Analyse de l'apprentissage humain dans la plateforme SIDES 3.0 : une approche basée sur la sémantique

Oscar Rodríguez Rocha, Catherine Faron Zucker

Analyse de l'apprentissage humain dans la plateforme SIDES 3.0 : une approche basée sur la sémantique ⋆

Oscar Rodríguez Rocha, Catherine Faron Zucker

University Côte d'Azur, CNRS, Inria, I3S, France
oscar.rodriguez-rocha@inria.fr,faron@unice.fr

Résumé : SIDES 3.0 est un projet national français visant à fournir aux étudiants en médecine des services intelligents pour soutenir l'apprentissage en ligne dans le Système Intelligent d'Enseignement en Santé 3.0 (SIDES). La plateforme SIDES contient un grand nombre de ressources d'apprentissage annotées, notamment des questions de formation et d'évaluation, et recueille les traces d'apprentissage des étudiants. Ces annotations de ressources et traces d'apprentissage ont été intégrées sous la forme d'un graphe RDF, et enrichies grâce à des ontologies. Cet article présente les résultats de l'analyse de l'apprentissage des étudiants dans la plateforme SIDES en exploitant le graphe de connaissances associé en reposant sur les technologies du Web sémantique. Cette analyse est préliminaire à la conception et mise en oeuvre des fonctionnalités visant à permettre un apprentissage personnalisé et adaptatif sur la plateforme.

Mots-clés : e-Education, eHealth, Web sémantique, Représentation des connaissances et Raisonnement

1 Introduction

Depuis 2013, les facultés de médecine en France utilisent une plateforme nationale commune qui permet à leurs enseignants de créer et d'appliquer des tests d'évaluation locaux, qui sont ensuite partagés entre les universités afin de constituer une base de données nationale de tests de formation. Cette plateforme Web a été baptisée *SIDES* [1], acronyme de Système Informatisé Distribué d'Évaluation en Santé. Elle permet de préparer des étudiants en médecine aux Épreuves Classantes Nationales Informatisées (ECNi) depuis 2016.

Le projet français national *SIDES 3.0* vise à faire évoluer la plateforme SIDES vers une solution innovante baptisée Système Intelligent d'Enseignement en Santé 3.0 (SIDES 3.0), offrant des services intelligents centrés sur l'utilisateur, tels que : le suivi individuel, des tableaux de bord enrichis, des recommandations personnalisées, des corrections augmentées pour l'auto-évaluation, un environnement numérique normalisé de partage du savoir.

Pour atteindre cet objectif, le développement de *SIDES 3.0* s'appuie sur les modèles et technologies du Web sémantique et sur l'utilisation systématique des normes internationales en vigueur pour les métadonnées sur les ressources pédagogiques (MLR) [2] et les traces d'apprentissage (xAPI) [3], en les intégrant et les enrichissant par des ontologies.

Dans cet article nous présentons les résultats d'une analyse des ressources et de l'activité des étudiants de la plateforme SIDES, qui repose sur la conception d'un ensemble de requêtes SPARQL permettant d'interroger le graphe de connaissances construit à partir des données de la plateforme, en tenant compte de leur sémantique. Cette analyse axée sur les ressources et l'activité des étudiants dans la base de connaissances OntoSIDES, constitue la base pour concevoir et mettre en œuvre des fonctionnalités orientées à permettre un apprentissage personnalisé et adaptatif sur la plateforme *SIDES 3.0*. Après une brève présentation de travaux

⋆. Ce travail a été réalisé dans le cadre du projet DUNE SIDES 3.0 soutenu par l'Agence Nationale de Recherche (ANR-16-DUNE-0002-02).

1. http://side-sante.org
2. https://www.iso.org/standard/62845.html
3. https://xapi.com

connexes, nous commençons par une présentation du graphe de connaissances OntoSIDES, puis nous décrivons les résultats de l'analyse que nous avons menée de la plateforme SIDES, en termes de ressources pédagogiques et d'activité des apprenants, en montrant chaque fois quels types de requêtes permettent d'obtenir les résultats.

2 Travaux connexes

L'analyse de l'apprentissage est définie comme la mesure, la collecte, l'analyse et la présentation de données sur les élèves et leurs contextes, afin de comprendre et d'optimiser leur apprentissage (Ferguson, 2012). A notre connaissance, aucun des travaux existants ne présente une approche basée sur la sémantique pour l'analyse de l'apprentissage dans le domaine de la formation médicale. Des travaux connexes peuvent être trouvés dans l'analyse de l'apprentissage en exploitant les technologies du Web sémantique.

Dans (d'Aquin & Jay, 2013), les auteurs présentent une méthode qui exploite les connaissances externes disponibles sur le LOD pour faciliter l'interprétation des résultats d'exploration de données non sémantiques, en créant automatiquement une structure de navigation et d'exploration dans les résultats. Les résultats de l'exploration de données sont présentés de manière compatible avec une représentation LOD, puis sont liés aux sources du LOD existantes de sorte que l'analyste peut facilement explorer les résultats enrichis. Comparée à cette approche, celle que nous proposons n'exploite pas le LOD pour l'interprétation mais directement des données sémantiques dans le graphe OntoSIDES et pour cette même raison, notre analyse de l'apprentissage ne repose pas sur un algorithme d'exploration de données, mais sur des requêtes SPARQL.

MeLOD (Fulantelli *et al.*, 2013) est un environnement mobile conçu pour prendre en charge, via l'utilisation d'appareils mobiles, les expériences d'apprentissage informelles lors de la visite d'une ville. MeLOD exploite les technologies du Web sémantique pour prendre en charge les expériences d'apprentissage mobile et soutient l'analyse des apprentissages en fournissant des outils spécifiques pour analyser les activités des élèves. Comparée à ces travaux, notre approche s'intéresse à l'apprentissage dans le domaine de la médecine, ce qui change complètement l'objectif et le but de l'analyse d'apprentissage.

Dans (Softic *et al.*, 2013), les auteurs présentent les résultats d'une analyse des activités d'apprentissage basée sur le comportement des utilisateurs dans leur environnement d'apprentissage personnel à l'Université de technologie de Graz. Ils utilisent les technologies du Web sémantique pour mettre en place un tableau de bord d'analyse d'apprentissage pour la visualisation de métriques. La création d'un tableau de bord avec des métriques d'apprentissage pour les étudiants n'est pas l'objectif final de notre travail, mais c'est une perspective que nous discuterons avec les médecins impliqués dans le projet car cela pourrait aider les étudiants à améliorer leur expérience d'apprentissage.

Dans (Dietze *et al.*, 2017) les auteurs présentent le jeu de données LAK en RDF, un corpus de travaux de recherche dans les domaines de Learning Analytics et Educational Data Mining qui permet l'investigation et l'analyse de l'évolution des disciplines scientifiques et la validation de méthodes et d'outils scientométriques.

3 OntoSIDES

OntoSIDES (Palombi *et al.*, 2019) est un graphe de connaissances qui comprend une ontologie de domaine et un ensemble de déclarations factuelles sur des entités manipulées par la plateforme SIDES, reliant celles-ci aux classes et propriétés de l'ontologie. L'ontologie de domaine est représentée en OWL et les connaissances factuelles dans le modèle RDF. Il est ainsi possible d'interroger OntoSIDES avec le langage de requête standard SPARQL. Le graphe de connaissances OntoSIDES a été généré automatiquement à partir de la base de données relationnelles de la plateforme SIDES, et en enrichissant ces données à l'aide de l'ontologie développée. La version actuelle de l'ontologie OntoSIDES contient 52 classes et 50 propriétés. Les classes suivantes sont centrales dans la modélisation :

Action (`sides:action`) la classe des actions possibles des étudiants lorsqu'ils interagissent avec les ressources pédagogiques de la plateforme SIDES. Par exemple, avec la sous-classe `sides:action_to_answer` il est possible de caractériser l'action de sélectionner la proposition d'une réponse à une question.

Content (`sides:content`) la classe racine de la hiérarchie des types de ressources disponibles dans la plateforme SIDES. La classe des questions (`sides:question`), celle des propositions de réponse à une question (`sides:proposal_of_answer`) et celle des réponses (`sides:answer`) d'un étudiant à une question sont des sous-classes de `sides:content`. La Figure 2 présente un graphe RDF décrivant une action de réponse à une question d'un étudiant, qui utilise ces trois classes.

Referential entity (`sides:referential_entity`) la classe des éléments de référence du programme d'éducation français en médecine publié par le Ministre de l'Enseignement Supérieur.

Medical schools (`sides:institute`) la classe des universités et facultés de médecine dans lesquelles des plateformes locales SIDES sont déployées.

Person (`sides:person`) la classe des personnes impliquées dans les études de médecine. Ses sous-classes correspondent aux rôles spécifiques des utilisateurs de la plateforme SIDES : par exemple, la classe `sides:student` est une sous-classe de `sides:person`.

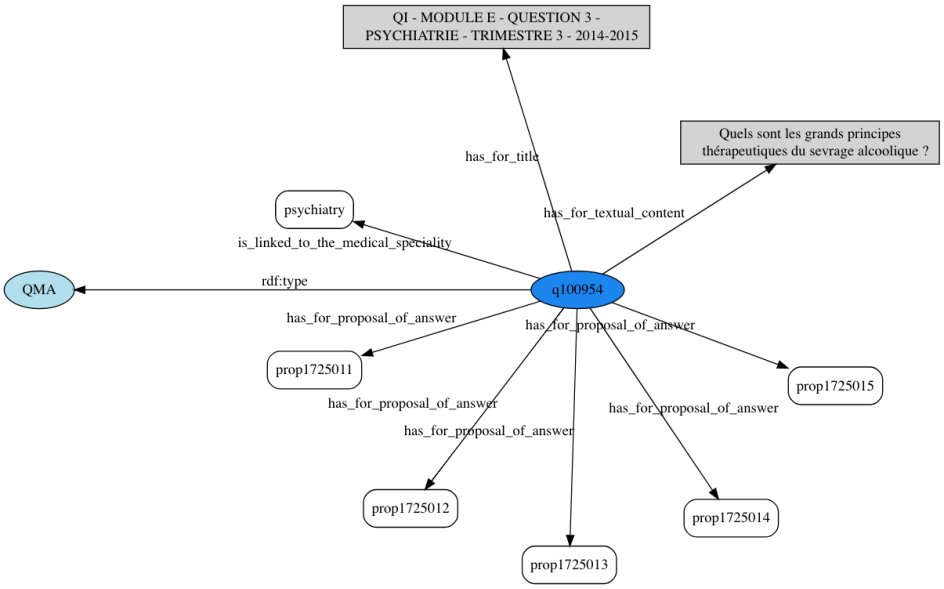

FIGURE *1 – Graphe RDF décrivant une question à réponse multiple (QMA)*

4 Analyse des ressources de la plateforme SIDES

Cette section vise à fournir une caractérisation du contexte de l'apprenant à travers des informations quantitatives et qualitatives sur les ressources présentes dans la plateforme SIDES. Ces informations ont toutes été calculées à l'aide de requêtes SPARQL appliquées sur le graphe OntoSIDES, dont certaines sont fournies dans la suite.

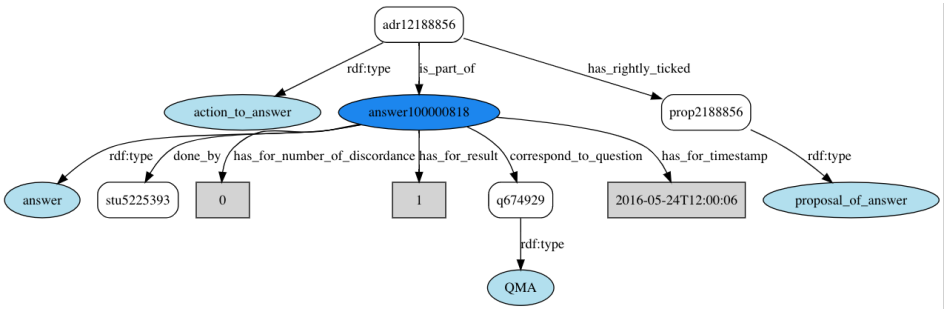

FIGURE *2 – Graphe RDF décrivant une réponse à une question*

4.1 Analyse des questions dans le graphe OntoSIDES

Le graphe de la plateforme SIDES contient actuellement un total de 590,654 questions différentes réparties en 4 catégories comme représenté sur la Figure 3. Cette distribution des questions a été calculée à l'aide de la requête 1. Comme cela apparaît clairement sur la Figure 3, la répartition des questions selon ces catégories n'est pas uniforme : Une grande majorité des questions dans la plateforme SIDES sont des questions à réponses multiples (QMA) : 467,498 questions, soit 79 ;1% ; il y a 81,155 questions à réponse unique (QUA), soit 13,7%, 40,249 questions rédactionnelles ouvertes courtes (QSOA), soit 6,8%, et seulement 1,752 questions de test de concordance de script (TCS), soit 0,3%.

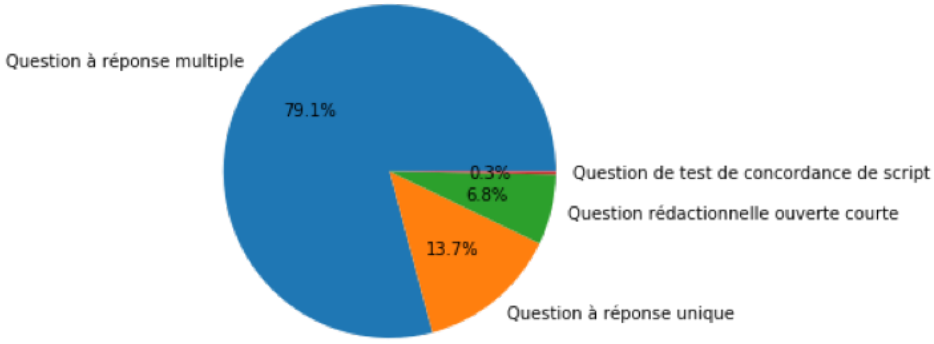

FIGURE *3 – Diagramme circulaire des catégories de questions*

```
SELECT ?type ?label (count(DISTINCT ?question) as ?count)
WHERE { ?question rdf:type ?type . ?type rdfs:subClassOf sides:question .
        ?type rdfs:label ?label FILTER (lang(?label) = 'fr')
} GROUP BY ?type ?label ORDER BY DESC(?count)
```

Query 1 – Requête SPARQL pour calculer la distribution des questions par type

Les descriptions de ces questions ne sont pas homogènes et souvent incomplètes. Notamment, les spécialités et objectifs d'apprentissage auxquelles les questions sont relatives qui permettent de bien caractériser le contexte d'apprentissage sont très peu mentionnés : 50,550 questions QMA, soit seulement 10.81%, sont liées à une spécialité médicale (i.e. décrites avec la propriété `sides:is_linked_to_speciality`), et 54,497 questions, soit seulement 11,66%, sont liées à un à un objectif d'apprentissage (i.e. décrites avec la propriété `sides:is_linked_to_ECN_referential_entity`). De façon similaire, 5,181 questions QUA, soit 6.38%, sont liées à une spécialité et 5,912 questions QUA, soit

7,28%, sont liées à un objectif d'apprentissage ; 1,034 questions QSOA, soit 2.57%, sont liées à une spécialité et 1,461 questions QSOA, soit 3,63% sont liées à un à un objectif d'apprentissage. Enfin, aucune question TCS n'est liée à une spécialité ou un objectif d'apprentissage.

4.2 Analyse des spécialités dans le graphe OntoSIDES

La plateforme SIDES contient actuellement des questions relatives à 31 spécialités médicales (instances de la classe `sides:speciality`). Le tableau 1 présente la répartition des questions par spécialité. Il a été construit à l'aide de la requête 2 qui calcule pour chaque spécialité le nombre de questions associées (une question peut être associée à plusieurs spécialités). Ce tableau montre notamment que la répartition des questions par spécialité n'est pas uniforme, variant de 189 pour la spécialité *Toxicologie* à 5440 pour la spécialité *Pédiatrie*, avec une moyenne de 2242 questions, et un écart type de 1283.

	QMA	QUA	QSOA	Total	%
Pédiatrie	4867	516	57	5440	7.83
Maladies infectieuses	4126	487	73	4686	6.74
Cardio-vasculaire	3550	300	86	3936	5.66
Endocrinologie - Métabolisme - Nutrition	3276	233	38	3547	5.10
Cancérologie - Radiothérapie	3042	300	75	3417	4.92
...					
Chirurgie maxillo-faciale	649	77	9	735	1.06
Neurochirurgie	553	51	13	617	0.89
Médecine du travail	496	59	8	563	0.81
Addictologie	286	43	9	338	0.49
Toxicologie	176	13	0	189	0.27

TABLE *1 – Nombre et proportion de questions associées à chaque spécialité*

```
SELECT ?speciality ?q_type (COUNT(DISTINCT ?question) AS ?questions)
WHERE { ?question rdf:type ?question_type .
        ?q_type rdfs:subClassOf sides:question .
        ?question sides:is_linked_to_the_medical_speciality ?speciality  .
} GROUP BY ?speciality ?q_type ORDER BY ASC(?speciality)
```

Query 2 – Requête SPARQL pour calculer le nombre de questions de chaque type associées à chaque spécialité

4.3 Analyse des objectifs d'apprentissage dans le graphe OntoSIDES

La plateforme SIDES contient actuellement des questions relatives à 362 objectifs d'apprentissage (instances de la classe `sides:ECN_learning_objective`) et 921 sous-objectifs d'apprentissage (instances de `sides:ECN_learning_sub_objective`). Ils ont été dénombrés avec la requête 3 qui prend en compte les relations de subsomption entre objectifs.

```
SELECT ?type ?label (count(DISTINCT ?lo) as ?count)
WHERE { ?lo rdf:type ?type . ?type rdfs:subClassOf* sides:ECN_referential_entity .
        ?type rdfs:label ?label FILTER (lang(?label) = 'fr')
} GROUP BY ?type ?label
```

Query 3 – Requête pour calculer le nb d'objectifs et sous objectifs d'apprentissage

Des requêtes similaires à celles utilisées pour l'analyse des spécialités permettent d'analyser la répartition des questions par objectifs d'apprentissage. De même que la répartition des questions par spécialité, celle des questions par objectif n'est pas uniforme, le nombre de questions associées à un objectif d'apprentissage variant de 7 pour l'objectif *Dopage* à 1651

pour l'objectif *Prescription et surveillance des classes de médicaments les plus courantes chez l'adulte et chez l'enfant*, avec une moyenne de 196 questions, et un écart type de 149.

5 Analyse de l'activité des étudiants dans le graphe OntoSIDES

Cette section fournit le résultat de l'analyse de l'activité des apprenants sur la plateforme SIDES réalisée en interrogeant le graphe OntoSIDES avec des requêtes SPARQL dédiées. Etant données les contraintes de longueur d'article imposée, nous nous sommes concentrés ici sur l'analyse de l'activité selon les spécialités. Un travail similaire a été conduit sur l'analyse de l'activité par objectif d'apprentissage.

5.1 Analyse de l'activité des étudiants par question

À ce jour, 64,957 étudiants (instances de la classe `sides:student`) sont identifiés sur la plateforme SIDES, mais seuls 41,442 étudiants on réalisé au moins une action, soit 63%. On constate que l'activité des étudiants actifs n'est pas uniforme, le nombre total de réponses données par étudiant variant de 1 à 62,015.

Le graphe OntoSIDES contient la description de 100,812,181 réponses à 456,854 questions, donc seules 77,34% des questions ont reçu au moins une réponse et chaque question a reçu en moyenne 221 réponses.

5.2 Analyse de l'activité des étudiants par spécialité

Toutes les spécialités ont été abordées par au moins un étudiant et en moyenne 17,600 étudiants ont abordé au moins une question de chaque spécialité. Cependant ce nombre n'est pas uniforme, variant de 6,111 pour *Toxicologie* à 24,461 pour *Maladies infectieuses*, avec un écart-type de 4,420.

Pour analyser plus finement l'activité des étudiants selon les spécialités, nous avons également calculé à l'aide de la requête 4 le nombre de réponses à des questions par spécialité et le nombre moyen de réponses à des questions par étudiant dans chaque spécialité. La figure 4 montre les résultats obtenus à partir de cette requête, les spécialités étant triées par ordre décroissant du nombre moyen de réponses par étudiant. On constate ainsi que *Maladies infectieuses* est la spécialité ayant reçu le plus grand nombre de réponses à des questions, mais que *Pédiatrie* est la spécialité avec le plus grand nombre de questions par étudiants. Une interprétation possible est que les étudiants intéressés par la *Pédiatrie* ont été les plus actifs dans cette spécialité. De même, les étudiants les moins actifs dans uns spécialité ont été ceux qui s'intéressaient à la *Toxicologie*.

```
SELECT ?label (COUNT(DISTINCT ?answer) AS ?answers) (COUNT(DISTINCT ?student) AS ?students)
WHERE { ?answer rdf:type sides:answer .
        ?answer sides:correspond_to_question ?question .
        ?answer sides:done_by ?student .
        ?question sides:is_linked_to_the_medical_speciality ?speciality  .
        { SELECT ?speciality (MIN(?duplicated_label) AS ?label)
          WHERE { ?speciality a sides:speciality .
                  ?speciality rdfs:label ?duplicated_label .
                  FILTER (lang(?duplicated_label) = "fr")
          } GROUP BY ?speciality ORDER BY ?label }
} GROUP BY ?label ORDER BY ?label
```

Query 4 – Requête pour calculer le nb d'étudiants et le nb de réponses par spécialité

Pour approfondir encore notre analyse, nous nous sommes également intéressés à la qualité des réponses des étudiants aux questions par spécialité. Nous nous sommes limités à l'analyse des réponses aux questions à réponse unique (QUA) dont le résultat est binaire (correct ou faux). Nous pourrions étendre l'analyse aux questions à réponses multiples (QMA), en considérant un seuil au-delà duquel considérer que le nombre d'options correctement sélectionnées pour une répondre à une question constitue une réponse correcte. La requête 5 permet de compter les nombres de réponses correctes et incorrectes à des questions QUA et la Figure 5 présente les résultats obtenus, les spécialités étant triées par ordre décroissant du

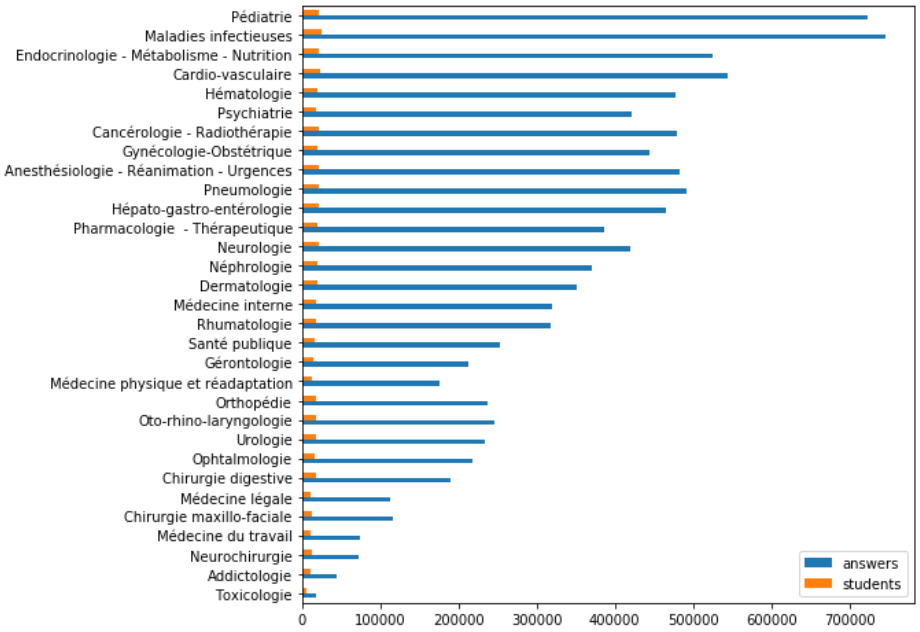

FIGURE *4 – Nombre de réponses et d'étudiants ayant répondu pour chaque spécialité. Les spécialités sont triées par ordre décroissant d'activité des étudiants mesuré par le nombre moyen de réponses par étudiant*

ratio entre le nombre de réponses correctes et le nombre de réponses incorrectes. On constate ainsi notamment que la spécialité *Maladies infectieuses* est celle ayant le plus grand nombre de réponses correctes, la spécialité *Médecine physique et réadaptation* est celle avec la plus grande proportion de réponses correctes, et la spécialité *Toxicologie* est celle avec la plus faible proportion de réponses correctes (il y a même davantage de réponses incorrectes que correctes). Ceci, combiné aux résultats de la requête précédente, montre que c'est la spécialité dans laquelle les étudiants répondent le moins et le moins bien. Une interprétation possible est que la spécialité *Toxicologie* est celle pour laquelle les concepts à acquérir sont les plus difficiles, tandis que la *Médecine physique et réadaptation* celle manipulant les concepts les plus simples. Une autre interprétation peut être donnée en terme de qualité des questions conçues pour chaque spécialité, les questions les moins bien répondues pouvant nécessiter une révision.

```
SELECT ?label (sum(if(?result = 1, 1, 0)) as ?corrects)
(sum(if(?result = 1, 0, 1)) as ?wrongs) (count(?answer) as ?answers)
WHERE { ?answer rdf:type sides:answer .
        ?answer sides:correspond_to_question ?question .
        ?question a sides:QUA .
        ?question sides:is_linked_to_the_medical_speciality ?speciality .
        ?answer sides:has_for_result ?result .
        { SELECT ?speciality (MIN(?duplicated_label) AS ?label)
          WHERE { ?speciality a sides:speciality .
                  ?speciality rdfs:label ?duplicated_label .
                  FILTER (lang(?duplicated_label) = "fr")
          } GROUP BY ?speciality ORDER BY ?label }
} ORDER BY DESC(?corrects)
```

Query 5 – Requête pour extraire le résultat des réponses aux questions pour chaque spécialité

6 Conclusions et perspectives

Nous avons présenté un premier travail d'analyse des ressources de la plateforme d'apprentissage SIDES et de l'activité des étudiants sur la plateforme, basé intégralement sur

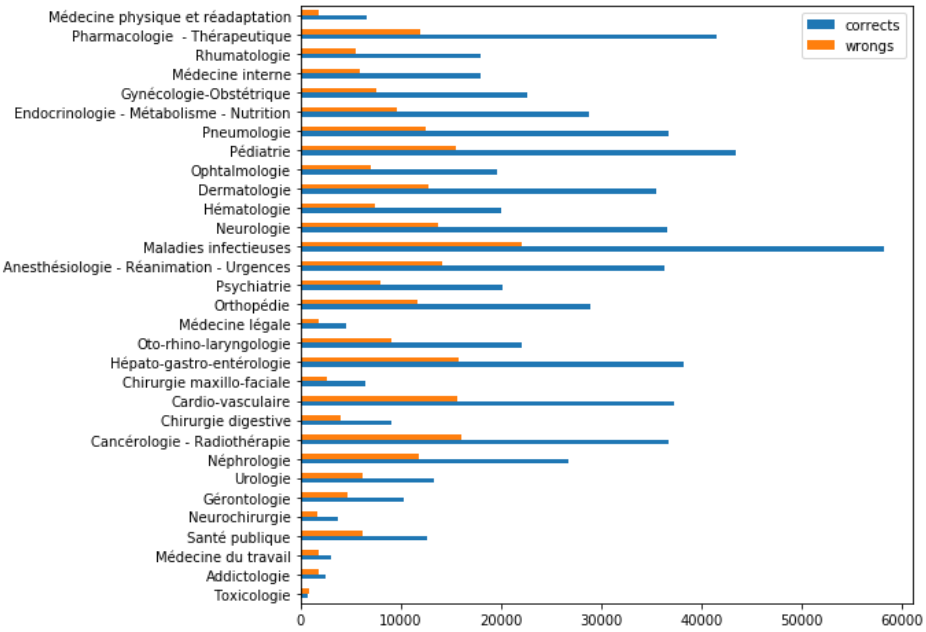

FIGURE *5 – Nombres de réponses correctes et incorrectes à des questions QUA pour chaque spécialité. Les spécialités sont triées par ordre décroissant de difficulté mesurée par le ratio entre les nombres de réponses correctes et incorrectes*

l'interrogation du graphe de connaissance OntoSIDES en RDF avec des requêtes SPARQL. Le résultat des analyses peut donner lieu à différentes interprétations possibles, en terme de suivi individuel des apprenants mais aussi en terme de retours sur la popularité, la difficulté, voire la qualité des ressources disponibles sur la plateforme. A court terme nous allons discuter et valider ces résultats et interprétations avec les médecins impliqués dans le projet. Les résultats de ce travail nous serviront pour concevoir et mettre en œuvre des fonctionnalités orientées vers un apprentissage adaptatif et personnalisé dans la plateforme, telles que la recommandation de questions en fonction du niveau de connaissances, des objectifs d'apprentissage et des spécialités médicales de l'étudiant.

Références

D'AQUIN M. & JAY N. (2013). Interpreting data mining results with linked data for learning analytics : Motivation, case study and directions. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, LAK '13, p. 155–164, New York, NY, USA : ACM.

DIETZE S., TAIBI D. & D'AQUIN M. (2017). Facilitating scientometrics in learning analytics and educational data mining - the lak dataset. *Semantic Web*, **8**, 395–403.

FERGUSON R. (2012). Learning analytics : drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, **4**(5/6), 304–317.

FULANTELLI G., TAIBI D. & ARRIGO M. (2013). A semantic approach to mobile learning analytics. In *Proceedings of the First International Conference on Technological Ecosystem for Enhancing Multiculturality*, TEEM '13, p. 287–292, New York, NY, USA : ACM.

PALOMBI O., JOUANOT F., NZIENGAM N., OMIDVAR-TEHRANI B., ROUSSET M.-C. & SANCHEZ A. (2019). Ontosides : Ontology-based student progress monitoring on the national evaluation system of french medical schools. *Artificial Intelligence in Medicine*, **96**, 59 – 67.

SOFTIC S., TARAGHI B., EBNER M., DE VOCHT L., MANNENS E. & VAN DE WALLE R. (2013). Monitoring learning activities in ple using semantic modelling of learner behaviour. In *Human Factors in Computing and Informatics*, p. 74–90 : Springer Berlin Heidelberg.