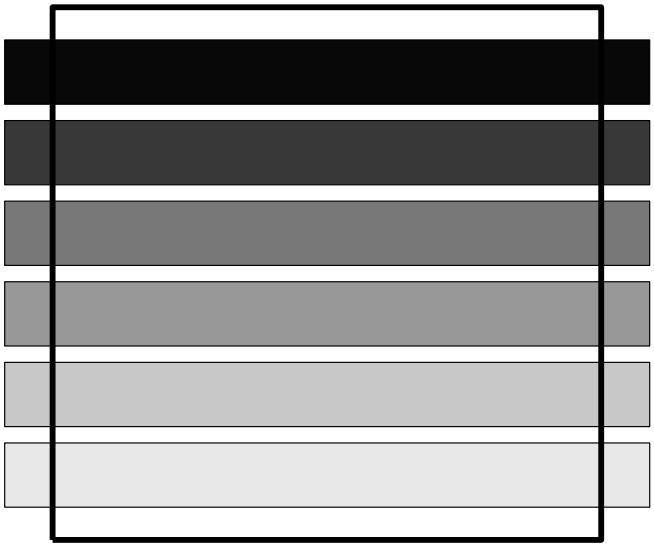
LES CAHIERS DU LANCI



DES MOTS POUR SE RETROUVER : RECHERCHE D'INFORMATION DANS L'ŒUVRE DE MAGRITTE À L'AIDE D'UN CORPUS DE DESCRIPTEURS SÉMIOTIQUES

Louis Chartrand, Jean-François Chartier et Jean-Guy Meunier

No 2013-01

UQÀM
Université du Québec à Montréal

Le Laboratoire d'ANalyse Cognitive de l'Information (LANCI) effectue des recherches sur le traitement cognitif de l'information. La recherche fondamentale porte sur les multiples conceptions de l'information. Elle s'intéresse plus particulièrement aux modèles cognitifs de la classification et de la catégorisation, tant dans une perspective symbolique que connexionniste.
La recherche appliquée explore les technologies informatiques qui manipulent l'information. Le territoire privilégié est celui du texte.
La recherche est de nature interdisciplinaire. Elle en appelle à la philosophie, à l'informatique, à la linguistique et à la psychologie.

Volume 9, Numéro 2013-01 – Janvier 2013
Publication du Laboratoire d'ANalyse Cognitive de l'Information
Directeurs : Luc Faucher, Jean-Guy Meunier, Serge Robert et Pierre Poirier
Université du Québec à Montréal
Document disponible en ligne à l'adresse suivante : www.lanci.uqam.ca
Tirage : 5 exemplaires

Dépôt légal – Bibliothèque Nationale du Canada
Dépôt légal – Bibliothèque Nationale du Québec
ISBN-10 : 2-922916-13-8
ISBN-13 : 978-2-922916-13-3

 Mise en page : Louis Chartrand

DES MOTS POUR SE RETROUVER : RECHERCHE D'INFORMATION DANS L'ŒUVRE DE MAGRITTE À L'AIDE D'UN CORPUS DE DESCRIPTEURS SÉMIOTIQUES

Louis Chartrand, Jean-François Chartier et Jean Guy Meunier
Université du Québec à Montréal

Résumé

Hébert et Trudel (2011) ont conçu une base de données regroupant toutes les œuvres de Magrittes indexées dans Sylvester et al. (1992-1997), qu'ils ont ensuite analysé suivant un protocole sémiotique rigoureux (Trudel et Hébert, 2011) conçu pour cette occasion. Cette analyse a produit un péritexte de descripteurs associés aux tableaux, que les auteurs ont intégré dans une base de donnée avec les métadonnées des œuvres. Cette dernière est accessible à travers une interface web qui permet la recherche par mots-clés. Afin d'évaluer ce péritexte, nous avons évalué son utilité dans le contexte de la recherche d'information pour lequel l'interface web a été conçue. L'analyse des données d'utilisation révèle que la recherche par mot-clé libre est impraticable, mais qu'elle devient possible (quoique laborieuse) avec l'aide d'un lexique. Nous concluons avec des suggestions pour l'amélioration de l'interface et des pistes d'utilisations pour le péritexte.

I. Introduction

L'interprétation des images est l'objet d'une longue tradition pluridisciplinaire et qui prend des formes différentes, par exemple, en histoire de l'art, en esthétique et sémiotique. Si le paradigme extérieur à la lumière duquel on tente de comprendre ces images varie, l'exercice reste à peu près le même : en restaurant à l'image un contexte historique, philosophique ou sémiotique, on rend son sens. En cela, l'interprétation des images ressemble à l'analyse des textes, telle que pratiquée dans plusieurs disciplines.

Depuis quelques années, l'essor de l'informatique offre des outils qui permettent d'assister les techniques d'analyse développées pour l'interprétation des textes et des images. Ces outils participent du renouvellement des méthodes et outils qu'apportent les *humanités digitales*. Ce

domaine de recherche comprend à la fois le développement d'outils d'analyse et l'édition de contenu numérique à partir des œuvres du patrimoine culturel. L'édition vise la valorisation du contenu, ce qui signifie qu'on doit produire un contenu dans un format polyvalent qui convienne à une grande variété d'utilisations. C'est pourquoi les éditeurs fournissent de plus en plus des métadonnées (titre, date, médium, etc.) sur les œuvres et sur ce qui les composent.

Dans ce contexte, l'image pose un problème particulier : contrairement au texte, son contenu semble inaccessible au traitement automatisé, que ce soit pour des fins de recherche, d'indexation ou de représentation. En effet, si certains outils informatiques permettent facilement de décrire les caractéristiques de bas niveau, c'est-à-dire les caractéristiques plastiques de l'image (couleur, texture, formes 2D ou 3D, ressemblance à une autre image, etc.), ils arrivent difficilement à faire de la reconnaissance de haut niveau, c'est-à-dire de lier ces représentations de bas niveau à des concepts (Hare et al., 2006). C'est ce qu'on appelle le « fossé sémantique » (*semantic gap*). Plusieurs tentatives ont été faites en vue de le combler (Liu et al., 2007), mais le problème subsiste : en effet, s'il existe des algorithmes qui parviennent à reconnaître des caractéristiques de bas niveau comme des formes (2D ou 3D) ou des régularités dans les couleurs, il est encore très difficile de les identifier aux catégories de haut niveau (e.g. chaise, lit, chien, etc) qui forment le langage courant.

C'est dans cette optique qu'Hébert et Trudel (2011) ont conçu une méthode de description rigoureuse (consignée dans le protocole Trudel et Hébert, 2011), et l'ont appliqué au corpus des œuvres du peintre surréaliste belge René Magritte, tel que catalogué par Sylvester et al. (1992-1997). Des descripteurs sont ainsi associés à 1 870 œuvres – peintures, sculptures, croquis, affiches, etc. Cette interprétation prend la forme de segments de texte, et on l'appellera, pour cette raison, *péritexte*[1]. Celui-ci est enregistré sur une base de données avec les métadonnées des tableaux (titre, année, médium, etc.), que les auteurs ont rendu accessible à travers une interface de recherche et de navigation sur le web.

Certes, la méthode de Trudel et Hébert n'est pas la première à permettre la recherche d'information, mais elle trace sa propre voie. D'une part, elle se distingue des « folksonomies » (descriptions par mot-clé produites par les utilisateurs d'une base de données) en ceci qu'elle produit des critères et des normes explicites pour la description, et qu'elle compte sur l'expertise d'annotateur·trices qui ont suivi une certaine formation en sémiotique. D'autre part, elle se

[1] Nous disons *péritexte*, car même si les descripteurs sont des lexèmes, ils sont des expressions synthétiques d'une phrases avec un contenu propositionnel. Par exemple les descripteurs « soleil » et « nuage » sont une abréviation de la phrase « Il y a dans le tableau un soleil et des nuages. ». Et la séquence de ces descripteurs forme un sorte de texte.

distingue d'approches plus structurées (e.g. Tam et Leung, 2001) qui attribuent aux descripteurs un rôle dans l'image et les mettent en relation autour d'évènements, d'actes, etc. en ceci qu'elle n'impose pas *a priori* à l'image une structure particulière. Elle allie donc la souplesse des folksonomies avec la rigueur d'une approche structurée.

Contrairement à la plupart des folksonomies, mais à l'instar des approches de description structurée, la méthode Trudel et Hébert compte sur des directives afin de répondre aux prérogatives d'une annotation précise et idoine. Cependant, là où la description structurée emploie une ontologie rigide à laquelle on ne peut déroger, l'approche Trudel et Hébert produit des critères explicites qui servent de guide aux personnes qui font la description, mais leur laisse plus d'initiative:

1. Les descripteurs doivent référer à un objet qui représente autre chose que soi-même (on notera la présence d'une pipe, mais pas la texture de la peinture).
2. Les descripteurs doivent référer à des signifiés iconiques figuratifs, c'est-à-dire à des entités (objets, actions, activités, phénomènes, etc.) qui se retrouvent dans l'image qui évoquent directement l'un des cinq sens. Cette condition, cependant, ne spécifie pas si l'objet du descripteur tient son sens des conventions du langage ou de celles de l'image : en effet, Magritte aimait à mêler mots et image dans le but de produire un certain effet (cf. Magritte, 1929). Hébert et Trudel ont donc noté le texte dans les œuvres de Magritte en le mettant entre guillemets; ils ont aussi ajouté certains caractères (« ~ », « ? ») afin de noter le caractère néologique de l'usage d'un mot ou un doute quand à la catégorisation d'un objet. De plus, afin de désambiguïser certains usages, on retrouve souvent, entre parenthèses, un mot qui vient ajouter de l'information sur l'assignation d'un descripteur (e.g. « feuille_(végétal) »).
3. Le critère fondamental de description est la saillance, et il vaut autant pour déterminer si quelque chose dans l'image doit recevoir un descripteur que pour déterminer le ou les descripteurs qu'on devra lui assigner. La saillance sémiotique se définit comme « l'intensité de présence d'une unité dans le champ perceptif ou dans le champ de présence » (Hébert et Dumont-Morin, 2012). Comme le notent Trudel et Hébert, on peut expliquer qu'un objet soit saillant de différentes façons – la « présence remarquable », les procédés ontologiques (taille, nombre, etc.), les procédés rhétoriques,

etc. Dès lors, bien que la saillance ne puisse être mesurée directement[2], elle ne se détermine pas non plus sur la base de la seule intuition de l'expert, puisque celui-ci doit pouvoir justifier ses choix des descriptions. Elle n'est donc pas soumise à l'arbitraire d'une perception incorrigible, et on peut espérer que la variation des descriptions, un problème fréquent dans les ontologies, soit ainsi limitée.

II. Problématique

Empruntant une voie moyenne entre la folksonomie et la description structurée, et le faisant à travers une méthode sémiotique, la méthode Trudel et Hébert innove, promettant une description rigoureuse et systématique sans recours à une ontologie lourde et dénaturante.

On peut spéculer sur les avantages de cette méthode. Par rapport à la folksonomie, elle promet d'être plus fine, et plus prévisible, parce qu'elle donne aux experts produisant le péritexte des critères d'exhaustivité et des instructions claires quant à ce qu'il faut décrire. Par rapport à la description structurée, elle promet d'être plus légère, plus facile à exploiter et de rendre plus adéquatement la structure propre à l'image, parce qu'elle n'alourdit pas le travail de description avec une structure *a priori*. De telles structures, ignorant la composition réelle de l'image, produisent des informations triviales qui sont par la suite difficiles à filtrer; en revanche, l'emphase de Trudel et Hébert sur la saillance pourrait permettre de donner une description plus fidèle à la structure originale de l'image.

Cependant, une évaluation rigoureuse du péritexte demande qu'on l'étudie dans le contexte de son utilisation réelle. Quels usages peut-il permettre? Hébert et Trudel ont leurs propres suggestions, qu'ils annoncent à la page d'accueil du site de la base de données Magritte :

> « Combien de toiles Magritte a-t-il peintes entre 1934 et 1944 ? Dans quelles œuvres du peintre trouve-t-on une tortue ? Une pomme ? Combien d'huiles, combien de gouaches a-t-il produites ? Quelles œuvres contiennent le mot « femme » dans leur titre ?
>
> Autant de questions auxquelles il est laborieux de répondre, même armé du fabuleux catalogue en cinq volumes de D. Sylvester. Autant de questions auxquelles notre de base de données répond instantanément. Mieux, une telle base, à notre connaissance une première en histoire de l'art, livre une vue d'ensemble quasi exhaustive des thèmes d'un peintre. »

[2] Ce qui ne nous empêche pas de croire qu'il est possible que des techniques utilisées en psychologie et neurosciences, comme l'oculométrie, puissent apporter des données empiriques indicatives de la saillance d'une entité ou d'un type d'entité.

Les questions que suggèrent les auteurs sont des questions de *recherche d'information*, en ceci qu'elles visent à retrouver des documents (dans le cadre de la base de données, un document est un tableau avec ses métadonnées et son péritexte) dans la base de données en fonction des critères de recherche propres à l'usager·e. Formellement, cette recherche d'information est une fonction qui prend comme paramètre une représentation des critères de l'usager·e et retourne les documents qui les rencontrent. Hébert et Trudel suggèrent donc que le péritexte permet de produire de telles fonctions, et donc qu'il permet la recherche d'information.

La référence à « une vue d'ensemble quasi exhaustive des thèmes » ne peut que faire allusion à la fonction de lexique, qui donne accès à la liste alphabétique de tous les descripteurs, lesquels donnent accès à une liste des œuvres qui les contiennent. Cette fonction, cependant, peut aussi être considéré comme recherche d'information: même si l'interface d'entrée est différente (liste plutôt que champs de texte), il s'agit quand même d'une entrée d'information de recherche.

D'autres usages sont possibles, en faisant intervenir de nouveaux outils – comme nous l'avons fait dans Chartrand, Chartier et Meunier (à paraître) – ou même un utilisant autrement les outils disponibles – par exemple, on peut analyser le lexique en soi, sans l'utiliser pour consulter les œuvres, et y faire des découvertes intéressantes. Cependant, la recherche d'information est au centre de la conception du site web, et cela se voit dans les détails : par exemple, le lexique est en ordre alphabétique, ce qui facilite la tâche de quelqu'un qui cherche un mot en particulier, mais qui complexifie celle de celui ou celle qui veut trouver les termes les plus utilisés.

Une évaluation complète du péritexte demanderait qu'on l'étudie dans tous ses contextes d'utilisation, mais on peut entamer ce projet en en choisissant un. Étant donné qu'on dispose d'une implémentation fonctionnelle et publique[3] du péritexte dans un contexte de recherche d'information, il est naturel que l'on commence notre évaluation par celle-ci. D'ailleurs, le site a été utilisé pendant plusieurs mois, et les gens qui l'ont utilisé ont laissé des traces sur le serveur web qui permettent de produire des données sur cette utilisation, lesquelles sont disponibles pour notre analyse.

Nous pouvons prendre pour hypothèse celle qui est manifestement soutenue par les auteurs de la base de donnée et du péritexte, à savoir que ce dernier est adéquat pour la recherche d'information. La question qui nous intéressera sera donc de savoir si cette hypothèse est vraie, et dans quelle mesure elle l'est – c'est-à-dire, dans quelle mesure le péritexte des œuvres de Magritte est approprié pour la recherche d'information.

[3] Dans une bonne mesure, du moins, car les images elles-mêmes ne sont pas accessibles pour des utilisateurs qui n'ont pas de codes d'accès. Ces derniers ont cependant été largement diffusés à travers un réseau de chercheurs universitaires

III. Méthode

Afin de répondre à cette question, nous analyserons donc les données dont nous disposons sur l'utilisation du site de la base de donnée Magritte.

À chaque fois qu'un·e usager·e accède à une page, son navigateur fait une requête au serveur pour celle-ci, puis pour tous les fichiers qu'elle contient (images, scripts, typographie, etc.), et celle-ci est archivée sur le serveur dans des journaux d'accès. Ceux-ci contiennent entre autre un identifiant de la machine qui fait la requête (adresse IP), le nom de la page demandée, l'heure et la date de la requête et, lorsque disponibles, des données de requête (par exemple, le texte recherché lorsqu'on soumet une recherche) et l'adresse de la page qui l'a référée. Ce sera là le substrat de notre analyse, à partir duquel on peut reconstituer le parcours des utilisateurs.

Il est donc aisé de retrouver les instances de recherche d'information parmi ces données. Évaluer leur succès et la contribution du péritexte à ce succès, par opposition à celle de l'algorithme de recherche, peut être plus difficile.

La méthode traditionnelle pour évaluer un système de recherche consiste en un calcul du *rappel* et de la *précision*. Pour ce faire, on demande à des usager·es de soumettre des recherches ainsi que, pour chacune, un ensemble E de résultats escomptés. On soumet la recherche au système, qui nous donne un ensemble R de résultats retournés; la précision :

$$\frac{R \cap E}{R}$$

est la proportion de résultats escomptés dans les résultats retournés, et le rappel :

$$\frac{R \cup E}{R}$$

est la proportion de résultats retournés parmi les résultats escomptés.

Malheureusement, les données dont nous disposons ne nous permettent pas de faire ces mesures : pour ce faire, il aurait fallu recruter des usager·es pour qu'ils nous donnent un corpus de recherches avec des résultats escomptés, ce qui aurait demandé, d'une part, des ressources plus importantes. Mais aussi, d'autre part, il nous aurait fallu avoir une meilleure idée du profil d'usager·e. En effet, l'utilité d'un système qui permet la recherche d'information dépend essentiellement de l'usage qu'en fait l'utilisateur. Et vu que celui-ci varie considérablement selon le genre de contenu qu'on peut trouver dans la base de données, il est impossible d'en faire un modèle formel. Impossible, donc, de calculer le rappel et la précision sans faire une enquête préalable auprès des utilisateur·trices.

On peut cependant tenter d'évaluer le système de recherche en lisant dans le comportement des usager·es des indices de leurs satisfaction quant aux résultats de leurs requêtes. En effet, à partir des données issues de journaux d'accès, il est possible de faire la mesure de ce que nous avons appelé le *taux de suivi* et la *proportion de suivi*. Le *taux de suivi* est la proportion des recherches où l'usager·e a cliqué sur au moins un résultat. Supposant qu'on clique plutôt sur les résultats qui nous intéressent, un haut taux de suivi devient un indice que les recherches ont donné des résultats satisfaisants. La *proportion de suivi* est la proportion des résultats cliqués sur le nombre total de résultats soumis à l'usager·e. Une basse proportion de suivi indique qu'en plus des résultats désirables, la recherche retourne beaucoup de résultats indésirables.

À partir du seul parcours des utilisateurs, qu'on obtient en analysant les journaux d'accès, on peut calculer le taux de suivi. En téléchargeant, pour chaque requête, une page de résultat, on peut aussi savoir combien de résultat elle a produite, et ainsi calculer la proportion de suivi.

Cependant cette méthode nous permet d'évaluer l'ensemble du système de recherche, incluant l'algorithme de recherche et les métadonnées et le péritexte sur lequels s'opère la recherche, or nous désirons n'évaluer que le péritexte.

Concernant les métadonnées, peu de recherches sont faites sur elles, de sorte qu'elles forment une quantité négligeable. Parmi plusieurs scripts de recherche sur le site web de la base de données Magritte, nous avons choisi de nous limiter aux recherches envoyées au script « rec_sem_simple.php » parce qu'elles seules étaient suffisamment nombreuses pour que les résultats puissent révéler des informations intéressantes. Or les mots-clés entrés suggéraient qu'elles semblaient se faire, dans une écrasante majorité, sur le péritexte plutôt que sur les titres ou les autres métadonnées[4].

Concernant le système de recherche, il est possible d'en savoir un peu plus sur le péritexte si l'on distingue entre recherches faites en entrant des mots-clés ou des locutions avec le clavier dans un champs de texte, comme c'est de coutume en cherchant sur Google, et celles faites en cliquant sur des mots-clés parmi une liste classée en ordre alphabétique. On dira du premier type de recherche qu'il se fait par le biais de mots-clés librement entrés, ou de *descripteurs libres*, alors que la recherche du second type se fait par le biais de *descripteurs canoniques*.

On suppose en effet que l'usage de descripteurs canoniques change considérablement la dynamique de recherche. D'une part, comme l'usager·e doit choisir parmi le lexique, celui-ci

[4] Il est difficile d'avoir de statistique précise à ce sujet. De fait, 70,5 % des requêtes sont formées à 100% de descripteurs utilisés dans le péritexte; mais une grande proportion du 29,5% restant est formé de recherches mal orthographiées, faites en langage naturel (e.g. « tableau où il y a des maisons ») ou utilisant des descripteurs que l'usager·e présume trouver dans le péritexte, mais qui n'y sont pas.

exerce une contrainte : on ne peut chercher qu'un terme à la fois, et on ne peut entrer de segments de mots ; et d'autre part, en lui présentant le lexique ou une partie du lexique, on informe l'usager·e sur le contenu du péritexte sur lequel se fait la recherche.

Étant donné que l'usager·e en sait plus sur le corpus sur lequel s'effectue la recherche, on s'attend à ce que les résultats soient plus adéquats. Cependant, la mesure dans laquelle les descripteurs canoniques améliorent les résultats nous indique quelque chose quant à la connaissance que l'utilisateur a *a priori* des mots employés qui sont dans le péritexte pour parler de l'œuvre de Magritte. Ou dit autrement : la mesure dans laquelle l'usager·e utilise spontanément les mêmes mots que le péritexte. Le taux de suivi et la proportion de suivi ne sont donc pas des mesures directes des performances du péritexte, mais si le vocabulaire du péritexte n'est pas familier à l'usager·e, leurs valeurs en seront affectées.

Bref, notre approche est de calculer le taux et la proportion de suivi à partir des journaux d'accès du site web, en distinguant entre recherche faites avec descripteurs libres et descripteurs canoniques.

IV. Résultats

Les journaux d'accès dont nous disposions ont référencé tous les accès depuis le 27 janvier 2012 au 15 juin 2012. Il y a eu, au total, 2033 requêtes de recherches envoyées au script « rec_sem_simple.php », dont 1034 ont été été faites par le biais de descripteurs canoniques. Les résultats sont illustrés dans le tableau 4.1, et leur évolution est illustrée dans la figure 4.1.

	Descripteurs canoniques (N=1034)	Descripteurs libres (N=999)	Total (N=2033)
Taux de suivi	51,55%	28,03%	39,99%
Proportion de suivi	17,25%	3,65%	10,92%

Tableau 4.1 : Taux et proportion de suivi en date du 15 juin 2012 à minuit

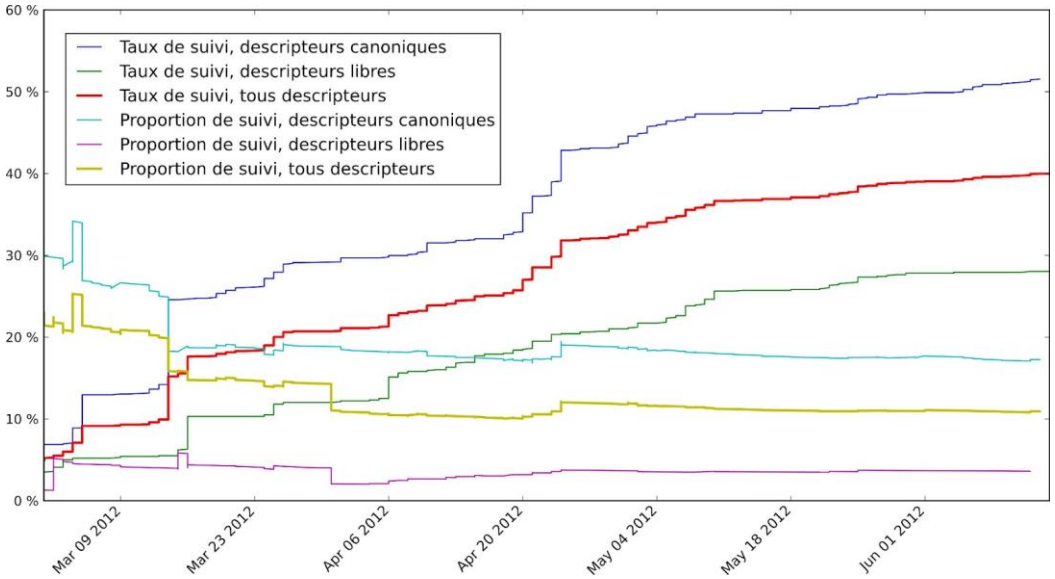


Figure 4.1 Évolution du taux et de la proportion de suivi

On constate que les taux de suivi tendent à monter alors que les proportions de suivi tendent à descendre – mais que ces statistiques tendent vers un certain équilibre à partir du mois de mai. Par ailleurs, on note que les proportions de suivi des descripteurs canoniques sont supérieures par plusieurs ordres de grandeur à celles des descripteurs libres, alors que les taux de suivi des premiers excèdent celui des seconds par une confortable marge de 15% à 20%.

V. Discussion

Évaluer le rappel d'information sur la base de données Magritte est un problème particulièrement délicat. Comme nous l'avons noté en section III, les mesures que nous avons faites sont un peu hors normes, de sorte qu'il est difficile de faire des comparaisons, sinon sur la base de simples intuitions. À notre connaissance, les moteurs de recherche traditionnels ne publient pas leurs taux et proportion de suivi.

Comme nous le disions plus haut, l'analyse des statistiques générales – i.e. indépendamment de la distinction *descripteur libre* vs. *canonique* – nous donne une idée globale de la performance du système, alors que l'analyse des différences entre les statistiques des recherches par descripteurs libres et des recherches par descripteurs canonique nous donne un indicateur qui peut imputer une partie des résultats au péritexte. C'est pourquoi nous les traiterons séparément, en commençant par les résultats généraux.

5.1 Résultats généraux

Les résultats du tableau 4.1 nous donnent un taux de suivi de près de 40%. C'est donc dire que dans une majorité de cas, soit l'usager·e trouve dans la page de résultats obtenue les informations qu'elle ou il cherchait et n'a donc pas besoin de suivre les liens qu'elle contient, soit les résultats escomptés ne sont pas au rendez-vous, et ne donnent donc pas satisfaction.

Si le taux de suivi était un signe certain que l'utilisateur a ou n'a pas trouvé un résultat qui le satisfasse, il serait assez facile de dire que ce résultat est peu satisfaisant. Cependant, beaucoup de facteurs peuvent amener l'utilisateur à ne pas cliquer sur un résultat satisfaisant ou à cliquer sur un résultat insatisfaisant.

Par exemple, il arrive que la page de résultats contienne toutes les informations recherchées : en effet, cette page montre certes les métadonnées de l'œuvre, un format très réduit de l'image[5] et les descripteurs correspondants à la recherche. Cependant, les autres descripteurs sont absents et le format réduit l'est généralement trop pour qu'on puisse y voir clairement davantage que quelques objets, de sorte que l'utilisateur·trice est forcé de consulter l'image en haute résolution ou la liste des descripteurs pour connaître son contenu. Étant donné que la base de donnée a été mise en place principalement pour répondre aux besoins de chercheurs et que les codes d'accès ont été diffusés dans un réseau d'universitaires, et on peut s'attendre à ce que les usager·es aient besoin d'informations précises sur le contenu des images. Ainsi, lorsque la page de résultat comprend les tableaux recherchés, on s'attend à ce que l'usager·e veuille en savoir davantage sur ceux-ci.

Par ailleurs, les « faux négatifs » (cas où l'usager·e trouve dans la page de résultats les informations recherchées et ne suit aucun lien) sont probablement mitigés par un phénomène de « faux positifs » mesurés, qui surviennent, par exemple, lorsque l'usager·e clique sur un résultat pour connaître son contenu, et s'aperçoit que ce n'est pas ce qui était recherché. Bref, il est difficile d'évaluer la proportion des uns et des autres, mais nous considérerons qu'ils s'annulent plus ou moins.

Nous estimons ainsi que, la majorité du temps (60%), la recherche ne donne pas les résultats escomptés. Le revers de cette médaille est que dans 40% des cas, elle a donné au moins un bon résultat. Bien que de tels résultats puissent justifier que l'utilisateur·trice continue à utiliser ce système de recherche, cela suppose de devoir souvent faire plusieurs recherches pour arriver aux résultats escomptés.

---

5 Les images étant sous droits d'auteurs, elles ne s'affichent cependant que pour les usagers qui ont un accès spécial, protégé par mot de passe.

La proportion de suivi, qui est de 13%, indique que pour chaque recherche, l'usager·e consulte un peu plus d'un résultat sur dix, ce qui suggère que l'énorme majorité des résultats ne sont pas utiles – supposant encore une fois que les quantités de faux positifs et de faux négatif soient comparable.

Qu'un seul résultat sur dix soit pertinent n'est pas un grand problème si n'a besoin que d'un seul résultat – regarder dix résultats pour trouver le bon n'est pas particulièrement pénible. Un algorithme de classement par pertinence peut palier à l'inconvénient lorsque ce chiffre est plus grand et que l'on doit chercher les entrées pertinentes dans une centaine de résultat, mais la base de données Magritte n'en contient pas. De plus, comme la base de donnée ne saurait disposer des données de recommandation et de citation qui font le succès d'algorithmes comme *PageRank*, il y a fort à parier que, quand même développerait-on un algorithme de classement, il ne saurait être aussi performant.

Par ailleurs, on ne fait pas de la recherche académique comme on navigue le web ; la rigueur exige qu'on prenne le temps d'analyser scrupuleusement au moins un bon échantillon des objets étudiés. Trouver un lien pertinent sur une liste de résultats est une chose, en trouver une dizaine est bien plus long, et l'absence de résultats indésirables en est d'autant plus appréciable.

On en conclut donc de cette proportion de suivi, quoiqu'encore une fois, elle ne démontre pas que la fonctionnalité de recherche soit complètement inutile, qu'elle dépeint son usage comme étant assez laborieux.

5.2 Descripteurs libres et canoniques

Les performances plutôt ordinaires des recherches sur le péritexte pourraient toujours être attribuables à l'algorithme de recherche. Cependant, si la différence entre les statistiques des recherches faites avec des descripteurs libres et celles faites avec des descripteurs canoniques est suffisamment grande, on peut suspecter que le vocabulaire du péritexte n'est pas familier aux usager·es, et que cela affecte la performance des recherches.

Dans les faits, la différence est très importante: 23,5% pour le taux de suivi (la moitié du taux de suivi moyen) et 13,6% pour la proportion de suivi (plus que la proportion de suivi moyenne). Il semble donc que le vocabulaire employé par Hébert et Trudel dans le péritexte est assez différent de celui qu'emploie spontanément un·e usager·e pour parler des œuvres de Magritte – un·e usager·e qui utiliserait les mêmes termes ne verrait pas le succès de ses recherches diminuer aussi drastiquement.

Figure 5.1: Fragment d'une peinture abandonnée (1954)

Pourtant, chez Magritte, le vocabulaire pictural est volontairement redondant et stéréotypé, de sorte qu'on peut facilement reconnaître que plusieurs tableaux représentent le même objet. Mais le stéréotype dans la représentation pictural n'a pas toujours son équivalent dans le langage – aussi la consigne du protocole, qui demande de privilégier les dénominations simples, si possible composées d'un seul mot, peut créer un vocabulaire qui, tout en étant compréhensible, a un caractère un peu idiosyncrasique. Par exemple, il est difficile de donner un nom à la figurine de la figure 5.1, qui dans le péritexte est appelée « canon », « balustre », « bilboquet » et « forme_anthropomorphe ». Cependant, cette pluralité n'épuise certainement pas les possibilités: d'autres termes auraient pu être appliqués (« forme_phallique », « pion_(échecs) », etc.).

Ici, la rigueur est peut-être ennemie de la recherche d'information. Les objets inhabituels ne sont pas rares chez Magritte, et ils sont souvent mis dans des contextes qui les aliènent de toute fonction propre qui viendrait spontanément à l'esprit. Les décrire demande un certain tâtonnement, et le langage qui sera utilisé ne saurait ressembler au langage mûr et uniformisé que préconise le protocole.

Enfin, on peut voir une confirmation de cette thèse dans le fait que le taux de suivi grimpe au fur et à mesure que les utilisateur·trices se familiarisent avec la base de données (figure 4.1). Mais que dire alors du fait que la proportion de suivi fasse le chemin inverse ? Plusieurs hypothèses peuvent être avancées. Par exemple, la hausse du succès dû à l'usage pourrait être mitigé par le fait qu'en avançant dans sa recherche, un·e chercheur·se acquière normalement une connaissance de base de son objet de recherche, de sorte que les retours à la base de donnée pourraient être davantage motivés par le besoin de confirmer ou de corriger son idée plutôt que celui d'acquérir une connaissance systématique d'un certain nombre d'œuvres. On pourrait aussi croire que la lassitude joue un rôle, et que devant la pénibilité de retrouver les œuvres qui l'intéressent dans une centaine de résultats, notre chercheur se contente d'en regarder un ou deux.

Pour cette raison, il semble que si le seul objectif était de faire de la recherche d'information, l'énergie requise pour produire le péritexte pourrait être mieux investie. Une folksonomie, par exemple, pourrait produire de meilleurs résultats, dans la mesure où les usagers qui produisent la description sont à peu près les mêmes qui utilisent la base de donnée, et où ils ont de bonne chance de décrire et de fouiller le corpus d'œuvres avec le même vocabulaire. De plus, les usager·es qui produisent les folksonomies les produisent directement dans le but de faciliter la recherche d'information, alors que la méthode Trudel et Hébert suggère plutôt de mettre d'abord l'emphase sur une description idoine, et espère ensuite que celle-ci sera appropriée pour la recherche d'information. Même si l'objectif intermédiaire peut aider à formuler des critères de description, on peut se demander s'il ne nuit pas à la fonction ultimement visée.

5.3 Confiance en les résultats

Comme nous le mentionnions en section II, la recherche d'information est un usage du péritexte parmi d'autres. Or, bien que la base de donnée ait été construite en partie sur cette prémisse, nous n'avons pas *a priori* de raison de croire que le péritexte soit particulièrement adapté à la recherche d'information.

Notre évaluation des résultats généraux – sur la valeur des taux et proportion de suivi relativement aux autres systèmes de recherche d'information disponibles sur le marché – montre que l'outil de recherche est assez laborieux à utiliser. Il semble donc raisonnable d'affirmer que ces deux conclusions ne sont pas indépendantes : les difficultés rencontrées par l'utilisateur·trice confronté·es à un système qui ne partage pas son vocabulaire doivent nécessairement entraîner une baisse de performance de celui-ci. Le problème du vocabulaire est donc un facteur qui explique la performance plutôt moyenne du système.

Il convient de rappeler que les résultats mentionnés ici ne nous donnent qu'une esquisse de la situation, puisque l'on ne connaît pas très bien les facteurs qui influencent le comportement des utilisateur·trices, en tout cas pas assez pour les quantifier et éclairer convenablement nos données. Cependant, plusieurs facteurs nous donnent particulièrement confiance en ce que nous avons observé. Premièrement, un échantillon de plus de 2000 recherches nous donne de bonnes garanties, et la stabilisation, voire amplification, des tendances qui nous intéressent nous rend confiant que les phénomènes observés sont réels et robustes. Deuxièmement, les effets observés sont assez importants. Que ce soit la différence entre les taux et proportions de suivi obtenus avec des descripteurs canoniques et des descripteurs libres ou la valeur assez basse des taux et proportions de suivi, il est difficile d'imaginer d'artefact qui explique l'un ou l'autre. Enfin,

comme les observations sont cohérentes, elles valent comme plusieurs évidences qui convergent vers nos conclusions.

Cependant, dans quelle mesure le problème du vocabulaire contribue-t-il aux performances ordinaires du système ? Si on améliorait le système informatique, pourrait-on avoir des résultats plus intéressants ?

Il serait effectivement possible de faire un calcul de pertinence, par exemple en évaluant quels tableaux sont les plus typiques d'un ensemble trouvé et/ou en faisant en sorte que le moteur de recherche apprenne du comportement des utilisateur·trices. Mieux encore, on pourrait concevoir un système de suggestion de mot-clé, ce qui permettrait d'augmenter les performances des requêtes faites avec des descripteurs libres. Cependant, il est possible qu'aucune amélioration au système informatique ne puisse combler le fossé entre le vocabulaire de description et le vocabulaire employé dans les recherches. Pour cette raison, à technologie égale, un système basé sur un péritexte conçu selon le protocole de Trudel et Hébert risque de continuer d'obtenir des résultats assez moyens comparativement à d'autres techniques d'annotation.

Par ailleurs, même si un système informatique en venait à rendre le péritexte aussi efficace, ou même un peu plus efficace qu'une folksonomie, cette dernière est beaucoup moins coûteuse à produire. Il faut des gains en performance pour justifier l'emploi de la méthode Trudel et Hébert, et il faut que ces gains soient suffisamment grands pour justifier la dépense qu'elle nécessite.

En somme, les données d'utilisation que nous avons utilisé pour cette étude ne nous permettent pas de contrôler tous les facteurs qui peuvent influencer les résultats, de sorte qu'on ne saurait avoir de conclusion rigoureuse sans étendre notre recherche, par exemple en étudiant directement les usager·es du site. Cependant, les données d'utilisation sont sans équivoque dans le portrait qu'elles font de la satisfaction des usager·es et des gains qu'apportent la recherche par descripteurs canoniques.

VI. Conclusion

L'analyse des données d'utilisation du site web de la base de données Magritte porte à croire 1) que le système de recherche de la base de données est utile, mais pénible à utiliser et 2) que cette pénibilité est en bonne partie due au péritexte, et en particulier au fait que les usager·e n'emploient pas le même vocabulaire pour les recherches que le péritexte pour la description des œuvres.

Nous pouvons cependant émettre quelques réserves. Il ne s'agit pas d'une étude comparative, donc il est fort possible qu'une comparaison rigoureuse avec, par exemple, les folksonomies, montre que ces dernières sont moins performantes que nous ne le pensons. Mais même si on en venait à cette surprenante conclusion, le coût très supérieur d'une description experte comme celle de Hébert et Trudel demande que la contrepartie soit appréciable.

C'est pourquoi nous croyons qu'il est plus raisonnable d'explorer d'autres usages qui seraient plus à même d'exploiter les forces de la méthode Trudel et Hébert.

Après tout, la méthode en est essentiellement une de description – ses conditions d'adéquation au moment de sa production se trouvent non pas dans une évaluation des recherches faites sur le corpus, mais sur l'adéquation entre le péritexte et l'image. De plus, son critère de saillance reflète le besoin de rendre compte de l'image telle qu'elle nous apparaît, telle qu'on en fait l'expérience – telle qu'elle est perçue. On pourrait profiter de cette caractéristique pour faire, par exemple, de l'analyse thématique ou conceptuelle sur de large corpus d'image.

Références

CHARTRAND, Louis, CHARTIER, Jean-François et MEUNIER, Jean-Guy, « Peindre Magritte avec des mots: analyse conceptuelle dans l'œuvre de Magritte à l'aide d'un corpus de descripteurs sémiotiques », à paraître.

GROUPE μ, *Traité du signe visuel: pour une rhétorique de l'image*, Seuil, 1992, 536 p.

HARE, Jonathon S., LEWIS, Paul H., ENSER, Peter G. B.[et al.], « Mind the Gap: Another look at the problem of the semantic gap in image retrieval », *Multimedia Content Analysis, Management and Retrieval 2006*, SPIE V, éds. Edward Y. Chang, Alan Hanjalic et Nicu Sebe, SPIE and IS&T, 2006, p. 607309-1, [En ligne : http://eprints.soton.ac.uk/261887/].

HÉBERT, Louis et DUMONT-MORIN, Guillaume, « Dictionnaire de sémiotique générale », in Louis Hébert, (éd.). *Signo*, éd. Louis Hébert, Rimouski, Québec, 2012, [En ligne : http://www.signosemio.com/documents/dictionnaire-semiotique-generale.pdf].

HÉBERT, Louis et TRUDEL, Éric, *Magritte: toutes les oeuvres, tous les thèmes*, Rimouski, Québec, 2011, [En ligne : http://www.signosemio.com/signodb/login.php].

MAGRITTE, René, « Les mots et les images », *La Révolution Surréaliste*, 1929.

SYLVESTER, D., WHITFIELD, S., RAEBURN, M.[et al.], *René Magritte: catalogue raisonné*, éd. D. Sylvester, Anvers, Fonds Mercator/Menil Foundation, 1992.

TAM, A. M. et LEUNG, C. H. C., « Structured natural-language descriptions for semantic content retrieval of visual materials », *Journal of the American Society for Information Science and Technology*, vol. 52 / 11, 2001, p. 930-937.

TRUDEL, Éric et HÉBERT, Louis, « Protocole d'analyse de la base de données « Magritte. Toutes les œuvres, tous les thèmes » », 2011, [En ligne : http://www.magrittedb.com/Protocole_d'analyse.pdf].