Crowdsourcing versus outsourcing : nouvelles technologies de la contribution pour l'indexation du patrimoine

Vincent Puig
Directeur Exécutif
IRI / Centre Pompidou
vincent.puig@centrepompidou.fr
Colloque « Patrimoine du Magreb à l'ère numérique », Alger, 28-30 avril 2013

L'avènement du Web 2.0, en d'autres termes du Web social ou Web de la contribution, rebat largement les cartes du paysage du patrimoine culturel dans un contexte de post-numérisation. En effet la numérisation généralisée des collections pose selon nous quatre nouveaux enjeux : 1) une nouvelle définition du patrimoine comme relation d'un lecteur à la mémoire collective et ceci de plus en plus fréquemment à travers la conservation des objets de sa propre histoire, 2) la nécessité par conséquent de penser l'indexation non plus seulement centrée objet ou centrée document mais orientée selon les catégories de la relation du lecteur au patrimoine, 3) le développement d'une culture amateur et par conséquent d'un Web plus critique, 4) la capacité pour les institutions culturelles à sortir du modèle couteux et hégémonique de l'*outsourcing* (sous-traitance intégrale de l'indexation de leurs fonds) pour développer des stratégies de *crowdsourcing* ou plus généralement de collaboration avec leurs publics.

I – Une nouvelle définition du patrimoine

La figure émergente de « l'amateur 2.0 » que nous connaissons aujourd'hui hérite d'une longue histoire, depuis l'amateur aristocrate de Académie royale de peinture de Versailles, en passant par la figure bourgeoise du collectionneur à l'époque de la révolution industrielle, et enfin aujourd'hui celle de l'amateur confronté à la société de consommation et aux industries culturelles, le contributeur passionné souvent opposé à tord au professionnel. De fait, les communautés d'amateurs, de part leur dynamisme sur les réseaux sociaux et leur production active de données personnelles et de métadonnées, sont à présent l'objet de convoitise de bien des industries mais sont aussi un immense défi pour l'avènement d'un patrimoine ouvert dans la ligne des programmes *Open science*. A l'Iri nous abordons cette question de la science ouverte sous l'angle plus général des *Digital Studies*[1], au sens où, au delà du mouvement des *Digital Humanities*, il ne s'agit pas premièrement d'équiper le patrimoine avec les outils du numérique, mais bien d'envisager comment ces outils posent de nouvelles questions épistémologiques. Elles se posent doublement au patrimoine dans la mesure où le numérique est une nouvelle « organologie »[2] qui a la particularité

[1] http://digital-studies.org
[2] http://www.arsindustrialis.org/vocabulaire-ars-industrialis/organologie-générale

d'être à présent totalement partagée par les institutions patrimoniales et les particuliers induisant d'une part une forte convergence des patrimoines personnels et collectifs et déplaçant selon nous l'enjeu du concept d'objet patrimonial vers le concept de relation patrimoniale. D'une certaine manière, ce déplacement a été confirmé en 1997 par l'Unesco avec la distinction opérée entre patrimoine matériel (monuments, objets) et patrimoine « immatériel » (pratiques, connaissances, traditions, …)[3]. Notion qui n'a d'ailleurs qu'un lien indirect avec le numérique (artefacts associés…) et entretient même une forte ambiguïté avec les formes nativement numérique du patrimoine (photos, films, installations, sites web, …). Il serait donc judicieux d'aller au delà de cette définition du patrimoine en tant qu'objet pour interroger le patrimoine comme relation et tout d'abord comme relation à la mémoire et à l'histoire.

L'acte de mémoire est ce qui fonde la constitution des patrimoines personnels qu'ils soient matériels (objets physiques ou numériques sachant qu'un objet numérique est aussi physique) ou immatériels : souvenirs, contes, traditions qui passent par des supports hypomnésiques[4] (aides-mémoires, langage, enregistrements, numérisations). Mais ce lien au patrimoine pour passer d'une relation personnelle à une relation collective pose la question de l'histoire. Une collection de photos personnelles va trouver un intérêt collectif, on dit va se « patrimonialiser » si l'histoire personnelle qu'elle retrace converge avec la grande Histoire, l'histoire collective. Dans les termes du philosophe Gilbert Simondon on dirait que le patrimoine est ici le support d'une individuation psychique et collective, un outil de transindividuation[5]. Cette nouvelle conception du patrimoine comme relation à la mémoire et à l'histoire est tout à fait liée aux projets *Open science*, où notamment dans les humanités numériques, la recherche va fondamentalement changer d'un point de vue épistémologique lorsqu'elle est conduite à l'aide d'une organologie adaptée à la contribution. Pour essayer de préciser ce contexte, nous présentons ici un exemple de recherche menée par Hélène Fleckinger, historienne du cinéma à Paris 8 en collaboration avec la Bnf et l'Iri dans le cadre du projet CineCast[6]. Déjà sensibilisée à la question organologique, Hélène Fleckinger interroge les rapports complexes entre les débuts de la vidéo amateur et les mouvements féministes dans les années 70. Son analyse porte d'abord sur un corpus restauré et conservé au département audiovisuel de la Bnf sous la direction d'Alain Carou et dénommé « bobines féministes ». Au cours de la recherche un déplacement épistémologique intéressant se produit lors de la mise en ligne de ces enregistrements dans un dispositif de contribution par annotation vocale conçu par l'Iri. En effet, la large mise à disposition de ces archives pose un problème de dénaturation de la mémoire de ces événements encore largement soutenue par le témoignage vivant des protagonistes et c'est la raison qui nous pousse à proposer précisément aux témoins de l'époque d'annoter ces archives par le biais d'un enregistreur vocal visant à restituer la part sensible qui pouvait disparaître dans un simple dispositif d'annotation textuel. L'organologie premièrement étudiée par la chercheuse (la vidéo amateur des années 70) est ici prolongée par une organologie numérique contributive qui vise à renforcer le caractère testimonial des vidéos et qui oblige, dans l'hypothèse où de nouveaux témoignages seraient déposés sur le dispositif, à devoir tenir le projet de recherche comme impossible à clore, exemplifiant d'une autre manière ce que l'on nomme aujourd'hui une expérience de science ouverte ou de science contributive.

---

[3] Selon Wikipedia, le seul patrimoine immatériel répertorié à l'Unesco pour l'Algérie est l'Ahellil du Gourara (2008)

[4] Voir la distinction anamnèse/hypomnèse sur http://www.arsindustrialis.org/anamnèse

[5] Gilbert Simondon, *L'individuation psychique et collective*, Aubier, 2007, préface de B. Stiegler sur la transindividuation

[6] Vidéo des premiers temps : Collectifs vidéo et expériences militantes (France, 1968-1981), séminaire INHA du 22 octobre 2012

Fig1 : Annotation vocale en ligne sur des films féministes conservés à la Bnf (projet FUI CineCast)

II – Métadonnées et catégories de la relation au patrimoine

Cette nouvelle conception du patrimoine à la fois comme outil de transindividuation et modalité de notre relation à l'Histoire est particulièrement intéressante si on tente d'étudier plus précisément l'organologie numérique qui la sous-tend et en premier lieu les métadonnées. Les métadonnées sont historiquement produites par un travail d'indexation descendant (top-down) qui en général s'appuie sur un thésaurus permettant de catégoriser le domaine, de le hiérarchiser et de définir les mots-clés à utiliser (ou taxinomies). Ce processus peut être soumis à des protocoles comme par exemple à la Bnf ou au Cndp[7]. Il va suivre éventuellement des modèles conceptuels d'indexation tels que les FRBR[8] qui proposent une catégorisation partant de l'œuvre, son expression, sa manifestation et les documents qui lui sont attachés mais aussi des attributs, tels que le titre, l'auteur, etc et enfin des relations sémantiques entre ces attributs (créé par, présenté à, …). L'indexation va par ailleurs spécifier des formats ou des normes d'encodage à privilégier tels que Marc, Unimarc, DublinCore, etc… Parallèlement à ces approches top-down, apparaissent notamment sur les documents mis en ligne des stratégies bottom-up qui laissent le champ totalement libre aux contributeurs pour procéder par « tagging » c'est à dire à l'aide de mots-clés ou d'expressions libres (y compris graphiques) souvent dénommées « folksonomies » par opposition aux taxinomies. Comment concilier ces deux approches ? C'est le premier enjeu de l'indexation contributive. Pour cela trois options. La première consiste à imposer aux contributeurs des procédures très formalisées par exemple recopier les noms visibles sur un manuscrit ou utiliser une taxinomie prédéfinie, c'est l'option choisie par exemple par les archives départementales de l'Ain[9] dans un domaine, la généalogie, où la motivation des contributeurs est bien compréhensible. La seconde consiste à développer des procédures informatiques permettant de relier les folksonomies aux taxinomies de l'archive, en général des listes d'équivalence. La troisième option consiste à utiliser le plus gros réservoir de métadonnées au monde : Wikipedia. C'est cette voie que nous avons explorée à l'Iri pour le portail Histoire des arts du Ministère de la Culture[10] avec l'objectif de faire converger les technologies du Web sémantique avec celles du Web social. En effet, le premier outil développé visait à permettre de rapprocher les mots-clés des notices du portail de tous les termes approchants

---

[7] http://www.cndp.fr/motbis/telechargement/guide_d_indexation.pdf
[8] http://fr.wikipedia.org/wiki/Sp%C3%A9cifications_fonctionnelles_des_notices_bibliographiques
[9] http://www.archives-numerisees.ain.fr/n/l-indexation-comment-faire/n:54, voir aussi http://www.archinoe.net/portail/
[10] http://www.histoiredesarts.culture.fr/ et http://hdalab.iri-research.org/hdalab/

dans Wikipedia et ceci grâce à la base DBpedia en français (Semanticpedia) développée à l'initiative de la Délégation à la Langue Française, de l'Inria et de Wikimedia France. DBpedia représente actuellement l'une des plus grosses bases d'index au niveau mondial, elle est d'accès gratuit et fournit tous les liens sémantiques entre les index, liens produits par les contributeurs eux-mêmes, au cours de leur travail d'édition des notices. Cette base propose également une fonction très puissante de liens vers toutes les langues utilisées sur Wikipedia ce qui permet de produire une traduction rapide des mots-clés des notices. Pour tirer parti de ce « backoffice », nous avons également développé une interface de recherche par facettes qui permet de naviguer dans les notices de manière simultanée par la période historique, la carte géographique, les disciplines artistiques et le nuage de tags (Fig2). La navigation bénéficie par conséquent de toute la puissance des liens sémantiques mais dans ce cas, non point déterminés par une institution mais bien par les contributeurs eux-mêmes.

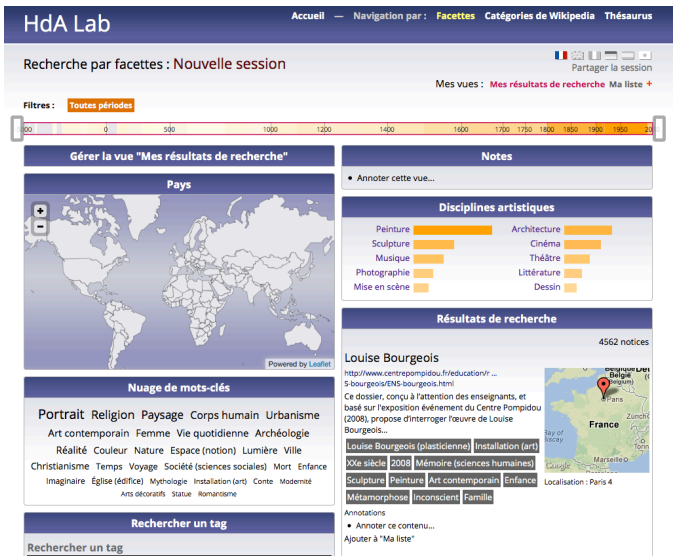


Fig2 : Cinq facettes de recherche, une fenêtre d'annotation et une fenêtre de présentation des résultats de recherche (HDALab)

Comme Google l'a bien compris en agrégeant nos lectures et en nous les revendant, on voit bien que ces outils vont se généraliser. Mais au delà de l'indexation des documents, puis comme on vient de le voir des relations sémantiques entre documents, l'enjeu nous semble à présent de réfléchir à l'indexation de notre relation au document et donc plus largement de notre relation au patrimoine[11]. Dans le cadre de son programme de recherche *Digital Studies*, l'Iri a mis en place en septembre 2012 un séminaire de recherche sur l'indexation et l'annotation[12] avec pour ambition de concevoir et de développer des méthodes et des outils de catégorisation de notre relation aux documents et pourquoi pas à terme de définir des normes et des recommandations. L'approche se veut empirique et procède par ateliers d'annotation de textes successifs. Les premiers ateliers font clairement apparaître la disparité des méthodes d'annotation (individuation) mais aussi la nécessité de faire converger ces méthodes pour pouvoir produire un travail contributif (transindividuation). On constate par exemple l'importance de l'intention visé par l'annotateur qui va conditionner les catégories d'annotation : interrogation, hiérarchisation, contestation, connexion, synthèse, traduction, … et les codes spatiaux utilisés pour les matérialiser : ponctuation, soulignage,

[11] C'est un des enjeux du groupe de travail animé par l'Iri, le Cnam et Cap Digital dans l'Atelier de Recherche Prospective sur le Patrimoine lancé par l'ANR (http://www.univ-paris1.fr/centres-de-recherche/eirest/projets-en-cours/arp-nouveaux-defis-pour-le-patrimoine-culturel/groupe-de-travail-gt-6/)

[12] Séminaire coordonné par Lanval Monrouzeau et Ariane Mayer (thèse en cours UTC-Iri)

surlignage, couleurs. Dans la perspective d'un crowdsourcing de l'indexation qui s'intéresserait au type de relation au patrimoine que nous souhaitons entretenir ou favoriser, ces travaux nous semblent d'une particulière importance, à tout le moins pour normaliser ces actions sinon au niveau international, du moins au niveau d'une institution ou d'une communauté de lecteurs.

III – Le développement d'une culture amateur, vers un Web critique rémunéré ?

Le contexte de science contributive que nous venons de décrire s'inscrit plus largement dans celui de « l'économie de la contribution » et pose notamment comme l'a montré récemment le rapport Collin et Colin[13], la question de la reconnaissance, voir de la rémunération de ce qu'ils nomment abusivement le « travail gratuit » des contributeurs. De fait, les amateurs constituent une dynamique qui « polinise » tous les domaines notamment par la production massive de métadonnées non contrôlées (*folksonomies*). Faire le pont entre cette dynamique ascendante et les taxinomies produites par les institutions culturelles ou académiques est donc un enjeu fondamental qui touche aussi le patrimoine. C'est pourquoi l'Iri a fait récemment évoluer son logiciel d'annotation vidéo *Lignes de temps* vers une plateforme contributive alimentée en « métadonnées sociales ». Le dispositif fonctionne en trois phases :

1) la proposition d'une syntaxe polémique destinée à impliquer les spectateurs d'une émission de télévision ou les participants à une conférence dans un dispositif éditorial qui met en valeur leur esprit critique en direct (fig3),

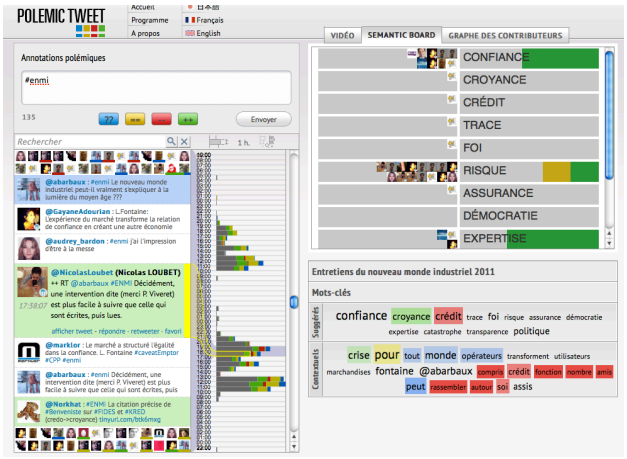


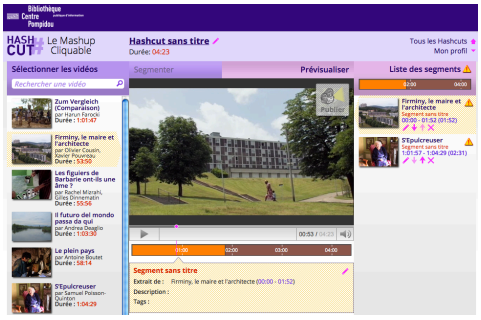
Fig3 : interface de contribution polemictweet : principaux mot-clés, visualisation des questions (tweets bleus) ou des adhésions (tweets verts)

[13] Pierre Collin et Nicolas Colin, *Mission d'expertise sur la fiscalité de l'économie numérique*, janvier 2013

2) la publication d'enregistrements enrichis des contributions polémiques précédemment décrites (fig4),

Fig4 : Publication de la vidéo enrichie des tweets alimentant le moteur de recherche intra-vidéo et proposant la poursuite du débat avec la même syntaxe polémique

et 3) de nouvelles formes d'éditorialisation contributive telles que le *mashup cliquable*, porte d'entrée créative à un fonds d'archives (fig5) et les cartes heuristiques (fig6).



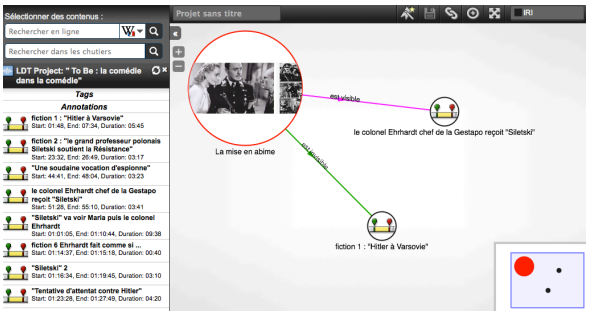


Fig5 : Mashups réalisés par les lecteurs de la BPI Fig6 : Carte heuristique sur le film To be or not to be *de Lubitsch*

Avec de telles interfaces destinées aux amateurs, nous voulons stimuler ce que nous avons appelé en 2003 à l'Ircam des « écoutes signées »[14] puis à l'Iri des « regards signés », c'est à dire des appropriations, ou éditorialisations de contenus au travers d'espaces critiques numériques et qui peuvent être « signées » par les contributeurs. De telles productions dépendent largement de la dynamique amateur qui repose souvent sur un refus de la rémunération. Comment les institutions patrimoniales peuvent s'associer à cette dynamique ? La rémunérer n'est ce pas la dénaturer, voir l'épuiser ? Toutes ces questions, que nous allons aborder dans notre dernière partie, sont précisément l'objet des travaux les plus récents menés par l'Iri notamment en collaboration avec HEC, l'Institut Télécom et le site Allociné et qui vise à concevoir un réseau social cinéphile autour d'un service de VoD où les contributeurs sont « rémunérés » de leur activité par une capacité accrue à offrir des films (fig7). A ce stade du projet[15], et le service n'étant pas encore ouvert, il est trop tôt pour savoir si l'on devra aller au delà de ce mécanisme de « bonus » et proposer un jour une rémunération aux contributeurs déjà considérés par Allo ciné comme des « éditeurs » du site (notamment les membres de ce que AlloCiné appelle le club 300, ses 300 contributeurs les plus actifs associés aussi bien aux avant-premières qu'aux développements expérimentaux tels que le projet CineGift.

---

[14] http://apm.ircam.fr/ecoutes_signees/

[15] Projet Investissement d'Avenir CineGift (2012-2013) associant également le Lip6 et NoDesign

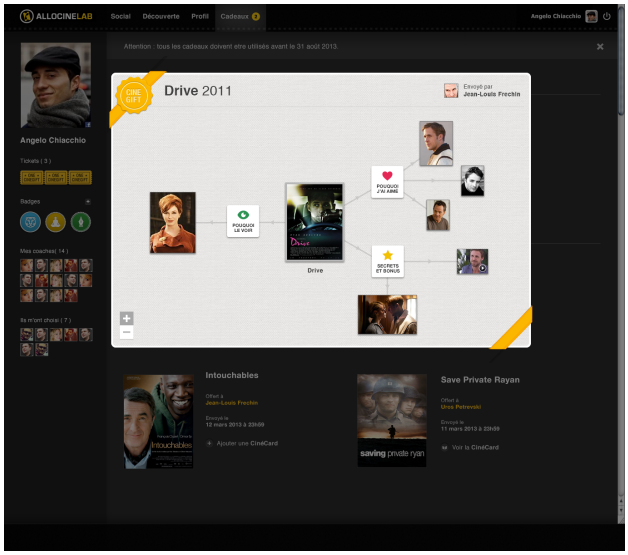


Fig7 : carte mentale associée à un don de film pour le service VoD d'AlloCiné (projet CineGift)

Les institutions culturelles du Spectacle vivant sont, elles aussi, en recherche d'un mode de collaboration avec les amateurs. Dans le projet Spectacle en Ligne(s)[16], nous étudions cette dynamique[17] en tentant d'associer le public amateur d'opéra et de théâtre aux séances de répétition qui, enregistrées, feront l'objet d'une publication annotable retraçant la génétique du spectacle.

IV – Crowdsourcing vs outsourcing, avec l'Open science, l'Open muséum ?

L'indexation numérique du patrimoine s'est généralisée à partir des années 80 parallèlement à la numérisation des fonds. L'ampleur et parfois l'urgence de la tâche a conduit à privilégier jusqu'à présent des objectifs de préservation et d'accès parfois au dépend du développement de nouvelles pratiques. De même, la priorité donnée à la numérisation a conduit à une recherche de productivité obligeant les institutions à recourir à de la sous-traitance massive (outsourcing) soit pour des raisons techniques (disponibilité d'une technologie particulière) soit pour des raisons de coûts. L'outsourcing est ainsi encore aujourd'hui largement pratiqué dans les pays à faible coût de main d'œuvre. Mais la prise de conscience de l'importance de l'indexation en parallèle de la numérisation est en train de modifier largement le paysage de l'outsourcing avec notamment pour les institutions françaises l'enjeu de la connaissance la langue. C'est ce qui explique le fort développement d'entreprises de numérisation et d'indexation notamment à Madagascar[18]. Mais dans un contexte où d'une part la majorité des grands fonds de référence sont à présent numérisés et surtout où l'enjeu se déplace de la numérisation vers l'indexation, l'*outsourcing* traditionnel est à présent confronté à la montée en puissance du *crowsourcing*. Le crowdsourcing peut prendre des formes inédites comme l'utilisation du re*Captcha*[19] pour faire travailler gratuitement les internautes à la reconnaissance de caractères manuscrits issus de la numérisation des livres ou encore la pratique, elle rémunéré, de l'appel à la main d'œuvre humaine par l'intermédiaire de services tels que Amazon mechanical turk[20].

---

[16] Projet ANR Corpus (Iri, Inria, Liris, Ubicast, Cerilac, Festival d'Aix, Théâtre des Célestins)
[17] Enquêtes menées par Joëlle Le Marec au laboratoire CERILAC de Paris 7
[18] L'entreprise DIADEIS est leader sur ce marché
[19] https://fr.wikipedia.org/wiki/ReCAPTCHA
[20] https://www.mturk.com/mturk/

Dans tous les cas, l'enjeu du crowdsourcing nous semble majeur pour les institutions culturelles, d'une part dans la mesure où elles n'ont pas toujours les moyens de recourir à l'outsourcing et d'autre part car, comme nous avons tenté de le démontrer précédemment, l'indexation collaborative est au cœur d'une nouvelle approche du patrimoine, permettant de faire converger les attentes du public et celles des chercheurs dans le contexte des projets de science contributive (ou *Open science*)[21]. C'est particulièrement frappant dans des projets comme *VigieNature* conduit par le Museum National d'Histoire naturelle, où les contributeurs sont étroitement associés par leurs observations de la nature aux progrès de la connaissance scientifique sur la biodiversité[22]. C'est aussi l'ambition de deux nouveaux projets auxquels l'Iri participe et qui visent chacun dans des domaines différents à organiser l'indexation et plus généralement l'enrichissement d'archives au bénéfice de l'institution patrimoniale et/ou d'une communauté d'amateurs. Dans le premier cas[23] il s'agit de mettre en place une dynamique d'enrichissement contributive autour de la plus importante base iconographique nationale, gérée par la Réunion de Musées Nationaux. Compte tenu de la masse critique de photographies (600.000), le projet eGonomy vise à stimuler la curiosité des amateurs par des modes de navigation non-prédictifs et par différentes modalités d'indexation (par analyse d'image, par utilisation des traces ou des parcours utilisateurs et par tagging). L'enjeu pour l'Iri est de monter que le tagging est d'autant plus pertinent et motivé s'il s'attache à un fragment d'image et non plus à la totalité. Les premiers tests ont montré l'intérêt de cette approche par exemple pour développer des parcours en littérature ou en histoire de l'art.

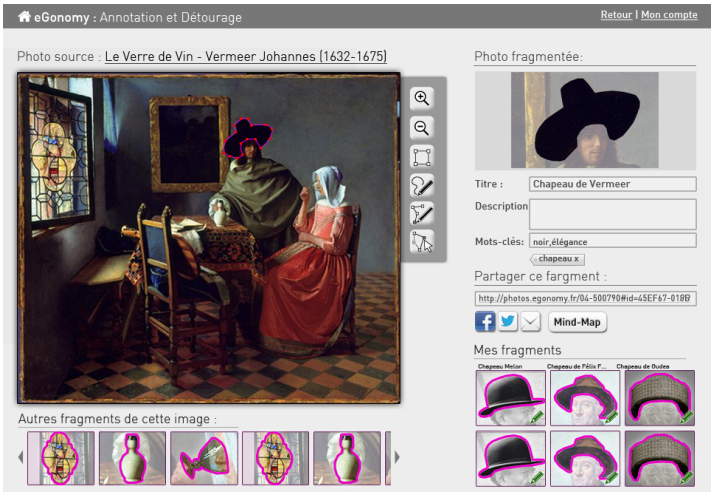


Fig8 : détourage d'un fragment d'image, tagging, partage et mashup (eGonomy)

Plus récemment, nous avons entamé un nouveau projet[24] auquel participent le Musée de la Musique et la Cité de l'Immigration et qui mise aussi sur le crowdsourcing mais principalement à l'occasion de la visite du musée. Les contributions attendues, et qui seront ensuite synchronisées aux archives du musée, portent dans ce cas sur des témoignages enregistrés vocalement, des photos ou des vidéos qui doivent passer par le filtre d'une segmentation et d'une indexation fine. Cette indexation est produite automatiquement grâce à un système très précis de géolocalisation fonctionnant à l'intérieur du bâtiment. Par défaut, on affecte à la contribution de l'utilisateur, les métadonnées correspondant à l'œuvre devant laquelle il se trouve mais il peut aussi modifier ces métadonnées ou

---

[21] E. Rojas et G. Vidal, *Le web 2.0 et les musées des sciences: quel accès à la culture scientifique* ? (http://ceur-ws.org/Vol-398/S2_RojasEtAl.pdf)

[22] http://vigienature.mnhn.fr/

[23] http://www.egonomy.net/

[24] Projet AMMICO, Assistant de visites de Musées Mobile Intelligent et Collaboratif (http://www.iri.centrepompidou.fr/projets/ammico/)

en ajouter. Ce type de dispositif avait pu être testé avec succès en 2008 pour l'exposition Traces du sacré au Centre Pompidou[25] mais à l'époque avec une synchronisation très grossière par salle, faute d'outil de géolocalisation suffisamment précis.

***

Peut on, dès lors que la production de métadonnées contributive est reconnue comme apportant un réel enrichissement aux œuvres, parler de musée ouvert (open museum), de la même manière que l'on reconnaît aujourd'hui l'intérêt des projets de science contributive (open science) ? Le concept est aujourd'hui loin d'être admis en dehors des musées de sciences (nous avons donné l'exemple du Museum National d'Histoire Naturelle). Nous avons soutenu dans cet article que la relation au patrimoine a évolué vers une relation entre histoire personnelle et histoire collective, entre individuation personnelle et individuation collective pour reprendre les concepts de Gilbert Simondon. Cette conviction est depuis longtemps partagée par Elisabeth Caillet depuis son expérience de l'exposition « Naissance », sujet sur lequel la participation du public était particulièrement réussie[26]. C'est donc dans le choix des sujets entre histoire personnelle et mémoire collective que doit venir se tisser l'outil numérique de transindividuation qui fera selon nous du musée, un musée « ouvert », contributif. Ce choix relève encore largement des conservateurs mais peut aussi émerger du public comme l'ont montré les expériences de crowdfunding d'exposition du Brooklyn Museum[27] ou plus récemment l'expérience MuseoMix[28]. Notre hypothèse, illustrée dans cet article, est que le musée ouvert doit être capable de faire communiquer les deux approches (top-down et bottom-up) y compris dans un contexte économique très mouvant où il doit rester vigilant aux modèles économiques émergents issus des « data ». Pourquoi ne pas imaginer bientôt des coopératives de producteurs de données, des cercles d'amateurs équipés pour produire leurs métadonnées et contrôler l'utilisation notamment de leurs données personnelles ? De tels modèles économiques « ascendants »[29] restent largement à inventer au delà des disciplines scientifiques, dans le champ des humanités et de la culture.

---

[25] http://web.iri.centrepompidou.fr/traces/forum/main/com et plus récemment la société Bobler vient d'annoncer la sortie de son dispositif d'annotation vocal géolocalisé (http://www.bobler.com).

[26] Elisabeth Caillet, *Accompagner les publics: L'exemple de l'exposition "Naissances" au Musée de l'Homme*, Ed L'Harmattan, 2007

[27] http://www.brooklynmuseum.org/exhibitions/click/

[28] http://www.museomix.com/

[29] Bernard Stiegler, *Le design de nos existences : à l'époque de l'innovation ascendante*, Mille et une nuits, 2008