FUTURS FANTASTIQUES

Vendredi 10 décembre 2021

Grand Auditorium

Yaniv Benhamou & Sarah Kenderdine ,AI & GLAM collections: legal & ethical challenges to sharing GLAMs collections

Yaniv :

Bien bonjour à toutes et tous, merci à la Bibliothèque nationale de France et à AI4LAM pour l'invitation à parler des enjeux juridiques et éthiques au partage des collections, en particulier sous l'angle de l'utilisation de l'intelligence artificielle avec les collections GLAM.

Intervention que je donnerai, comme ça a été indiqué, en binôme : tout à l'heure, d'ici 25 minutes ma collègue professeure Sarah Kenderdine, à distance, interviendra depuis Lausanne pour donner un aperçu des questions technologiques, muséologiques. Pour ma part je donnerai un aperçu des enjeux juridiques éthiques vraiment sous un aspect très légal et éthique. Grand plaisir pour moi d'être ici avec vous, en particulier en présentiel, aussi sachant que je travaille en tant que professeur de droit du numérique beaucoup sur les questions de droits d'auteur, protection des données et musées et puis, par extension, de plus en plus avec les questions GLAM. Ce qui m'a amené à être expert international pour l'OMPI à Genève (l'Organisation Mondiale de Propriété Intellectuelle) pour ces questions, et membre des affaires juridiques à ICOM international donc je suis peut-être le M de cette journée si j'ose dire avec d'autres collègues de ICOM s'ils sont aussi dans la salle ou en distanciel.

Pour ma part donc j'ai 20/25 minutes pour donner une cartographie des enjeux juridiques et éthiques. Avant cela, je veux juste rappeler l'importance de l'intelligence artificielle et GLAM dans la valorisation du patrimoine culturel – lien car elles offrent des possibilités intéressantes pour les humanités numériques, vous l'avez vu tous ces jours. Je pense par exemple à la reconnaissance d'images et de textes qui permet de recréer des villes à l'image de Time Venice qui a recréé Venise, de cartographier la circulation des images ou même d'analyser l'émotion des artistes et des visiteurs dans les salles d'exposition.

Les GLAM pour leur part ont aussi un rôle très important à jouer parce qu'ils traitent et valorisent les données de qualité avec leurs collections qui peuvent ainsi alimenter, irriguer les intelligences artificielles comme je disais avec des données de qualité. Avec ces opportunités technologiques, il existe toutefois une complexité juridique qui peut faire obstacle au partage des collections en raison de nombreuses questions dont la licéité du *text and data mining,* dont la question de titularité lorsque plusieurs personnes, voire la machine co-créent une nouvelle œuvre, et puis enfin peut-être des questions de réappropriation du domaine public, de sorte qu'il est essentiel à mon sens de vraiment maîtriser ces notions juridiques et éthiques pour ensuite valoriser et permettre un partage des collections. En 20 minutes maintenant pour les questions de droit, pour ma part, j'ai comme modeste objectif de cartographier quelques défis légaux en particulier dans les trois domaines que sont le droit d'auteur, la protection des données et l'éthique. Je précise que je ne prétends à aucune exhaustivité évidemment en 20 minutes, je vais faire des considérations plutôt générales, j'ai

compris aussi qu'il n'y avait pas beaucoup d'experts en droit ou d'avocats, alors je vais essayer de rester aussi accessible et général que possible. Pour ceux qui souhaitent approfondir beaucoup de questions juridiques je profite de mentionner juste cette référence qui est un *policy paper* que j'ai co-dirigé avec différents experts de droit qui cartographie et adresse les questions juridiques et éthiques avec notamment un code de conduite pour les musées mais par extension pour les GLAM, vous avez ici la table des matières il est en libre accès avec le lien du site internet.

Le premier domaine que j'ai indiqué, le droit d'auteur, est en particulier pertinent et complexe notamment lorsqu'on veut faire du big data culturel, d'abord s'agissant de la licéité d'utiliser des *inputs* (des données entrant) pour arriver avec des données sortantes (*output*) et puis ensuite de l'autre côté avec les données sortantes (*output*) des questions de protégeabilité des résultats. Alors comme premier défi (ça concerne l'*input*, les données entrantes), il s'agit de savoir s'il est licite d'utiliser les collections protégées par le droit d'auteur comme données d'entraînement. Selon une interprétation stricte du droit d'auteur, une telle utilisation simplement pour faire du big data culturel requiert l'autorisation des ayants droit puisque avec l'interprétation stricte droits d'auteur on déclencherait le droit d'auteur peu importe que les données d'entraînement soient reconnaissables ou non dans le résultat. Cette interprétation stricte semble hélas, si j'ose dire, prévaloir dans plusieurs juridictions dont a priori le droit européen qui considère le droit d'auteur comme couvrant, je cite, "toute reproduction qu'elle soit directe ou indirecte, provisoire ou permanente, par quelconque moyen et sous quelque forme que ce soit", donc on se retrouve avec une problématique que le droit d'auteur serait déclenché avec l'utilisation des données d'entraînement. Il existe des échappatoires, des solutions, par exemple s'appuyer sur les exceptions du droit d'auteur dont le *fair use* en droit américain ou l'exception européenne dite de fouille de textes et de données *text and data mining,* exception prévue à l'article 3 de la directive de droits d'auteur comme exception académique qui permet aux institutions culturelles et de recherche de faire de la fouille et de l'extraction de données, et puis à l'article 4 comme une exception générale de *text and data mining*. Comme autre solution en delà des exceptions on peut aussi s'appuyer sur des solutions contractuelles : je crois que ça a été déjà illustré ce matin, par exemple utiliser les images sous forme de *Creative Commons* ou autre licence ouverte standardisée, et puis finalement une autre solution serait de renverser l'interprétation stricte du droit d'auteur pour considérer que le droit d'auteur ne s'applique pas aux données d'entraînement puisque la donnée est finalement méconnaissable dans la plupart des données sortantes des résultats et puis, en plus, que seul le contenu informationnel de la donnée est utilisé, et pas la donnée en tant qu'oeuvre protégée du droit d'auteur. Ça, c'est une interprétation plus souple pour essayer de renverser l'interprétation stricte et là, peut-être que les GLAM ont aussi un rôle à jouer pour favoriser ce genre d'interprétation. Malgré l'existence d'exceptions dans certains pays, de nombreuses questions subsistent encore. D'abord il y a incertitude quant à savoir quelle est l'étendue de l'exception de *text and data mining* en particulier savoir quel est le cercle de bénéficiaires de l'exception. Par exemple, pour l'exception académique, si la BnF ou d'autres GLAM font des partenariats public-privé, la question va se poser à partir de quand l'influence de l'entreprise privée dans le partenariat est tellement importante qu'on sort de cette exception, de même il y a des questions type dans quelle mesure les mesures techniques de verrouillage TPM ou DRM ainsi que les conditions générales de sites web qui seraient en libre accès pourraient venir restreindre voire évincer cette exception. Pour terminer sur l'exception de *text and data mining*, je rappelle que la date de transposition de la directive dans les lois nationales était fixée au 7

juin 2021. A ma connaissance, la plupart des pays européens sont encore en train de faire le travail de transposition, et, pour information et référence, la mission du CSPLA a rendu il y a un peu moins d'un an le récent rapport pour transposer l'exception et puis ce rapport donne de nombreuses réponses et des recommandations dont une loi de transposition avec un décret et puis finalement des chartes de code de conduite que les GLAM pourraient édicter pour régler toutes les questions de modalités, par exemple tarifer la prestation lorsque la BnF met à disposition sa collection. Finalement encore mentionner ici que les incertitudes quant à l'étendue de l'exception sont rendues encore plus complexes en cas d' activités transfrontières parce qu'on le sait l'IA (et ça a été évoqué avec le cosmolocal), l'IA est forcément rapidement globale pour que bénéficient d'autres personnes d'autres institutions à l'étranger et dès qu'on a un effet global la difficulté c'est que le droit reste fragmenté, national, local et donc ça renforce la difficulté. Exemple : si la BnF met à disposition sa collection pour le *text and data mining* à mes institutions culturelles dont je dépends en Suisse, par ailleurs à des institutions en Angleterre et aux Etats-Unis, on pourra voir une pluralité, une démultiplication des droits applicables et puis, par exemple le droit Suisse, nous avons une exception *text and data mining* mais beaucoup plus limitée par exemple elle peut être évincée par des conditions générales et/ou des mesures techniques de protection.

Deuxième défi dans le droit d'auteur ça concerne l'*output* : là, il s'agit de savoir si les résultats générés par le big data culturel méritent protection du droit d'auteur. La protection du résultat je dirais pour synthétiser n'est pas systématique, parce que la plupart des juridictions exigent la double condition d'originalité et d'intervention humaine. Pour la première condition d'originalité cela suppose en droit européen, je cite, « une création intellectuelle propre à son auteur » laquelle sera refusée, je cite, « lorsque la création est dictée uniquement par des contraintes techniques ne laissant pas de place pour la liberté créative » donc dans beaucoup de cas si on admet qu'une création est uniquement dictée, avec l'IA, par des considérations techniques, eh bien on ne peut plus remplir cette première condition, pas de protection et donc dans le domaine public. Et puis la deuxième condition d'intervention humaine cela suppose une contribution créative apportée par des développeurs ou des créateurs humains et donc de distinguer entre des oeuvres uniquement assistées par un ordinateur (*assisted generated works*) qui vont rester considérées comme humaines et donc protégées par le droit d'auteur, par opposition aux créations générées uniquement par ordinateur type *computer generated works* qui seront considérées comme des oeuvres dépourvues de protection et donc dans le domaine public. Malgré une réponse d'apparence simple, en droit de nombreuses questions subsistent encore. Une première question est déjà d'attribuer les les droits d'auteur aux différents intervenants participant à l'IA. On le sait maintenant, il y a un nombre impressionnant d'intervenants dans l'IA que soit le concepteur, designer, développeurs, utilisateurs… eh bien toutes ces personnes pourraient potentiellement avoir des droits d'auteur soit en co-création et donc oeuvre commune soit sous l'angle d'une oeuvre dérivée : cela dépendra du degré de conservation et d'intention des personnes participant à l'IA pour co-créer simultanément une œuvre. Et puis deuxième question, c'est de savoir si la copie numérique d'une collection (par exemple le *digital twin*, un double culturel réalisé grâce à des technologies de pointe) est une simple reproduction servile du sous-jacent tangible même si le sous-jacent est une oeuvre dans le domaine public, ou au contraire si cela crée une nouvelle oeuvre originale mais avec le risque cette fois d'utiliser le droit d'auteur pour se réapproprier les oeuvres tombées dans le domaine public. Ici je vous mentionne juste quelques jurisprudences pour être bref, cette question a été posée à différents tribunaux encore récemment, par exemple aux Etats-Unis

lorsque la bibliothèque numérique Bridgeman Art Library a numérisé des peintures anciennes du domaine public mais ensuite a poursuivi pour violation de droit d'auteur l'entreprise Corel Corporation pour la vente d'images dont elle serait la seule à détenir le fichier numérique dont cette peinture, vous voyez, de 1624 de Frans Hals. Ce qui est intéressant ici c'est que le tribunal de New York a considéré que même si la majorité des photographies (donc des doubles numériques) a un certain degré d'originalité (cette première condition pour bénéficier la protection), Bridgeman (donc la bibliothèque) a de son propre aveu effectué, je cite, « une copie servile du sous-jacent » pour reproduire le plus fidèlement possible le sous-jacent, de sorte, selon tribunal, qu'il n'y a pas de protection de droits d'auteur. Donc copie numérique avec fidélité absolue a priori pas de protection de droit d'auteur. Ça c'est au tribunal américain qui a tranché. Et puis dans l'union européenne je mentionne ici que le législateur européen, pour éviter une incertitude et des réappropriations du domaine public, a adopté dans la directive droits d'auteur un nouvel article 14 qui prévoit que tout acte de reproduction d'une oeuvre du domaine public ne peut pas être soumis au droit d'auteur ni aux droits voisins, comme je disais, afin d'éviter tout réappropriation du domaine public et puis de favoriser la circulation du patrimoine européen.

Deuxième domaine juridique (et là je serai plus bref), c'est la protection des données. Protection des données parce que c'est un domaine qui pose de nombreuses questions également au partage des collections en particulier puisque les images ou les données personnelles d'individus sont de plus en plus utilisées, par exemple sur les réseaux sociaux lorsqu'une bibliothèque ou des GLAM vont utiliser les données de réseaux sociaux également avec des outils de reconnaissance d'image (je mentionnais avant la reconnaissance de Art Emotion, reconnaissance d'émotion face à des images artistiques), on a toujours des données personnelles et donc ça déclenche des questions de protection des données. Premier défi (j'en ai aussi deux ici), premier défi, cela concerne l'archivage numérique. Je pense par exemple à l'archivage du web que différentes institutions culturelles ont comme mission. Eh bien l'archivage du web peut bien évidemment contenir de nombreuses données personnelles, aussi des images donc comme je le disais celles sur les réseaux sociaux et autres sites librement accessibles et cela va mettre alors en tension, d'une part la tâche d'archivage du web incombant à l'institution, et d'autre part la protection des données personnelles et en particulier le droit à l'oubli lorsqu'une personne va réclamer l'effacement de ses données. Cette tension va être finalement arbitrée, s'il y a un litige parce qu'une personne réclame la surpression, par un tribunal entre les différents intérêts soit non seulement l'intérêt d'une personne s'estimant lésée par l'archivage et demandant suppression de ses données au nom du droit à l'oubli,(lequel peut être invoqué au nom du respect de la vie privée dans la charte des droits fondamentaux au sens du nouveau règlement européen RGPD et des lois de protection des données),ça c'est d'une part, mais il faut aussi tenir compte de l'intérêt de l'institution dont l'archivage repose souvent sur des bases légales par exemple dépôt légal et où les lois de protection des données qui permettent avec des motifs justificatifs de quand même traiter des données nonobstant l'objection de l'individu. Cela, hélas, peut donner lieu à des décisions contradictoires malgré des états de faits similaires : à l'image, ici, de la Cour de cassation (j'ai la référence : Cour de cassation belge en 2016) qui a imposé à un journal belge ayant archivé un article, imposé l'anonymisation des données d'une personne s'estimant lésée parce que cela relatait un accident de voiture dans lequel il avait blessé plusieurs et même tué plusieurs personnes 20 ans auparavant. Même état de fait en France avec la Cour de cassation française qui au contraire refuse une telle anonymisation au motif qu'il s'agirait d'une ingérence à l'archivage

et la Cour de cassation française fait cette fois-ci prévaloir l'intérêt à la liberté d'expression et de la presse. Deuxième défi un peu plus prospectif si j'ose dire, cela concerne les données de personnes décédées qui sont de plus en plus utilisées, à l'image des données Facebook : j'entends dire que d'ici cinq ans, Facebook aura plus de personnes décédées que de personnes vivantes, et puis de plus en plus on peut vouloir avoir accès aux données Facebook après le décès de l'utilisateur d'un compte. Par exemple en Allemagne vous avez peut-être entendu parler de cette mère qui a assigné en justice Facebook à Berlin pour accéder aux messages privés de sa fille décédée dans un métro à Berlin pour comprendre le contexte du décès. Je mentionne aussi les images d'hologramme qui permettent de faire revivre certaines personnes, par exemple le site Deep nostalgia, qui permet de faire revivre même des personnes avec des images d'archives en mouvement (type, on peut alimenter Deep nostalgia d'images de nos aïeux puis ça va faire revivre cette personne). Dernier exemple que je mentionnerai brièvement, c'est cette mère coréenne dont vous avez peut-être entendu parler, qui correspond avec un avatar de sa fille décédée : c'est des images choquantes mais impressionnantes et alors voilà, cela pose les questions d'utilisation de données personnelles avec la protection des données. La problématique ici, c'est que la protection des données personnelles en principe expire au décès de la personne et donc pour que les héritiers ou les proches fassent valoir des intérêts soit à la suppression soit dicter dans quelle mesure l'image peut être réutilisée eh bien là il va falloir recourir à d'autres lois et d'autres intérêts/ Typiquement, par exemple en France, il y a maintenant la loi informatique et libertés qui règlemente d'une certaine manière le droit des personnes décédées mais pas dans le RGPD, règlement européen.

Troisième et dernier défi l'éthique qui je le rappelle vient souvent s'ajouter comme un remède – remède palliatif ou remède supplémentaire – aux normes juridiques contraignantes qui sont parfois jugées insuffisantes en matière de droits fondamentaux, je me réfère ici à la définition pour les questions IA et éthique aux différents travaux européens et au conseil de l'Europe vous avez différentes définitions de qu'est-ce qu'une IA éthique. Alors le premier défi (de nouveau deux défis), premier défi dans l'éthique concerne les biais des algorithmes et les discriminations puisqu'on le sait, l'IA peut faire apparaître des biais à toutes les étapes du processus, de la création au déploiement par les utilisateurs et puis ce phénomène de biais algorithmique va être même amplifié lorsqu'il y a une *black box* puisqu'on sait ce qui entre dans l'IA mais on ne sait pas, on ne comprend pas quand ça ressort quel était le fonctionnement interne. Alors afin de s'assurer de l'absence de biais et puis s'assurer d'une construction IA éthique, il paraît à mon sens nécessaire d'avoir à chaque étape, de la conception au déploiement de l'IA, des validations humaines, des contrôles humains et puis chaque fois bien s'assurer de la qualité des données. Je pense que c'est là que les GLAM peuvent jouer un rôle crucial du fait de leur longue expérience en matière d'éthique et de valorisation des données qui pourront ainsi assurer une IA éthique et respectueuse des droits fondamentaux. Deuxième défi : on se pose la question lorsque les GLAM vont utiliser, traiter du patrimoine culturel indigène : là il va y avoir aussi une tension entre d'une part l'intention des GLAM de partager le matériel indigène (qui est pour la plupart d'ailleurs, je le précise, dans le domaine public parce que en vertu des droits d'auteur traditionnels il n'y a pas de protection de patrimoine culturel indigène au sens droit d'auteur strict). Donc d'une part il y a un intérêt des GLAM et du public à accéder à ce patrimoine et puis d'autre part il y a les obligations éthiques de respecter ce matériel et la volonté des communautés indigènes. Par exemple je ne peux pas utiliser certaines données qui vont être sensibles et sacrées du point de vue des communautés indigènes. A l'image ici, tout en bas, vous avez un exemple

que ma collègue va présenter de Kung Fu, de culture du Kung Fu qui a été numérisée dans la gestuelle et puis là elle nous indiquera dans quelle mesure elle était en contact régulier, elle consultait des communautés indigènes pour s'assurer d'une utilisation éthique. Pour terminer sur l'éthique cela signifie que la prudence est de mise pour refléter correctement les conditions associées au partage ouvert du patrimoine culturel indigène, telles que la reconnaissance de la communauté de manière appropriée : lorsqu'on va taguer les données, taguer la collection on va mentionner, reconnaître la communauté de manière appropriée et puis comme je le disais s'assurer préalablement d'une consultation pour tout type d'utilisation.

Pour conclure ma cartographie des trois défis juridiques, je propose deux questions et puis deux débuts de réponse. Première question à mon sens c'est de savoir quelle solution légale pour le partage des collections GLAM sachant que sinon le risque est de perpétuer la tendance actuelle selon laquelle seule une fraction des collections est exposée et ensuite mise à disposition sur le numérique. Par exemple au Smithsonian, mais je crois qu'il y a des personnes ici qui pourront confirmer, j'ai compris que c'était 2% de la collection qui est en libre accès et au British Museum (aussi s'il y a des collègues pour confirmer) j'ai compris que c'était 0.4 pour cent de ces millions d'objets qui sont mis en libre accès. Alors parmi les solutions, je disais première question début de réponse, comme solution il y a les exceptions que j'évoquais en faveur desquelles les GLAM peuvent œuvrer en particulier pour clarifier leurs modalités, par exemple à travers du *policy making* (et à Genève on essaye de le faire à l'organisation mondiale de la propriété intellectuelle pour le droit d'auteur) et puis de plus en plus avec des codes de conduite et des bonnes pratiques. J'ai été interpellé par la notion de standardisation : effectivement est-ce que la standardisation est une bonne idée ou au contraire est-ce qu'il faut pas continuer de réfléchir de manière cosmolocale, question quand même ouverte mais voilà le rôle que les GLAM peuvent jouer y compris à l'échelle internationale afin d'encourager les usages transfrontières. Et puis deuxième solution en plus des exceptions ou de la standardisation, c'est les solutions contractuelles. Pour le contrat j'aimerais juste indiquer qu'il y a effectivement les outils type licence ouverte par exemple *Creative Commons* qui sont des licences standardisées qui jouent un rôle clé au partage des collections. Par exemple le Louvre et le Smithsonian auraient mis un nombre impressionnant d'images sous *Creative Commons* mais je précise qu'il faut quand même utiliser ces contrats à bon escient pour éviter des situations de blocage de projets en particulier l'exemple de la numérisation de Venise semble-t-il avec un consortium européen et les EPF en Suisse a été bloqué parce que contractuellement les partenaires n'avaient pas prévu qui détenait les données et donc le projet, à ma connaissance, est pour l'instant suspendu avec ce risque de réappropriation des données ou même de perte des données donc outils contractuels à maîtriser et anticiper. Et puis deuxième et dernière question et je passerai ensuite la parole à ma collègue c'est quel rôle les GLAM peuvent jouer dans l'écosystème de l'intelligence artificielle, comme je le disais pour garantir une IA fiable et éthique. Comme début de réponse, j'ai indiqué à mon sens que les GLAM jouent un rôle crucial du fait de leur longue expérience en matière d'éthique et de valorisation des données ce qui assurerait, assurera que les IA réutilisent des données fiables et éthiques. Voilà ma cartographie j'aurai si on a cinq minutes après l'occasion de poser des questions à ma collègue et d'échanger avec vous si vous en avez.

I will switch into English because my colleague from Lausanne, Switzerland will now interving in English. Thank you very much Sarah for your intervention in the 5 minutes left, excuse me you have 15 minutes I give you the floor.

Sarah:

Thank you so much Yaniv. I'll just share my screen for everybody and just let me know that it all good from your side. You should be seeing a video right now. So thanks so much Yaniv and thank you to the audience for your forbearance if I continue in English. We decided for the further discussion to make a small intervention at this point with real world GLAM projects.

These projects come from the lab for experimental museology at the EPFL with the title “Computational Museology”. This term – computational museology – is a scaffold that unites machine intelligence with data curation, ontology with visualization, and communities of publics and practitioners with embodied participation through immersive and interactive interfaces, and this is at the heart of the work that we're doing at M+ and the two topics that I'll touch on today: cultural big data and embodied knowledge systems. Each of these domains have a host of copyright, ethical and privacy issues. But once, which one presented will act as a source for the conversation to follow.

Keywords in this section that embraces cultural and archival big data, our access, copyright, metadata, machine intelligence, emergent narrative and serendipitous browsing. Audio-visual archives are the major records of the 20th and 21st century, however the sheer size and temporal nature of audio-visual material present as custodians to access, and meaningful engagement with these vast archives is of primary concern. Jazz luminaries is based on a constellation of jazz greats from EPFL, CNL, UNESCO “Memory of the World” Montreux Jazz archives digitization project. This archive, like so many, has significant copyright restrictions and the only mode of access is on the campus or at a venue of the festival. For those lucky enough to encounter it, the installation cuts, remixes and replays 5400 artists and 13,000 videos from a total archives of 11,000 hours of videos. The neural net like image that you see here is based on the social network of the artist, the clustering is based on the numbers of times, and artist played with another artist. And for those of you who are curious BB King lies at the very centre of this dense network. Visitors lie under the dome and uses spherial interface to navigate this constellation emulating the hemisphere in which the full dome is staged. The search paradigm is akin to tuning a radio, it's search by listening, not by typing and it circumvents a lack of public knowledge into who actually the jazz greats are. So hearing what you like drives the design to unfold in three layers : first, a whole series of samples, then all songs and then a full song. And I'll just show you a bit of the video sequence that comes from here.

So recent advances in biotech engineering and computer science mean that one can now transform digital data into synthetic DNA that can in turn be retrieved to reconstruct the original. Miles Davis *Tutu* from the Montreux Jazz archives and also *Smoke on the water* by Deep Purple were encoded into DNA from the Montreux Jazz archive and deep encoded in the results you can now listen to. So the entire archive, once encoded like this will fit in a grain of sand. This is a medium that’s considered to be stable over 5000 years but of course that’s yet unproven. In such experiments, it’s been suggested that all the data of the world could be stored in a single suitcase, initiating a massive shift in archival and retrieval

practices in museums, libraries and archives. A radical repositioning in the transmission of objects in time and space, and these long trajectory's of data storage have profound legal implications going forward.

This project lays the basis for another large scale research effort funded by the S&F to bring access to 120,000 hours of copyright encrusted video from major museum and archive collections. Using machine learning, visual analytics, visualization and human interaction intended for public consumption in situated settings. And one of the most important things in this work is how GLAM data informs machine learning tools, which are potentially transformative for industry. For example, one of these archives has been hand tagged with 15 fields of emotion, providing the machine learning scientists with unprecedented opportunities to improve algorithms that can detect emotion and tools for automatically tagging film materials with these attributes. Years of investment by the GLAM sector in moving image archives will have very important part to play in the development of new tools and hopefully with a much better ethical underpinning.

The keywords for embodied knowledge systems are, of course, intangible heritage and living traditions, transmission for the knowledge systems, motion ontologies, ethics, privacy, ownership, consent, inclusiveness versus representation and not least heritage at risk. Hong Kong is an extraordinary reservoir of intangible heritage, primarily of Hakka people but with globalization, urbanization and dwindling, numbers of practitioners this living heritage made internationally so famous by Jackie Chan, Gorgon Liu and Hong Kong cinema, is now in danger of becoming lost. This photograph is a typical record from the 70s of a clan meeting in Hong Kong but today we see this scene is quite different. Here we have master Lam in the center of the room with his European novices surrounding him. The Hong Kong Martial Arts living archive examines this intangible heritage and the processes of its digital documentation, reproduction and transmission into museological contexts. This goes to the heart of computational museology with a whole of environment encoding approach. The four day typology is based on extensive motion capture, a host of green screen technologies, high speed video methods and so on. Currently, we've collected 130 sets of empty hand and weapon sequences. And the archive represents 19 styles by 33 elite masters, and it's 53% of the total repertoire of this tradition so far. This is Oscar from the wire frame motion capture data, we can of course start to abstract it to motion analytics. And after 10 years of work, we're still recording and analyzing this material. One aspect of this research by my PhD student, you mean how her motion archive study has created a search and retrieval algorithm for different motions based purely on visual filmic and 3D mocap data as well as describing martial art movements through motion ontologies. We've had 9 exhibitions worldwide, including Hong Kong, China, Australia, Macau and two in Europe. Out of the 20 odd installations, this one is a 6-sided rare projected system that allows visitors to walk around and view the masters from any point of view at one to one scale. Through an interactive application, the various Kung Fu moves are reinterpreted as a series of motion over time analysis that I generated before. And we can talk about corporeal, ask gestural haptic writing when motion capture is a continuous topological model that allows us easier retainment for the effective qualities of movement. And it's performance theorist Diana Taylor sayed embodied and performed acts generate record and transmit knowledge.

And just to conclude, here is Lam Sai Wing, he's the first Kung Fu master to systematically use photography in studio practicein the first half of the 20th century. He comes from the Lam family, and they continue to be at the forefront of technical innovation for the transmission of

teaching practices. Referencing these early studio photographs of Lam Sai Wing and also a book of hand drawings which you see the animated GIF from, we use motion capture of his great great grand-nephew who's also an elite Kung Fu master, to create a very strange archival object. We have in the galleries life, for life scale combined a digital model of Lam Sai Wing with this capture data, it's a kind of strange DNA of his great great grand-nephew and these videos stand life- size in the galleries. They return us to a point that Yaniv made earlier on the use of our materials on people that are deceased, and from here I'd like to hand back to Yaniv.

Thank you.

Yaniv

Thank you very much Sarah.

I didn't know the last video, so pleasure to see it. Greetings from Paris to snowing Lausanne. I think we have still 5 minutes to address a few questions. If I may, I will get back to my concluding remarks or rather questions that I addressed at the end to know and to ask you unless there are other questions in the audience. What solutions would you advise to the sharing of GLAM collections? For instance, you mentioned the Montreux Jazz archives and you said luckily you were to have access to this important database so what solutions would you suggest to access to important collections? And second question, I think maybe the most important one, what role would you see for GLAM in the AI ecosystems for instance, to build an ethical and truthworthy AI? So two questions. Happy if you can answer, unless there are other questions, maybe we can already take one question Sarah before you answer.

We address all of them together.

Questions

Gautier Poupeau:

Bonjour, vous pouvez remettre vos questions peut-être… : c'est à vos questions je vais essayer de répondre - Gautier Poupeau de l'institut national de l'audiovisuel. Sur la première question, j'ai pas la solution puisque l'essentiel de ma collection est sous droits donc ça se pose pas. Sur la deuxième par contre je voudrais juste faire état d'une expérience : en l'occurrence, on est en train de mettre en place un système d'identification de visages à partir de nos collections et à partir de nos référentiels, et je voulais limiter le fait que vous disiez que grâce à nos jolies data qui sont super propres, on allait pouvoir limiter les problèmes de biais. En fait c'est l'inverse puisque nos collections sont déjà biaisées, biaisées par l'actualité elle-même et puis par la manière de gérer cette actualité puisque évidemment nos collections sont produites essentiellement par la manière de voir cette actualité : c'est la société à travers les médias. Et en l'occurrence ce dont on s'est aperçu, c'est que sur les visages racisés notre algo, parce que notre collection est biaisée, est biaisé. Donc il va falloir qu'on fasse un travail particulier dessus, réellement, par contre voilà c'était pour vous dire que non, nos collections ne sont pas naturellement non biaisées : elles le sont, et donc il va

falloir qu'on fasse un travail spécifique de reprise pour pouvoir limiter les problèmes qu'on peut rencontrer avec ces algos.

Yaniv :

Merci beaucoup pour ce commentaire intéressant. Donc juste pour rebondir, je ne disais surtout pas que les collections étaient sans biais aucun : c'est plutôt le rôle que peuvent jouer les GLAm dans l'identification et le nettoyage des biais si tant est que c'est possible. Et vous indiquez que vous avez identifié qu'il existe des biais racisés, ce qui n'est probablement pas le cas de beaucoup d'autres institutions ou entreprises qui travaillent aussi avec des données existantes, or vous, vous avez déjà identifié qu'il y a des biais : c'est déjà une partie du travail. Alors certes maintenant il y a tout le reste qui va venir à faire mais au moins je pense qu'il y a dans les institutions culturelles une expertise pour déjà identifier et ensuite voir comment nuancer ou temporiser ces questions des biais. C'était ça l'idée.

Gautie Poupeau :

Je comprends mieux. En même temps c'est pas très compliqué : quand vous avez visage asiatique qui est toujours donné à la même personne, on finit bien par comprendre qu'il y a un souci, c'est assez rapide de voir mais je pense que les grands acteurs aujourd'hui en sont conscients et je pense que vous avez vu la polémique qu'il y a eu autour de l'éthique chez Google, et le départ de la responsable éthique chez Google il y a quelques semaines montre que cette question est au cœur de l'ensemble des réflexions aujourd'hui, et les nôtres et aussi des grands acteurs.

Yaniv :

Alors que je ne sais pas s'il y a une autre question peut-être dans l'audience puis après on peut voir avec Sarah pour répondre aux deux. Oui.

Aurélia Rostaing :

Ce n'est pas une question juste une remarque : les deux exemples que vous avez pris pour illustrer le droit à l'oubli en Belgique et en France concernent des organes de presse, pas des institutions patrimoniales. Or au titre du code du patrimoine, la BnF, les services d'archives, l'Institut national de l'audiovisuel ont pour mission de collecter des œuvres et des archives et les conserver dans leur intégrité, donc la question en fait que je me pose en entendant c'est si l'ayant droit avait demandé à la Cour de cassation de Belgique d'effacer son nom des serveurs de la collecte du dépôt légal belge ou de ne pas afficher le nom sur les postes de lecture dédiés sur place à la bibliothèque royale par exemple. Selon vous qu'est-ce que la Cour de cassation aurait dû répondre ? C'est un peu une question provocatrice, mais je pense que vous comprenez ce que je veux dire : il y a une distinction entre ce que les journaux peuvent diffuser en ligne et la mission de conservation des institutions patrimoniales.

Yaniv :

Alors je ne m'aventurerai probablement pas pour me prononcer pour du droit français (je suis plutôt expert de droit suisse et droit international), mais ce que je peux dire, c'est

qu'effectivement c'est une autre base légale, un autre fondement sur lequel reposerait l'archivage du web pour les bibliothèques dont la BnF, et là, le tribunal devra arbitrer en deux droits différents et puis c'est toujours la question quelle est l'ingérence admissible au droit à l'oubli de la part de la BnF qui repose sur une base légale (type dépôt légal) pour faire de l'archivage, et puis là l'ingérence c'est comme a fait le tribunal belge c'est de dire on va faire au niveau proportionnalité le moins de dégâts possible pour la victime (enfin, le « requérant ») et puis peut-être que dans ce cas un tribunal français demanderait d'anonymiser spécifiquement le nom de la personne mais quand même de garder l'intégralité, le reste de l'article. Mais de nouveau sans être expert de droit français pour moi c'est difficile : c'est toujours cette question d'arbitrage entre deux droits et puis, comme vous l'avez dit justement, l'archivage du web des bibliothèques repose sur un autre fondement que des organes de presse.

Je repasse à l'anglais. Maybe, we can address the questions to you Sarah, so that you have the floor. Would you comment on these two questions?

Sarah:

The first questions you asked me? So what advice do I have for GLAM workers in relation to these archives? I think there are two aspects to it. Our aspiration to share far outstrips the legal mechanisms that we have. And so if the heart of museological process is really about this sharing, we need to make significant work. But going forward, as we acquire these archives, we have to look more significantly at the modalities in which we might want to share them. And it is for sure that not all archives should be shared. I'm definitely a believer because I work in a range of indigenous collections which are completely secured from public view. It's not for us to share them to the world, and it's only where the custodians of this material give permission that these archives should be shared from museums, and so we have to create databases that allow for this nuance sharing. It's not about the idea you put everything into the commons, which I know is a way of thinking about patrimony, but I'm really a believer that a lot of this sacred knowledge is not sharable. I think that a number of people that have presented today are also out of the research sector. They're not out of the GLAM sector per se, so libraries are better equipped than museums are to actually store and activate these archives. And we have, I think, a crisis coming where all these digital humanities projects have no archival long term storage place. The rights are not well defined and if they should be so lucky to end up in a museum context where there will be a whole host of problems and I think that looking at these issues of the legal aspects is really fundamental to these new kinds of datasets that we're creating. That could be one comment.

And if we take the Hong Kong Kung Fu as an example, it's the Hong Kong government now have created a museum and will sponsor an archive which hosts the archive. The archive is owned by the custodians of these traditions, so the international Goushu Association are the custodians for it. And I think this is a very good model for its transfer into the future. How GLAM can support ethics in AI? It's mainly because we have hugely diversified datasets that the typical machine learning people do not have access to. They're using online available materials and what GLAM collections have in them is a huge range of diversity and I think this is really a very important card for the GLAM sector to play into the technology sector at this time to help diversify the training datasets that are used by the technology sector.

So thank you.