Le patrimoine numérique, entre enjeux matériels et immatériels

PAR
Valérie Schafer

NIVEAU DE LECTURE
Facile ● ● ●

PUBLIÉ LE
12/06/2017

Nous passons tous les jours du temps sur Internet, mais nous avons déjà oublié les sites que nous visitions assidûment il y a dix ans. À quoi ressemblaient-ils ? Pour s'en souvenir, nous pouvons nous plonger dans les archives du Web. La question de l'archivage du patrimoine numérique a d'ailleurs pris de l'ampleur ces dernières années…

Photo Jason Scott [CC BY 2.0], via Wikimedia Commons.

« Internet Archive redonne vie au Macintosh de 1984 », « Internet Archive : testez le Macintosh de 1984 dans votre navigateur », pouvait-on lire en ligne à la mi-avril 2017, sur des sites spécialisés dans les contenus informatiques et numériques. La fondation Internet Archive annonçait en effet la sortie d'un émulateur permettant de retrouver l'environnement des premiers ordinateurs personnels et avec lui des logiciels comme MacWrite, MacPaint ou des jeux tels que Dark Castle et Space Invaders.

Cette annonce médiatisée, le succès d'expositions consacrées aux jeux vidéo ces dernières années ou celle consacrée aux gifs par le Museum of the Moving Image de New York en 2014, le dynamisme d'institutions comme le Computer History Museum aux États-Unis, ou encore l'organisation en juin 2017 à Londres d'une semaine consacrée aux archives du Web (Web Archiving Week), sont autant de signes d'un intérêt pour le patrimoine numérique sous toutes ses formes. C'est ce patrimoine varié et hétérogène, conjuguant aspects matériels et immatériels et réunissant de nombreuses parties prenantes que nous vous invitons à découvrir, mais aussi les enjeux sous-jacents de cette patrimonialisation. En effet, la volonté de conserver les documents et traces numériques, d'archiver le Web, de transmettre aux générations futures un patrimoine informatique, si elle s'inscrit dans la continuité d'initiatives de patrimonialisation à la fois technique, scientifique et industrielle, devient aussi une patrimonialisation de la communication et par son ampleur acquiert un statut particulier, reconnu en 2003 par l'Unesco : celui de patrimoine numérique.

Les différentes facettes du patrimoine numérique

En octobre 2003, le patrimoine numérique est reconnu — et ainsi son existence et sa valeur pleinement légitimées — par une Charte de l'Unesco qui met sous un même chapeau, tout en les distinguant, patrimoine numérisé et patrimoine nativement numérique (ce que les Anglo-Saxons appellent *Born-Digital Heritage*) :

« Le patrimoine numérique se compose de ressources uniques dans les domaines de la connaissance et de l'expression humaine, qu'elles soient d'ordre culturel, éducatif, scientifique et administratif ou qu'elles contiennent des informations techniques, juridiques, médicales ou d'autres sortes, créées numériquement ou converties sous forme numérique à partir de ressources analogiques existantes. Lorsque des ressources sont "d'origine numérique", c'est qu'elles existent uniquement sous leur forme numérique initiale », note ainsi la Charte. Celle-ci énumère quelques-uns de ces documents nativement numériques qui peuvent être des textes, des bases de données, des images fixes et animées, des documents sonores et graphiques, des logiciels et des pages Web.

Si ce patrimoine partage bien des points communs avec le patrimoine culturel immatériel défini par l'Unesco la même année, une troisième forme de patrimoine, que nous qualifierons de patrimoine du numérique pour le distinguer des précédents, apparaît aussi en filigrane. Ainsi, la Déclaration de Vancouver sur le numérique de 2012 — La Mémoire du monde à l'ère du numérique : numérisation et conservation — souligne à quel point les enjeux matériels sont prégnants pour la sauvegarde d'un patrimoine numérique risquant d'être perdu en cas d'obsolescence rapide du matériel et des logiciels qui servent à le créer.

La conservation du matériel a certainement été l'enjeu le mieux identifié et le plus rapidement dans le cadre de la poursuite des projets de conservation d'un patrimoine technique, industriel et scientifique. Elle n'a pas attendu le numérique pour être prise en charge par de multiples acteurs de la patrimonialisation.

Depuis la fermeture en 2010 du musée de l'informatique installé à la Défense, il n'existe plus de lieu fédérateur unique pour les collections françaises, alors dispersées entre différentes associations et musées dont celui des Arts et Métiers. Mais un mouvement est actuellement entrepris pour la réalisation d'un projet global s'appuyant sur des matériels, logiciels, documentations techniques et histoires orales, déjà préservés par plusieurs partenaires et acteurs de la gestion du patrimoine du numérique sur l'ensemble du territoire français (l'ACONIT, AMISA, le Cnam et son musée, la FEB, Homo Calculus, ou encore l'Espace Turing).

Outre la préservation indispensable des matériels, le patrimoine numérique doit absolument être associé à une réflexion sur les éléments de documentation divers (guides et modes d'emploi, Cd-Roms, kits de connexion, etc.), qui permettent de le recontextualiser, mais aussi de retrouver un patrimoine interactif. En effet, l'émulation, la préservation de consoles, d'ordinateurs, d'interfaces de programmation applicative (API), contribuent à les maintenir vivants au sein de leur écosystème. Brewster Kahle l'avait relevé dès 1997 dans Archiving the Internet, notant que *« alors qu'il est possible de lire un livre ancien de 400 ans imprimé par Gutenberg, il est souvent difficile de lire une disquette informatique qui a 15 ans »*. Celui qui dès 1996 bouleverse le patrimoine numérique en se lançant par la création d'Internet Archive dans l'entreprise titanesque d'archiver le Web mondial soulignait déjà des enjeux que relèvent aujourd'hui en partie sa fondation et une pluralité d'autres acteurs, institutionnels et scientifiques, parmi lesquels le récent projet Software Heritage soutenu par Inria.

Le patrimoine nativement numérique : d'Internet Archive à Software Heritage

La Charte de l'Unesco en 2003, en insistant sur le patrimoine dit « d'origine numérique » (mentionné dans les articles 1 et 7) au même titre que le patrimoine numérisé, reconnaît la valeur de documents qui n'existent qu'en format numérique, mais aussi les efforts de préservation et de patrimonialisation engagés en amont de cette Charte.

Parmi les pionniers dans ce domaine, la fondation Internet Archive est lancée en 1996 par Brewster Kahle en s'appuyant sur son entreprise Alexa (créée en 1996 et vendue à Amazon en 1999), spécialisée dans l'analyse de flux et la recommandation de sites. Dès 2001, la Wayback Machine permet aux internautes de parcourir la Toile du passé (aujourd'hui 286 milliards de pages archivées).

Figurines en céramique de Ted Nelson, Mary Austin et Brewster Kahle présentes dans la grande salle d'Internet Archive à San Francisco. Photo Jason Scott [CC BY-SA 2.0], via Wikimedia Commons.

En parallèle, d'autres initiatives se manifestent, par exemple au sein des bibliothèques nationales canadiennes et australiennes. Des projets précoces dans les pays scandinaves visent aussi dans la seconde moitié de la décennie 1990 à étendre le périmètre du dépôt légal au Web, tandis qu'est lancé le projet AOLA (*Austrian On-Line Archive*) au début des années 2000 pour développer un archivage du Web autrichien.

Toutes ces démarches font écho aux évolutions qu' a connues le patrimoine au cours des dernières décennies, à une patrimonialisation de plus en plus sensible à de nouveaux objets, mais aussi à l'ascension du numérique, qui prend place dans des aspects de plus en plus étendus et variés de nos vies professionnelles, économiques, sociales et personnelles.

Le mouvement est suivi dans la décennie 2000 par de nombreux pays européens, la France inscrivant l'archivage du Web dans le dépôt légal en 2006. Déjà dotée d'une expérience de conservation des vidéogrammes et documents multimédia composites depuis 1975 puis des multimédias, logiciels et bases de données depuis 1992, la Bibliothèque nationale de France (BnF) prend alors en charge cette mission avec l'Institut national de l'audiovisuel (Ina) qui se voit confier les sites Web relevant du périmètre audiovisuel. Au-delà de ces initiatives nationales, des initiatives transnationales peuvent être évoquées, par exemple le lancement en 2008 du projet LiWA (*Living Web Archives*).

En 2009, le projet Memento du Los Alamos National Laboratory Research Library a par ailleurs permis de réaliser un outil libre, offrant aux internautes un accès aux versions précédentes d'une page web grâce à un plug-in à ajouter au navigateur. Dans le même esprit, le projet « 404-no-more » porté par Firefox et Internet Archive vise à éliminer les « erreurs 404 » en redirigeant automatiquement vers une version archivée de la page demandée.

Outre les archives du Web, les archives des Newsgroups, espaces de discussion de la communauté Usenet (réseau né à l'extrême fin des années 1970), méritent aussi notre attention : gérées depuis 2001 au sein du service de forum Google Groups, elles *« ont accompagné les efforts de légitimation de l'entreprise auprès des publics d'utilisateurs, à une époque où Google était en phase de développement et de diversification de ses activités »*, rappelle Camille Paloque-Berges dans son article. *« Google, alors en train de gagner la guerre de moteurs de recherche, s'est érigé par ce geste en protecteur du passé du réseau, ainsi qu'en candidat à sa propre reconnaissance au sein de cette histoire. »*

Les communications et usages numériques les plus récents n'échappent pas non plus à cette patrimonialisation, à l'instar de l'archivage de Twitter, pris en charge par la Bibliothèque du Congrès américaine en vertu d'un accord avec Twitter depuis 2009 ou encore, avec un périmètre beaucoup plus restreint, le suivi par l'Ina et la BnF de quelques centaines de comptes Twitter et mots-dièses précis.

Enfin, parmi les derniers venus, avec des ambitions complémentaires des autres et spécifique à un champ jusque-là peu préservé, le projet Software Heritage lancé en 2016 complète ce paysage en plein essor. Comme le note Roberto di Cosmo, un des principaux instigateurs et porteurs de cette initiative : *« [...] Archiver du code source pose des problèmes spécifiques qu'on ne rencontre pas dans d'autres domaines. [...] La préservation du code source avec ses spécificités n'était vraiment au cœur de la mission de personne : on préservait des logiciels exécutables, jouables, des jeux vidéo, c'était notamment fait par Internet Archive qui a une grosse sélection de jeux vidéo. On préservait des pages web qui parlaient de logiciels et de codes sources. Mais les codes sources, comme objet noble, non. »*

L'articulation entre patrimoines et publics

Public scientifique, experts, amateurs et grand public, monde des médias, industriels, étudiants et enseignants, les publics potentiels du patrimoine numérique sont nombreux et les usages de celui-ci encore largement à explorer, favoriser, stimuler, inventer. Ainsi Roberto Di Cosmo espère que le projet Software Heritage intéressera les acteurs du patrimoine scientifique et technique ainsi que ses publics, mais aussi le monde de la recherche scientifique, qui pourra y trouver une archive de référence, ou encore le monde industriel.

Cependant, pour réunir et accueillir pleinement les publics, plusieurs défis sont encore à relever, car la vocation d'ouverture et de participation n'a pas toujours été pensée au préalable : bien sûr, il y a des questions d'accessibilité des données, notamment dans le cadre du dépôt légal, qui limite la consultation des archives du Web *in situ* en France à la BnF et quelques bibliothèques en région. Mais les enjeux concernent aussi l'interopérabilité, qui se pose par exemple à l'échelle européenne, car les fonds d'archives du Web sont imperméables entre les différents pays. L'accessibilité doit aussi être cognitive et pose le problème de l'accompagnement dans la découverte de ces sources, de la maîtrise des outils de traitement, de la littératie numérique, du substrat de culture informatique et numérique nécessaire (sujet d'actualité autour de l'apprentissage du code dans le secondaire). Enfin, des enjeux éthiques ne peuvent manquer de se manifester. Reste également à penser davantage la place de ces publics en amont même des réalisations. Comme le notaient en 2011 Hafizur Rahaman et Beng-Kiang Tan dans leur article :

« Les projets actuels de patrimonialisation numérique se concentrent surtout sur le "processus" ou sur le "produit", mais ne considèrent que rarement les "utilisateurs" [...]. Pour une meilleure interprétation et expérience d'un site relevant du patrimoine numérique, il nous faut une méthode d'interprétation inclusive, qui devrait tenir compte de la variété de compétences des utilisateurs, dépasser la linéarité de la narration et la subjectivité dans la création des contenus. » (traduction : Francesca Musiani)

Si en quelques années la situation a déjà beaucoup évolué, notamment sous l'effet d'échanges de plus en plus féconds et nombreux entre le monde des archives, des bibliothèques et des chercheurs, elle peut aller encore plus loin pour pleinement inscrire dans cette dynamique les producteurs et publics, notamment les « publics ordinaires ». Ceux-ci restent souvent simples spectateurs de choix qui ne sont au demeurant pas le seul fait des institutions patrimoniales, mais aussi de plus en plus souvent des grandes entreprises de communication.

Des objets de recherche, des objets au service de la recherche

Alors qu'à ses débuts, le patrimoine nativement numérique concernait essentiellement le monde des bibliothèques et des archives, les chercheurs commencent à s'y intéresser sérieusement depuis quelques années, l'envisageant à la fois comme objet de recherche propre et objets-sources au service de leurs recherches.

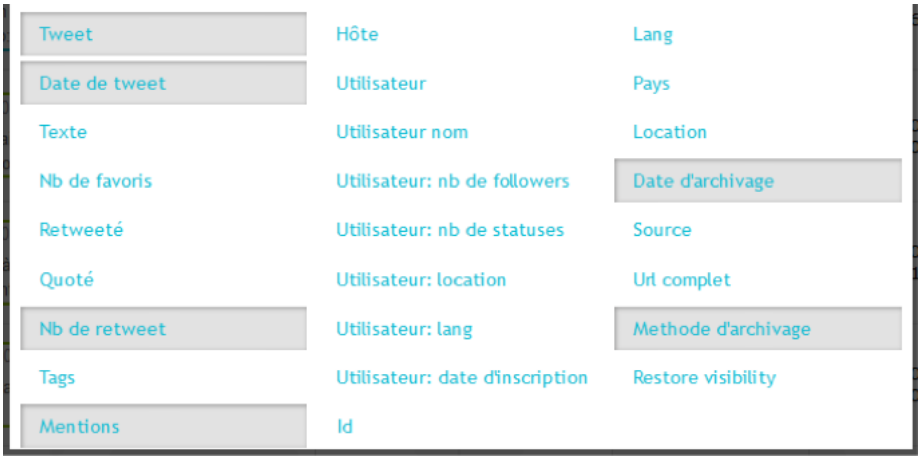
La réflexion a d'abord porté sur le patrimoine numérisé, que ce soit dans le champ de l'histoire ou des sciences de l'information et de la communication, mais des initiatives comme les ateliers du Dépôt Légal du Web à l'Ina, sont un jalon important en France dans l'implication des communautés de recherche autour des archives du Web. Comme le relevait Louise Merzeau, coorganisatrice des ateliers, dans son article : *« Bien sûr, ce déploiement d'une vue stratifiée du réseau ne nous est pas familier, et il nous faudra apprendre à la manipuler. Comme outil de représentation, de navigation et de compilation, c'est l'archive elle-même qui produira ces nouveaux usages. De la même manière que l'archivage des sources audiovisuelles a rendu possibles quantité de recherches sur la radio et la télévision qu'on ne pouvait auparavant formaliser, le dépôt légal du Web est une condition de sa conversion en fait de culture. »*

Les historiens du monde contemporain se convertissent aujourd'hui pour certains avec enthousiasme à ces nouvelles sources. Au sein de ces approches, l'importance des réflexions épistémologiques et méthodologiques est notable : sans rompre avec les méthodes historiennes antérieures, les chercheurs sont conscients de l'importance de bien comprendre ces sources avant de les exploiter. Nous avons notamment pu souligner avec Francesca Musiani et Marguerite Borelli dans notre article *« Negotiating the Web of the Past »* l'importance d'ouvrir les boîtes noires des archives du Web pour en saisir les biais et les multiples médiations subies au cours de l'archivage. Nous n'en rappellerons ici que quelques rapides éléments afin d'insister sur le fait que, comme l'avait noté l'historien danois Niels Brügger en 2012 dans la revue Le Temps des Médias, l'archive du Web est rarement une copie parfaite du site Web dans son aspect originel sur le Web vivant. Enchâssée dans des interfaces de consultation contemporaines, transformée sous l'effet de la perte de documents (des publicités, des images dans les années 1990, etc.), une page subit de nombreux changements. Ceux-ci sont encore amplifiés à l'échelle d'un site, par la remise en hypertextualité, quand certains hyperliens introduisent des sauts temporels entre plusieurs pages archivées à des dates différentes, mènent parfois à des impasses (les pages ne sont pas toutes archivées, et un site est

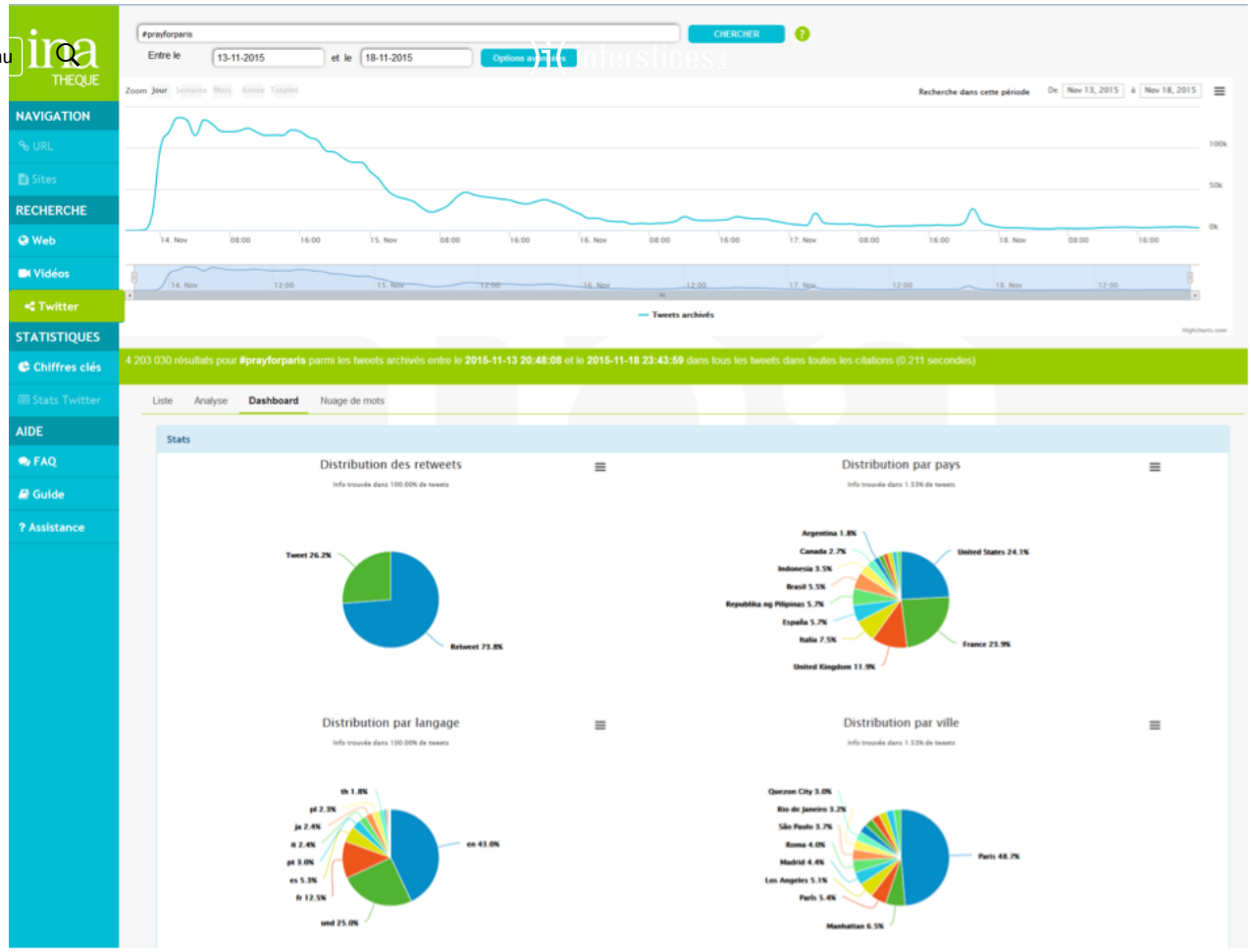
rarement archivé à plus de deux ou trois clics de profondeur), etc. Comprendre les techniques, périmètres, choix de conservation opérés par les institutions est un préalable à une création raisonnée de corpus, quand bien même le chercheur semble avoir à disposition suffisamment, voire trop, de données à étudier.

L'archivage de Twitter au moment des attentats parisiens de 2015, opéré par la BnF et l'Ina, en témoigne. Ainsi, si l'Ina a pu collecter au moment de ces attentats environ 11 millions de tweets, reste que cette collection pléthorique comporte nécessairement certains biais et lacunes, notamment par le choix des mots-dièses archivés (dont la sélection a été faite en temps réel, au cours des événements) ou encore par des pertes de tweets au moment de la collecte via l'API publique de Twitter (celle-ci limite en effet la collecte gratuite à 1% du flux mondial à un instant donné. Or les flux Twitter consacrés aux attentats ont parfois représenté plus de 1% du total de tweets émis au niveau mondial, faisant perdre partie d'entre eux).

De cette masse de données découle aussi une autre piste de réflexion, sur la nature des outils permettant d'exploiter ces vastes gisements. Comme le souligne Thomas Drugeon, responsable du dépôt légal du Web à l'Ina, lors de notre entretien — et la question se pose à l'identique côté BnF —, le chercheur ne peut emporter avec lui les données, pour leur offrir le traitement appareillé par les outils informatiques de son choix. Les règles du dépôt légal le contraignent à traiter ces documents dans les enceintes des institutions. Aussi le monde des archives du Web développe-t-il de plus en plus des outils destinés à accompagner les chercheurs, permettant notamment dans le cas de l'Ina la réalisation de *timelines* ou de nuages de mots, le suivi de la circulation et de la popularité d'images, ou encore le croisement de nombreuses métadonnées, dont témoignent quelques-unes des figures suivantes.



Possibilité de croiser les données et métadonnées au cours de l'exploration des tweets et mots-dièse dans l'interface Ina. © Ina



Timeline et statistiques d'une recherche sur #prayforparis dans l'interface Ina. © Ina

Possibilité de générer un nuage de mots à partir d'une recherche, ici sur #jesuisahmed, dans l'interface Ina. © Ina

La BnF, en implémentant également dans ses archives des attentats de 2015 une recherche plein texte qui permet de croiser de multiples facettes, offre une entrée facilitée dans les données, non sans questionner également le chercheur sur les biais que ces outils peuvent induire dans la recherche qu'il va mener et la manière dont il va aborder ces masses de données.

Menu

Recherche plein texte et possibilité d'affiner les résultats à l'aide de facettes dans les archives du Web des attentats de 2015. © BnF

Conclusion

« Toute personne qui travaille avec des archives du Web s'est rapidement habituée au fait que la plupart des gens n'en ont même jamais entendu parler — et encore moins comprennent ce qu'elles sont et comment y accéder. En 2016 cependant, il semble que les archives du Web ont commencé à pénétrer la conscience du public, à passer des pages Technologies de la presse aux sections politiques et même culturelles », notait Jane Winters dans son article en début d'année. L'année 2016 aura-t-elle été celle des archives du Web, comme le suggère l'historienne britannique, familière de ces matériaux depuis plusieurs années ? Et ce succès de visibilité ne risque-t-il pas de se faire au détriment d'autres patrimoines numériques, moins valorisés actuellement, mais tout aussi importants (conservation des banques de données par exemple) ?

Dans tous les cas, en France comme dans le monde anglo-saxon, ce sujet, jusque-là plutôt confidentiel, aura fait l'objet d'une plus large couverture médiatique, notamment de la part du Monde, de Libération ou encore de L'Express, à la faveur des vingt ans de la fondation Internet Archive et des dix ans du dépôt légal du Web en France. Ainsi, les 22 et 23 novembre 2016, au cours du colloque « Il était une fois dans le Web. 20 ans d'archives de l'Internet en France », se réunissaient de multiples acteurs intéressés par ce patrimoine, professionnels de l'archivage et des bibliothèques, des médias, journalistes et chercheurs. Tous les intervenants témoignaient avec passion des défis techniques, mais aussi politiques et culturels passés et à venir de ce patrimoine nativement numérique. De plus en plus pléthorique, ce patrimoine mettra également au défi l'écriture de l'histoire, non seulement celle du numérique mais celle de nos sociétés contemporaines dans toutes ses facettes.

Repères bibliographiques

facile à lire (affiché : facile)

Menu

totalement accessible avec un léger bagage scientifique (facile)

accessible en grande partie avec un léger bagage scientifique (intermédiaire)

accessible avec des connaissances scientifiques (avancé)

difficile d'accès (avancé)

Si vous souhaitez expliquer votre choix, vous pouvez ajouter un commentaire (qui ne sera pas publié).

Enregistrer

RECOMMANDATIONS

L'apprentissage profond : une idée à creuser ?

Regard sur « Le temps des algorithmes »

Idée reçue : C'est la faute à l'ordinateur !