Un océan d'images : établir un catalogue raisonné d'estampes à l'ère du numérique

Johanna Daniel

Un océan d'images : établir un catalogue raisonné d'estampes à l'ère du numérique

Johanna Daniel

Introduction

1. À partir de l'exemple concret d'une thèse en cours[1], l'objet est d'illustrer le potentiel offert par l'ouverture des données patrimoniales pour la recherche en histoire de l'art[2] tout en soulignant la complexité de la mise en œuvre de projets individuels sur de gros corpus. De l'acquisition des données à leur analyse et leur partage, en passant par le nettoyage, le traitement et l'enrichissement, il s'agit de donner à voir la chaîne de traitement d'un corpus de recherche et de relever quelques enjeux de la collaboration entre chercheurs et professionnels des institutions patrimoniales dans la dynamique des humanités numériques.

1. Daniel, Johanna. « Les Vues d'optique, une production européenne d'estampes semi-fines (1740-1830) ». Thèse de doctorat, Université Lyon 3. http://www.theses.fr/s220481.
2. Le lecteur pourra se référer au glossaire pour plus d'informations en matière d'Open access à l'entrée correspondante.

« Un océan d'images » : étudier la vue d'optique à l'aide du numérique

2. Comme dans les autres disciplines des SHS, les pratiques des historiens de l'art ont été profondément transformées par le numérique : les photographies haute définition permettent aujourd'hui d'explorer d'infimes détails d'une œuvre d'art là où nos prédécesseurs se contentaient de reproductions imprimées d'inégales qualités ; les catalogues des musées, de plus en plus informatisés, s'interrogent depuis n'importe quel point de la planète ; de nombreuses sources textuelles (archives, imprimés) sont désormais numérisées et peuvent être, elles aussi, consultées à distance.

3. La numérisation des œuvres et leur mise en ligne, couplées au développement d'outils informatiques performants et accessibles, ouvrent, pour les historiens et historiennes de l'art, de nouvelles possibilités : faire émerger des questions de recherches inédites, changer d'échelle en travaillant sur des corpus de taille plus imposante, etc.

4. Mon objet de recherche, la vue d'optique[3], se prête justement bien à ces changements d'échelles. Les vues d'op-

3. Pour une synthèse en français sur la vue d'optique, on consultera le catalogue d'exposition disponible en *open access* : Aressy, Lorraine, Bertrand Caron, Henri De Colbert, Morgane Didier et Hélène Palouzié. 2014. *Le Monde en perspective : vues et récréations d'optique au siècle des Lumières. Les collections montpelliéraines de vues d'optique au château de Flaugergues*. Montpellier, France : DRAC du Languedoc-Roussillon. https://www.biu-montpellier.fr/sites/default/files/2019-11/duo_vues_%20doptiques.pdf.

tique sont des estampes destinées à être visionnées à travers une lentille, qui déforme l'image, donnant au spectateur l'illusion d'une profondeur de l'image (figure 1). La production de ces images a été massive, à la hauteur de leur succès commercial : entre 1740 et 1830, une quarantaine d'éditeurs, à Londres, Paris, Augsbourg et Bassano del Grappa, inonde le marché de plusieurs centaines de milliers de feuilles.

Figure 1. L'usage d'un zograscope pour visionner des vues d'optique

J. F. Cazenave d'après Louis-Léopold Boilly, *L'optique*, gravure au pointillé, vers 1794, Amsterdam, Rijksmuseum, RP-P-2015-26-2079 (CC0) https://www.rijksmuseum.nl/nl/collectie/RP-P-2015-26-2079.

5. L'un de mes axes de recherche[4] consiste à comprendre les stratégies commerciales mises en œuvre par les éditeurs, en étudiant les motifs figurés dans les vues, la relation entre l'image et le texte, les choix linguistiques effectués pour la lettre, etc. Il s'agit notamment de retracer la circulation des motifs, car une même vue a souvent été proposée par plusieurs éditeurs, soit qu'ils se copiaient les uns les autres en fonction de la demande (phénomène de contrefaçon), soit que les éléments d'impression circulaient de main en main, au rythme des mariages, associations, successions et ventes après décès.

6. Pour répondre à ces questions, mon approche se veut sérielle et doit englober, autant que possible, toute la production européenne de la seconde moitié du XVIII^e siècle et des premières décennies du XIX^e siècle. En l'absence de catalogue raisonné des vues d'optique, il faut au préalable constituer un corpus le plus large possible, de plusieurs milliers de pièces.

7. Fort heureusement, et en dépit de leur caractère modeste, les vues d'optique ont été relativement bien conservées, et de nombreuses institutions en possèdent. Elles sont souvent regroupées en portefeuilles, rassemblant généralement entre cent et deux cents vues, parfois jusqu'à mille[5]. Une partie d'entre elles sont cataloguées, numéri-

4. Sur les axes de recherche développés, voir mon carnet de recherche *Isidore & Ganesh* : https://ig.hypotheses.org/.

5. À ce jour, j'ai identifié plus de 13 000 tirages de vues d'optique conservées dans une soixantaine de bibliothèques et musées européens.

sées et publiées en ligne. C'est le cas notamment des vues d'optique du département des estampes de la Bibliothèque nationale de France (635 vues), du Rijksmuseum (1038 vues) ou encore du musée De Lakenhal à Leyde (677 vues).

Automatiser l'acquisition d'un corpus

8. Pour constituer mon corpus de recherche, j'aurais pu procéder de différentes façons. La première option consistait à ne travailler qu'à partir de quelques fonds institutionnels préalablement sélectionnés sur des critères objectifs (importance matérielle de la collection, représentativité de la production, accessibilité des œuvres), au risque de lacunes et de biais. La seconde était de constituer un corpus, en sélectionnant un à un le « meilleur » exemplaire de chaque vue d'optique portée à ma connaissance, c'est-à-dire le plus représentatif et le mieux conservé, comme on sélectionne une édition particulière d'un texte. Tirant parti des possibilités offertes par le numérique, j'ai opté pour une troisième méthode : accumuler, dans une base de données, un maximum de vues d'optique, sans sélection préalable au sein des séries constituées par les institutions. Le but étant, dans un second temps, de traiter ces milliers d'images pour regrouper les exemplaires similaires (les doubles sont nécessairement nombreux) et de distinguer les versions différentes d'un même motif.

9. La première étape consiste à récupérer les métadonnées* de catalogage (transcription des titres et mentions de responsabilité, indexation des éditeurs et sujets représentés, cotes des exemplaires, etc.) et les images numérisées disponibles sur les sites des institutions conservant des vues d'optique. Effectué manuellement, ce travail de copier-coller vers ma base de données personnelle aurait été fastidieux. Heureusement, il existe diverses solutions pour automatiser l'acquisition des données : interrogation via des API* (*Application Programming Interface* – interface de programmation applicative), téléchargement en CSV* ou encore recours au *web scraping**.

10. Certaines institutions proposent, en plus de l'interface classique de consultation des collections, l'accès à une API qui permet d'interroger et récupérer en masse des données. Précurseur, le Rijksmuseum a été le premier musée en Europe à proposer un tel service dès 2012. L'API de la Bibliothèque nationale de France (BNF), quant à elle, est en service depuis l'automne 2017. Dans les deux institutions, l'API permet non seulement d'extraire les métadonnées de catalogage mais également de télécharger, en haute définition, les images numérisées[6].

6. Dans le cas du Rijksmuseum, données et images sont placées sous licence Creative Commons CCo (https://www.rijksmuseum.nl/en/data/terms-of-use, consulté le 30 mars 2020). En ce qui concerne les métadonnées, la BNF a opté pour la licence Etalab 2.0 (http://api.bnf.fr/conditions-generales-dutilisation-du-site-bnf-api-et-jeux-de-donnees, consulté le 30 mars 2020). Les images quant à elles sont gratuitement réutilisables pour un usage non-commercial et, depuis octobre 2019, pour des usages académiques ou scientifiques. Les usages commerciaux sont soumis à licence

11. Pour le chercheur qui veut se confronter à ces API, le chemin est cependant semé d'embûches, car leur usage nécessite quelques préalables techniques. Il faut ainsi formuler sa requête dans une syntaxe particulière (CQL* pour l'API de la BNF), dont la documentation[7] peut, à juste titre, paraître assez aride. Nous sommes ici loin du confort ergonomique offert par les interfaces graphiques de « recherche avancée ».

12. Une fois passée cette première étape et la requête envoyée, l'usager doit aussi être en mesure de traiter la réponse formulée par l'API, le plus souvent un fichier au format JSON* ou XML* : leur manipulation pour extraire les données souhaitées exige la maîtrise d'un langage informatique (par exemple Python*) bien qu'il soit possible de transformer le fichier en tableur avec des outils plus abordables comme OpenRefine, qui dispose d'une interface graphique.

13. Certaines institutions ont compris que l'usage d'une API – qui répond à certains besoins spécifiques – n'était pas à la portée de tous les publics qu'elles visaient, du fait de cette barrière technique. Aussi quelques-unes proposent des exports de leurs données dans un fichier tabulaire au format CSV, plus maniable. La structure des informations y est moins précise que dans un fichier JSON, mais le contenu est immédiatement exploitable sans compétences informatiques particulières (Bardiot 2018).

14. C'est le choix qui a été fait sur le portail *open data* du Ministère de la Culture, où l'internaute peut interroger une extraction de la base Joconde[8]. Une interface graphique permet de formuler la requête (recherche simple et filtres), dont le résultat est présenté sous forme de tableur téléchargeable. Interrogée sur le terme « vue d'optique », la base Joconde renvoie environ 300 notices, correspondant aux collections du MUCEM, du musée des Beaux-Arts de Bernay et du musée municipal de La Roche-sur-Yon.

15. Dans le cas précis de la base Joconde, cependant, l'usager est confronté à un autre problème : la complétude des données. En effet, la plateforme *open data* du ministère de la Culture ne propose pas un accès exhaustif à la base Joconde, mais seulement un extrait. Sont notamment exclus des champs requêtables et téléchargeables les champs « commentaire » et « inscriptions »[9]. L'argument avancé est le droit d'auteur, certaines descriptions rédigées par les catalogueurs pouvant en effet relever de la propriété intellectuelle lorsqu'il s'agit d'un contenu original et non d'une simple transcription ou d'une

spécifique et payante (https://www.bnf.fr/fr/reproduction-des-documents, consulté le 30 mars 2020).

7. La documentation de l'API BNF est accessible à l'adresse : http://api.bnf.fr/api-gallica-de-recherche.

8. La base Joconde est le nom donné au catalogue collectif des collections des musées de France. Y sont décrites 600 000 notices d'artefacts conservés dans les collections publiques. Depuis 2019, cette base est interrogeable via le portail POP. Le jeu de données publié sur le portail *open data* ministériel est accessible à l'adresse : https://data.culture.gouv.fr/explore/dataset/base-joconde-extrait/information/.

9. Notons également que le téléchargement des images n'est pas prévu par l'outil.

description factuelle[10] (cependant, sur les fiches consultées dans la base Joconde, jamais le nom du catalogueur n'apparaît). Dans le cas qui m'intéresse, les informations figurant dans « inscriptions » et « commentaires » sont justement parmi les plus précieuses : c'est là que les catalogueurs ont recopié manuellement la lettre des estampes où apparaissent le nom et l'adresse des éditeurs des vues d'optique, autant d'éléments indispensables à l'histoire de l'estampe.

16. A contrario des exemples jusqu'ici développés, la plupart des sites et bases institutionnels, ne propose aucun outil d'export des résultats, sinon un téléchargement, une à une, des images numérisées (quand le clic droit n'est tout simplement pas bloqué), et éventuellement de la notice individuelle (au format TXT ou PDF). Cela s'explique parfois par un manque de moyens (interfaces vieillissantes, solutions propriétaires peu performantes) ou par une non-appréhension du besoin (seule la consultation une à une des notices a été envisagée).

17. On objectera – à juste titre – que l'usager peut directement s'adresser à l'institution pour obtenir par retour d'e-mail un export de la base de données. Si dans beaucoup de cas la réponse est positive une fois les motivations exposées, il arrive que certaines institutions ne donnent pas suite, faute de moyens humains suffisants pour traiter la demande. Il arrive aussi, mais c'est heureusement rare, que la requête suscite des crispations, l'institution étant réticence à confier à des tiers des données et des fichiers dont l'usage, lui semble-t-il, pourrait lui échapper[11].

18. Dans ces cas, le chercheur n'aura d'autres choix que de se tourner vers la recopie manuelle du catalogue en ligne qu'il consulte, ou, solution nettement plus efficace mais toujours à la limite de la légalité, le recours à un outil de *scraping* de site web, tel que WebScraper[12].

19. Grâce à ce programme, qui étend les fonctionnalités du navigateur, il est possible d'extraire des données d'une ou plusieurs pages. Dans le cas d'un usage sur un catalogue en ligne de collection, il faut d'abord repérer sur une page test les informations à extraire, puis fournir au logiciel la liste des URL à traiter, pour récupérer, à la sortie, un tableur contenant toutes les données souhaitées. Assez abordable techniquement, cette solution demeure néanmoins tributaire de la qualité du code du site à scraper : si celui-ci est mal structuré, le *scraping* sera incomplet voire inexploitable[13].

10. Sur l'ouverture de la base Joconde, voir le wiki Ouvre-Boîte (https://wiki.ouvre-boite.org/index.php?title=Joconde, consulté le 19 avril 2020) et les lettres d'informations Joconde n° 32 (mars 2018, http://www2.culture.gouv.fr/documentation/joconde/fr/apropos/lettre_info_32.pdf) ; et n° 33 (juin 2018, http://www2.culture.gouv.fr/documentation/joconde/fr/apropos/lettre_info_33.pdf).

11. Maria Vlachou a exposé le cas pour les reproductions numériques d'œuvres d'art (Vlachou 2018).

12. Logiciel propriétaire (avec version limitée gratuite), disponible à l'adresse : https://webscraper.io.

13. L'outil s'appuie sur les balises HTML pour repérer les informations à extraire. Pour un résultat satisfaisant, il est indispensable que les pages et les champs à récolter présentent une certaine homogénéité dans leur structure.

20. Au cours de ma première année de thèse, j'ai collecté sur internet près de 5000 vues d'optique provenant d'une dizaine d'institutions patrimoniales européennes[14]. Dans chaque cas, y compris lorsqu'une API était mise à disposition, il a fallu mettre en place une chaîne de traitement particulière s'adaptant aux spécificités de chaque site fournisseur pour acquérir les données et les numérisations. Autant d'étapes très chronophages qu'il convient maintenant de détailler.

Des données exploitables en l'état ? Un nettoyage nécessaire

21. Une fois les données extraites des sites institutionnels, le chercheur pourrait penser être au bout de ses peines : il dispose de données structurées (ici des tableurs ou des fichiers JSON) qu'il n'a plus qu'à intégrer dans sa propre base de données. Cependant, il lui reste encore des manipulations à faire pour disposer de données réellement exploitables. En se plongeant dans les données collectées, il apparaît vite qu'elles sont très hétérogènes, c'est-à-dire que les pratiques de catalogage varient fortement d'une institution à l'autre.

22. Prenons un cas concret, celui de trois tirages d'une même vue d'optique, la vue de la chapelle de Versailles éditée par Georg Balthazar Probst (figure 2), cataloguée par trois institutions différentes : la Bibliothèque nationale de France (figure 3), le musée De Lakenhal à Leyde (figure 4) et la bibliothèque numérique de Valenciennes (figure 5).

Figure 2. Vue d'optique de la chapelle de Versailles, éditée par Probst

Johann Friedrich Leizelt, *Vüe particulière de la Chapelle du Chateau de Versailles, du côté de la Cour*, vue d'optique éditée par Probst, seconde moitié du XVIII[e] siècle, eau-forte coloriée, Paris, Bibliothèque nationale de France, département des estampes et de la photographie, LI-72 (3)-FOL. https://gallica.bnf.fr/ark:/12148/btv1b6949131n.

23. L'estampe porte quatre titres, en français, allemand, italien et latin. La BNF a transcrit les quatre, choisissant le titre français comme titre principal et les trois autres comme titres alternatifs. Sur la bibliothèque numérique

14. Dans le même temps, 1600 notices descriptives m'ont été fournies par 5 institutions sous format tableur ou PDF (collections non disponibles en ligne) et j'ai moi-même catalogué manuellement 1300 vues d'optique conservées dans 10 institutions françaises et italiennes.

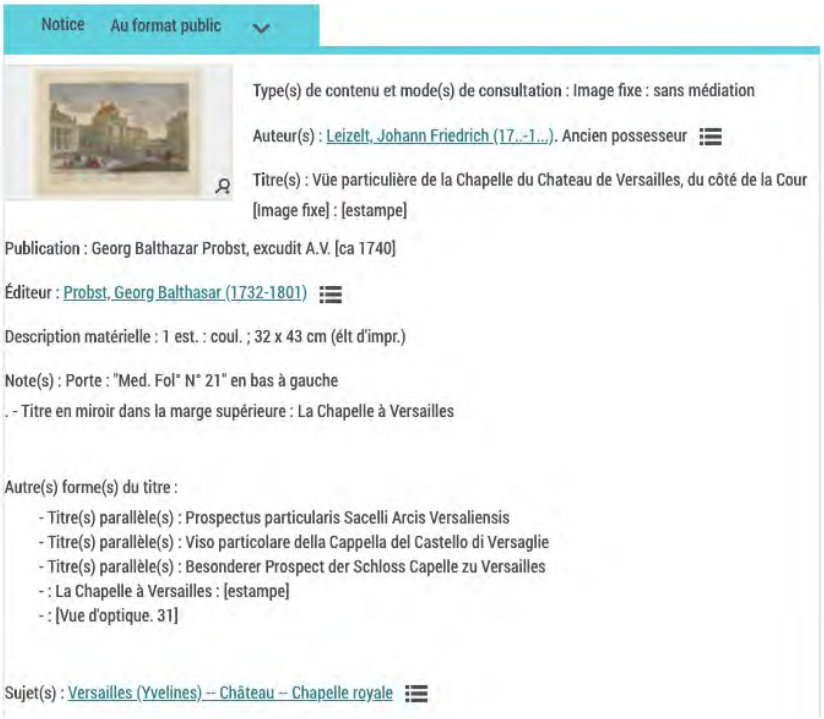

Figure 3. Capture d'écran du catalogue de la Bibliothèque nationale de France

Notice décrivant la vue d'optique *Vüe particulière de la Chapelle du Chateau de Versailles (…)*, éditée par Probst. Catalogue de la BNF : https://catalogue.bnf.fr/ark:/12148/cb41445839k.

Figure 4. Capture d'écran du catalogue des collections du musée De Lakenhal

Notice décrivant la vue d'optique *Vüe particulière de la Chapelle du Chateau de Versailles (…)*, éditée par Probst. Collections en ligne du musée De Lakenhal : https://www.lakenhal.nl/nl/collectie/3121-87.

Figure 5. Capture d'écran de la bibliothèque numérique de Valenciennes

Notice décrivant la vue d'optique *Vüe particulière de la Chapelle du Chateau de Versailles (…)*, éditée par Probst. Bibliothèque numérique de Valenciennes : https://patrimoine-numerique.ville-valenciennes.fr/ark:/29755/B_596066101_G-A18PRO0002.

de Valenciennes, seuls les titres en latin et français ont été retenus. Ils sont indiqués ensemble comme titre principal, séparés par un signe égal « = ». Dans l'interface des collections du musée De Lakenhal, c'est un titre en néerlandais qui apparaît : il s'agit d'une traduction du catalogueur, mais aucun élément ne permet de l'identifier comme un titre forgé. La lettre en quatre langues est bien transcrite, mais de façon non structurée, dans la description. Si l'on s'intéresse maintenant aux mentions de responsabilités (ici, la signature du graveur et l'adresse de l'éditeur), elles ne sont pas transcrites dans le cas de la bibliothèque de Valenciennes. Elles le sont exhaustivement sur le site du musée De Lakenhal, dans le champ « *signatuur* ». Quant à la BNF, le catalogueur a retenu la mention de l'éditeur (champ « publication »), mais pas celle du graveur. Sur cette base, les différents acteurs ont été indexés : la bibliothèque municipale de Valenciennes indique Probst comme auteur et ne fait pas mention du graveur Leizelt. Dans le catalogue de la BNF, les deux sont indiqués dans deux champs différents : Probst comme éditeur et Leizelt comme auteur, ce qui est juste. Enfin, dans le catalogue du musée De Lakenhal, les deux sont indiqués dans un même champ, intitulé « *makers* ». Leur rôle est précisé : graveur pour Leizel et « *Uitgever/drukker* » (éditeur) pour Probst. On notera ici des divergences dans l'écriture du nom du graveur (avec ou sans « t ») et dans l'indexation « Balthasar Friedrich Leizel » pour Lakenhal et « Leizelt, Johann Friedrich » à la BNF.

24. Plusieurs difficultés sont donc à relever ici. Chaque institution utilise des schémas de catalogage différents, avec des champs divergents qui ne se recoupent pas toujours. Ils dépendent de pratiques métiers parfois éloignées (bibliothèques, musées), mais aussi des solutions logicielles employées. Les choix de transcriptions, à la fois dans la nature des éléments à relever (tous les titres, un seul titre ?) et dans l'orthographe (abréviation, modernisation) varient également en fonction des institutions. Quant à l'indexation, elle repose sur des référentiels différents. Ces divergences sont accentuées lorsque l'on compare les données produites dans deux pays différents, qui ne partagent pas la même langue. Enfin, les datations qui apparaissent dans les catalogues (vers 1740 pour la BNF et Valenciennes, deuxième moitié du XVIII[e] pour Lakenhal) n'ont pas été relevées sur les estampes[15], mais sont tirées d'autres sources, non précisées, probablement bibliographiques. L'absence de source pose la question pour le chercheur de la fiabilité de ces informations.

25. Pour rendre exploitables ces données hétérogènes de provenance diverses, le chercheur doit s'atteler à un fastidieux travail de nettoyage, d'uniformisation et enfin, éventuellement, d'enrichissement. Au préalable, il est indispensable de conduire un véritable exercice d'analyse et de critique des métadonnées produites par les institutions, en s'interrogeant, notamment sur :

- la granularité du catalogage (notice sommaire ou détaillée)
- la structuration des métadonnées

15. Rares sont les vues d'optique qui portent une date gravée dans la matrice. Il faut généralement s'appuyer sur les périodes d'activité des différents éditeurs pour estimer une période de publication approximative.

- les choix de transcription (littérale ou modernisée, exhaustive ou partiel)
- les référentiels employés (thésaurus, etc.)
- la fiabilité et la traçabilité de l'information

26. Cela nécessite de comprendre le travail du catalogueur et donc, souvent, de maîtriser soi-même les bases des pratiques métiers. En effet, rares sont les institutions qui documentent clairement et précisément sur leur site les choix de catalogage, choix qui ont d'ailleurs pu évoluer dans le temps[16]. Il faut donc s'atteler à les reconstituer, à les redocumenter par une analyse critique des notices. Ici, l'aide des agents issus de ces institutions est particulièrement précieuse.

27. Une fois l'analyse des métadonnées réalisées, je commence le nettoyage de mes extractions : il s'agit d'abord de désagréger l'information (séparer deux informations qui ont pu être rassemblées, par exemple des titres alternatifs), puis de toiletter les données, en supprimant celles jugées non pertinentes ou non fiables (telles que les dates non sourcées). Il faut ensuite faire concorder les champs avec mon propre schéma de description et uniformiser les données en fonction des règles de transcription et des référentiels d'indexation que j'ai adoptés pour ma base. À ce stade, il est possible d'enrichir la notice, en transcrivant manuellement une information non relevée ou en indexant le sujet. Pour ce faire, il est souvent nécessaire de revenir à l'estampe elle-même, ce que la numérisation des œuvres facilite grandement. Cependant, encore faut-il avoir pu télécharger l'image numérisée et que cette dernière soit de qualité suffisante pour lire le texte gravé. On ne peut ici que déplorer les choix effectués par certaines institutions qui s'opposent à la mise en ligne de fichiers images en haute définition[17] ou à leur récupération par les internautes : elles rendent alors difficile, sinon impossible, le travail de recherche[18].

28. Toute cette étape de nettoyage, d'harmonisation et d'enrichissement peut en partie être automatisée, à condition de maîtriser un langage de programmation adapté. Il est aussi possible d'optimiser le traitement avec des logiciels à interface graphique, tel qu'OpenRefine[19]. C'est le choix que j'ai fait.

16. Les pages professionnelles de la BNF, ainsi que la bibliothèque numérique de l'ENSSIB offrent à tous un accès précieux à la documentation métier.

17. Certaines institutions imposent encore des limitations de quelques centaines de pixels de largeur et de hauteur pour des images mises en ligne sur le Web. Ce genre de pratiques avait une raison d'être technique à l'époque où les écrans d'ordinateur offraient une faible résolution et les fournisseurs d'accès à internet des débits limités. Aujourd'hui, ces restrictions visent surtout à empêcher une diffusion des clichés hors du contrôle de l'institution. À titre d'exemple, une des pages de documentation de Joconde, aujourd'hui hors ligne, indiquait en 2018, « Si la qualité des clichés numériques reversés sur Joconde tend à s'améliorer, bon nombre d'entre eux ne sont pas de haute qualité, notamment pour éviter le piratage éditorial ». Voir la capture de la page au 12 février 2018 sur la Wayback Machine : https://web.archive.org/web/20180212152649/http://www2.culture.gouv.fr/documentation/joconde/fr/apropos/pres_det_joc.htm.

18. À propos de l'ouverture des images numériques des institutions patrimoniales et l'impact de l'*Open Content* sur la recherche, voir (Denoyelle *et al.* 2018 ; Petermann 2018).

19. OpenRefine est un logiciel libre de nettoyage et de mise en forme de données. Il permet notamment d'effectuer des modifications en série sur des tableurs importants, de façon bien plus fine que les logiciels de tableurs traditionnels, comme Excel ou Calc. Pour télécharger OpenRefine : https://openrefine.org/. En français, on trouvera

Traiter, exploiter et analyser

29. Une fois les données acquises, nettoyées, structurées, elles sont enfin prêtes à être versées dans ma base de données. Vient alors le temps du traitement, de l'exploitation et de l'analyse. Une première opération, dans le cas de mon corpus de vues d'optique, a été de repérer combien de sujets différents apparaissent, et pour chacun de ces motifs (que j'appelle « vue type »), combien il en existe de versions différentes (c'est-à-dire de tirages distincts par des éditeurs différents – ici nommés « version »).

30. Reprenons le cas de la chapelle de Versailles (figure 6). Le motif est copié d'une estampe dite « savante » gravée par Jacques Rigaud vers 1740-1750[20] pour sa série des *Maisons royales de France* (figure 7). Jacques-Gabriel Huquier l'a reprise en vue d'optique (figure 8) vers 1760 (Version AA : exemplaires conservés à la BNF, à l'INHA et à la bibliothèque municipale d'Évreux). Probablement au moment de son émigration vers l'Angleterre, à la fin de la décennie, la matrice est passée aux mains de Basset, qui a remplacé l'adresse de son prédécesseur par la sienne (Version AB : exemplaire conservé à la BNF). Autour de 1790, la même plaque est exploitée par Jacques Chereau, après que ce dernier ait à son tour gratté l'adresse de Basset pour indiquer « A Paris, chez Chereau, rue St. Jacques (…) aux 2 colonnes, N° 257 » (Version AC : exemplaires conservés à la BNF

Figure 6. Différentes versions de la vue d'optique de la chapelle du Château de Versailles

Figure 7. Vue de la chapelle de Versailles par Jacques Rigaud

Jacques Rigaud, *Vüe particulière de la Chapelle du Chateau de Versailles, du côté de la Cour*, vers 1740-1750, eau-forte.

Cliché provenant de Wikimedia Commons, publié en CC0

d'excellents supports de formations réalisés par Matthieu Saby sur : https://msaby.gitlab.io/formation-openrefine-BULAC/.

20. La datation est proposée par Bentz et Ringot (2009).

et à Lakenhal). La version de l'éditeur allemand, Probst, évoquée plus haut, a été imprimée à l'aide d'une autre matrice. Si l'on se fie au système de numérotation de l'éditeur, cette estampe peut être placée au début de sa production de vues d'optique, vers 1765-1770. Il en existe enfin une troisième version publiée à Londres par John Tinney, probablement avant 1761, année de la mort de cet éditeur (un exemplaire est conservé au Rijksmuseum).

Figure 8. Comparaison des lettres de trois états de la matrice de la vue d'optique de la chapelle du Château de Versailles

Modification du texte gravé sur la matrice, état de Huquier (version AA), état de Basset (version AB) et état de Chereau (Version AC).

31. Pour chaque version et vue type identifiée, une nouvelle notice descriptive est générée dans la base de données. Exemplaires, versions et vues types sont liés entre elles, ce qui permet de facilement visualiser et comparer tous les tirages d'une version et toutes les versions d'une vue (figure 9).

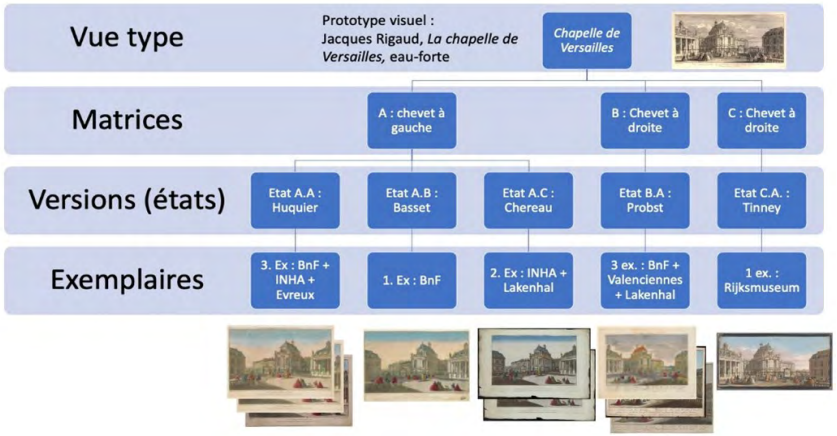

Figure 9. Modélisation des versions de la vue d'optique de la chapelle du Château de Versailles

Crédit : Johanna Daniel

32. Le traitement de chacun des tirages, à l'échelle d'une base de données contenant plusieurs milliers de notices, est complexe. Fort heureusement, le recours aux outils numériques (ici encore OpenRefine) facilite l'automatisation de certains rapprochements, par comparaison des métadonnées.

33. Une seconde phase d'opérations, actuellement en cours, consiste à appliquer des traitements statistiques pour quantifier des phénomènes, visualiser des répartitions, identifier des tendances (quels sont les motifs les plus repris par les éditeurs ? Y a-t-il des spécificités nationales dans le choix des sujets ? etc.). Par recoupement d'informations il est alors possible de produire de nouvelles connaissances sur ces vues d'optique : de reconstituer la

circulation d'un motif entre divers éditeurs, et d'affiner la datation d'un tirage sur la base des dates d'activités des éditeurs et des circulations des éléments d'impression. Ces nouvelles informations viennent enrichir la base de données et, dans un processus itératif, permettent d'affiner les traitements statistiques.

Lecture distante, traitement computationnel et regard outillé

34. Un tel travail sur un corpus d'une taille inhabituelle en histoire de l'art, qui emprunte aux méthodes quantitatives et qui revendique un traitement computationnel pourrait susciter de légitimes questionnements au sein d'une communauté dans laquelle le regard et l'approche visuelle sont au cœur de la pratique disciplinaire. Pourtant, il ne s'agit en aucun cas de « ne plus regarder les œuvres » mais d'outiller le regard, de l'enrichir de nouvelles focales. La lecture distante (*distant reading**), le traitement computationnel, l'approche statistique forment ici un point de départ et non un point d'arrivée. Ils permettent d'effectuer des observations, de relever des phénomènes, qu'il convient ensuite d'examiner, d'interroger, d'analyser, en faisant d'ailleurs appel le plus souvent aux outils « traditionnels » et éprouvés de la discipline, soit « l'œil et l'archive » (Passini 2017).

35. Regard proche et regard distant se nourrissent mutuellement, tout comme les instruments numériques ne sont qu'un enrichissement de la boîte à outils de l'histoire de l'art. S'il est beaucoup question, dans cet exposé, d'automatisation et d'obstacles au traitement computationnel, soulignons que certaines des difficultés rencontrées peuvent s'avérer profitables au travail de recherche. Le temps passé à critiquer les données, à les nettoyer, à les corriger est du temps investi à manipuler le corpus, à s'y confronter. C'est pourquoi ces tâches, certes chronophages et rébarbatives, ne doivent pas être systématiquement déléguées à des « petites mains » : c'est dans ce véritable travail de fourmi que se façonne une familiarité avec le corpus, qu'émergent de nouveaux questionnements et que se forgent des hypothèses.

Publier, partager, offrir à la réutilisation

36. Il serait dommage que ce corpus, patiemment constitué au cours des années de doctorat, dorme par la suite sur un disque dur. Aussi, je souhaite rendre publique ma base de données à l'issue de mon travail de recherche. Dans ce but, j'utilise le CMS* Omeka[21], logiciel *open source* conçu pour créer des bibliothèques numériques sur le Web.

21. Omeka est un *content management system* (CMS), c'est-à-dire un logiciel qui permet de créer et mettre à jour des sites web. Développé depuis 2008 par le Roy Rosenzweig Center for History and New Media, Omeka est *open source*. Il a été spécifiquement conçu pour mettre en ligne des bibliothèques numériques et des petits jeux de données. En France, il est notamment utilisé pour les bibliothèques patrimoniales de l'École nationale des Ponts et Chaussées (https://patrimoine.enpc.fr) et de l'École des Mines (https://patrimoine.mines-paristech.fr), mais aussi par l'Institut national d'histoire de l'art dans le cadre du projet *Digital Muret* (https://digitalmuret.inha.fr). Il est particulièrement adapté pour les corpus d'histoire de l'art, du fait de son interface graphique qui se prête bien à la présentation d'images. Sur le choix d'Omeka comme

37. Il s'agit d'abord de fournir une « annexe numérique » à la thèse qui permette aux examinateurs et aux lecteurs d'explorer le corpus, en profitant de fonctionnalités bien plus riches que celles offertes par un catalogue imprimé, nécessairement figé dans un ordre préétabli. Il sera par exemple possible d'effectuer une recherche multicritère dans le corpus ou encore de zoomer dans une image numérisée pour en discerner les détails. Il est possible d'aller plus loin et de fournir un « corpus outillé », c'est-à-dire de donner accès aux outils qui ont permis certaines visualisations cartographiques ou statistiques, et donc permettre au lecteur de « rejouer » certains calculs, d'en modifier les variables, afin de discuter des résultats obtenus, voire d'interroger le corpus à l'aune de ses propres préoccupations et objets.

38. Au-delà du cadre de ma recherche universitaire, la publication numérique du corpus est aussi pensée comme la mise à disposition d'un instrument de travail sur la vue d'optique, utile à ceux qui s'intéressent à ces artefacts. Fonctionnant comme un catalogue raisonné, il facilitera l'identification d'une vue d'optique, sa datation ou encore l'évaluation de sa rareté. Il s'agit ici, entre autres, de faciliter tout futur travail de catalogage de collections de vues d'optique. Il m'importe aussi – et surtout – de partager les résultats de mes travaux avec les institutions dont j'ai mobilisé les collections, afin qu'elles puissent en bénéficier.

CMS pour un corpus de thèse, voir mon carnet Hypothèses *Isidore & Ganesh* : https://ig.hypotheses.org/.

39. Au cours du traitement des données, de leur analyse et de leur enrichissement, j'ai en effet pu corriger ponctuellement une erreur de catalogage, compléter une transcription partielle, proposer une identification, une datation ou encore une attribution, etc. Ces informations intéressent l'institution, toujours encline à améliorer le catalogage de ses œuvres. C'est d'ailleurs l'un des arguments mis en avant dans les politiques de numérisation et d'ouverture des données patrimoniales : si la mise en ligne doit rendre accessibles à un large public les collections, elle doit aussi stimuler la recherche et favoriser la production de nouvelles connaissances. Plusieurs rapports en faveur de l'Open Data culturel[22] soulignent le potentiel offert par l'ouverture des données, qui permettrait la production de nouvelles connaissances, à leur tour réintégrables dans les systèmes d'information des institutions. *Crowdsourcing** et recherche scientifique se confondent ici dans un idéal de collaboration entre l'institution patrimoniale et ses publics (amateurs et chercheurs).

40. Ce vœu est-il cependant, d'un point de vue pragmatique, véritablement réalisable ? Dans le cas précis de ma recherche, il me semble que plusieurs obstacles existent. Le premier est technique : même si ma base de données disposera à terme d'une API, il faudra, pour les professionnels des musées, s'y adapter de la même manière que je l'ai fait avec celle fournie par leur institution. À leur tour, ils devront aligner mes référentiels sur les leurs. Autant d'opérations fort chronophages pour un gain

22. L'un des plus complets est celui de Domange (2013).

relativement faible à l'échelle de l'institution, sinon celle de la démonstration de faisabilité.

41. Comment, tout de même, porter à l'institution ces enrichissements ? Deux solutions peuvent être envisagées : la première est d'extraire de ma base, au format tableur, le jeu de données enrichies correspondant à chaque institution. Certes, les données ne seront, en l'état, pas réintégrables au sein du système d'information de la bibliothèque ou du musée, mais le fichier, intégré à leur documentation, pourra trouver des usages ponctuels. L'autre voie, plus ambitieuse, est celle du versement de ma base de données sur Wikidata : intégrées à l'écosystème global du Web sémantique*, les données produites dans le cadre de ma thèse seront facilement interrogeables et réintégrables selon des standards largement partagés par la communauté[23].

Comment mieux collaborer ?

42. L'objet de ce texte était d'illustrer, par un exemple concret, le potentiel offert à la fois par la mise en ligne et l'ouverture des données patrimoniales et par l'usage des technologies numériques dans la recherche en histoire de l'art. Le chemin est encore semé d'obstacles, notamment techniques, et je suis persuadée que c'est par la collaboration que nous les dépasserons. Aussi, j'aimerais, en conclusion de cet article, formuler trois vœux.

43. Le premier est que nous menions, collectivement et plus largement, une réflexion sur les usages possibles et autorisés des collections numérisées et des données mises en ligne. Encore trop souvent, seule la consultation « rapprochée » des documents, c'est-à-dire la lecture l'une après l'autre de chaque notice est envisagée. La lecture distante, le traitement computationnel sont en général rendus impossibles par l'absence de fonctionnalités permettant l'interrogation et le téléchargement en masse des métadonnées et images numérisées. Si l'implémentation de tels outils se révèle complexe et onéreuse, il est possible de mettre en place une solution intermédiaire : fournir des extractions CSV, indiquer clairement aux usagers que l'institution peut envoyer les données à ceux qui en font la demande. Encore trop souvent, y compris lorsque les solutions techniques existent, les démarches de lecture distante sont perçues avec méfiance par les agents (peur de la perte de contrôle des données), voire sont considérées comme non légitimes.

44. Pour faciliter l'exploitation des données produites par l'institution, il faut ensuite veiller à documenter les choix, les pratiques métiers, afin de donner à l'usager les outils pour évaluer la qualité des informations qu'il manipule. Cela passe notamment par la formation et le décloisonnement des métiers. Il est à mon sens essentiel de former les étudiants en histoire et histoire de l'art non seulement à l'usage des bibliothèques numériques et col-

23. Sur le Web sémantique et ses usages dans la recherche et les bibliothèques, voir les travaux de Gautier Poupeau, et notamment son blog *Les Petites Cases* : http://www.lespetitescases.net.

lections en ligne – ce qui est déjà le cas – mais aussi de leur donner à voir toute la chaîne de production, qui, de l'étagère des réserves à la publication sur Internet, leur permet de consulter en ligne les documents et données nécessaires à leurs recherches. Décloisonner, « donner à voir » : c'est ainsi que nous pouvons avoir des usagers véritablement informés et critiques face aux outils et données qu'ils mobilisent.

45. Enfin, continuer à favoriser l'interdisciplinarité, qui est le terreau fertile des humanités numériques. Si les réutilisations, les enrichissements de données via les API et l'exposition sur le Web de données sont souhaités par de nombreux acteurs, les réalisations concrètes de tels échanges sont encore rares. C'est pourquoi il est indispensable de les soutenir par des collaborations intéressées, afin de multiplier les preuves de concept qui ne pourront qu'insuffler une dynamique générale.

Humanités numériques et science ouverte
Les collectifs

Clarisse Bardiot, Esther Dehoux, Émilien Ruiz (dir.)

La fabrique numérique des corpus en sciences humaines et sociales

Centrales pour toutes les disciplines relevant des arts, des lettres et des sciences humaines et sociales, les questions relatives à l'identification, la sélection, le classement, les modalités d'exploitation et de diffusion des matériaux nécessaires à la production de connaissances ne sont pas nées avec l'ère dite « numérique ». En histoire, pour prendre l'exemple qui nous est le plus familier, le rapport à la documentation fut ainsi d'emblée au cœur des réflexions méthodologiques qui ont accompagné la construction des savoirs historiques en discipline et l'émergence du métier d'historien. Dès 1898, dans leur *Introduction aux études historiques*, Charles-Victor Langlois et Charles Seignobos formalisent les opérations qui composent ——

Septentrion
PRESSES UNIVERSITAIRES

Humanités numériques et science ouverte

Les collectifs

La collection Humanités numériques et science ouverte (HNSO), co-dirigée par Clarisse Bardiot et Émilien Ruiz, est financée par le Fonds national pour la science ouverte et portée par la Maison Européenne des Sciences de l'Homme et de la Société (MESHS) et les Presses universitaires du Septentrion (PUS).

Elle a pour objectif de publier en *open access* des monographies et des ouvrages collectifs ainsi que les données associées. Contribuant ainsi à l'ouverture et à la diffusion des données, la collection se veut aussi un terrain d'expérimentation et de réflexion en pratique sur ce que la science ouverte fait aux SHS.
Défendant une conception pluraliste des humanités numériques, cette collection s'adresse aux spécialistes des diverses disciplines des sciences humaines et sociales qui inscrivent leurs travaux dans une démarche empirique et accordent une attention particulière à la constitution, la structuration, l'exploitation et à la visualisation de leurs données ; sans exclusive concernant les types de sources, les méthodes employées ou les tailles de corpus mobilisés.

Cet ouvrage a été financé par le Fonds national pour la science ouverte.

ISBN : 978-2-7574-3610-3
ISSN : en cours

Ouvrage composé par
Jonas Mazot, Chloé Gaillard & Sarah Bouchez

Ouvrage réalisé avec
La chaîne d'édition XML-TEI Métopes
Méthodes et outils pour l'édition structurée

Dépôt légal
décembre 2022

2 108e volume édité par
les Presses universitaires du Septentrion
Villeneuve d'Ascq – France