Utilisation des propriétés sémantiques de la terminologie CISMeF pour la catégorisation de ressources de santé

Aurélie Névéol[1,2], Lina F. Soualmia[1,2], Alexandrina Rogozan[1], Magaly Douyère[2], Stéfan J. Darmoni[1,2]

[1] Laboratoire PSI – FRE 2645 CNRS – INSA de Rouen.
[2] CISMeF, CHU de Rouen et L@stics, Faculté de Médecine de Rouen.

Abstract:
The MeSH and the CISMeF terminology display medical concepts and the network of relationships existing between them. We present a method combining this type of semantic information with domain expert knowledge for resource classification purposes. We introduce a two-step classification consisting of 1/Mapping resource keywords to CISMeF metaterms and 2/Ranking metaterms by decreasing coverage in the resource. We evaluate this algorithm on a random set of 123 resources extracted from the CISMeF catalogue. Our gold standard for this evaluation is the manual classification provided by a librarian of the team. With 80,75% precision and 93,41% recall, the algorithm presented in this paper shows very good performance. A thorough analysis of the results has enabled us to find lacks in the knowledge representation modelized by the CISMeF terminology. We have made the necessary adjustments, and we are currently using the algorithm in CISMeF for resource classification. We are now planning to launch a second test phase in order to evaluate the impact of our changes in the terminology, and we will also compare the performance of our algorithm to those of classification with a statistical compression method.

Keywords:
Abstracting and Indexing ; Cataloguing ; Automatic Classification ; Controlled Vocabulary ; Database ; Information Storage and Retrieval ; Internet ; MeSH ; Subject Headings

1 Introduction

A l'heure actuelle, Internet est une source d'information importante dans tous les domaines, et en particulier celui de la santé. Les utilisateurs rencontrent d'énormes difficultés pour trouver précisément ce qu'ils cherchent dans la foule de documents mis à leur disposition. Les moteurs de recherche généralistes restent impuissants à résoudre ce problème car ils proposent souvent une sélection de documents trop large, ou encore mal ciblée.
Dans ce contexte, le catalogue CISMeF (Catalogue et Index des Sites Médicaux Francophones) créé en 1995, répertorie et indexe les ressources d'information institutionnelle de santé en langue française afin d'y permettre un accès rapide et précis [1].
Les ressources indexées par CISMeF sont d'une grande diversité, tant au niveau des types de documents sélectionnés (recommandations de pratique clinique, textes de loi touchant à la santé, cours, informations pour les patients, …) que de leur format (page Web, site entier, document pdf, …). Le catalogue contient à l'heure actuelle un peu plus de 10,000 ressources, et il est mis à jour au rythme de 50 nouvelles ressources en moyenne indexées chaque

semaine. L'ajout d'une nouvelle ressource au catalogue s'effectue en quatre étapes : le recensement des ressources potentielles par une veille stratégique quotidienne, la sélection des ressources selon des critères de qualité fondés sur le NetScoring (cf. http://www.chu-rouen.fr/netscoring/), l'indexation, et l'ajout définitif au catalogue par la mise en ligne de notices descriptives.
CISMeF a pour objectif de faciliter l'accès aux informations, et pour cela, nous souhaitons enrichir les notices grâce à un nouveau type de catégorisation.

L'objectif de ce travail est d'établir une catégorisation pertinente par spécialités biomédicales, pour les ressources indexées dans CISMeF. La catégorisation que nous proposons vise à faire ressortir le sens global du document, c'est-à-dire à dégager les principaux thèmes abordés et à les ordonner par importance croissante dans le document. Depuis 1995, nous n'avons cessé de progresser dans ce sens. Ainsi, les documents étaient au départ indexés par des mots clefs du thesaurus MeSH (Medical Subject Headings) [2] associés (ou non) à des qualificatifs. Puis, nous avons introduit la notion de mot-clé "majeur" (signalé dans CISMeF par une étoile) lorsque le thème désigné par ce mot-clé était abordé de façon prépondérante dans le document, vs "mineur" lorsque le thème désigné par ce mot-clé était abordé dans une partie du document seulement. Ensuite, comme le nombre mots-clés indexant un document est parfois important (plusieurs dizaines), et comme ces mots-clés renvoient à des notions spécifiques souvent pointues, il nous a paru important d'établir une catégorisation à un niveau plus large, afin de renseigner l'utilisateur sur les spécialités traitées par une ressource.

Le travail présenté dans cet article a pour but de tester la validité d'un algorithme permettant d'obtenir une catégorisation qui présente la liste des spécialités par ordre d'importance dans le document. La catégorisation proposée par cet algorithme est comparée à celle établie manuellement par un documentaliste de l'équipe CISMeF, qui sera considéré comme référence.

2 Matériel et méthodes

2.1 Terminologie CISMeF

Afin de bien comprendre la méthode de catégorisation des ressources dans CISMeF, il est important de connaître la structure de la terminologie CISMeF [3].
La description des ressources par CISMeF s'appuie sur plusieurs concepts :

- une liste de mots-clés MeSH associés (ou non) avec des qualificatifs. Le thesaurus MeSH est développé par la NLM (National Library of Medicine). Dans sa version 2003, le MeSH compte plus de 22,000 mots clés, et 84 qualificatifs. Le thesaurus MeSH permet de décrire les ressources à l'aide de mots clés et de qualificatifs qui peuvent leur être associés (ou non).
- une liste de types de ressources [1] (n=133). Un type ressource précise la nature de l'information véhiculée par le site ou le document (par exemple *cours*). La liste des types de ressource utilisés par CISMeF est disponible à l'URL: http://www.chu-rouen.fr/documed/typeressource.html.

Ces trois types d'information (mots clé MeSH, qualificatifs MeSH et type de ressource) sont structurés selon la hiérarchie du MeSH pour les mots clés et qualificatifs MeSH, ou selon la hiérarchie établie par l'équipe CISMeF pour les types de ressource.
Pour organiser l'accès aux ressources, les « métatermes » [4] ont été incorporés à la terminologie CISMeF. Ils correspondent en général à des spécialités biologiques ou médicales (par exemple, *cardiologie*, ou *bactériologie*) sélectionnées par le documentaliste expert de l'équipe CISMeF. Pour chaque métaterme (N=85) on définit une série de liens

sémantiques avec un ou plusieurs mots-clés MeSH, qualificatifs et/ou types de ressource. Ainsi, par exemple, le métaterme *psychiatrie* est associé (entre autres) aux mots clés MeSH *psychiatrie* et *hôpital psychiatrique* qui appartiennent à deux arborescences différentes du MeSH. Il ne comporte pas de lien avec des qualificatifs, mais il est associé avec le type de ressource *dispensaire hygiène mentale*.
A l'origine, les métatermes ont été introduits afin d'optimiser la recherche d'information dans CISMeF, et de mettre en évidence des relations entre certains termes dont le MeSH ne rendait pas compte. En effet, les requêtes "lignes directrices en cardiologie" et "bases de données en psychiatrie" où cardiologie et psychiatrie ne sont que des mots clés MeSH ne donnent que peu ou pas de réponses. L'introduction des métatermes *cardiologie* et *psychiatrie* se révèle une stratégie efficace pour obtenir plus de résultats pertinents car au lieu d'exploser un seul arbre MeSH (par exemple, l'arbre correspondant au mot clé *psychiatrie*), l'utilisation des métatermes permet d'étendre la recherche en explosant également les autres arborescences en rapport avec le métaterme en question. Pour la requête correspondant au métaterme *psychiatrie,* on explosera donc les arborescences correspondant à *psychiatrie, hôpital psychiatrique, dispensaire hygiène mentale,* etc. La liste des métatermes est disponible à l'URL http://www.chu-rouen.fr/ssf/santspe.html.

2.2 Algorithme à base de règles utilisé dans CISMeF.

L'algorithme de catégorisation proposé ici est fondé sur le savoir-faire des documentalistes CISMeF et exploite l'indexation manuelle des ressources disponibles dans la base de données CISMeF. Cet algorithme de catégorisation utilise les liens sémantiques existant entre les métatermes et les mots clés MeSH, les qualificatifs et les types de ressource afin d'extraire une liste de métatermes qui sera associée à une ressource donnée. Cette liste de métatermes est alors ordonnée à l'aide d'une série de règles établies par des experts du domaine, les documentalistes CISMeF.
Comme nous l'avons vu précédemment, chaque ressource recensée dans CISMeF est indexée par une liste de mots clés MeSH, associés ou non à des qualificatifs, et par une liste de types de ressource. Par l'intermédiaire des liens sémantiques de la terminologie CISMeF, l'algorithme associe chaque élément de ces listes à un ou plusieurs métatermes. Ainsi, si un terme (mot clé, qualificatif ou type de ressource) est lié à plusieurs métatermes, chacun de ces métatermes sera retenu pour la catégorisation. Par exemple, le mot clé *alcoolisme* nous conduira à retenir les métatermes *psychiatrie* et *toxicologie* tandis que le mot clé *pouce* conduira à retenir le seul métaterme *anatomie*.
Par ailleurs, pour obtenir la catégorisation finale, l'algorithme calcule deux scores pour chaque métaterme retenu: un score « mineur », et un score « majeur ». Le score « mineur » correspond au nombre de mots clés et de couples (mot clé/qualificatif) mineurs à partir desquels le métaterme considéré a été retenu. De même, le score « majeur » correspond au nombre de types de ressource, de mots clés majeurs et de couples (mots clés/qualificatifs) majeurs à partir desquels le métaterme considéré a été retenu. Ainsi, pour obtenir la catégorisation recherchée, les métatermes peuvent être classés par ordre de scores « majeurs » décroissants, les scores « mineurs » permettant de départager les éventuels *ex-aequo.*
Les métatermes ayant un score majeur non nul sont dit "majeurs" et ils sont représentés avec un nombre d'étoiles correspondant à ce score. Ainsi, si le métaterme *psychiatrie* est retenu à partir du mot clé majeur *hôpital psychiatrique*, du mot clé mineur *psychiatrie*, et du type de ressource *dispensaire hygiène mentale*, deux étoiles lui seront attribuées.

2.3 Evaluation

Nous avons testé notre algorithme sur un échantillon de 123 ressources choisies aléatoirement dans le catalogue CISMeF. Afin d'évaluer la pertinence des résultats obtenus, les catégorisations proposées par l'algorithme ont été comparées avec les catégorisations établies par un documentaliste CISMeF sur les mêmes ressources. Le tableau 1 représente le nombre de spécialités à extraire pour chaque ressource, selon la documentaliste.

Tableau 1 - Nombre de spécialités à extraire par ressource.

Nombre de spécialités	**Nombre de ressources**	**%**
Au plus 1 spécialité	31	25,20 %
2 spécialités	32	26,02 %
3 spécialités	32	26,02 %
4 spécialités et plus	28	22,76 %
Total	123	100 %

La catégorisation manuelle, considérée comme notre référence (c'est à dire, fournissant la liste des spécialités pertinentes correctement ordonnées) a été réalisée après avoir pris connaissance des résultats de l'algorithme.
Nous nous sommes attachés à mettre en évidence la pertinence de la catégorisation proposée par l'algorithme, et nous avons distingué trois niveaux : la catégorisation « sans faute » pour laquelle l'algorithme fournit exactement la même catégorisation que le documentaliste, la catégorisation « pertinente » pour laquelle l'algorithme fournit une catégorisation qui présente beaucoup de points communs avec celle du documentaliste, mais n'est cependant pas exacte, et la catégorisation « non pertinente » qui n'a rien en commun avec la catégorisation du documentaliste.
Les écarts observés entre les deux catégorisations (manuelle et automatique) sont dus à quatre types d'erreurs : proposition par l'algorithme d'une spécialité non pertinente, oubli d'une spécialité pertinente, erreur d'ordre dans le classement des spécialités et erreur de pondération majeur/mineur correspondant à ces spécialités.

3 Résultats

La figure 1 est un exemple de catégorisation obtenue par notre algorithme pour une ressource répertoriée dans le catalogue CISMeF.

L'évaluation de la catégorisation réalisée sur 123 ressources choisies aléatoirement dans le catalogue CISMeF donne une précision de 80,75% (au total 298 spécialités correctes sur les 369 proposées), soit un bruit de 19,25%. Le rappel est de 93,41% (au total 298 spécialités correctes sur les 319 attendues), ce qui correspond à un silence global de 6,59%.
Ces résultats sont très satisfaisants, et montrent que la catégorisation automatique des ressources est presque exhaustive.
Le tableau 2 présente la pertinence de la catégorisation proposée par l'algorithme sur l'ensemble des ressources, et le tableau 3 présente la répartition des types d'erreurs décrits au paragraphe précédent.

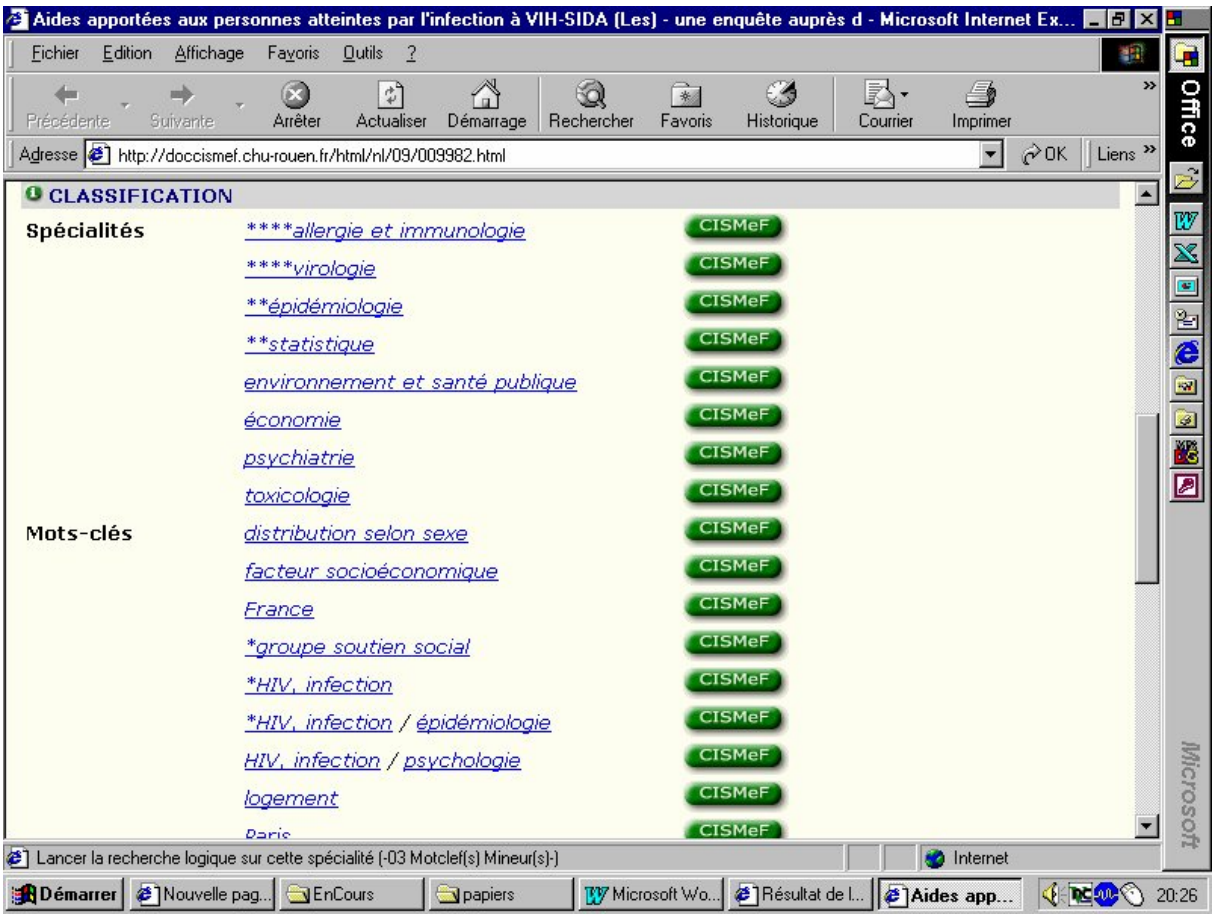

Figure 1 : Catégorisation de la ressource « Aides apportées aux personnes atteintes par l'infection à VIH-SIDA » (http://www.sante.gouv.fr/drees/etude-resultat/er-pdf/er203.pdf)

Tableau 2 – Pertinence de la catégorisation.

Pertinence	**Nombre de ressources**	**%**
Catégorisation sans faute	45	36.58 %
Catégorisation pertinente	70	56,91 %
Catégorisation non pertinente	8	6.50 %
Total	123	100 %

Tableau 3 - Erreurs de catégorisation observées.

Type d'erreur	**Nombre de ressources**
Spécialité inappropriée (bruit)	36
Spécialité omise (silence)	21
Erreur d'ordre	39
Erreur de pondération : Majeur au lieu de mineur Mineur au lieu de majeur	 42 7

4 Discussion

Les principaux résultats de ce travail (précision de 80,75% et rappel de 93,41%) sont très encourageants.
De plus, on constate que l'algorithme propose la catégorisation exacte (c'est à dire préconisée par le documentaliste) dans plus d'un tiers des cas. Les cas de catégorisation « non pertinente » sont généralement obtenus sur les ressources devant être indexées par une ou deux spécialités pour lesquelles un bruit ou un silence important intervient. Dans beaucoup des cas que nous qualifions de « pertinents », les erreurs observées sont très légères. Quelques cas seulement cumulent tous les types d'erreurs.
Notre méthode d'évaluation, qui a le désavantage de ne pas être réalisée en aveugle, permet de juger plus précisément la pertinence de chaque élément de la catégorisation automatique, afin de mettre en évidence le bruit et le silence de l'algorithme.

Le silence mis en évidence par les résultats porte sur certains domaines spécifiques de la terminologie. Ainsi, une ressource indexée avec les mots clés *voyage* et *médecine tropicale*, et avec le type de ressource *information patient et grand public* ne permet de retenir aucun métaterme car ces mots clés ne sont liés à aucun métaterme. Ainsi, une analyse des résultats permet de dégager une liste de termes de la terminologie CISMeF pour lesquels il est nécessaire d'instaurer des liens vers des métatermes existants, ou encore de créer des métatermes adaptés à cet effet. Dans notre exemple, *médecine tropicale* est une spécialité médicale, et il faut donc créer un métaterme *médecine tropicale* qui sera lié au mot clé *médecine tropicale*. La création d'un tel métaterme enrichira la terminologie CISMeF, et permettra d'améliorer les performances de l'algorithme de catégorisation, mais n'aura pas d'incidence sur la recherche d'information dans CISMeF.
Malgré ces manques dans la terminologie, qui ont été comblés depuis, le silence de la catégorisation est très faible, sans doute car les lacunes touchaient principalement des spécialités peu abordées dans CISMeF (par exemple, *biologie cellulaire, biochimie* ou *chirurgie esthétique*).

En revanche, il est impossible de réduire le bruit observé sans affecter les performances de la recherche d'information. En effet, pour réduire le bruit, il faudrait identifier les liens provoquant une catégorisation « bruitée », et les supprimer. Il faut également remarquer que les métatermes proposés par l'algorithme de catégorisation et jugés superflus par le documentaliste ne constituent pas une mauvaise description du document en soi, mais il nuisent à l'aspect synthétique de la catégorisation.

Pour ce qui est de l'identification majeur/mineur des métatermes de la catégorisation, on observe six fois plus d'erreurs pour les mots clés « mineurs » identifiés comme « majeurs » (42 erreurs sur 123) que l'inverse (7 erreurs sur 123). Ceci résulte en fait de l'importance accordée au type de ressource qui n'est pas pondéré à l'origine, mais que nous considérons comme un terme « majeur » dans la constitution de la catégorisation.

Dans le domaine de la catégorisation de ressources textuelles, une étude réalisée par le département d'informatique médicale de l'Université de Columbia [5], recense les nombreuses méthodes qui peuvent être utilisées. Des travaux récents de Teahan et Harper [6] montrent que les modèles statistiques de compression PPM (Prediction by Partial Match) ont des performances intéressantes, tout en relevant d'une approche globale ne nécessitant aucune extraction de mots clés *a-priori*. On a également observé récemment une émergence des

Machines à Vecteurs Support [7] qui semblent très efficaces, même sur certains problèmes multi-classe. Comme la plupart des autres méthodes de catégorisation, que ce soit par *apprentissage inductif* (arbres de décision, classifieurs naïfs de Bayes, réseaux bayésiens) ou par *règles d'association*, les SVM sont fondés sur une approche analytique [8]. Le processus d'apprentissage des classes nécessite une paramétrisation préalable, c'est à dire l'extraction de vecteurs de données constitués des termes représentatifs des spécialités à l'étude. Ainsi, les méthodes à base de règles sont très performantes, à condition de pouvoir modéliser les principes de catégorisation avec l'aide d'experts du domaine. Dans ce contexte, nous avons choisi de formaliser les connaissances métier des documentalistes de l'équipe pour définir les règles propres à la catégorisation de documents dans CISMeF

Les résultats montrent clairement que les performances de notre algorithme sont directement affectées par les manques qui peuvent exister dans notre représentation des connaissances. Comme le souligne Bodenreider dans des travaux similaires [9], l'implémentation de notre méthode de catégorisation nous permet d'optimiser notre terminologie. En effet, alors que les relations sémantiques de l'UMLS (Unified Medical Language System) constituaient une barrière pour améliorer les performances [9], nous avons la possibilité d'étendre la couverture de notre terminologie. Suite à l'évaluation de l'algorithme que nous décrivons, nous avons introduit 18 métatermes (ainsi qu'une série de liens sémantiques associés) dans la terminologie CISMeF en décembre 2002.

Afin d'améliorer la représentation des connaissances modélisée par la terminologie CISMeF, nous allons réaliser une seconde évaluation de notre algorithme sur un ensemble de 100 nouvelles ressources. Cela nous permettra de déterminer d'une part s'il est nécessaire de créer de nouveaux liens sémantiques ou de nouveaux métatermes, et d'autre part de vérifier l'influence des liens et des métatermes nouvellement introduits sur le *bruit* de la catégorisation.
Par ailleurs, nous nous fixons comme perspective d'effectuer une catégorisation des ressources CISMeF fondée sur des modèles de compression PPM afin de comparer les deux méthodes.

5 Conclusion

Nous avons présenté un algorithme élaboré par l'équipe CISMeF afin d'établir une catégorisation synoptique de spécialités médicales reflétant les thèmes abordés dans les ressources du catalogue CISMeF avec une pondération majeur/mineur. La catégorisation automatique que nous proposons est fondée sur l'indexation manuelle des ressources par des couples (mots clés/qualificatif) MeSH et des types de ressources. Elle exploite également les liens sémantiques existant entre les éléments de la terminologie CISMeF. Une évaluation sur 123 ressources choisies au hasard fournit des résultats très satisfaisants, qui ont permis d'une part d'enrichir la terminologie CISMeF, et d'autre part de montrer qu'il était tout à fait pertinent de mettre l'algorithme introduit en production dans le catalogue CISMeF.

Références

[1] Darmoni SJ, Leroy JP, Thirion B, Baudic F, Douyère M, Piot J. CISMeF: a structured Health resource guide. - *Methods of Information in Medicine* 2000; Jan;39(1) 30-

[2] Nelson SJ, Johnston D, Humphreys BL. Relationships in Medical Subject Headings. In : Bean CA, Green R. (eds.). *Relationships in the organization of knowledge*. New York: Kluwer Academic Publishers; 2001. p.171-84.

[3] Soualmia LF, Barry-Greboval C, Abdulrab H, Darmoni SJ. Modélisation et représentation des connaissances dans un catalogue de santé. *Journées Francophones d'Ingénierie des Connaissances*; IC'2002, Rouen, France pp 139-149.
[4] Thirion B, Darmoni SJ. Simplified access to MeSH Tree Structures on CISMeF. *Bull Med Libr Assoc*, 1999; *4, 480-481, 87.*
[5] Wilcox A, Hripcsak G, Classification Algorithms Applied to Narrative Reports, *Proc AMIA 1999*; Symp. :455-9.
[6] Teahan, Harper. Using compression based language models for text categorization. In: Callan J, Croft B and Lafferty J, (eds.), *Workshop on Language Modelling and Information Retrieval* 2001. pp 83-8.
[7] Dumais S et al. Using SVMs for text categorization. In: Hearst M, (ed), *IEEE Intelligent Systems Magazine, Trends and Controversies*, 13(4).
[8] Han et Kamber, Mining Text Databases, In Data Mining – *Concepts and Techniques, Morgan Kaufman Series in Data Management Systems* 2001 ; pp. 428-44
[9] Bodenreider O. Using UMLS semantics for classification purposes. *Proceedings of AMIA* 2000, Annual Symposium 2000:86-90.

Adresse de correspondance

Adresse Postale: Aurélie NEVEOL
Direction Informatique et Réseaux, CHU de Rouen
1, rue de Germont – 76031 Rouen

Adresse électronique: aurelie.neveol@chu-rouen.fr