



Recherche / Archives :  
numériser les images, et après ?

18 / 19 / 20 novembre 09

Research/Archives:  
Digitizing images, then what?

## INDEXATION DE MASSE ET NOUVEAUX OUTILS DE REPRESENTATION

### La recherche en représentation de contenus iconiques

**Jean-Marc OGIER,**  
*Université de La Rochelle.*

Bonjour. Je remercie tout d'abord l'équipe d'Archimages de m'avoir invité. Je me sens tout petit à côté de ce qui vient d'être présenté. Les universitaires sont microscopiques à côté des dimensions des équipes de recherche, et de la masse du consortium. C'est néanmoins une excellente entrée en matière par rapport à ce que je vais vous présenter. Vous allez retrouver un peu le même concept présenté par mes collègues de l'INA, mais appliqué à du document papier qui a été dématérialisé.

Je vais vous parler principalement d'un projet nommé Navidomass financé par l'Agence Nationale pour la Recherche. Navidomass signifie Navigation Into Document Masses. C'est un projet qui regroupe différents partenaires universitaires en France, y compris les institutionnels de la recherche en informatique, comme l'INRIA Lorraine.

Vous voyez que le sujet aborde en fait deux projets : Navidomass et Madame. Navidomass fait suite à un projet appelé Madame, comme Masse de Données Appliquées à la Numérisation du Patrimoine, traitant finalement du même sujet. L'idée générale est que de nombreuses institutions (musées, bibliothèques) ont engagé des processus de numérisation de leurs collections (les ouvrages, les livres). La numérisation est d'actualité partout dans le monde en ce moment. La problématique qui se pose pour ces organisations est qu'elles numérisent, et que cela génère des images. Il faut pouvoir se déplacer, naviguer et rechercher de l'information dans ces images. Il faut savoir que nous n'avons pas toujours tous les outils adaptés pour interpréter précisément et reconnaître le contenu de ces images. L'objet de ces projets Navidomass et Madame est donc d'apporter des services permettant de rechercher de l'information, et de naviguer dans ces images.

Vous êtes tous des spécialistes de la numérisation. Vous connaissez les intérêts de la numérisation bien mieux que moi. Il s'agit d'une protection des originaux et d'une consultation de plusieurs personnes à distance, et de la possibilité d'un accès simultané à distance. Je vous montre un petit calendrier où il apparaît que notre aventure a démarré en 2003, dans le cadre du projet Madame. Ce dernier se nommait à l'époque les ACI (Actions Concertées Incitatives), financées par le ministère de la Recherche. Les ACI sont devenues des projets ANR depuis la création de l'Agence Nationale pour la Recherche. Plusieurs laboratoires en France travaillent depuis 2003 sur ce sujet commun, en visant à développer des services de navigation dans les documents.

Il est utile de localiser ces laboratoires pour avoir une vision des intervenants sur ce sujet. Nous n'avons pas les mêmes réseaux que les collègues qui ont fait leur présentation tout à l'heure. Je

pense qu'il serait intéressant de les connecter. Vous retrouvez sur la carte de France les différents laboratoires. Je viens personnellement d'en bas à gauche : laboratoire L3I à l'Université de La Rochelle, qui est porteur du projet ANR, dont je suis en train de parler, Navidomass. Vous avez aussi différents laboratoires d'informatique en général : l'IRISA à Rennes ; le LORIA à Nancy ; le CRIM5 à Paris V ; l'ITIS à Rouen ; le LI (Laboratoire d'Informatique) à Tours et le LIRIS à Lyon...

Tous ces collègues universitaires travaillent ensemble pour développer ces techniques. L'origine de toutes ces personnes est le traitement automatique de l'écrit. Nous travaillons au départ sur l'analyse de documents avant d'arriver à ces problématiques de numérisation du patrimoine, qui n'étaient pas des documents du patrimoine historique et culturel, mais d'autres types, à commencer par la reconnaissance automatique des chèques. Dans les années 90, un certain nombre de banques ont voulu commencer à dématérialiser automatiquement leurs chèques. La reconnaissance automatique des adresses sur les enveloppes postales a également été l'objet de réflexions. Nous développons des outils de reconnaissance de formes et d'images sur ces dispositifs. Il ne s'agit pas seulement de documents écrits, mais aussi de documents graphiques, tel ce bout de plan de cadastre se trouvant sur le transparent. L'idée était précisément de contribuer à la dématérialisation et à la vectorisation automatique des plans de cadastre. Nous avons d'ailleurs un autre projet ANR que je vais citer à la fin. Il consiste à travailler sur la vectorisation automatique des plans de cadastre de l'Atlas Vasserot (XIX<sup>e</sup> siècle), en partenariat avec le laboratoire LAMOP de la Sorbonne. Nous pourrions en dire quelques mots après.

Les gens œuvrant sur l'écrit et le document travaillent sur tous ces documents plus ou moins structurés, c'est-à-dire respectant une certaine organisation spatiale telle une page de magazine comportant des paragraphes. Un chèque a également toujours la même structure. Une image a beaucoup moins de structure. Comme l'expliquaient les collègues précédemment. Ils travaillent également sur les nouveaux documents. Nous avons aussi des projets européens qui visent à essayer de développer de nouveaux services sur la reconnaissance de documents qualifiés de « online », soit de l'écriture en ligne. Nous venons de cette communauté. Il s'agit juste de donner les sources.

Les différentes étapes du traitement de l'écrit sont en général à croiser avec ce que les collègues ont présenté tout à l'heure. Il y a tout d'abord une acquisition de l'image. Nous allons ensuite améliorer la qualité de l'image pour les traitements ultérieurs suivant le contexte : si l'image est fortement dégradée, si elle a vieilli, ou si elle contient des problèmes de capture. Nous allons essayer ensuite d'extraire un peu d'information physique sur cette image. Si nous nous positionnons sur les magazines, il s'agit par exemple des paragraphes. Dans le cas de documents anciens, ce seront les zones graphiques et les zones textuelles. Tous ces traitements relèvent de l'imagerie numérique. Nous sommes donc au niveau du traitement de l'image.

La partie basse du transparent concerne des problématiques de décision et de reconnaissance des formes. Toute l'information du bas niveau est injectée dans des processus d'un niveau plus élevé, qui vont commencer à interpréter le contenu. Ils vont essayer ensuite de contribuer à son interprétation pour faciliter sa reconnaissance ou son indexation un peu plus tard. Cet étage de décision et de reconnaissance de formes alimente lui un étage encore supérieur relevant plutôt de l'intelligence artificielle. Nous allons intégrer ici des connaissances du contexte. Comme les collègues le disaient tout à l'heure, quand nous naviguons dans des corpus, il arrive un moment où il faut intégrer de la connaissance liée à notre recherche. C'est à ce moment-là que ce type de connaissances est introduit dans nos chaînes, pour arriver finalement à une interprétation adaptée aux souhaits de l'utilisateur.

Vous pouvez donc constater que ce domaine est pluridisciplinaire. Nous avons des techniques de traitement d'images, des techniques de reconnaissance de formes et des techniques d'intelligence artificielle. Nous sommes surtout obligés de discuter avec les utilisateurs pour concevoir des systèmes qui leur sont adaptés.

Les ambitions du traitement automatique de l'écrit sont exactement les mêmes problématiques que celles évoquées par vos collègues, comme combler les fossés sémantiques. Dans le cadre du projet de Navidomass qui nous concerne, notre objectif est de donner un accès à des éléments de contenu

pertinents par rapport à des requêtes qui peuvent être soit des requêtes textuelles, soit des requêtes de graphique, soit des requêtes de structure. Nous souhaitons aussi faire rupture avec les OCR (système de reconnaissance de caractères) traditionnels. En effet, les OCR ne fonctionnent pas dans certains cas, surtout en présence de documents très anciens. Nous sommes alors amenés à utiliser des processus de reconnaissance de formes qui ne sont pas des OCR pour naviguer dans les documents.

Si j'essaye de faire une petite synthèse de tout cela, les problématiques de Navidomass sont très nombreuses et variées. Un projet particulier correspond à chaque institution qui numérise. Vous allez voir différents objets, différents objectifs et différents processus. Les usages sont très diversifiés. Nous coopérons avec des historiens du Centre d'Études Supérieures de la Renaissance de Tours qui sont en train de numériser leur patrimoine documentaire. Ils ont des objectifs de recherche et de navigation très particuliers dans leur corpus documentaire. Nous sommes amenés dans ce contexte à développer des choses un peu spécifiques, en introduisant des connaissances liées au domaine dans lequel nous nous trouvons. Nous développons donc des outils pour l'indexation et la navigation dans les masses d'images.

Je vous propose quelques exemples pour illustrer mes propos. Nous voyons ici le Vésale, une encyclopédie médicale du XVI<sup>e</sup> siècle. Une fois numérisé, nous souhaitons passer d'un paragraphe à son suivant en détectant un certain nombre d'indices graphiques qui sont les lettrines. Le but est de pouvoir naviguer rapidement, comme en hyper textuel sur le Web par exemple, d'un document à son suivant. Il faut pour cela pouvoir analyser le contenu de la page, détecter les éléments graphiques qui relèvent de la lettrine, des autres qui n'en sont pas, repérer les zones textuelles et ensuite pouvoir proposer des outils de navigation adaptés. Une thèse s'est terminée sur ce sujet, tandis qu'une nouvelle revient sur ce thème actuellement.

Un autre exemple concerne la requête sur la structure. Le collègue historien était intéressé dans certains cas pour pouvoir naviguer en recherchant des images dont la structure est proche. Vous voyez que toutes ces pages, en comparaison avec le transparent précédent, n'ont pas de ressemblance structurelle. Dans le cadre affiché, nous avons une image et nous souhaitons rechercher d'autres représentations dont la structure physique ressemble à celle-ci. Le système doit alors être capable d'aller rechercher des images ayant des propriétés à peu près similaires à celles qui correspondent à la requête. Vous voyez qu'il apparaît un cadre graphique autour, du texte au milieu et quelques graphiques. Pour pouvoir faire ce genre de chose, il faut être capable de segmenter et d'extraire les différentes informations, celles qui relèvent du dessin, du graphique et du texte, en utilisant des techniques très proches de celles présentées par les collègues. La seule différence avec leur présentation est que nos images sont très particulières. Ce sont des images imprimées, et des images de traits. Les caractéristiques extraites sur les images sont de ce fait spécifiques à ce contexte.

Un nouvel exemple traité par le collègue de Rennes concerne le registre de matricules au XIX<sup>e</sup> siècle, contenant des fiches d'incorporation militaire. La problématique ici est de pouvoir retrouver des fiches individuelles en tapant le nom de la personne. Au centre, vous voyez qu'il est écrit Brochard sur celle du milieu. La possibilité existe de rechercher, parmi des centaines de milliers de fiches qui ont été numérisées, quelle est celle de M. Brochard, et ensuite d'accéder à des informations lui correspondant. Pour effectuer ce genre d'action, il faut analyser les documents et leur structure, localiser les cellules de différents champs, reconnaître l'écriture manuscrite cursive qui date d'il y a très longtemps et qui est très variable. Il s'agit d'absorber la variabilité que nous pouvons avoir dans ces contextes-là.

Je reprends à présent le même exemple que celui qui a été présenté sur les logos concernant la recherche sur des requêtes graphiques. Comme vous pouvez le voir, nous avons ici zoomé sur une lettrine, réalisée à l'époque à l'aide d'un tampon en bois. Les collègues historiens sont parfois intéressés pour savoir comment circulait ce tampon en Europe au moment où les documents ont été imprimés. Il faut être en mesure de retrouver dans des ouvrages numérisés en Italie ou ailleurs, par rapport à un document qui est en France, là où on retrouve la même lettrine.

Je vous présente maintenant un autre exemple un peu plus compliqué. C'est la recherche d'une petite face, exactement comme pour le logo de tout à l'heure. Vous vous souvenez du cas des Jeux olympiques. C'est exactement la même chose. Nous recherchons des objets ressemblant à la requête qui est graphique ici, avec pour particularité le fait que nos images sont des images de traits. C'est la différence avec la présentation précédente des collègues.

Les collègues de Rouen travaillent sur des problématiques d'authentification et de transcription de manuscrits. Nous retrouvons là un manuscrit de Flaubert. Il s'agit alors d'authentifier cette œuvre, et de s'assurer qu'elle est bien du fait de son auteur. Comment ceci est-il réalisable ? Des caractéristiques sont liées au scripteur, comme la manière dont il écrit, mais aussi la façon dont il annote le document, et la manière dont il est raturé spatialement. Chaque individu a ses propres caractéristiques dans sa manière d'annoter les documents. Les collègues de Rouen travaillent précisément sur la définition de signatures caractéristiques de Flaubert, permettant d'authentifier le document et sa véracité.

Les problématiques scientifiques sont encore une fois les mêmes que celles présentées par mes collègues. Il existe beaucoup de redondance entre nos exposés, mais cela donne un autre domaine d'application. Les documents sont très diversifiés. Nous retrouvons des documents du patrimoine à gauche. Ces derniers peuvent être structurés, imprimés, anciens, manuscrits, graphiques... À partir de l'analyse de ces images, l'objectif est d'être capable de proposer des services de navigation en allant interpréter, rechercher suivant le but visé, l'information requise par l'utilisateur. La complexité se situe exactement au même niveau que celle exprimée par les collègues, c'est-à-dire dans la définition de caractéristiques pertinentes pour signer les images, et dans la mesure de similarité entre images. Vu l'existence de masses d'images, si nous voulons des traitements efficaces, nous devons structurer les espaces des caractéristiques que nous extrayons pour pouvoir naviguer rapidement dans ces masses. La problématique est donc rigoureusement équivalente.

Une banque d'images de lettrines apparaît en bas à gauche à titre d'exemple. Nous réalisons en back-office le calcul d'un grand nombre de signatures sur ces images. Celles-ci sont stockées sous forme de points dans un espace de représentation. Lorsque nous avons une image requête et que nous recherchons une image similaire, nous projetons cette image dans cet espace de points, et nous allons mesurer la distance entre deux points pour connaître la distance la plus proche.

Nous avons différentes contributions, différentes plateformes qui ont été élaborées dans notre consortium, comme vous avez pu le voir au travers des quelques exemples précités. Certaines sont visibles sur les sites de recherche du réseau Navidomass. L'ensemble va être intégré sur une plateforme unique pendant l'année.

Nous allons revenir en détail sur ce que nous venons de voir. Comment fonctionne l'image requête ? Nous venons extraire tout ce qui correspond à des zones de textes et à des zones autres que textuelles, en extrayant des propriétés spatiales et radiométriques des différents pixels des images. Nous regardons comment ils sont organisés les uns avec les autres. En fonction de ces propriétés, nous arrivons à classer chacun des pixels en disant : « Toi tu es du texte ; toi tu n'en es pas ; toi tu es du graphique ; toi tu n'en es pas. » À partir de ces différentes classifications, nous pouvons commencer à structurer nos masses de documents, de la même façon que les masses d'images présentées par les collègues où les paysages apparaissaient en haut à droite. Là, c'est la même chose. Nous avons les images fortement graphiques complètement en bas à droite, et les images fortement textuelles en bas à gauche. Nous naviguons entre les deux pour développer des moteurs qui permettent de naviguer et de retrouver rapidement, grâce à cette structuration, des images par rapport à une requête.

Je vous apporte à présent quelques détails sur l'exemple des registres d'incorporation militaire. La difficulté traitée par les collègues de Rennes est en fait que ces registres sont des tableaux qui ont été très fortement et très souvent modifiés. L'équivalent d'une sorte de post-it apparaît sur la partie gauche. Le post-it existe déjà depuis très longtemps contrairement à ce que l'on croit. Nous sommes venus coller des morceaux de page par-dessus le document. Le fait de rajouter de l'information vient

perturber les outils d'analyse d'images. Ces derniers vont aller chercher une certaine structure qui est elle-même modifiée par cet ajout d'informations.

Une des problématiques est donc de retrouver cette structure. Malgré la présence du post-it, il faut être capable de retrouver la structure du document, et de pouvoir ensuite extraire la cellule. Nous devons dans certains cas en cacher une partie, notamment si nous voulons la mettre en ligne sur le Web, et qu'il existe des informations confidentielles, etc. Ces éléments donnent la volumétrie. La France dispose de millions de fiches sur toutes ces questions. Tout est numérisé.

J'ai vu qu'il y avait les archives départementales des Yvelines qui allaient intervenir demain. C'est un partenariat avec eux entre autres qui est traité par le collègue de l'Université de Rennes. Ils vont donc se retrouver ici.

Ce sont les exemples que je donnais tout à l'heure. Nous devons être en mesure d'aller sur les bonnes cases dans cette structure une fois celle-ci localisée, et de déclencher des moteurs de reconnaissance de caractères qui permettent de rechercher par nom dans cette masse documentaire. Il existe des tas de techniques venant de la reconnaissance de l'écriture, les chèques, etc. dont je parlais précédemment.

Nous avons des cas très difficiles à traiter, comme la superposition de tampons sur l'écriture manuscrite. Une grande partie des masses d'images sont cependant traitées dans de bonnes proportions pour satisfaire des usages.

Nous allons aborder très rapidement le sujet de la reconnaissance graphique. L'idée est la même que celle présentée par les collègues. Nous avons une banque d'images de lettrines quelque part, et nous voulons savoir où nous pouvons en retrouver une en particulier au sein de cette banque d'images en Europe, dans le monde ou peu importe. La problématique est d'être en mesure de retrouver ces banques d'images. Nous allons caractériser ces lettrines de la même manière, par des propriétés spatiales et pixels qui les constituent. Nous pouvons commencer grâce à ce procédé à caractériser le contenu de cette image pour pouvoir retrouver de l'information équivalente. Nous avons donc un exemple de requête qui est une lettrine. Vous retrouvez un moteur de recherche du type Google, sauf que la requête n'est pas d'ordre textuel, mais une image. Le système renvoie alors toutes les images ressemblantes à la requête. Dans les outils que nous avons développés, il est intéressant de savoir ce que recherchent les gens dans un cas comme cela, sur la première lettrine. Si la requête est en haut à gauche, les gens cherchent-ils à retrouver un fond similaire ? Recherchent-ils la lettre A ? Je vous cite ici deux exemples d'intentions différentes auxquelles il faut s'adapter en fonction des objectifs de l'utilisateur.

Il me paraît indispensable d'insérer l'utilisateur dans la boucle pour réaliser ce genre de dispositifs, et ce, à plusieurs niveaux. Nous l'intégrons tout d'abord dans des outils de modélisation de la connaissance du domaine, appelés dans notre jargon des ontologies du domaine. Nous essayons d'y décrire (sous forme informatique) la nature de l'information manipulée par les historiens. Cette connaissance informatique se décline sous forme d'outils d'analyse pour pouvoir y adapter les traitements. Nous développons également par ailleurs des outils permettant aux utilisateurs naïfs de construire des scénarios de manière interactive, en fonction d'un corpus d'images à leur disposition. Ils ont des batteries d'outils, et ils les essayent les uns après les autres. Quand cela marche, ils continuent. Sinon, ils effacent le scénario, et ils recommencent à une étape précédente, etc. Le but est de construire un scénario adapté à leurs usages.

La dernière partie de l'interaction avec l'utilisateur est les problèmes de bouclage de pertinence. Quand un système renvoie un résultat (vous avez vu les lettrines avec la requête, avec le A et les différentes lettrines qui étaient renvoyées) l'utilisateur vient dire « Non, je ne suis pas d'accord avec le retour. Je pense que cette lettrine-là correspond davantage à ce que je recherche. » L'utilisateur vient modifier le résultat proposé par le système automatiquement, et vient inférer sur la manière dont le système doit répondre à la fois prochaine pour cet utilisateur-là. Il y a donc une forme de spécialisation et d'enrichissement de la connaissance par rapport à l'utilisateur.

Les problématiques scientifiques que nous développons sont des outils d'analyse d'images, d'indexation, de classification, et de modélisation de connaissances, comme nous venons de le voir, dans un contexte de masses de données. La grosse difficulté est d'être en mesure de fournir des services réalistes pour un utilisateur cherchant à avoir un résultat quasiment en temps réel, sachant que nous manipulons des grandes volumétries d'informations.

Les difficultés auxquelles nous sommes confrontés concernent le fait que les utilisateurs ne sont pas toujours en mesure d'exprimer leurs besoins. Quand nous travaillons notamment avec des historiens, ils nous disent qu'ils ne savent pas ce qu'ils peuvent faire avec les images. Nous leur amenons alors quelques services numériques et quelques développements. Ils disent alors que c'est intéressant, et commencent du coup à réfléchir en se disant qu'il serait peut-être possible d'enrichir ces services en travaillant sur ce sujet. Chacune des communautés, l'informatique et les historiens, progressent dans leurs propres recherches, tout simplement par interaction. Notre difficulté est de développer des méthodes qui soient à la fois génériques, et qui s'adaptent en même temps à des usages particuliers. Chaque utilisateur a ses propres usages. Tout en concevant des outils génériques, il faut pouvoir les rendre de plus en plus flexibles pour s'adapter à l'utilisateur.

Les interactions avec les sciences humaines et sociales sont très intéressantes et nouvelles pour nous. Nous travaillions précédemment avec l'industrie, comme pour la reconnaissance de chèques. Les objectifs n'étaient pas les mêmes. Je pense que nous sommes assez bons en France sur ces problématiques. Il faut le faire savoir. Sur les conférences internationales de niveau mondial, la communauté française est toujours majoritaire. Ce sont pourtant des conférences sélectives. Il serait certainement opportun de construire quelque chose en France qui mêle industrie et recherche, et qui permettrait de développer une industrie numérique française sur ces sujets-là.

Je vous propose juste pour terminer un transparent pour parler du projet Alpage, qui est commun avec les collègues du LAMOP. Nous travaillons ensemble sur la vectorisation des plans de cadastre anciens au XIX<sup>e</sup> siècle. Notre idée est de vectoriser automatiquement des plans de cadastre qui ont été numérisés, et de développer des services numériques pour les historiens. Il pourrait s'agir par exemple de recherche d'objets comme la recherche de puits, d'orientations pour savoir comment étaient construites les villes.

\* \* \* \* \*

*Suivi éditorial : Lorraine Pereira – chargée de mission pour le patrimoine cinématographique / INP.*