les enjeux d'accès dans le temps à leurs informations numériques,*
*Vitam propose une solution logicielle libre d'archivage, évolutive, simple et facilement interfaçable,*
*Qui permet la gestion unitaire et sécurisée de milliards d'objets*
*Et vise son adoption par le plus grand nombre d'acteurs publics.*

La carte des acteurs début 2020 :




Des premières implémentations mises en production :

- 2018 : Culture (Adamant)
- 2019 : Affaires Etrangères (Saphir) et Xelians-Locarchives

Le développement logiciel :

- Rythme Agile

  - Sprint 3 semaines, Release tous les 3 mois, Versions majeures tous les ans

(4) Phrase résumant l'essentiel d'un projet dans les standards de développement Agile.

- V1 début 2018, V2 début 2019, V3 début 2020
- Jusqu'à 4 équipes Scrum en parallèle, plus de 10000 commits

- Architecture :

  - Microservices full REST
  - Souches « BigData et schema-less » : MongoDB, ElasticSearch
  - Stockage « agnostique et extensible » : File system, Objet (SWIFT, S3), Bande (LTO)
  - Preuves : Chaînage de journaux, arbre de Merkle, horodatage sûr de lots d'opérations (proche des concepts de Blockchain)
  - Multi-site, multi-tenant, de 3 à 75 VM, voire plus.
  - Automatisé par Ansible
  - Langages et OS supportés : Java 8 et 11, Debian et CentOS

- Performances

  - Conçue pour des milliards d'archives, manipulables unitairement, constituant des pétaoctets.
  - Tenue en charge testée régulièrement sur plus de 400 millions d'objets
  - Exemple de test d'injection : 160 000 objets bureautiques pris en compte pour archivage (donc avec toutes les garanties de conservation légale) en 18 min.

- Communauté

  - Une souche principale et des satellites (bibliothèques, griffons, IHM) libres (CECILL) sur GitHub
  - Plus de 15 partenaires publics utilisateurs (implémentation propre ou mutualisée), dont 2 contributeurs actifs (CEA, CINES)
  - Plus de 10 ESN, et des plus grands, formés sur VITAM
  - Le leader de l'archivage privé en France, Xelians, a construit son offre numérique sur VITAM et contribue au code

## Un programme volontairement mené de manière inhabituelle...

Tirant les leçons de l'échec de l'Office National de Paye et utilisant VITAM comme démonstrateur d'une « autre façon » de faire et de réussir les grands projets, la DISIC a exigé que le projet :

1. soit mené en Agile,
2. associe plusieurs prestataires pour éviter d'être sous le contrôle d'un seul dominant,
3. implique des agents publics experts fonctionnels et techniques maîtrisant les choix,
4. et que les projets d'implémentation soient intégrés au programme autour du projet de développement logiciel, chacun ayant du coup des comptes à rendre aux autres.

Outre les exigences de la DISIC, vinrent s'ajouter les exigences du CGI. Prodiguant des fonds qui ont vocation à soutenir le développement économique, le CGI a demandé une approche au moins de développement de l'usage, et si possible, d'activité économique autour de la solution.

Fort de ces deux injonctions atypiques dans les projets de l'époque, le programme VITAM est donc devenu un programme Agile, dont les choix étaient maîtrisés par l'administration et dont la priorité était les besoins des utilisateurs, garants d'une large réutilisation. Une approche fortement collaborative a été mise en place avec des archivistes de tous horizons (ministères, collectivités…) pour capter au mieux leurs réels besoins (et non envies), mais aussi avec des informaticiens pour garantir le réalisme technique des solutions proposées dans leur environnement technique.

Ainsi le programme VITAM n'était plus un projet informatique, mais une opération de construction d'un produit à l'usage d'une communauté professionnelle investie. Gagnant sa crédibilité, il a aussi su attirer les acteurs du monde économique.

## Un travail collaboratif de transformation...

Si le programme a pu exister, c'est parce qu'il répondait à un besoin fort de toute une communauté professionnelle obligée d'évoluer face à la révolution numérique. L'information changeant de média, il fallait de nouveaux outils pour la manipuler et assurer sa pérennité. Mais qu'en était-il des pratiques archivistiques ?

L'essentiel des notions séculaires utilisées dans la gestion des archives papier avaient une raison d'être, bien au-delà de la forme de l'information, et avaient toujours un sens dans le numérique. Quel que soit le média, l'objectif de l'archivage reste de récupérer l'information utile à conserver, dans des conditions assurant sa validité dans le temps et la capacité à la retrouver au moment utile même des années, des dizaines d'années ou des siècles plus tard.

Toutefois, au fur et à mesure des travaux, des notions essentielles furent interrogées par les moyens numériques.

Par exemple, l'archivage du nativement numérique a ouvert la voie à un traitement à l'échelle du document, voire de l'information élémentaire. Cela était inenvisageable en papier en s'appuyant sur une gestion manuelle. Tout l'intérêt du travail conjoint archiviste/informaticien fut de trouver la bonne façon d'utiliser cette puissance du numérique. Fallait-il s'en remettre complètement à de la recherche sur du vrac comme savent le faire les grands moteurs de recherche ? En fait, une vision plus riche, mariant vision hiérarchique et métadonnées sélectionnées à tous les points de la hiérarchie jusqu'à l'information finale, fut retenue. Les moteurs de recherche donnaient ainsi toute leur puissance sur des données de qualité.

D'autres notions, comme la prise en charge, la destruction, le changement de dépositaire, la gestion du cycle de vie aussi, furent revues et, à chaque fois, il s'avérait que la pratique pouvait être enrichie par le numérique, plutôt que remise en cause ou abolie.

Certains concepts nouveaux ont été ajoutés et nécessitent encore de s'enrichir avec le temps. Le plus complexe est celui de la pérennisation de l'information. Un vieux papier conservé proprement, quitte à prendre des cours de paléographie, on peut toujours le lire. Mais un vieux fichier ? Qui n'a jamais fait l'expérience d'un document, d'une photo, voire d'un ancien mémoire, retrouvé sur un disque qu'on n'arrive plus à ouvrir ou dont l'ouverture avec les applications actuelles donne une forme presque illisible ?

Le travail effectué avec la communauté des archivistes dans le cadre du programme VITAM a permis de faire monter en maturité un grand nombre d'archivistes sur l'adaptation au numérique des processus d'archivage. Il a aussi permis aux informaticiens qui construisaient la solution de comprendre les fondements et les vrais besoins, de partager les contraintes et d'apporter les pistes d'évolution et de transformation ouvertes par le numérique.

Une conclusion s'impose : la compétence des archivistes sur la gouvernance de l'information de manière probante et sur le temps long a bien toujours sa raison d'être, et peut-être encore plus dans le monde volatile du numérique.

## Un programme qui continue…

La phase initiale du programme VITAM était prévue jusqu'à la V3. Cette version riche de toutes les fonctionnalités essentielles pour l'archivage numérique est sortie début 2020. Mais le programme ne s'arrête pas là :

- D'une part, il a été décidé d'assurer la maintenance et l'amélioration continue de la solution logicielle, avec le maintien d'une équipe et de moyens apportés par les trois ministères d'origine

(Culture, Défense, Affaires étrangères), mais aussi au travers de l'animation d'une communauté d'usagers et de contributeurs ;

- D'autre part, a été lancé un projet interministériel de mutualisation d'une plateforme d'archivage en mode service, « VITAM Accessible en Service » (VAS), embarquant à ce jour les ministères de la Culture (pour ses besoins non patrimoniaux), de l'Écologie et des Affaires sociales. Ce projet vient compléter le développement du *back-office* VITAM par un enrichissement fort des IHM.

Une réelle communauté d'usage est constituée autour de VITAM, elle continue à porter et à faire vivre le produit. Celui-ci pourra ainsi continuer à évoluer, en lien avec la transformation numérique du métier et des pratiques des archivistes.

J'ai eu la très grande chance d'être à la barre de cette aventure de mi-2015 à fin 2019. J'y ai découvert le monde des archives avec des professionnels extraordinaires, porteurs d'une mission de service public sous-estimée et pourtant essentielle.

*Je ne peux malheureusement pas citer toute l'équipe VITAM mais je tiens à souligner l'action d'Edouard Vasseur* [5] *et Frédéric Brégier, pour la réalisation collaborative de la solution, et de Mélanie Rebours, pour sa diffusion* [6]*. Je tiens aussi à saluer l'équipe qui a repris le flambeau sur VITAM, notamment Emmanuel Laborde, qui en a repris la tête et qui était déjà directeur technique depuis 2018, et Alice Grippon, en charge de la diffusion. Toute une équipe d'agents publics et de prestataires a relevé ce défi technologique et humain, et peut être sincèrement fière du résultat.*
*Merci à tous.*

Pour tout savoir sur VITAM et son actualité : www.programmevitam.fr ou Twitter : @progvitam

(5) Maintenant professeur à l'Ecole Nationale des Chartes.
(6) Et la cohésion de l'équipe.

# Résumés

croissant du nombre de plateformes digitales dans la sphère éducative renvoie à une pluralité de stratégies au levier de l'exploitation des données recueillies et de leur médiation algorithmique. Cet article vise à en analyser les spécificités, à travers l'identification de leurs attributs techno-stratégiques différenciateurs.

# Abstracts



Given the semantic and institutional dimensions of the media, a complex industrial organization has arisen in which ecosystems comprising several firms interact to enhance their offer of contents. Digitization has introduced in these ecosystems new media that change not just production and distribution but also the very structure and aggregation of contents. The pricing schemes that used to serve to finance creative productions have been upended. To adapt, the historical media must thoroughly review their offers and pricing methods.



SVOD platforms (Subscription Video On Demand), like Netflix, have set off a revolution in the consumption of television and movies and introduced series as a new mixed genre between the two. This change is probably the most familiar. However the success of these platforms has wrought a second, less visible change: a total overhaul of television, whether for pay or for free, with an unprecedented consolidation of the big groups in the television and movie industries. While some telecommunication operators are buying up some of these firms and launching new platforms, others are merging to launch their own platforms. Meanwhile, the Internet giants are increasingly penetrating the sector if content production, thus making competition more intense with ever more sophisticated platforms. This article tries to interpret these changes and their effects on contents: on the one hand, more creativity, a freer tone of expression, the race to come up with original, exclusive contents, and the increasingly crucial role of recommendation systems for managing this hyperchoice; but on the other hand, the flight of art films, the industrialization of production, the war of talents (which is fueling a financial bubble), and contents that have to prove their worth very fast.



Platforms play a key role in the architecture of various markets. The market of online training is typical. Platforms have sprung up following a new business model based on strategies for using digital technology to capture and create value. Business models in the training market have developed around two-sided platforms, which enable them to make offers that simultaneously attract categories of users on each side of the market. Nonetheless, the online platforms in education, their numbers swelling, have a plurality of strategies for gaining leverage so as to process collected data and use algorithms. The characteristics of these strategies are analyzed by identifying their technostrategic attributes.

## 66 VITAM, an interministerial archival program

Jean-Séverin LAIR

Archiving is indispensable, even in the digital realm. Public archives are a key to the continuity of the state's activities. Without public digital archives, the state risks losing its memory. Digital archives have to harmonize the nanosecond with the century, flows with stability. VITAM, the interministerial program for a digital archives launched in 2015, hinges on developing a scalable, simple and easily interfacable freeware for archiving activities that can be used to securely manage billions of "objects". Initially borne by the ministries of Culture, Foreign Affairs and the Armed Forces, this program has attracted many users, public and private, to become involved in a shared approach. The first phase, 2015-2019, has come up with a product that, supported by a user community, will evolve and be enhanced.

# Ont contribué à ce numéro

**François ACQUATELLA** est Maître de conférences en Sciences de Gestion et du Management à l'École Universitaire de Management de Limoges et membre du laboratoire de recherche CREOP-EA 4332. Il est détenteur d'un Doctorat de Télécom Paris sur l'analyse stratégique du marché de la formation en ligne. Ses travaux de recherche actuels portent sur les écosystèmes digitaux et les *business models* qui y sont associés. Il est l'auteur de plusieurs articles sur le rôle et les incidences de l'intelligence artificielle sur la trajectoire stratégique des plateformes.
→ ***La dynamique stratégique des plateformes digitales : analyse du marché de la formation en ligne***

**Arnaud BEAUFORT** est Ingénieur général des Mines, chevalier des Arts et des Lettres et chevalier de la Légion d'honneur, ancien élève de l'École polytechnique. Il a commencé sa carrière au sein du groupe La Poste (1993-1996) puis a rejoint, en 1997, le Service juridique et technique de l'information (SJTI) en charge du bureau des affaires économiques. Il a pris part, en tant que rapporteur, à la mission de Patrice Martin-Lalande (« L'Internet, un vrai défi pour la France », 1997). En 2002, il est sous-directeur Produits, services et diffusion de la Documentation française où il est chargé des éditions, des sites internet (dont www.service-public.fr), de la promotion et de la diffusion. En 2007, il rejoint la Bibliothèque nationale de France comme Directeur des services et des réseaux et Directeur général adjoint en charge du numérique. Il a reçu en 2010 le prix du manager public de l'année.
→ ***Gallica, mine d'or et source de culture***

**Olivier BOMSEL** est professeur d'économie et directeur de la Chaire d'économie des médias et des marques à MINES ParisTech. Il est l'auteur de nombreux articles et ouvrages sur l'économie numérique, les médias, et le rôle des institutions dans l'économie. Son dernier livre, *La nouvelle économie politique*, est paru en 2017 chez Folio.
→ ***Mutation des écosystèmes de médias***

**Philippe BOUQUILLION** est professeur de sciences de la communication à l'université Sorbonne Paris nord et chercheur au Labex « Industries culturelles et création artistique » ainsi qu'au Laboratoire des sciences de l'information et de la communication, où il dirige la thématique « Industries culturelles, éducatives et créatives ». Il est coordonnateur de l'axe « Industries culturelles et arts » à la Maison des Sciences de l'Homme Paris Nord. Ses travaux portent sur l'économie politique de la communication et en particulier sur les mouvements de concentration et de financiarisation au sein des industries culturelles et de la communication ainsi que sur les enjeux des tournants créatifs et numériques. Plusieurs de ses recherches ont été conduites dans le cadre de contrats de recherche, notamment avec le Département des études, de la prospective et des statistiques du ministère de la Culture et de la Communication, l'Agence Nationale de la Recherche et la région Île-de-France. Ses recherches les plus récentes sont relatives au déploiement des plateformes numériques dans l'audiovisuel en Inde et en Europe, qu'il étudie du point de vue des enjeux industriels, réglementaires et des transformations de la mondialisation.
Publications récentes :
(2020), "Platforms at the Heart of Capitalism: Industrial and Financial Structures of OTTs in India", ATHIQUE A. & PARTHASARATHI V., *Platform Capitalism in India*, Palgrave, IAMCR Series. (à paraître)
(2019), "Digital Audiovisual Platforms, Between Transnational Flows and National Frameworks »,

in George Eric, *Digitalization of Society and Socio-political Issues 1. Digital, Communication and Culture*, ISTE Edition, 107-116.
(2019) "Cultural Diversity in the Country of Cultural Exception", in L. A. ALBORNOZ, M. T. GARCIA LEIVA, *Audio-visual industries and Diversity. Economics and Politics in the Digital Era*, Routledge.
(2018), BOUQUILLION P., MOREAU F., (co-editors), *Digital Platforms and Cultural Industries*, Brussels, Peter Lang.
→ ***Les stratégies de visibilité, le rôle des plateformes***

**Alexandre CHAUTEMPS** est chef du service Dépôt légal numérique à la Bibliothèque nationale de France. Après une première carrière en bibliothèques territoriales (à la Ville de Paris puis dans le Pays Châtelleraudais), il est pendant trois ans chef de projet métier chez un prestataire informatique travaillant avec les bibliothèques et les musées. Il intègre la Bibliothèque nationale de France en 2010 et y exerce, successivement, différentes fonctions :

- Coordonnateur des Signets de la BnF (répertoire sélectif de sites web),
- Correspondant pour le Labo BNF, qui était un espace de médiation centré sur technologies émergentes et leurs usages dans les domaines de la lecture et de la valorisation du patrimoine,
- Associé à différents projets d'édition numérique enrichie conduits par l'établissement : *Candide* et *Au bonheur des dames*
- Chargé de collections au Pôle de ressources et d'information sur le monde de l'entreprise (PRISME),
- Chef du service Dépôt légal numérique depuis 2018.

Il est particulièrement intéressé par les questions touchant à l'informatisation des bibliothèques, la préservation des documents numériques, l'édition numérique et les usages d'Internet.
Bibliographie sélective :
- Nicholas Cronk et Alexandre Chautemps, « Candide : éditer et (re)lire les classiques en numérique », *Revue de la Bibliothèque nationale de France*, 2012/3 (n° 42), pp. 29-35,
- Alexandre Chautemps, « Les Plans départementaux de développement de la lecture », *Transversales*, n° 77-78, juillet 2000, pp. 19-37
→ ***L'archivage du Web ou le Web comme mémoire des sociétés contemporaines***

Ancien élève de la Scuola Normale Superiore de Pise, **Roberto DI COSMO** a obtenu son doctorat en Informatique à l'Université de Pise. Après avoir enseigné plusieurs années à l'Ecole Normale Supérieure de Paris, il est devenu Professeur d'Informatique à l'Université Paris Diderot, où il a été directeur adjoint pour l'Informatique dans l'école doctorale ED 386 de 2005 à 2009. Membre du conseil scientifique et du conseil d'administration d'IMDEA Software, membre du comité d'orientation pour la Science Ouverte en France, il est actuellement détaché chez Inria. Il a une longue histoire de contributions à la recherche en Informatique, dans des domaines allant de la réécriture à la logique et de la programmation fonctionnelle à la programmation parallèle. Il s'intéresse maintenant aux problèmes nouveaux posés par l'essor du Logiciel Libre, et en particulier à l'analyse statique de grandes masses de code. Il a publié plus de 20 articles dans des revues internationales, et 50 articles dans des conférences internationales. En 2008, il a créé et coordonné le projet de recherche européen Mancoosi, avec un budget de 4.4 M€ et 10 partenaires réunis pour travailler à améliorer la qualité des systèmes logiciels à base de paquets. Suivant de près l'impact de l'Informatique sur la société, il prône depuis longtemps l'adoption du Logiciel Libre, notamment à partir de la publication du best-seller *Le Hold-Up Planétaire* en 1998. Il a créé le groupe thématique Logiciel Libre dans le Pôle de compétitivité Systematic à Paris, qui a financé plus de 40 projets de R&D depuis 2007, et il a dirigé de 2010 à 2019 l'IRILL, une structure

de recherche de pointe sur le Logiciel Libre. En 2015, il a été à l'origine de Software Heritage, une initiative qui vise à construire l'archive universelle de tout le code source publiquement disponible, en partenariat avec l'UNESCO.
→ ***Software Heritage : l'archive universelle des codes sources du logiciel***

Dans sa thèse soutenue en 2002, **Fabien GANDON** a été pionnier de l'intégration de l'IA distribuée et du Web sémantique. Chef de projet de recherche à l'Université Carnegie Mellon, il a ensuite étendu ces approches pour respecter la vie privée des utilisateurs lors des accès et des raisonnements sur leurs données personnelles. Chercheur Inria à partir de 2004, il étudie des modèles et algorithmes pour l'analyse sémantique des médias sociaux sur le Web. En 2012, il fonde Wimmics, une équipe de recherche commune (UCA, Inria, CNRS, I3S), qui étudie le rapprochement des sémantiques sociale et formelle sur le Web avec les méthodes d'IA. A présent directeur de recherche, il est aussi délégué scientifique adjoint du centre Inria Sophia Antipolis – Méditerranée, directeur du laboratoire commun QWANT-Inria, représentant d'Inria au World Wide Web Consortium (W3C), et responsable de la convention Ministère de la Culture/ Inria.
→ ***Une toile de fond pour le Web : lier les données et lier leurs vocabulaires sur la toile, pour un Web plus accessible aux machines***

**Jean-Séverin LAIR** est Ingénieur Général des Mines (92). Après un début dans les techniques de sécurité informatique, notamment au ministère des armées et dans le privé, il a fait sa carrière dans le numérique des services publics. En charge des services aux usagers à l'Agence du développement de l'administration électronique, en 2003, puis évoluant au sein de la structure pour devenir adjoint du service, en 2006, il a été responsable de la stratégie numérique au niveau interministériel. Il a notamment été porteur du premier cadre de collaboration des ministères pour la transformation numérique en 2007. Travaillant avec la Direction de la Réforme du Budget, puis intégré dans la Direction Générale de la Modernisation de l'État, il a aussi participé à des travaux sur les aspects modernisation/transformation macro-processus budgétaires, achat, archives, logistique… Il a été retenu en 2008 sur le poste de DSI au ministère de la Culture avec une mission de forte amélioration de l'efficacité des équipes projets et des infrastructures. Il a rempli ses objectifs, notamment en revoyant l'organisation numérique et en impliquant fortement dans le pilotage les métiers jusqu'aux directeurs. En 2015, il a pris la direction du programme VITAM, qui fait l'objet de l'article, dont l'objectif était de donner le moyen aux grandes entités de l'État d'assurer leur archivage numérique. Ce programme porté en forte autonomie a été mené en méthode Agile, tant au niveau du développement informatique que du pilotage global. Il est reconnu comme une vraie réussite, non seulement au niveau technologique, mais aussi au niveau de l'accompagnement de la transformation de la profession d'archiviste public. Enfin, depuis début 2020, Jean-Séverin Lair dirige le programme d'accélération de la transformation numérique TECH.GOUV, à la Direction Interministérielle du Numérique au sein des Services du Premier ministre.
→ ***Le Programme interministériel d'archivage VITAM***

**Valery MICHAUX** est titulaire d'un doctorat (Prix de thèse FNEGE) et d'une habilitation à diriger les recherches. Elle rejoint Neoma Business School en 2004 après seize années d'expérience professionnelle acquise dans le secteur privé puis le secteur public. Entre 2012 et 2015, elle devient directrice de la recherche de Neoma Business School. Elle a publié de nombreux articles de recherche et deux ouvrages. Elle a étudié durant de nombreuses années les compétences organisationnelles et interorganisationnelles, les dynamiques territoriales (l'effet *cluster*) et les politiques locales concertées. Ces dernières années, elle a centré ses recherches sur les transformations digitales, technologiques et sociétales qui traversent certains secteurs et provoquent ainsi de véritables mutations. Elle s'intéresse

à la fois aux méthodes de prospective permettant d'anticiper ces évolutions, aux répercussions de ces transformations sur les entreprises et sur leurs stratégies, mais également aux changements internes induits par ces mutations qui peuvent parfois être complexes à gérer. Elle est professeur de stratégie dans le département stratégie et entrepreneuriat de Neoma Business School.
→ ***Nouvelles plateformes entre télévision et cinéma : quelles mutations en cours ? Quels impacts sur les contenus ?***

**Michel SCHMITT** est membre du Conseil Général de l'Economie. Titulaire d'un doctorat et d'une habilitation à diriger les recherches en Morphologie Mathématique, il a successivement occupé des postes dans l'industrie (Laboratoire Central de Recherche de Thalès) et dans l'enseignement supérieur (directeur de la recherche de Mines ParisTech, Vice-président numérique de Paris Sciences et Lettres). Ses centres d'intérêt concernent le numérique et le traitement des données au sens large, probabilités, analyse d'image, intelligence artificielle, bio-informatique ainsi que ses interactions avec l'enseignement supérieur.
→ ***Introduction***

**Dr. VAZIRGIANNIS** is a Distinguished Professor at LIX, Ecole Polytechnique in France. He has conducted research in Frauenhofer and Max Planck-MPI (Germany), in INRIA/FUTURS (Paris). He has been teaching in AUEB (Greece), Ecole Polytechnique, Telecom-Paristech, ENS (France), Tsinghua, Jiaotong Shanghai (China) and in Deusto University (Spain). His current research interests deal with deep and machine learning for Graph analysis (including community detection, graph classification, clustering and embeddings, influence maximization), Text mining including Graph of Words, deep learning for word embeddings with applications to web advertising and marketing, event detection and summarization. He has active cooperation with industrial partners in the area of data analytics and machine learning for large scale data repositories in different application domains. He has supervised twenty completed PhD theses. He has published three books and more than 200 papers in international refereed journals and conferences and received best paper awards in ACM CIKM2013 and IJCAI2018. He has organized large scale conferences in the area of Data Mining and Machine Learning (such as ECML/PKDD) while he participates in the senior PC of AI and ML conferences – such as AAAI and IJCAI. He has received the ERCIM and the Marie Curie EU fellowships, the Rhino-Bird International Academic Expert Award by Tencent. He leads the AXA Data Science chair (2015-18) and the ANR-HELAS chair (2020-24). https://scholar.google.fr/citations?user=aWGJYcMAAAAJ&hl=en
→ ***Artificial Intelligence – challenges for the future***

**Emmanuel VINCENT** est ingénieur de recherche à l'École des hautes études en sciences sociales (EHESS), où il est chargé de mettre en œuvre la politique éditoriale et numérique de l'établissement. Littéraire de formation, il a publié des articles et collaboré à des ouvrages en tant que chercheur associé du CÉRÉDi (Centre d'études et de recherches éditer-interpréter, université de Rouen). Il a ensuite été éditeur, d'abord aux PURH (Publications de l'université de Rouen et du Havre, 2005-2008) puis aux Éditions de l'EHESS (2008-2015). Depuis 2016, il est responsable éditorial multisupport, pilote le projet du déploiement de Métopes au sein de l'EHESS, et s'est positionné comme référent auprès de structures éditoriales partenaires. Outre des actions de formation aux bonnes pratiques éditoriales, aux outils PAO et à la chaîne d'édition multisupport, la cellule Métopes qu'il anime propose aux utilisateurs une assistance technique et fonctionnelle sur la solution.
→ ***Métopes, édition et diffusion multisupports : un exemple de déploiement à l'EHESS***

UNE SÉRIE DES
ANNALES DES MINES
FONDÉES EN 1794

# ENJEUX NUMÉRIQUES

Série trimestrielle • N°10 - Juin 2020

**Rédaction**
Conseil général de l'Économie,
ministère de l'Économie et des Finances
120, rue de Bercy - Télédoc 797
75572 PARIS Cedex 12
Tél. : 01 53 18 52 68
http://www.annales.org

**François Valérian**
Rédacteur en chef

**Gérard Comby**
Secrétaire général

**Delphine Mantienne**
Secrétaire générale adjointe

**Liliane Crapanzano**
Correctrice

**Myriam Michaux**
Webmestre et maquettiste



**Photo de couverture :**
Robert Delaunay (1885-1941), *Rythme n°2*,
huile sur toile, 1938.
Paris, musée d'Art moderne.
Photo ©Musée d'Art moderne/Roger-Viollet.

**Iconographie**
Christine de Coninck

**Abonnements et ventes**
COM & COM
Bâtiment Copernic - 20, avenue Édouard-Herriot
92350 LE PLESSIS-ROBINSON
Alain Bruel
Tél. : 01 40 94 22 22 - Fax : 01 40 94 22 32
a.bruel@cometcom.fr

Mise en page : Nadine Namer
Impression : Printcorp
N° ISSN : 2607-9984

Éditeur délégué :
FFE – 15, rue des Sablons - 75116 PARIS -
www.ffe.fr

**Régie publicitaire : Belvédère Com**
Fabrication : Aïda Pereira
aida.pereira@belvederecom.fr
Tél. : 01 53 36 20 46
Directeur de la publicité : Bruno Slama
Tél. : 01 40 09 66 17
bruno.slama@belvederecom.fr

Le sigle « D. R. » en regard de certaines illustrations correspond à des documents ou photographies pour lesquels nos recherches d'ayants droit ou d'héritiers se sont avérées infructueuses.