



## Élaboration d'un moteur de génération automatique de descripteurs pour les images biomédicales

**Sujin Kim**

Assistant Professor, School of Library and Information Science  
and Department of Pathology and Laboratory Medicine,  
University of Kentucky  
Lexington, USA

**Aswathnarayanan Sadagopan**

Graduate Research Assistant, Department of Computer  
Science, University of Kentucky  
Lexington, USA

*Traduction :*

*André Allard*

*Chef de service*

*Centre de documentation du CHUM*

*(Centre hospitalier de l'Université de Montréal)*

**Meeting:**

**180. Audiovisual and Multimedia and Bibliographic Control**

*WORLD LIBRARY AND INFORMATION CONGRESS: 75TH IFLA GENERAL CONFERENCE AND COUNCIL*

23-27 August 2009, Milan, Italy

<http://www.ifla.org/annual-conference/ifla75/index.htm>

### **Résumé**

*Contexte : Les groupes de recherche rassemblent souvent de nombreuses images de pathologie et les rendent disponibles sur le Web sous forme de bibliothèques numériques pour le bénéfice de la communauté académique. Toutefois, le repérage de ces images par les scientifiques et autres universitaires, en soutien à leurs travaux, n'est pas toujours facile.*

*Objectifs : Dans le cadre de ce projet pilote, les auteurs proposent un moteur d'indexation automatique qui produit des descripteurs pour les clichés microscopiques numériques (publiés?) publiés sur le Web. Le système met en place un module de cartographie qui range les mots-clés dans des thésaurus à l'aide de l'algorithme MetaMap Transfer (MMTx) mis au point dans le cadre du projet Unified Medical Language System (UMLS) (National Library of Medicine, USA)*

*Méthode : Les participants à ce projet pilote ont mis au point un fureteur qui repère les sites web proposant des clichés microscopiques accompagnés de commentaires. Les textes d'accompagnement sont soumis au générateur automatique de descripteurs pour obtenir des (set of topical descriptors) mots-clés.*

*Résultats : Le moteur de recherche dédié au repérage d'images est capable d'associer les termes d'indexation au métathésaurus UMLS. Les divers essais effectués démontrent que les*

*mots-clés obtenus à partir des légendes (sous-titres) analysées sont utiles pour bien décrire les clichés, facilitant ainsi le repérage. La pondération sémantique obtenue par MMTx pour chaque mot-clé, s'avère fort utile pour classer les descripteurs*

*Conclusion : Le mécanisme de cartographie sémantique des images biomédicales mis en lumière par la présente étude contribue d'un contrôle bibliographique relatif au web, condition préalable incontournable à un meilleur repérage des documents requis.*

## **Introduction**

La contribution d'un support visuel est une composante vitale à la pratique médicale, tant clinique que fondamentale. Les éducateurs, les professionnels de la santé et le grand public affichent de plus en plus d'images biomédicales sur le Web. Celles-ci représentent une source d'information importante pour l'éducation, les soins et la recherche médicale. Le diagnostic en anatomo-pathologie, par exemple, repose sur l'analyse cellulaire rigoureuse d'images des coupes histologiques. Le progrès en imagerie numérique a imposé comme standard la microscopie électronique pour les communications et les publications scientifiques, la formation des résidents et les consultations en histologie. Dans les grands centres hospitaliers universitaires, la majorité des départements de pathologie développe des banques de données de cas histologiquement documentés comme outils d'apprentissage.

Les cliniciens et les chercheurs consultent le Web à la recherche de données probantes supportant leurs résultats. Il est toutefois fort malaisé de localiser les images publiées dans les publications scientifiques.

L'indexation de ces articles dans la banque de données PubMed ne permet pas la recherche textuelle des commentaires juxtaposés aux images. Conséquemment, le potentiel d'information contenu dans les images demeure hors de portée. Plus encore, l'indexation avec les descripteurs sujets (MeSH) néglige la granularité de l'information accompagnant le support visuel au profit d'une analyse plus générale de la portée de l'article. La légende (sous-titre) étant le lien essentiel entre l'image et le message général de l'article publié, c'est là que se trouveront les termes d'indexation pour l'information non-textuelle contenue dans les publications savantes. Des études partielles ont tenté d'apprécier l'utilité de mots-clés extraits des légendes (sous-titre) accompagnant les images biomédicales. Aucune n'a tenté de cerner les caractéristiques essentielles de celles-ci.

Dans un premier temps, nous avons mis au point le prototype d'un moteur de recherche permettant l'indexation, la cartographie et le repérage des images biomédicales publiées sur le Web, plus précisément dans deux sources : GoogleImage et dans les banques de données de PubMed Central. Les caractéristiques des mots-clés obtenus et leur identification (intégration) aux thésaurus sont discutés sommairement dans la section des résultats.

## **Études réalisées à ce jour**

### **Clichés numériques en pathologie sur le Web**

La pratique quotidienne en laboratoire de pathologie implique l'examen, avec un microscope électronique, de cellules selon leur structure, leur forme et leur grosseur à l'aide de tissus fixés entre deux lames de verre. La numérisation des images obtenues a permis l'utilisation du Web à fins d'éducation, de recherche ou de diagnostic. Les clichés obtenus peuvent être

numérisés, éliminant ainsi les risques de bris, d'égratignures et de pertes de définition. Grâce aux progrès obtenus dans l'analyse d'image par ordinateur, en immunochimie et en microscopie fluorescente, la pathologie numérique est promise à un brillant avenir dans l'arsenal de la médecine moléculaire. (1-3)

Plus récemment, les séquences animées en mode streaming ont retenu l'attention. Ceci permet aux clichés d'être numérisés comme des ensembles plutôt que comme des items spécifiques et séquentiels. (4,5). Il y a des progrès majeurs à accomplir dans la description même grossière du matériel pathologique. Une description minimale s'obtient du nom du fichier généré par le numériseur. Toutefois l'information nominale contenue dans le dossier patient limite son utilité pour des raisons de confidentialité.

Un traitement et une analyse documentaire adéquate facilitent le repérage et la dissémination de l'information, générant ainsi une amélioration dans la facilitation du processus d'échanges d'images dans le milieu de la santé. Quelque chose d'aussi simple qu'une convention sur la description du système de capture par le microscope électronique et encore moins sur un ensemble de métadonnées distinctes, n'existe pas pour la pathologie numérique. Le repérage d'images par descripteurs-sujets est bien sûr impossible sans que ces documents de nature biomédicale déposés sur le web disposent d'une analyse approfondie (granular level of figure discription) des images comprises dans les articles savants. La situation se détériore encore si l'on considère la portion des items retenus qui sont simplement publiés puisqu'il n'existe pas de manière de régulariser l'appropriation de l'analyse documentaire par la communauté WEB. Le 'social tagging' est une avenue intéressante à la mesure de l'implication de l'individu. (?) Toutefois, devant une obligation de résultats dans un contexte de soins de santé, une meilleure prise en charge des documents numériques cliniques facilite leur repérage au besoin.

### **Légendes descriptives ( caption-based images descriptions)**

Les légendes sont de brefs commentaires attirant l'attention sur le contenu des images soumises en support à l'argumentation des articles publiés. Plusieurs études rapportent que les mots-clés trouvés dans les légendes présentent une opportunité sous-estimée, même plus productive que la recherche habituelle dans les mots des titres et des résumés (7-11). Heart et al. ont aussi démontré l'importance des légendes d'images pour témoigner d'avancées expérimentales. Ainsi, la recherche avec le terme "Western Blot" dans les légendes permet d'identifier plus de mille documents tandis que la recherche de la même expression dans les titres et les résumés du même corpus donnera des résultats beaucoup plus fragmentaires. (6) L'importance des légendes descriptives pour catégoriser les documents biomédicaux a aussi été précédemment soulignée ailleurs. Dans un projet pilote, Murphy et al. (2004) et Hua et al. (2007) ont mis au point un système pour identifier les images de microscopie fluorescente dans la littérature savante en analysant la fréquence des tons de gris et autres indicateurs de proximité (k-nearest neighbor classifiers) (8,9).

D'autres chercheurs s'appliquent à cataloguer les documents à l'aide de mots-clés pré-autorisés extraits des légendes. Sneiderman et al. (2008) présente un système de repérage qui indexe automatiquement les images biomédicales à partir de mots-clés obtenus dans les légendes d'images et commentaires de celles-ci dans un corpus d'images de dermatologie (10). Les résultats du processus d'extraction automatique furent plutôt décevants puisque seulement 26% des descripteurs UMLS obtenus dans les légendes sont pertinents à l'analyse des contenus iconographiques. Gay (2005) publie des résultats plus prometteurs quant à l'utilité des légendes pour l'indexation automatique de la littérature biomédicale à l'aide du

Medical Text Indexer (MTI) (14). Kahn (2008) va plus loin dans l'analyse des légendes en utilisant l'information contenue pour filtrer le repérage par âge, genre et modalité d'image dans sa banque de données iconographiques, un corpus de clichés radiologiques est construit à partir de 5 journaux reconnus de cette discipline. Conséquemment, la classification d'images par caractéristiques d'âge, de sexe et de méthode de saisie mérite une attention soutenue dans (un futur travail ?) une étude prospective. Les résultats obtenus tel que discutés dans cette section, nous incitent à contribuer à l'évaluation des descripteurs obtenus de différentes sources textuelles afin d'optimiser le repérage d'images.

## Méthode

Questions de recherche (QRs)

A partir du corpus de GoogleImage et des banques de données de PubMed Central :

**Q1 :** Quelles sont fonctions essentielles et les extrants de trois modules développés, nommément iRetrieve, iIndex et iTransfer ?

**Q2.** Quelles sont les caractéristiques des mots-clés utilisés par iIndex ?

**Q3.** Quelles sont les caractéristiques des inscriptions méta-cartographiées par MMTx pour les mots-clés échantillonnés par iIndex ?

## Conception et architecture du système

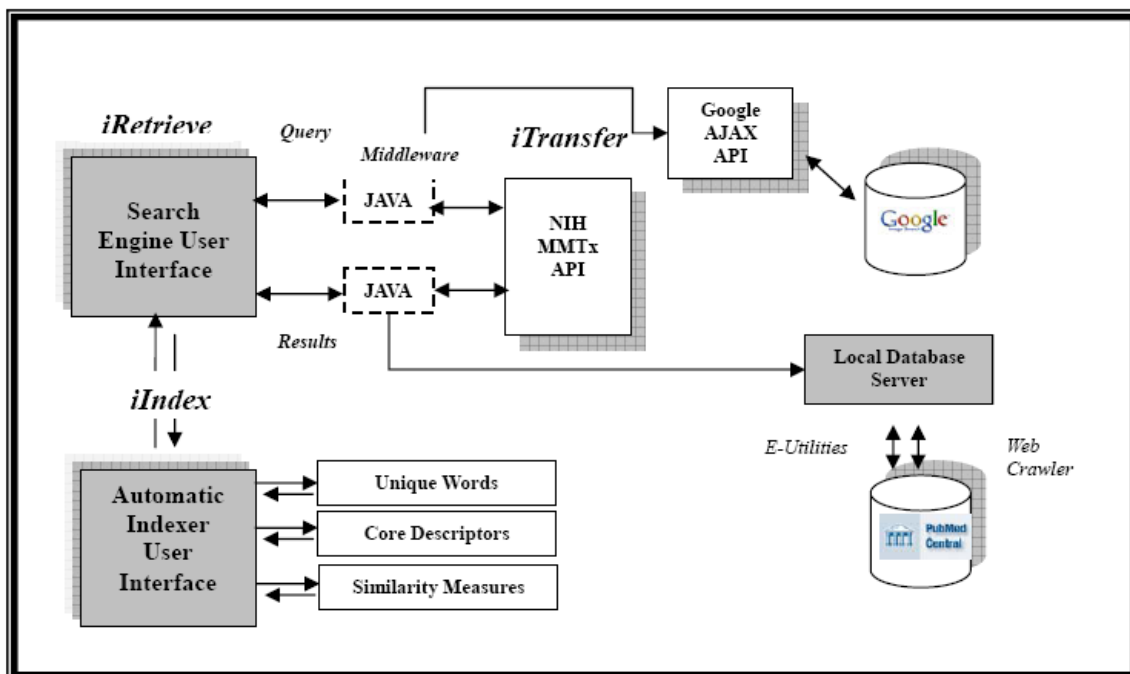


Tableau 1 : Architecture du moteur d'indexation et de recherche

L'architecture du système proposé est illustrée dans le tableau 1. Les principaux outils technologiques nécessaires pour mettre au point le moteur de recherche d'images biomédicales à partir du Web sont : (1) une interface pour la recherche et l'indexation (2) un programme Java pour l'accès et la cartographie (3) les banques de données fournissant les images et autres éléments bibliographiques (4) un débuseur Web et les utilitaires requis pour enregistrer les notices Medline et les images comme telles et (5) un serveur interne qui conservent localement les images (e.g. Uniform Resource Location) ainsi que leur description bibliographique. Trois modules ont été développés à cette fin. Le premier est un programme d'indexation ilindex mis au point pour choisir et extraire des mots-clés à partir des légendes, sous-titres, résumés et titres d'articles publiés recensés. Le second est iRetrieve qui traite, raffine et affiche les résultats pour l'utilisateur. Le dernier module propose une méta-cartographie automatique, établie selon les thésaurus de la National Library of Medicine avec le MetaMap Transfer Engine (MMTx). L'algorithme MMTx traite la requête soumise en 5 temps soit le parsing, la "variant generation", le candidate retrieval, le candidat evaluation et la production de la cartographie.(16). Le tableau 1 illustre et décrit les connexions principales du système proposé. Une description détaillée avec exposé sur la qualité des extraits sera faite lors de la discussion. Cet exposé du détail des fonctions termine notre réponse à la question de recherche no 1.

### **Moteur de cartographie et corpus de données**

Les données initiales obtenues du corpus d'images déterminés sont associées (linked) par des APIs java spécifiques utilisées pour chacun des deux serveurs, selon leur provenance. La Banque de données GoogleImage utilise un api Ajax, version beta d'un nouveau moteur chez Google. La requête Google se limite actuellement à chercher les mots dans les noms de fichiers dans le texte lié à l'image et les autres textes adjacents (17). Tandis que GoogleI fournissait les collections d'images web, les requêtes à PubMed Central tentaient de recouper ces images dans les publications disponibles chez ce serveur public. La banque de données PubMed Central est un produit offert par la National Library of Medicine qui contribue à l'accès libre au texte intégral de nombreux articles médicaux. Pour développer et optimiser le moteur de recherche, nous y avons téléchargé une partie de notre corpus à l'aide des E-Utilities de la NLM.

MetaMap Transfer Engine (MMTx), un algorithme java de cartographie de mots-clés, a été le principal outil de travail. A partir d'une requête en langage naturel, le MMTx génère une liste de mots-clés potentiels qui sont liés sémantiquement au vocabulaire de l'émetteur. Ces candidats MetaMP sont classés par score (le score des candidats est obtenu par 4 paramètres (metrics) viz. (centrality, variant, coverage and cohesiveness). Si le candidat n'est pas reconnu comme la forme prédominante du concept du métathésaurus, la forme préférée retenue est signalée dans le même écran par les parenthèses tandis que le genre sémantique auquel appartient la requête est signifié par des crochets carrés. Si la représentation générale du concept ne s'obtient pas spécifiquement avec le thésaurus, la cartographie conséquemment proposée fournira les concepts périphériques qui délimitent le concept de la requête. MetaMap offre à l'utilisateur une flexibilité réelle mais limitée pour raffiner les cartographies.

## Résultats

### Collecte de données pour l'essai

Les fonctionnalités des applicatifs proposés ont été mis à l'épreuve à partir du corpus d'images obtenues dans des journaux en accès libre signalés dans PMC Open Access Subset, complétée par une moisson dans les bases de données de GoogleImage. La révision manuelle complète a été réalisée pour sélectionner les articles signalés par PubMed Central fournissant des clichés microscopiques (microscopic images). Le cancer du sein a été la maladie retenue pour l'exercice, cette pathologie réduisant la dispersion sémantique dans la dénomination d'une même discipline. Les légendes, sous-titres, commentaires et résumés associés sont traités par l'applicatif ilindex afin d'obtenir automatiquement des mots-clés à partir du corpus d'articles supportés par des éléments iconographiques microscopiques.

### Intentions, finalités et technologies des applicatifs iRetrieve, ilindex et iTransfer (Q1)

iRetrieve lit la requête à travers quatre champs de recherche précédemment mentionnés soit légende, titre, résumé et descripteurs MeSH et génère une réponse à la suite. Des paramètres supplémentaires sont offerts pour préciser la recherche par méthode d'imagerie ( x-ray,ct, lames microscopiques) indicateurs démographiques(âge,sexe,humain ou animal), type d'activité de laboratoire ( staining, bioassay, etc), et autres mesures qualitatives et quantitatives.( 100x,200x,10%,20%,etc.)

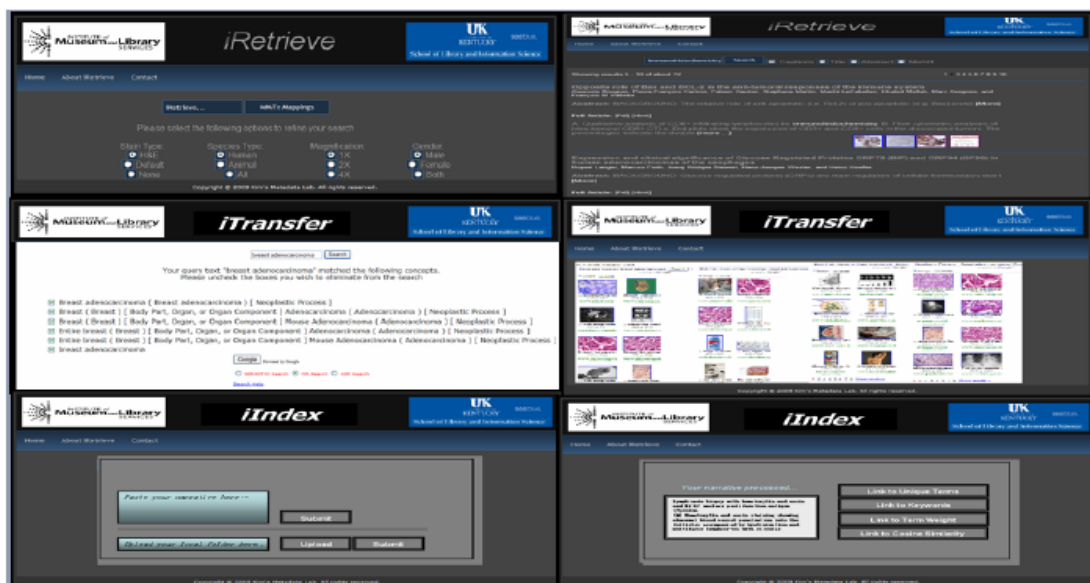


Figure 2 : Captures d'écran des modules iRetrieve,ilindex et iTransfer

A partir de l'analyse des légendes (sous-titres) d'une image en mode lot ( pour un ensemble de documents) ou à la pièce comme requête individuelle, ilindex génère plusieurs fichiers incluant un listing d'unitermes, , une pondération de ceux-ci menant à confection d'un listing de mots-clés principaux et des mesures de cohérence (similarity) associées.

La pondération et le tri effectué par *ilindex* s’obtient à partir de l’analyse de la fréquence de répétition du terme et de sa “cosine similarity”. Le traitement terminé, les résultats sont transférés en format Excel.

*iTransfer* prend en charge le pairing de tout mot-clé, nonobstant sa provenance, *ilindex* inclus ainsi que son traitement dans le continuum d’analyse pour optimiser le repérage. Les extraits de *iTransfer* sont des suggestions pour le MetaMapping, MetaCandidates, Matching scores, termes choisis, nombre de proposition pour candidat et catégories de genre(semantic types)

Table 1 : fonctions principales et extraits du système proposé

MODULES	CORE FUNCTIONS	OUTPUTS
<i>iRetrieve</i>	• Reading user-given search word	Images
	• Restricting search by captions or abstracts or MeSHs	Figures
	• Displaying retrieved images and descriptions	Captions
	• Refining search results by setting up search limits	Medline Records
	• Collecting locations of images and Medline records	Article URLs
	• Linking to the other two modules	User Interfaces
<i>ilindex</i>	• Reading user-given narrative image description	Unique words
	• Selecting core descriptors based on word frequency	3 Keywords
	• Calculating cosine similarity for a core descriptor	Similarity scores
	• Exporting aggregate automatic indexing results	Excel tables
<i>iTransfer</i>	• Paring user-given search keywords	MetaMapping
	• Generating word variations	MetaCandidates
	• Retrieving mapping candidates	Matching scores
	• Scoring mapping candidates	Preferred terms
	• Selecting final mapping suggestion	No of Candidates
	• Adding semantic types	Semantic Types

## Q2. Spécificité des mots clés indexés par *ilindex*

Répondre à cette question permettra de décrire le processus d’identification automatique effectué par *ilindex* et de tenter d’identifier les concordances entre le résumé et la légende pour optimiser la détermination des descripteurs. La table 2 présente les meilleurs 30 mots-clés. En accord avec la thématique retenue, le cancer du sein, les mots clés top comprennent les termes breast, cancer carcinoma, tumor, etc. Quelque 30% des mots clés automatiquement produits coïncide avec les 30 mots-clés les plus recommandés à l’interne pour l’indexation du corpus établi pour cette étude. L’étude permet aussi de remarquer qu’il existe peu de mots-clés obtenus uniquement des légendes(sous-titres) (page 8) Le nombre de mots-clés obtenus uniquement des légendes, hors les résumés, est petit. Parmi les mots-clés obtenus uniquement des légendes et sous-titres, nous retrouvons magnification, immunohistochemistry, hematoxylin and eosin, etc. termes principalement relatifs aux

procédures de laboratoire, peu informatif de la spécificité des résultats scientifiques présentés dans les résumés. L'ajout des mot-clés obtenus des (sous-titres) légendes à ceux des résumés fournit une liste de mots-clé pertinents au repérage de l'iconographie microscopique de la pathologie mammaire.

Table 2 : mots clés obtenus dans les légendes (sous-titres) des images et dans les résumés d'articles par ilindex

Abstract-based				Caption-based			
Keywords	Count	%	Sum (%)	Keywords	Count	%	Sum (%)
breast	1490	6.35		Cells	569	5.01	
cancer	907	3.87	10.22	Breast	270	2.38	7.39
cells	746	3.18	13.40	Carcinoma	229	2.02	9.77
expression	568	2.42	15.83	Expression	200	1.76	11.79
cell	412	1.76	17.58	Ductal	170	1.50	13.55
patients	360	1.54	19.12	Magnification	159	1.40	15.05
tumor	345	1.47	20.59	Tumor	136	1.20	16.45
carcinoma	234	1.00	21.59	Normal	134	1.18	17.65
human	219	0.93	22.52	Cell	103	0.91	18.83
growth	191	0.81	23.34	Figure	103	0.91	19.73
tumors	171	0.73	24.06	Cancer	101	0.89	20.64
gene	146	0.62	24.69	Tissue	92	0.81	21.53
invasive	137	0.58	25.27	Invasive	90	0.79	22.34
cases	120	0.51	25.78	Original	79	0.70	23.13
carcinomas	110	0.47	26.25	Positive	77	0.68	23.83
dcis	108	0.46	26.71	Tumour	72	0.63	24.51
normal	108	0.46	27.17	Mammary	63	0.56	25.14
receptor	99	0.42	27.60	Negative	63	0.56	25.70
survival	99	0.42	28.02	Antibody	61	0.54	26.25
treatment	99	0.42	28.44	Epithelial	61	0.54	26.79
ductal	97	0.41	28.85	Panel	56	0.49	27.33
protein	92	0.39	29.25	Nuclear	54	0.48	27.82
results	92	0.39	29.64	Sections	54	0.48	28.30
lines	90	0.38	30.02	Tumors	53	0.47	28.77
levels	89	0.38	30.40	Dcis	51	0.45	29.24
metastasis	88	0.38	30.78	representative	48	0.42	29.69
tumour	87	0.37	31.15	Control	47	0.41	30.11
women	87	0.37	31.52	Lobular	47	0.41	30.53
primary	86	0.37	31.89	Nuclei	47	0.41	30.94
mammary	85	0.36	32.25	Situ	47	0.41	31.35

## Spécificités des méta-catégories générées par MMTx

Notre troisième objet de recherche questionnait les modalités du processus de sélection automatique des méta-candidats par iTransfer. Ceux-ci sont les mots-clés retenus en raison de leur lien sémantique avec les mots-clés du document d'origine. Ainsi tel que présenté à la table 3, les formes retenues dans le métathésarus autour de l'expression (source keyword) écrite antibody (anticorps) sont antibodies et antigen binding, présentés entre parenthèses. Les catégories sémantiques qui complètent la définition des mots-clés retenus sont signalées par les crochets carrés (square brackets). Amino Acid, Peptide, or PROTEIN, Immunologic factor, Indicator, Reagent, or Diagnostic Aid and Molecular Function sont les deux catégories sémantiques principales identifiées pour le mot-clé texte du corpus antibody. La liste des candidats suggérés est présentée par leurs scores spécifiques tels 1000, 944, 916 et 900, en liaison avec la présence du mot breast dans les documents du corpus retenu. Ce type de classement est utile pour la sélection des candidats.

Table 3 : métag-candidats (descripteurs) échantillonnés par MMTx

carcinoma	Meta Candidates (5) 1000 Carcinoma [Neoplastic Process] 1000 Carcinoma (Carcinoma of the Mouse Prostate Gland) [Neoplastic Process] 1000 Carcinoma (Mouse Carcinoma) [Neoplastic Process] 900 carcinogen (Carcinogens) [Hazardous or Poisonous Substance] 900 Carcinogenicity [Neoplastic Process]
chemotherapy	Meta Candidates (3) 1000 Chemotherapy (Pharmacotherapy) [Therapeutic or Preventive Procedure] 1000 chemotherapy (pharmacotherapeutic) [Functional Concept] 1000 Chemotherapy (Chemotherapy-Oncologic Procedure) [Therapeutic or Preventive Procedure]
dcis	Meta Candidates (1) 1000 DCIS (Carcinoma, Intraductal) [Neoplastic Process]
ductal	Meta Candidates (4) 1000 Ductal [Qualitative Concept] 1000 Ductal (Ductal Hypoplasia of the Mouse Mammary Gland) [Disease or Syndrome] 928 Duct [Body Part, Organ, or Organ Component] 928 Duct (Entire duct) [Body Part, Organ, or Organ Component]

## Discussion/conclusion

Le système proposé a été développé en fonction des besoins des chercheurs en matière d'imagerie biomédicale. Il identifie des mots-clés recommandés, les redirige et éventuellement les reformule pour optimiser le repérage dans GoogleImage et PubMedCentral. Dans les publications scientifiques, l'imagerie biomédicale est une composante importante pour caractériser l'information transmise dans celles-ci. Les descripteurs MeSH usuels n'ont pas été pris en compte pour faciliter le repérage. Néanmoins, la combinaison des termes MeSH et les termes suggérés par l'analyse des légendes (captions) aident à obtenir une description fine des contenus iconographiques. Conséquemment ilindex, tel que développé, améliorera l'efficacité du repérage des images accompagnant les articles publiés à partir de nouvelles inhabituelles mais non moins pertinentes sources telles les légendes (sous-titres).

Nos prochains efforts tendront à vérifier l'efficacité du module iRetrieve avec des scénarios d'utilisation en mode réel. Nous prévoyons aussi prendre en compte des autres (multiples) modalités d'imagerie biomédicale (TDM,IRM,TEP,RX,etc.).

La faiblesse principale constatée lors de notre essai a été l'indexation uniterme générée par ilindex. Afin de pallier à cette carence, nous opterons à l'avenir pour une solution open source, Apache Lucene, qui assurera une meilleure qualité de l'indexation et du repérage grâce au classement des termes d'indexation en synergie le module de cartographie dans iTransfer. Le corpus de données utilisées pour notre étude était restreint au cancer du sein et l'iconographie retenue était uniquement microscopique. Nous prévoyons étendre notre investigation en y incluant d'autres modalités d'imagerie, en radiologie, en tomographie, etc. De plus, ayant observé un meilleur repérage par la prise en charge des descripteurs MeSH (principalement les étiquettes) en combinaison avec les descripteurs générés automatiquement, nous projetons évaluer plus rigoureusement les paramètres de l'efficacité du repérage des documents de PubMedCentral et de la banque de GoogleImage.

Tel qu'évoqué précédemment, les progrès obtenus dans l'imagerie médicale rendent son iconographie (ses clichés très importants) très importante pour le diagnostic clinique, les activités de recherche et l'enseignement. Notre étude proposera un mécanisme de cartographie sémantique pour contribuer à la compréhension générale du contrôle bibliographique des images web. Dans le champ de l'organisation et le repérage de l'information, l'intérêt pour les outils et méthodes pour le repérage de l'information iconographique est proportionnelle à la croissance du web. C'est la responsabilité des bibliothèques et autres centres de documentation de promouvoir et supporter l'intérêt en regard des opportunités et des défis proposés par l'imagerie biomédicale.

## References

1. Leong FJ & Leong AS. Digital imaging applications in anatomic pathology. *Advances in Anatomic Pathology*, 2003. 10(2):88-95.
  2. Pritt BS, Gibson PC, and Cooper K. Digital imaging guidelines for pathology: a proposal for general and academic use. *Advances in Anatomic Pathology*, 2003. 10(2):96-100.
  3. Montalto MC. Pathology RE-imagined: the history of digital radiology and the future of anatomic pathology. *Archives of Pathology and Laboratory Medicine*, 2008. 132(5):764-5.
  4. Dee FR. Virtual microscopy for comparative pathology. *Toxicologic Pathology*, 2006. 34(7): 966-7.
  5. Li XX., et al. A feasibility study of virtual slides in surgical pathology in China. *Human Pathology*, 2007. 38(12): 1842-1848.
  6. Shatkay H, Chen C, and Blostein D. Integrating image data into biomedical text categorization, *Bioinformatics*, 2006. 22(14): e446-53.
  7. Xu S, McCusker J, and Krauthammer M. Yale Image Finder (YIF): a new search engine for retrieving biomedical images, *Bioinformatics*, 2008. 24(17):1968-70.
  8. Murphy RF, et al. Extracting and Structuring Subcellular Location Information from On-line Journal Articles: The Subcellular Location Image Finder. Proceedings of the IASTED International Conference on Knowledge Sharing and Collaborative Engineering 2004. KSCE 2004:109-114.
  9. Hua J, et al. Identifying Fluorescence Microscope Images in Online Journal Articles Using Both Image and Text Features. Proceedings of the 2007 IEEE International Symposium on Biomedical Imaging, 2007. ISBI 2007: 1224-1227.
  10. Sneiderman CA., et al. [Web Document]. UMLS-based Automatic Image Indexing. AMIA Annual Symposium proceedings AMIA Symposium, 2008. p.1141. [Accessed on March 15, 2009]. Available at: <[http://archive.nlm.nih.gov/pubs/pubPDFs/Sneiderman\\_et\\_al\\_AMIA\\_2008.pdf](http://archive.nlm.nih.gov/pubs/pubPDFs/Sneiderman_et_al_AMIA_2008.pdf)>
  11. Yeh AS, Hirschman L, and Morgan AA. Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics*, 2003. 19: p. i331-9.
  12. Hearst, M.A., et al., BioText Search Engine: beyond abstract search. *Bioinformatics*, 2007. 23(16): p. 2196-7.
  13. Hearst MA & Wooldridge MA. Exploring the Efficacy of Caption Search for Bioscience Journal Search Interfaces. BioNLP 2007: Biological, translation, and clinical language processing, 2007:73-80.
  14. Gay CW, Kayaalp M, and Aronson AR. Semi-automatic indexing of full text biomedical articles. AMIA Annual Symposium proceedings AMIA Symposium, 2005. [Accessed on March 15, 2009]. Available at: <<http://ii.nlm.nih.gov/resources/amia05.fulltext.w.footer.pdf>>
  15. Kahn CE. Effective metadata discovery for dynamic filtering of queries to a radiology image search engine. *Journal of Digital Imaging*, 2008. 21(3):269-73.
  16. Aronson AR. The MetaMap Mapping Algorithm. [Accessed on March 15, 2009]. Available at: <<http://skr.nlm.nih.gov/papers/references/mm.mapping.pdf>>
- Google Image Search retrieved from Wikipedia. [Accessed on March 15, 2009]. Available at: [http://en.wikipedia.org/wiki/Google\\_image](http://en.wikipedia.org/wiki/Google_image)

**About the Authors:**

Sujin Kim, Ph.D., Assistant Professor  
School of Library and Information Science and Department of Pathology and  
Laboratory Medicine, University of Kentucky  
339 Lucille Little Fine Art Library Building  
Lexington, Kentucky 40506-0224, USA  
[sujinkim@uky.edu](mailto:sujinkim@uky.edu)  
(+1) 859-257-8657 (Tel), (+1) 859-257-4205 (Fax)

Aswathnarayanan Sadagopan, MS, Graduate Research Assistant  
Department of Computer Science, University of Kentucky  
350 Lucille Little Fine Art Library Building  
Lexington, Kentucky 40506-0224, USA  
[aswathn.s@uky.edu](mailto:aswathn.s@uky.edu)  
(+1) 859-257-2335 (Tel), (+1) 859-257-4205 (Fax)