



HAL
open science

Les outils de l'apprentissage profond au service de l'évaluation et de la conservation des archives

Valérie Lee, Julien Longhi, David Picard, Camille Simon Chane, Michel
Jordan

► To cite this version:

Valérie Lee, Julien Longhi, David Picard, Camille Simon Chane, Michel Jordan. Les outils de l'apprentissage profond au service de l'évaluation et de la conservation des archives. *Conservation-restauration des biens culturels*, 2022, 38, pp.7-16. hal-03620940

HAL Id: hal-03620940

<https://hal.science/hal-03620940v1>

Submitted on 15 Apr 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

LES OUTILS DE L'APPRENTISSAGE PROFOND AU SERVICE DE L'ÉVALUATION ET DE LA CONSERVATION DES ARCHIVES

Valérie Lee, Julien Longhi, David Picard, Camille Simon Chane, Michel Jordan

Résumé Cet article présente les résultats de la première année de recherche doctorale sur l'utilisation d'une technologie de l'intelligence artificielle, l'apprentissage profond, pour évaluer l'état de conservation des archives reliées du Parlement de Paris à partir de photographies et d'une manière automatique. Ce doctorat est mené à l'université de CY Cergy Paris Université en partenariat avec les Archives nationales et le laboratoire ETIS.

Abstract This article presents the results of the first year of a doctoral research on the use of an artificial intelligence technology, deep learning, to evaluate from photographs the condition state of the binded archives of the Parlement de Paris (parliament of Paris), with an automatic mode. This doctoral thesis is carried out at the university CY Cergy Paris Université in partnership with the Archives nationales (national archives) and the ETIS laboratory.

Resumen Este artículo presenta los resultados del primer año de investigación del doctorado sobre el uso de una tecnología de inteligencia artificial, el aprendizaje profundo, para evaluar el estado de conservación de los archivos encuadernados del Parlamento de París por medio de fotografías y de manera automática. Este doctorado se realiza en la universidad CY Cergy Paris Université en asociación con los Archivos nacionales y el laboratorio de ETIS.

Mots-clés apprentissage profond, réseaux neuronaux, archives, registres, Parlement de Paris, conservation

Une des missions essentielles des Archives nationales est de conserver ses fonds pour les transmettre aux générations futures. Les 370 kilomètres linéaires de livres et de documents sur papier, qui ne cessent d'augmenter au fil des versements, rendent cette tâche particulièrement complexe. En effet, il est difficile de planifier des traitements de stabilisation des reliures et des papiers sans connaître en détail leur l'état. L'intelligence artificielle offre aujourd'hui la possibilité de traiter des masses de données importantes en un temps réduit. Elle est utilisée dans le domaine patrimonial pour classifier, entre autres, les styles (Crowley, 2014; Seguin, 2018), les matériaux et la mise en œuvre des peintures (Li *et al.*, 2011), pour reconnaître des écritures manuscrites (Muehlberger *et al.*, 2018) ou pour modéliser en 3D des monuments à partir d'anciens plans (Jordan, Ringot, 2020). Elle est aussi à la base du développement d'outils de détection d'altérations sur des bâtiments en pierre (Kwon, Yu, 2019) ou de classification de fragments de céramiques archéologiques (Ostertag, Beurton-Aimar, 2020). Cependant, nous n'avons

pas trouvé d'exemple de son utilisation dans le domaine spécifique de la conservation des archives. Dans le cadre d'une thèse à l'école doctorale Arts, Humanités, Sciences sociales de CY Cergy Paris Université, nous proposons d'appliquer une technologie de l'intelligence artificielle, la segmentation sémantique, afin de connaître l'état des reliures des registres du Parlement de Paris conservés aux Archives nationales. Les objectifs, le protocole ainsi que les résultats de cette première année de recherche (2020-2021) seront présentés dans cet article.

Les méthodes d'évaluation de l'état des collections d'archives et de bibliothèques

Dès les années cinquante, les problèmes particuliers de conservation des documents dans les bibliothèques ont été discutés et étudiés, en particulier dans les pays anglo-saxons (Barron, 1959). Afin d'estimer l'état de leur collection et le nombre de livres comportant des papiers acides, les responsables des archives et bibliothèques se sont tournés vers les outils statistiques pour les évaluations des états de conservation. De nombreux tests ont été entrepris pour mettre en place des protocoles (Drott, 1969; Walker *et al.*, 1985; Chrzastowski *et al.*, 1989; Eden, 1998).

Les années 2000 marquent une avancée notable pour la gestion des collections, avec le développement des outils informatiques. Une étude d'envergure appelée *Knowing the need* est conduite en Angleterre entre 2001 et 2005 sur deux cents collections de bibliothèques et d'archives par le National Preservation Office. Elle aboutira à la mise en place de méthodologies de constat d'état fondées sur les statistiques, avec des méthodes de tirage aléatoire, de tirage systématique ou de tirage stratifié (Walker, Foster, 2006). L'intervalle de confiance de ces méthodes est estimé à 5 % (Walker, 2009). Les outils de collecte sont en majorité des feuilles de calcul Excel® ou des bases de données construites avec Access® ou FileMaker-pro®. Récemment, il y a eu un mouvement d'intérêt pour un langage de programmation et un logiciel libre de statistique, le logiciel R. Celui-ci permet d'obtenir des évaluations qui limitent les risques de modifier accidentellement les données (Elie, 2018, p. 138). En 2005, en France, un groupe de professionnels du patrimoine et de statisticiens mettent au point une méthode d'évaluation de l'état physique des fonds d'archives et de bibliothèques qui s'applique uniquement aux documents graphiques (norme AFNOR NFZ40-011). Elle est aussi fondée sur une méthode de sondage.

Si les méthodes statistiques permettent d'avoir une idée assez précise de l'état de conservation d'un fonds, elles n'offrent pas la possibilité de connaître le niveau d'altération de chaque document. Cela pose un problème pour les prêts ou la numérisation des documents, puisque l'information reste partielle. Il nous a donc semblé pertinent d'explorer les nouvelles technologies informatiques, en particulier celles de l'intelligence artificielle, pour obtenir des données précises sur l'état de chaque document, tout en réduisant le temps de travail. Notre étude s'est tournée vers les fonds reliés, car les altérations mécaniques qui rendent les prêts et la numérisation difficiles sont visibles de l'extérieur, sur le dos de la reliure, et ne demandent pas de manipulation.

Le fonds des archives du Parlement de Paris, Archives nationales, site de Paris



Figure 1 Une partie des réserves des registres du Parlement de Paris. © Eric Laforest, Archives nationales.

Pour ce projet de recherche, notre choix s'est porté sur la collection des archives du Parlement de Paris, qui est exceptionnelle par son histoire, son volume et son homogénéité. Le Parlement de Paris a été actif depuis le règne de Saint-Louis jusqu'à la Révolution française. Il a reçu des appels, au civil et au criminel, des juridictions ordinaires inférieures, et instruit directement les procès concernant les personnages les plus importants du royaume : roi, ducs et pairs (Hildesheimer, 2011).

Les archives datent du XIV^e au XVIII^e siècle et sont regroupées dans la série X (**fig. 1**). Elles sont composées de vingt-six mille articles organisés en minutes et registres. Les registres sont au nombre de onze mille six cent cinquante-neuf. Ils contiennent des arrêts qui sont écrits à l'encre ferro-galique sur du parchemin de bonne qualité. Les reliures sont homogènes sur l'ensemble de la collection : elles sont en cuir de veau mégis sur le dos avec des plats cartons parfois recouverts de parchemin de veau. Les corps d'ouvrage sont en bon état mais les reliures présentent toutes, à des degrés différents, des altérations mécaniques créées par leur usage ou leur conditionnement. On peut ainsi observer des pertes de maintien et de cohésion du corps d'ouvrage causées par le décollement de l'apprêture, des ruptures de couture et des abrasions éten-

duës sur le cuir des dos. De nombreuses reliures sont trop fragilisées pour être données en consultation (**fig. 2**). Pourtant, ce fonds est régulièrement étudié par des chercheurs car il présente un important intérêt historique.

Le protocole de la recherche

La recherche se concentre sur la zone que l'on voit lorsque les registres sont placés sur les rayonnages de la réserve. Elle correspond au dos, qui est structurellement déterminant pour permettre l'ouverture et la fermeture du registre. Trois parties sont particulièrement sensibles aux accidents de manipulation : la charnière, la coiffe et la tranchefile. En partenariat avec les responsables du fonds, messieurs les conservateurs Michel Ollion et Amable Sablon



Figure 2 Registres du Parlement de Paris sur une étagère. © Eric Laforest, Archives nationales.

du Corail ainsi qu'avec monsieur Eric Laforest, chef de travaux d'art, restaurateur de documents graphiques et de livres aux Archives nationales, six classes d'altération ont été choisies :

- charnières extérieures fendues;
- lacunes du cuir sur le dos;
- cahiers apparents sur le dos;
- coiffes lacunaires;
- tranchefiles lacunaires;
- sangles qui retiennent les reliures très endommagées;
- reliures placées dans des boîtes ou dans des jaquettes en papier à cause de leur état de dégradation.



Figure 3 Carte des zones altérées sur six registres. Chaque couleur correspond à une classe d'altération. © Valérie Lee.

L'objectif de la recherche est d'obtenir une carte des zones altérées qui se superposera à la photographie du registre (**fig. 3**). Cette carte sera ensuite transcrite en un texte qui décrira les altérations ainsi que leur niveau (état critique, moyen, bon). Ce texte pourra être facilement intégré à la base de données des Archives nationales (**fig. 4**).

X/1a/8773 : non communicable, état grave. Charnières extérieures fendues/ critique.
 Dos lacunes de cuir/critique. Cahiers apparent/critique. Tranchefiles/critique. Coiffes/critique. Sangle.

Figure 4 Exemple de texte généré à partir de la carte des zones altérées.

La segmentation sémantique

La technologie que nous avons choisie pour cette recherche est une opération de traitement des images appelée segmentation sémantique. Elle utilise des algorithmes de haut niveau, les réseaux de neurones, qui ont pour but de rassembler les pixels des images dans des régions en suivant un critère d'homogénéité (Goodfellow *et al.*, 2016). Ces réseaux de neurones doivent être entraînés avec des bases de données d'images annotées pour reconnaître les images et les classifier. Prenons par exemple l'image d'un chat et entrons cette image dans un réseau de neurones qui aura été entraîné à reconnaître la différence entre une image de chat et une image de chien. Le réseau va analyser les pixels des images à l'aide de filtres spécialisés dans un certain type de forme (coins, contours, aplats, etc.), et trier les parties analysées en classe « chat » et classe « non chat ». À la sortie du réseau, on obtiendra un pourcentage de probabilité indiquant que l'image en entrée est celle d'un chat (**fig. 5**). Plus le nombre d'images entrées dans le réseau sera grand, plus les performances de reconnaissance seront bonnes. En effet, les réseaux de neurones possèdent une capacité d'apprentissage automatique, c'est-à-dire qu'à chaque étape de l'entraînement les mauvaises réponses sont corrigées et le modèle mathématique ajusté. Au fur et à mesure, le programme réorganise les informations en blocs de plus en plus complexes : c'est ce que l'on appelle l'apprentissage profond. Ces réseaux de neurones sont utilisés très largement dans notre quotidien avec, par exemple, la classification d'images sur internet, la reconnaissance faciale pour les téléphones portables ou le traitement des clichés dans l'imagerie médicale.

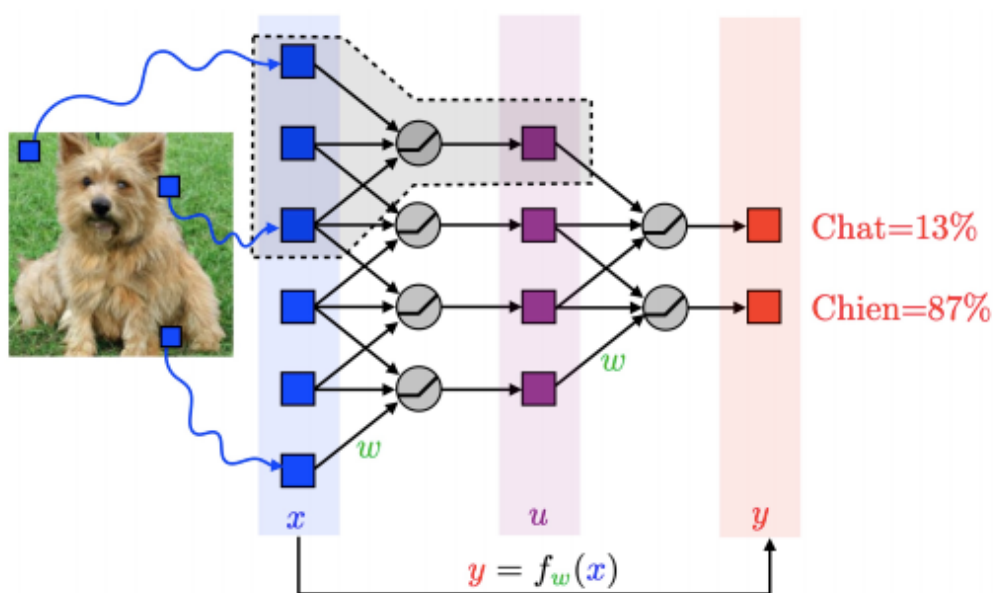


Figure 5 Exemple d'un réseau de neurones. © Gabriel Peyré.

Pour notre recherche, des réseaux neuronaux généralistes ont été adaptés au contexte particulier des registres afin de réduire la complexité de l'apprentissage. Tout d'abord, une base de données de deux cents photographies comportant en tout mille deux cents registres sur leur étagère a été segmentée : les altérations dangereuses situées sur le dos des reliures ont été relevées directement sur les photographies par des traits ou des aplats de couleur avec le logiciel Adobe Photoshop® (fig. 6). Trois mille masques ont été créés. Le but de l'entraînement du réseau de neurones est d'obtenir automatiquement des masques signalant les surfaces altérées à partir de photographies non annotées. Ainsi, avec une base de données de mille deux cents registres annotés, les informations sur l'état de conservation des dix mille quatre cent cinquante-neuf registres restants seront générés automatiquement.



Figure 6 Segmentation des images de registres. © Valérie Lee.

Premiers résultats

La base de données a été confiée au laboratoire ETIS UMR 8051 (CY Cergy Paris Université, ENSEA, CNRS). Sous la direction de David Picard, puis de Camille Simon Chane et de Michel Jordan, deux étudiants en Master 2 à l'ENSEA, Florian Zeni et Lahcen Yamoun ont testé différents réseaux neuronaux généralistes comme U-Net (Long, 2015) et Deeplab (Chen, 2018). C'est le réseau U-Net pré-entraîné qui a donné les meilleurs résultats avec, pour certaines altérations comme les boîtes ou les sangles, une reconnaissance à 96 % (fig. 7). Par ailleurs, les registres ont pu être identifiés individuellement sur les étagères par le réseau Mask RCNN (He *et al.*, 2017) avec comme *backbone* ResNet101+FPN (fig. 8). Il reste encore de nombreux tests à conduire avant d'être en mesure de publier les résultats mais ceux-ci sont déjà prometteurs.



Figure 7 Zones de lacunes de cuir en bleu détectées automatiquement avec le réseau U-Net pré-entraîné. © Florian Zeni.



Figure 8 Registres détectés automatiquement sur leur étagère avec le réseau Mask RCNN. © Lahcen Yamoun.

Les recherches vont continuer au laboratoire ETIS pour identifier avec une technologie de reconnaissance optique des caractères (OCR) les cotes des registres sur les photographies et lier la cote avec les altérations détectées. Nous souhaitons aussi calculer les surfaces d'altération par rapport à la surface du dos de chaque registre pour déterminer la gravité de l'état d'altération. Le passage des données en format Excel sera aussi étudié pour permettre le tri et l'entrée des informations dans la base de données des Archives nationales.

Les données de cette recherche (images des étagères de registres et images annotées pour l'apprentissage automatique) seront hébergées sur les serveurs informatiques d'ETIS et accessibles à tous les partenaires du projet. Les infrastructures d'ETIS garderont également les résultats obtenus par les algorithmes d'apprentissage sous la forme d'un site web ou d'un carnet de recherches présentant la méthodologie du projet et ses avancées.

Conclusion

Nous espérons que cet outil, une fois développé, pourra être utilisé pour d'autres types de reliures que celles, très spécifiques, des registres du Parlement de Paris. Il devrait très certainement aider les bibliothèques patrimoniales, présentes sur l'ensemble du territoire français, à conserver leur fonds précieux. En effet, les employés de ces bibliothèques manquent de temps et de connaissances pour relever les problèmes d'altération sur les reliures. Les photographies des rayonnages seront aisément prises par les employés puis envoyées au laboratoire. Ensuite, un premier constat sera généré par l'outil. Ainsi, les bibliothèques pourront être alertées sur l'état de leur fonds et solliciter de l'aide pour la restauration. L'outil permettra aussi de surveiller l'état d'un fonds d'archives ou de bibliothèque à un instant T. En effet, à cause du très grand nombre de volumes conservés, les responsables des fonds n'ont pas de vision détaillée de l'état de leurs collections. De plus, il arrive régulièrement que des études partielles sur l'état d'un fonds soient faites sur plusieurs années et deviennent inutilisables quand des décisions de restauration doivent être prises à un instant T. C'est dans l'optique d'obtenir un outil utile adapté à la réalité du terrain que nous continuerons cette recherche doctorale dans les deux années à venir.

Remerciements

Nous tenons à remercier Monsieur Michel Ollion et Monsieur Sablon du Corail pour nous avoir donné accès à la collection des registres du Parlement de Paris. Nous sommes reconnaissants à Monsieur Eric Laforest pour nous avoir indiqué ce fonds exceptionnel et pour ses nombreux conseils avisés.

Le stage de Lahcen Yamoun a été financé par la Fondation des sciences du patrimoine (appel à projets 2020 bis) sous le nom de *Utilisation d'outils de deep learning pour l'évaluation de l'état sanitaire des registres du parlement de Paris (Archives nationales de France)*.

Références bibliographiques

- Andulla W.** (2018), *Splash of color : Instance segmentation with maskrcnn and tensorflow*, [en ligne]. Disponible sur : <<https://engineering.matterport.com/splash-of-color-instance-segmentation-with-maskrcnn-and-tensorflow-7c761e238b46>> (consulté le 20 janvier 2021).
- Barnier J.** (2013), *Introduction à R*, Centre Max Weber CNRS-UMR 5283, [en ligne]. Disponible sur : <https://cran.r-project.org/doc/contrib/Barnier-intro_R.pdf> (consulté le 5 janvier 2021).
- Barrow J.** (1959), *Deterioration of book stock, causes and remedies. Two studies on the permanence of book paper conducted by W.J.Barrow*, Richmond, Virginia State Library, 72 p.
- Chague A., Terriel L., Romary L.** (2020), « Des images au texte : LECTAUREP, un projet de reconnaissance automatique d'écriture », dans *DHNord2020, The Measurement of images : computational approaches in the history and theory of the arts, Nov 2020*, Lille. Disponible sur : <<https://hal.archives-ouvertes.fr/hal-03008579>> (consulté le 5 janvier 2021).
- Chen L., Zhu Y., Papandreou G., Schroff F., Adam H.** (2018), « Encoder-decoder with atrous separable convolution for semantic image segmentation », dans *Proceedings of the European conference on computer vision (ECCV)*, p. 801-818. Disponible sur : <<https://arxiv.org/abs/1802.02611>> (consulté le 6 janvier 2021).
- Chrzastowski T., Cobb D., Davis N.** (1989), « Library collection deterioration : a study at the University of Illinois at Urbana-Champaign », *College and research libraries*, Vol. 50, N° 5, p. 577-584.
- Crowley E., Zisserman A.** (2015), « In Search of art », *Computer Vision - ECCV 2014 Workshops. Lecture Notes in Computer Science*, Vol. 8925, p. 54-70.
- Drott M.** (1969), « Random sampling : a toll for Library research », *College and Research Libraries*, Vol. 30, N° 2, p. 119-125.
- Eden P.** (1998), *A model for assessing preservation needs in libraries*, London, British Library research and innovation center, 110 p.
- Elie P.** (2018), *12/3061, Etude et conservation-restauration des 207 pièces d'archives rassemblées dans les registres BNF, MSS, Pièces originales (PO) 12 (XVI^e-XVII^e siècles; Paris, Bibliothèques nationale de France); Proposition d'une démarche de préparation à la numérisation réalisable sur les 30161 registres PO. Évaluation statistique de l'état du fonds : mise en place d'une méthode par sondage*, mémoire de diplôme de restaurateur, INP département des restaurateurs.
- Feutry D.** (2013), *Plumes de fer et robes de papier, logiques institutionnelles et pratiques politiques du Parlement de Paris au XVIII^e siècle, 1715-1790*, Paris, Fondation Varenne, 674 p.
- Goodfellow I., Bengio Y., Courville A.** (2016), *Deep Learning*, Cambridge, MIT press, 800 p.
- Grün A.** (1863), « Notice sur les archives du Parlement de Paris », dans *Actes du Parlement de Paris*, Paris, t. 1, p. I-CCXC.
- He K., Gkioxari G., Dollár P., Girshick R.** (2017), « Mask R-CNN », dans ICCV (ed), *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, p. 2961-2969. Disponible sur : <<https://ieeexplore.ieee.org/document/8237584>> (consulté le 6 janvier 2021).
- Hildesheimer F., Morgat-Bonnet M.** (2011) *État méthodique des archives du Parlement de Paris*, Paris, Archives nationales, 242 p.
- Kwon D., Yu J.** (2019) « Automatic damage detection of stone cultural property based on deep learning algorithm », *The International Archives of the photogrammetry, remote sensing and spatial information sciences*, Vol. XLII-2/W15, p. 639-643. Disponible sur : <<https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-2-W15/639/2019/isprs-archives-XLII-2-W15-639-2019.pdf>> (consulté le 2 mars 2021)
- Jordan M., Ringot B.** (2020) « Le projet VERSPERA », dans *Les archives contemporaines (éd.) Les enjeux du numérique en sciences sociales et humaines*, p. 145-158. Disponible sur : <<https://hal.archives-ouvertes.fr/hal-03194808>> (consulté le 5 juillet 2021).
- Kim J., Jun J., Hong M., Shim H., Ahn J.** (2019) « Classification of oil painting using machine learning with visualized depth information », dans *The International Archives of the photogrammetry, remote sensing and spatial information sciences*, Vol. XLII-2/W15, p. 617-623. Disponible sur : <<https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-2-W15/617/2019>> (consulté le 7 janvier 2021).
- Kwon D., Yu J.** (2019), « Automatic damage detection of stone cultural property based on deep learning algorithm », dans *The International Archives of the photogrammetry, remote sensing and spatial information sciences*, Vol. XLII-2/W15, p. 639-643. Disponible sur :

<<http://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-2-W15/639/2019/isprs-archives-XLII-2-W15-639-2019.pdf>> (consulté le 7 janvier 2021).

Li J., Yao L., Hendricks E., Wang Z. (2011), « Rhythmic brushstrokes distinguish Van Gogh from his contemporaries : Findings via automated brushstroke extraction », *IEEE Transactions on pattern analysis and machine intelligence*, Vol. 34, N° 6, p. 1159-1176.

Long J., Shelhamer E., Darrell T. (2015), « Fully convolutional networks for semantic segmentation », dans IEEE Computer Society (ed.), *Proceedings of the IEEE conference on computer vision and pattern recognition*, Santiago, Chile, IEEE Computer Society, p. 3431-3440.

Maugis D. (1913-1916), *Histoire du Parlement de Paris. De l'avènement des rois Valois à la mort d'Henri IV*, Paris, Picard, t. 1, t. 2, t. 3.

Muehlberger G., et al. (54 auteurs) (2019), « Transforming scholarship in archives through handwritten text recognition : Transkrius as a case study », *Journal of Documentation*, Vol. 75, N° 5, p. 954-976.

Ostertag C., Beurton-Aimar M. (2020), « Matching ostraca fragments using a siamese neural network », *Pattern recognition letters*, Vol. 131, p. 336-340.

Peyré G. (2021), *Les mathématiques des réseaux de neurones*, CNRS et DMA, PSL, École Normale supérieure. Disponible sur : <<https://mathematical-tours.github.io/book-basics-sources/neural-networks/NeuralNetworksFR.pdf>> (consulté le 5 juillet 2021).

Picard D., Gosselin P. (2015), « Évaluation de descripteurs visuels pour l'annotation automatique d'images

patrimoniales », dans GRETSI (ed.), *GRETSI 2015*, Sep. 2015, Lyon. Disponible sur : <<https://hal.archives-ouvertes.fr/hal-01277486>> (consulté le 8 janvier 2021).

Picard D., Gosselin P., Gaspard M. (2015), « Challenges in content-based image indexing of cultural heritage collections », *IEEE Signal processing magazine*, Vol. 32, N° 4, p. 95-102. Disponible sur : <<https://hal.archives-ouvertes.fr/hal-01164409>> (consulté le 8 janvier 2021).

Seguin B. (2018), *Making large art history photo archives searchable*, thèse EPFL, laboratoire d'Humanités digitales, Lausanne, École polytechnique fédérale de Lausanne. Disponible sur : <<https://infoscience.epfl.ch/record/261212>> (consulté le 20 février 2021).

Walker A., Foster J. (2006), *Knowing the need, A report on the emerging picture of preservation need in libraries and archives in the UK*, London, British Library, National Preservation Office.

Walker A. (2009), « Statistical tools for the evaluation of preservation need », dans IFLA (ed.), *World library and information congress : 75th IFLA general conference and council, 23-27 August 2009, Milan, Italy*. Disponible sur : <<https://www.ifla.org/past-wlic/2009/92-walker-en.pdf>> (consulté le 5 janvier 2021).

Walker G., Greenfield J., Fox J., Simonoff J. (1985), « The Yale Survey : a large-scale study of book deterioration in the Yale university library », *College and research libraries*, Vol. 46, N° 2, p. 111-132.