Extraction de relations complexes. Application à des résultats expérimentaux en physiologie rénale

Anne-Lyse Minard, Brigitte Grau, Anne-Laure Ligozat, Stephen Randall Thomas

Extraction de relations complexes : application à des résultats expérimentaux en physiologie rénale

Anne-Lyse Minard[*,**] — **Brigitte Grau**[*,***] — **Anne-Laure Ligozat**[*,***] — **Stephen Randall Thomas**[**,****]

[*] *LIMSI-CNRS, rue John von Neumann, 91400 Orsay, France*
[**] *Université Paris-Sud, 91400 Orsay, France*
[***] *ENSIIE, 1 square de la résistance, 91000 Évry, France*
[****] *IR4M CNRS UMR 8081, 94805 Villejuif, France*
prenom.nom@limsi.fr, stephen-randall.thomas@u-psud.fr

RÉSUMÉ. *Dans le domaine biologique, des bases de données ont été créées pour que les données expérimentales contenues dans les articles scientifiques puissent être retrouvées et utilisées plus facilement. Dans cet article, nous présentons un modèle d'extraction d'information permettant d'alimenter semi-automatiquement une base de données avec des résultats d'expérimentations, avec comme cadre d'application la physiologie rénale quantitative. L'extraction d'information est guidée par une ressource termino-ontologique qui exploite des patrons d'extraction et un lexique. Elle a été intégrée dans un assistant guidant l'expert pour remplir la base de données. Nous avons évalué le système d'extraction d'information intrinsèquement et extrinsèquement, et montrons que le peuplement de la base de données par les experts est rendu plus rapide et plus complet grâce à l'outil proposé.*

ABSTRACT. *Manual extraction of relevant information for populating a database is very-time consuming and error-prone. In this paper, we present an information extraction model to help data entry into a database of experimental results. The domain is quantitative renal physiology. Information extraction is driven by an Ontological and Terminological Resource and uses patterns and identification of domain terms and their variations. It was integrated into an assistant, which helps experts to populate the database. We evaluated the system both intrinsically and extrinsically, by evaluating the assistant. The evaluation shows that populating of the database by experts is faster and more complete using our assistant.*

MOTS-CLÉS : *extraction d'information, ontologie, peuplement d'une base de données*

KEYWORDS: *information extraction, ontology, populating a database*

1. Introduction

Une grande part des connaissances en biologie se trouve dans des articles scientifiques et c'est en particulier le cas des résultats d'expérimentations. Ces résultats sont exploités par des modèles mathématiques pour explorer des hypothèses complexes ou, dans un but plus heuristique, pour donner une représentation cohérente de l'état des connaissances. La mise en place de modèles mathématiques, leur calibration et leur validation s'appuient sur une connaissance exhaustive des mesures quantitatives expérimentales ; or, du fait de l'explosion de la quantité d'informations disponibles dans la littérature, la maîtrise de cette information est très difficile.

Devant ce constat, nous nous sommes intéressés au problème du peuplement de bases de données de résultats expérimentaux : nous avons mis en œuvre des méthodes d'extraction d'information, et les avons intégrées dans une interface d'aide au peuplement de bases de données.

Notre cadre d'application est la base de données du domaine de la physiologie rénale Quantitative Kidney DataBase, ou QKDB (Dzodic *et al.*, 2004), qui a été créée dans le contexte du projet Physiome international (Hunter *et al.*, 2010). Elle s'adresse à l'ensemble de la communauté de physiologie rénale et de néphrologie, et vise à rassembler les résultats expérimentaux quantitatifs utiles pour la modélisation à partir de leurs publications dans les articles de physiologie rénale. Une version générique (QxDB, voir (Ribba *et al.*, 2006)), épurée de toute spécificité du domaine du rein, a aussi été développée et commence à être adoptée dans d'autres domaines tels que la croissance tumorale.

Construite en 2004, QKDB a été peuplée manuellement ; à l'heure actuelle environ 8 500 résultats ont été enregistrés dans la base, provenant de quelques 300 articles. Toutefois, et malgré l'utilité d'une telle ressource et la mise en place d'une interface web conviviale, la participation espérée des chercheurs du domaine au remplissage de la base est restée faible à cause du temps nécessaire pour entrer les valeurs et les commenter. Ce sérieux problème de *curation*, c'est-à-dire du peuplement manuel de la base de données, est commun à beaucoup d'autres bases de données dans le domaine de la biologie. Nous avons donc conçu un système d'extraction automatique des informations qui décrivent des résultats expérimentaux.

Nous appelons *résultat expérimental*, un résultat quantitatif obtenu suite à une expérience et mis en relation avec les informations permettant de décrire cette expérience. Pour illustrer ce qu'est un résultat expérimental, voici un exemple provenant d'un article scientifique :

Apical membrane P_f averaged (in cm/s) **9.37 +- 0.77 e-4** (n = 5) at 20°C

La phrase indique que l'expérience consistait à mesurer la perméabilité (P_f) de la membrane apicale, à une température de 20°C sur 5 individus (l'espèce n'est pas précisée dans la phrase). Le résultat de cette expérience est exprimé en cm/s. L'ensemble de ces informations sera appelé résultat expérimental. Les informations décrivant un résultat expérimental peuvent être concentrées au sein d'une seule phrase mais aussi être

renseignées dans différentes parties de l'article (*Methods*, *Results*, etc.), dans le texte ou dans un tableau. Le système d'extraction doit donc analyser les articles complets. Peu de systèmes traitent d'articles complets, et ceux qui existent, comme BioRAT (Corney *et al.*, 2004) et Pharmspresso (Garten *et al.*, 2009), extraient des informations en relation au sein d'une même phrase.

Nous formalisons un résultat expérimental par une relation n-aire. Une relation n-aire est une relation entre plus de deux entités (une relation entre deux entités est appelée relation binaire). Les entités en relation constituent les arguments de la relation. La reconnaissance de relation n-aire repose sur l'identification d'un élément pivot auquel sont associés les différents descripteurs. Dans le cas de résultats expérimentaux, la relation elle-même ainsi que les liens avec les différents descripteurs ne sont généralement pas formulée explicitement dans les textes. La reconnaissance de ces relations consistera à identifier des termes descripteurs de la relation puis à les relier.

Cette reconnaissance est guidée par les connaissances du domaine, nous situant ainsi dans le cadre de l'extraction d'information guidée par une ontologie (voir (Wimalasuriya *et al.*, 2010) pour une synthèse). La description d'expérimentations fait appel à des concepts relevant d'ontologies de domaines différents et nous ne disposons pas d'une ontologie unifiée.

Nous avons formalisé la relation n-aire du domaine sous forme d'une ressource termino-ontologique à partir des concepts choisis lors de la conception de la base de données, et des termes qui leur avaient été associés. Nous avons dû compléter les lexiques afin de mieux couvrir la variabilité d'expression linguistique présente dans les articles.

Enfin, l'objectif de ce travail étant d'aider les experts du domaine à peupler la base de données, nous avons conçu un assistant qui utilise le système d'extraction d'information. L'assistant permet de choisir le texte sur lequel on veut travailler, propose des valeurs à insérer dans la base suite à l'application du processus d'extraction, et permet de modifier celles-ci en maintenant toujours la relation au texte. De tels outils permettent d'accélérer le travail de l'expert qui procède à l'analyse d'un article et facilitent le travail de curation qui a lieu avant insertion définitive dans la base ((Van Auken *et al.*, 2009), (Alex *et al.*, 2008), (Donaldson *et al.*, 2003)). Nous montrons l'apport de ce travail par une évaluation auprès d'utilisateurs.

Dans cet article, notre contribution porte sur différents points, liés à la biologie et à l'extraction d'information. Dans le domaine de la biologie, nous avons créé une aide au peuplement d'une base de données, cette aide ayant donné lieu à une évaluation utilisateur. Du point de vue de l'extraction, nous avons défini un système d'identification de relations n-aires guidée par une ressource termino-ontologique et associée à un processus de reconnaissance de variantes terminologiques.

Après avoir présenté l'état de l'art dans la section 2, nous présenterons section 3 le modèle que nous avons défini pour formaliser la notion de résultat expérimental. Nous décrirons ensuite les méthodes utilisées pour extraire les résultats d'expérimentations

(section 4) et l'architecture de l'assistant d'aide au peuplement de la base de données (section 5). Dans la section 6, nous présenterons les corpus et lexique utilisés, puis les résultats de l'évaluation du système d'extraction et de l'assistant. Nous terminerons par une discussion sur l'apport de l'outil développé et son adaptation possible à d'autres domaines (section 7).

2. État de l'art

Notre travail se situant à la fois dans le domaine de l'extraction d'information, avec la reconnaissance de relations complexes et de variantes de termes, et de l'aide au peuplement de bases de données, nous présenterons les travaux existants selon ces différents points de vue.

2.1. *Extraction de relations complexes et d'événements*

L'extraction de relations n-aires ou complexes est proche de la détection d'événements. Dans le domaine médical et biologique, la notion d'événement est utilisée pour désigner par exemple le changement d'état d'une molécule en biologie, ou encore l'ensemble des informations concernant l'administration d'un traitement en médecine, qui relient alors des termes du domaine. Dans le domaine général, beaucoup de travaux ont été proposés dans le cadre de la conférence ACE 2005 [1]. Les événements y sont décrits comme des structures complexes reliant des arguments qui sont eux mêmes complexes. Dans le corpus ACE 2005, huit types d'événements sont annotés. Par exemple, l'événement LIFE peut avoir plusieurs arguments, et a cinq sous-types : be-born, marry, divorce, injure et die. Une étape importante concerne l'identification du déclencheur ; (Ahn, 2006) la traite comme une tâche de classification de mots et effectue une classification binaire pour détecter les mots déclencheurs, puis une classification multi-classes pour déterminer le type de l'événement.

(McDonald *et al.*, 2005) proposent une méthode d'extraction de relations complexes, et plus précisément d'extraction d'événements de variations génomiques, qui se décompose en deux étapes. Premièrement, ils détectent des paires d'entités en relation en faisant une classification binaire (grâce à un classifieur basé sur le modèle d'entropie maximale), ensuite ils reconstruisent les relations complexes en sélectionnant les cliques maximales dans le graphe des relations.

La tâche de BioNLP'09 a porté sur l'extraction d'événements bio-moléculaires, événements qui décrivent le changement d'état d'une bio-molécule [2]. Dans l'exemple suivant, trois événements sont décrits : « phosphorylation » (avec comme argument TRAF2), « binding » (avec comme arguments « TRAF2 » et « CD40 ») et « nega-

1. http ://www.itl.nist.gov/iad/mig/tests/ace/2005/
2. http ://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/

tive regulation » (indiqué par « inhibits » avec pour arguments les deux événements précédents).

« In this study we hypothesized that the phosphorylation of TRAF2 inhibits binding to the CD40 cytoplasmic domain. »

Trois sous-tâches étaient proposées : a) la détection du déclencheur de l'événement ainsi que son typage, et la reconnaissance de la protéine (ou le gène) qui subit ce changement d'état, b) la reconnaissance du deuxième argument, et c) la détection d'un modifieur de l'événement, par exemple si l'événement est à la forme négative dans le texte. (Björne *et al.*, 2010) associent de l'apprentissage supervisé et l'utilisation de règles pour la détection des événements, et (Nguyen *et al.*, 2010) font de l'apprentissage de patrons. (Buyko *et al.*, 2009), qui sont arrivés deuxièmes pour la tâche 1 de BioNLP'09 avec une F-mesure [3] de 0,46 %, décrivent la tâche d'extraction d'événement en six sous-tâches (entre parenthèses, nous indiquons le résultat de l'extraction de l'événement de l'exemple donné précédemment) :

1) l'identification du déclencheur de l'événement (*gene expression*) ;

2) la désambiguïsation du déclencheur ;

3) le typage de l'événement : classification sémantique du déclencheur et attribution de la catégorie de l'événement (l'événement est du type *GeneExpression*) ;

4) l'identification des arguments de l'événement (*jun* et *fos*) ;

5) le typage des arguments (les deux arguments sont des protéines) ;

6) l'attribution des rôles des arguments à l'intérieur d'un événement (les deux arguments ont le même rôle de *Theme* dans l'événement).

Quatre de ces sous-tâches sont présentes dans notre travail. Dans un premier temps, nous avons à identifier la valeur numérique du résultat expérimental : le déclencheur de l'événement (1). Dans un deuxième temps, nous repérons les arguments de cet événement (4), c'est-à-dire les descripteurs du résultat expérimental, que nous associons à un type (*Species*, *Parameter*, etc.) (5). Et la dernière étape est la mise en relation des descripteurs et de la valeur numérique déclencheur de l'événement (6). Cependant dans les travaux d'extraction d'événements biologiques, la détection des événements porte sur des mots déclencheurs (et non des chiffres) et les arguments sont reliés à ces déclencheurs par des relations syntaxiques. D'autres travaux effectués dans le domaine des interactions protéines-protéines, comme (Ananiadou *et al.*, 2010) ou (Wattarujeekrit *et al.*, 2004), extraient des événements caractérisés par des verbes (ou des substantifs verbaux).

Dans le domaine médical, un événement de médication, comme défini par (Gold *et al.*, 2008), est décrit par un médicament administré à un patient, avec sa dose, sa fréquence, la nécessité de prendre ce traitement, etc. L'extraction d'événements médicaux a fait l'objet de la troisième campagne d'évaluation d'i2b2 2009 (Uzuner *et al.*, 2010). La tâche consistait à annoter, dans des textes cliniques, les médicaments

3. Mesure d'exactitude associant rappel et précision.

administrés à un patient, et pour chaque médicament à donner son dosage, sa fréquence, son mode d'administration, la durée du traitement et la raison du traitement. (Grouin *et al.*, 2010), comme la majorité des équipes participantes, ont utilisé des listes pour la reconnaissance des différents attributs, et des règles pour la mise en relation d'un médicament et des autres attributs. Ils ont obtenu une F-mesure globale de 0,773. (Patrick *et al.*, 2010) utilisent des CRF (Conditional Random Field) pour l'annotation des entités et des SVM (Support Vector Machine) pour la mise en relation des attributs. Ils sont arrivés premier du challenge avec une F-mesure globale de 0,857, mais l'utilisation des techniques d'apprentissage nécessite un corpus annoté de grande taille, ce dont nous ne disposons pas. La structure cherchée est assez analogue à notre problème, mais les attributs relèvent plus de formes figées et une difficulté importante dans cette tâche consistait à repérer les médicaments, qui agissent comme déclencheurs.

2.2. *Reconnaissance des termes*

La reconnaissance de termes constitue une tâche cruciale en extraction d'information. Un premier type d'approche consiste à extraire des termes de manière ascendante, c'est-à-dire à reconnaître des termes en corpus et les regrouper en classes. Si beaucoup de travaux ont proposé des extracteurs de termes, peu permettent de regrouper les variations d'un même terme sous un même concept. Ainsi, (Nenadic *et al.*, 2004) collectent des termes selon une mesure statistique et cherchent à les normaliser en s'appuyant sur des critères syntaxiques. Toutefois les variantes lexicales de synonymie ne sont pas traitées. De plus, les approches statistiques permettent d'acquérir des termes, mais à condition de disposer d'un corpus suffisant, couvrant toute la terminologie voulue.

La connaissance *a priori* du domaine permet d'opérer de manière descendante, guidée par les concepts, et projeter dans le texte les termes associés aux concepts (Wimalasuriya *et al.*, 2010). Les méthodes les plus courantes pour la reconnaissance de termes sont basées sur des terminologies (Aronson *et al.*, 2004), sur des règles linguistiques (Ananiadou, 1994), sur de l'apprentissage (Bodenreider *et al.*, 2002), ou encore sur des mésures statistiques (Dinh *et al.*, 2011).

Le problème qui se pose est de reconnaître toutes les variations linguistiques pour dénommer un concept. L'approche retenue dans Fastr (Jacquemin, 1996), qui permet de reconnaître des variantes de termes multi-mots de différents types (morphologique, syntaxique et sémantique) est fondée sur l'instanciation de métarègles appliquées par un analyseur de surface. Il repose sur des lexiques indiquant les variations au niveau des mots simples. Il reconnaît ainsi un nombre important de variations, tout en restant fiable. Avec SARDINE, (Maynard *et al.*, 2009) se proposent d'étendre une ontologie par la donnée de termes initiaux et l'application de patrons d'extraction de relations ontologiques afin de typer des relations entre termes et extraire de nouvelles classes. Notre corpus ne se prête pas à l'extraction de relations ontologiques, mais nous avons procédé de manière analogue en nous appuyant sur l'expression de relations pour re-

connaître des termes et leur classe de référence (Minard *et al.*, 2010). Ces relations n'étant pas explicites, nous avons eu recours à une approche non supervisée pour acquérir des patrons. (Bodenreider *et al.*, 2002) visent aussi à étendre une terminologie par une approche non supervisée en s'appuyant sur la présence de modifieurs de termes (complément du nom, adjectif, etc.), mais ce travail porte sur la reconnaissance d'hyponymes.

2.3. *Interface d'aide au peuplement d'une base de données*

L'extraction d'information en domaine biomédical a fait l'objet de nombreux travaux, mais la façon d'intégrer les solutions proposées dans un système complet d'annotation a été moins étudiée. Ainsi, les organisateurs de la campagne d'évaluation KDD Challenge Cup 2002, de fouille de textes scientifiques biologiques dans l'optique du remplissage de la base de données FlyBase, estiment la tâche possible au vu des résultats. Mais ils indiquent également qu'il reste à évaluer si les systèmes de fouille améliorent effectivement le remplissage des bases (Yeh *et al.*, 2003).

(Cohen *et al.*, 2005) notent que beaucoup de travail reste à faire pour l'évaluation en contexte réel des systèmes de fouille de textes. Mais cette évaluation est importante car les bases de données existantes sont incomplètes et (Baumgartner *et al.*, 2007) montrent que la complétion manuelle ne sera jamais suffisante.

Les systèmes de fouille de textes peuvent intervenir à plusieurs niveaux : ils peuvent notamment sélectionner des documents ou des passages de documents pertinents, ou pré-annoter les documents avec les informations pertinentes. Dans les deux cas, une mesure simple d'efficacité de l'assistance consiste à mesurer le temps passé par les curateurs pour annoter les documents. (Donaldson *et al.*, 2003) ont calculé le temps passé à remplir une base d'interactions entre protéines, avec ou sans leur système de sélection de documents, et estiment que le gain de temps est de 70%. (Karamanis *et al.*, 2006) ont évalué le temps mis par un curateur pour remplir les champs de la base de données FlyBase à partir d'articles scientifiques, et constatent que ce temps est réduit avec l'aide de leur outil.

Certains travaux ont effectué des analyses un peu plus complètes de l'aide apportée par un outil d'aide au remplissage des bases. (Alex *et al.*, 2008) extraient de l'information sur des interactions entre protéines pour remplir semi-automatiquement une base de données. Ils veulent étudier si les techniques de traitement automatique des langues aident réellement au remplissage, et quels doivent être les paramètres d'un outil d'assistance. Les expériences menées consistent à placer 4 curateurs dans trois conditions expérimentales pour la lecture de 4 articles : sans assistance, avec l'annotation de référence des articles, et avec l'annotation automatique. Ils montrent que l'annotation préalable des articles permet de réduire le temps de remplissage et d'augmenter le nombre d'interactions annotées ; par ailleurs, les curateurs préfèrent une annotation la plus complète possible (pas de résolution des ambiguïtés d'annotation, même si cela réduit le temps de remplissage), et un rappel maximum, i.e. tous les éléments intéres-

sants sont annotés quitte à annoter des éléments non pertinents. Ils montrent aussi que le temps de remplissage ne peut être le seul facteur à considérer pour l'évaluation de tels systèmes.

(Van Auken *et al.*, 2009) utilisent le système de recherche et d'extraction d'information Textpresso pour évaluer son apport au remplissage de l'ontologie Gene Ontology (Ashburner *et al.*, 2000), d'un point de vue de la sélection de documents puis de phrases pertinentes. Ils comparent également le remplissage par trois curateurs avec ou sans annotation du système, sur un ensemble de 20 articles. Les résultats montrent que la vitesse de remplissage est multipliée par 8 à 15 selon le curateur.

En résumé, plusieurs paramètres peuvent être pris en compte pour évaluer l'outil d'aide à l'annotation, le plus simple à utiliser étant le temps mis par un curateur pour remplir une entrée de la base, ainsi que le nombre d'entrées annotées, qui augmente généralement fortement avec une pré-annotation. Il convient également de ne pas négliger le ressenti des curateurs vis-à-vis de l'outil, qui peut contredire les résultats précédents. Enfin, il est important de maximiser le rappel de l'outil afin d'obtenir les meilleurs performances possibles d'annotation.

3. Modèle de connaissances

Dans ce travail, nous nous sommes intéressés à la reconnaissance de résultats expérimentaux dans des articles scientifiques en physiologie rénale. Notre méthode s'appuie sur une modélisation représentée par une base de données générique (QKDB). Cependant, cette base de données n'explicite pas toutes les informations, aussi nous avons formalisé ce modèle sous forme d'une ontologie associée à une terminologie. Nous présenterons d'abord ce modèle, puis expliquerons sa traduction en schéma relationnel.

3.1. *La ressource termino-ontologique*

Une ontologie est généralement composée d'une composante générique, représentant des concepts généraux indépendants du domaine, complétée par une ontologie du domaine, plus éventuellement des ontologies décrivant la tâche et l'application (Guarino, 1998). Une ressource termino-ontologique (RTO) met en relation les concepts de l'ontologie avec leurs dénominations dans la langue, les termes.

Notre objectif était de définir un modèle générique pour représenter un résultat expérimental. Un résultat expérimental est décrit dans les textes par un résultat quantitatif et les différents descripteurs de l'expérimentation qui ont permis de l'obtenir, et peut donc être vu comme une relation n-aire. Les descripteurs forment les concepts du domaine, en l'occurrence ceux de la physiologie rénale.

Les recommandations du W3C (Noy *et al.*, 2006) pour représenter les relations n-aires amènent à représenter une relation par un concept, et à rattacher les éléments

mis en relation par des propriétés. C'est par exemple le choix fait par (Touhami *et al.*, 2011), qui porte sur l'extraction de relations n-aires en microbiologie. En suivant ce type de modélisation, il faudrait créer un concept-relation par expérimentation lorsque le domaine change. Or, les descripteurs de l'expérimentation jouent tous le même rôle vis-à-vis de la relation. De ce fait, il est possible de représenter une relation par un concept générique, lié à un et un seul résultat quantitatif et à un seul type de concept représentant l'ensemble des descripteurs. Ce concept est ensuite précisé par les concepts du domaine.

La figure 1 illustre ce choix : un résultat expérimental est représenté par un concept-relation *ExperimentalResult*. Celui-ci est relié au concept *QuantitativeResult* qui correspond notamment à la valeur numérique du résultat, en précisant qu'il y a un et un seul concept possible, et au concept *ExperimentConcept* qui correspond aux descripteurs du résultat (espèce ou organe concernés par exemple), avec la restriction qu'il y a au moins une valeur de ce concept. Cette modélisation permet de décrire un résultat d'expérimentation, dans quelque domaine qu'il soit, et correspond donc à la partie générique de l'ontologie.

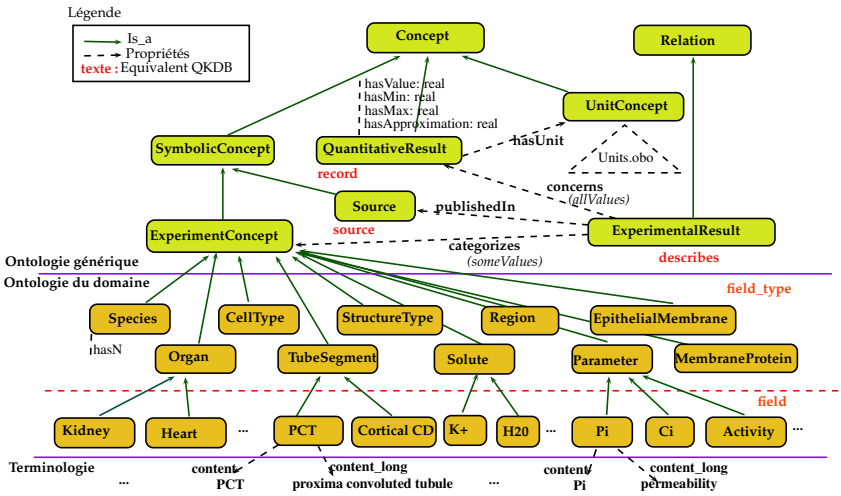

Figure 1. *La ressource termino-ontologique*

Un *résultat quantitatif* (*QuantitativeResult*) est quant à lui décrit par des propriétés permettant d'en donner sa valeur (*hasValue*), ses valeurs minimale et maximale (*hasMin* et *hasMax*), et l'approximation faite (*hasApproximation*).

Chaque résultat étant extrait d'un article scientifique, ce dernier est aussi décrit dans l'ontologie générique (concept *Source*). De la même manière, les résultats quantitatifs sont associés à une unité (concept *UnitConcept*), appartenant aussi à la partie générique.

Cette partie générique est complétée par une ontologie du domaine. En effet, un résultat quantitatif n'a de sens que si les données expérimentales associées sont précisées : espèce animale, organe, soluté... Ces données associées à un résultat seront appelées *descripteurs* du résultat, ces descripteurs appartenant à un *type* particulier.

Ils se traduisent dans l'ontologie de domaine par des concepts à deux niveaux de hiérarchie. Le premier niveau correspond au type des descripteurs qui peuvent décrire un résultat expérimental (*Species*, *Organ*...), et le second niveau aux descripteurs eux-mêmes (*Kidney*, *Heart*...). Tous ces concepts ne possèdent pas de propriétés spécifiques, sauf le concept *Species*, auquel on associe le nombre d'individus, *hasN*, sur lesquels a été faite l'expérimentation.

La partie terminologique consiste enfin à associer à chaque concept feuille un terme préféré [4], et éventuellement des variantes, qui peuvent être des synonymes, des hyponymes, des abréviations, des acronymes ou des symboles. Les termes peuvent être des mots simples, des abréviations et acronymes et sont aussi souvent formés de plusieurs mots que nous appelons termes complexes.

3.2. *Représentation de l'ontologie par la base de données*

3.2.1. *Correspondance entre l'ontologie et la base de données*

Le modèle que nous venons de décrire est représenté par la base de données qui permet de stocker les instances trouvées dans les articles. Ainsi qu'il a déjà été dit, le schéma de la base de données a été conçu pour faciliter la comparaison de données mesurées sur différentes espèces et dans des conditions expérimentales variées, mais aussi pour être facilement extensible et généralisable à d'autres domaines.

La figure 2 présente un schéma partiel de la base (voir la figure 12 en annexe pour le schéma complet). La formalisation faite sous forme d'ontologie se retrouve bien dans la base de données. Les concepts génériques se traduisent par des tables : les concepts *ExperimentalResult* et *Source* sont représentés par les tables *record*, et *source*. La table *record* contient les attributs suivants :

– la valeur numérique du résultat (*result_value*) ;

– les unités du résultat, qui qualifient la valeur numérique (*units*) ;

– une précision, qui indique généralement l'erreur standard de la mesure (*precision*) ;

– le nombre d'animaux observés (*n_animals*) ;

– des données qui décrivent qualitativement le résultat (*qualitative_data*) ;

– un commentaire, qui donne des informations complémentaires sur les techniques expérimentales (*comment*).

4. Un terme préféré est un terme choisi par les experts du domaine lors de la construction de la terminologie pour désigner le concept.

On peut noter que QKDB ne stocke pas les unités possibles sous forme de table, l'unité étant un simple attribut de la table *record*. Notons aussi que le nombre d'individus sur lesquels l'expérimentation est faite est représenté au sein de la table *record*, ce qui est un choix cohérent avec le fait qu'il n'y a qu'un résultat quantitatif et qu'une espèce par expérimentation.

Les concepts du domaine sont représentés par deux tables : *field* et *field_type*. *field* contient les nœuds feuille de l'ontologie, les termes correspondant étant stockés dans les champs *content* (variante préférentielle) et *content_long* (liste de termes constituant des variantes). Chaque *field* est lié au nœud père correspondant de l'ontologie, représenté dans une table *field_type*. Ainsi, le concept *Pi* de l'ontologie correspond à une entrée de la table *field* dont l'attribut *content* est *Pi*, et l'attribut *content_long* est *permeability* ; cette entrée est reliée par une clé étrangère à une entrée de la table *field_type*, dont l'attribut *type* est *parameter*, ce qui traduit le fait que la perméabilité est un descripteur de type paramètre. Les descripteurs généralement utilisés pour un résultat expérimental en physiologie rénale sont les suivants (correspondant donc à des entrées de la table *field_type*) :

– l'espèce sur laquelle l'expérience a été menée ;

– l'organe, la région, le segment, le compartiment et éventuellement le type de cellule, qui représentent les endroits sur lesquels l'expérience a été menée ;

– le type de paramètre, qui indique la propriété qui a été mesurée, comme le poids, la perméabilité, le diamètre ou la concentration ;

– le soluté, qui précise ce qui a été mesuré, par exemple K^+ si la concentration mesurée concerne ce soluté.

Seuls les deux niveaux les plus bas de l'ontologie ont été jugés nécessaires à traduire dans la base de données, mais cette structure pourrait être étendue en ajoutant un lien récursif sous la forme d'une clé étrangère dans la table *field_type*.

La relation n-aire, quant à elle, est représentée par la table *describes* qui stocke chaque occurrence de la relation trouvée dans un article par l'ensemble des couples qui lient l'occurrence de résultat (*record*) avec chaque descripteur trouvé. Ce sont ces tables qui seront complétées lors du processus d'extraction.

Un travail important a porté sur la définition des ensembles de termes associés aux différents concepts. Certains ensembles sont exhaustifs, alors que d'autres peuvent être complétés en fonction de ce qui est trouvé dans les textes. Les concepts qui précisent les détails anatomiques dans QKDB ont servi pour enrichir l'ontologie FMA (Foundational Model of Anatomy (Rosse *et al.*, 2003)). Pour les paramètres et autres listes de descripteurs, des travaux sont en cours pour développer des ontologies de référence.

3.2.2. *Exemple de résultat expérimental*

Voici un exemple de phrase contenant un résultat expérimental stocké dans QKDB : « In controls versus PAN **rats**, **Na+/K+-ATPase activities** were (**pmol**

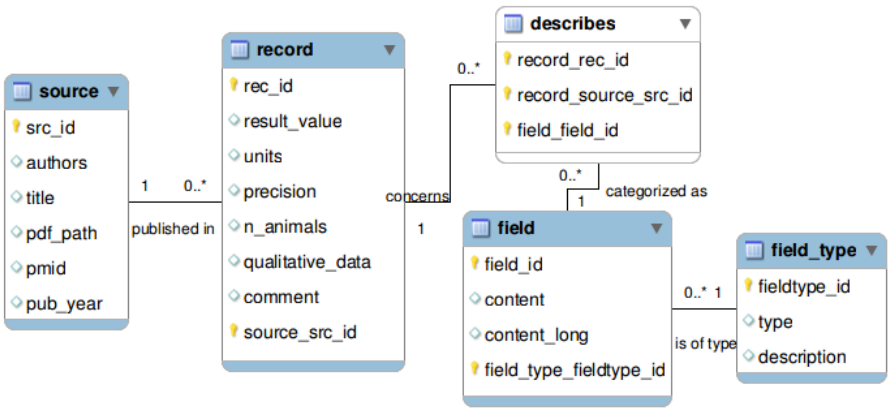

Figure 2. *Schéma UML simplifié de la base de données QKDB*

ATP/mm/h) : **proximal convoluted tubule**, **2954+-369** vs 2769+-230 ; thick ascending limb, 5352+-711 vs 5239+-803 ; and cortical collecting duct, 363+-96 vs 848+-194 (P<0.01), respectively. ».

Les informations stockées dans QKDB concernant le premier résultat cité dans la phrase (en gras) sont représentées dans les tableaux 1 et 2. On observe dans cet exemple que certains descripteurs ne sont pas dans la même phrase que la valeur numérique, comme l'organe ou le commentaire qui donne une information supplémentaire sur l'espèce étudiée. D'autres descripteurs ne sont pas exactement sous la même forme dans la base et dans le texte, par exemple le paramètre *activity* est au singulier dans la base mais au pluriel dans le texte.

table	record			
attribut	result_value	precision	units	comment
valeur	2954	369	pmol ATP/mm/h	Male wistar rats (100-130g)

Tableau 1. *Informations liées à un résultat expérimental dans QKDB (table record)*

Dans les exemples de la base, le nombre de descripteurs pour un résultat expérimental varie de 5 à 12.

4. Extraction des informations

À partir du modèle de connaissances et d'une analyse de corpus, nous avons développé un système d'extraction d'information pour repérer les résultats expérimentaux répartis dans les articles, que nous présentons maintenant.

table	field_type	field
attribut	type (type de descripteur)	content (descripteur)
valeurs	espèce	rat
	organe	kidney
	tube	PCT
	paramètre	activity
	protéine membranaire	Na-K-ATPase

Tableau 2. *Informations liées à un résultat expérimental dans QKDB (tables field_type et field)*

4.1. *Principe*

Rappelons que nous avons défini un *résultat expérimental* comme étant une relation n-aire entre :

– un *résultat quantitatif* composé notamment d'une *valeur numérique*, associée à certains *attributs* comme l'unité, la précision...

– et des *descripteurs*, qui précisent les conditions expérimentales, comme l'espèce ou l'organe considéré (le nombre de descripteurs peut varier de 3 (espèce, organe et paramètre) à 8).

Nous proposons une méthode en trois étapes pour la reconnaissance de ces résultats :

– détection d'une valeur numérique, qui joue le rôle de déclencheur et constituera l'élément pivot du résultat. En effet, nous considérons qu'une expérience n'est pertinente que si un résultat quantitatif est fourni. La reconnaissance de la valeur numérique s'apparente à une reconnaissance d'entité numérique. Les autres attributs du résultat quantitatif comme le nombre d'animaux étudiés doivent aussi être reconnus, ce qui peut se faire assez simplement avec des expressions régulières ;

– reconnaissance de descripteurs d'expérience. Il s'agit alors d'une reconnaissance terminologique puisqu'il faut faire le lien entre les concepts de la base de données et les termes de l'article. Pour cela, nous utilisons les lexiques de notre ressource termino-ontologique et identifions les variantes de ces termes (flexionnelles, dérivationnelles, etc.) ;

– mise en relation des attributs et des descripteurs avec la valeur numérique. Cette mise en relation prend en compte la distance entre la valeur numérique du résultat et les attributs ou descripteurs, ainsi que des critères de fréquence.

4.2. *Architecture*

Le schéma de la figure 3 présente les différents modules de notre système d'extraction. L'article XML est d'abord étiqueté par le TreeTagger (Schmid, 1994), ce qui

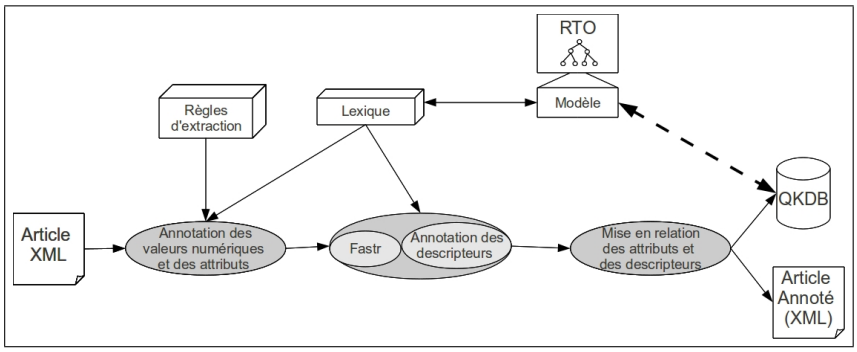

Figure 3. *Schéma de fonctionnement du système d'extraction*

permet de disposer des formes canoniques des mots (i.e. les lemmes) et de leur catégorie, ainsi que de procéder à un découpage en phrases selon l'étiquetage des signes de ponctuation. L'article étiqueté est ensuite fourni au module d'annotation des valeurs numériques, puis des descripteurs (qui utilise Fastr (Jacquemin, 1996) pour détecter les variantes des termes du lexique). Le dernier module effectue la mise en relation des résultats numériques avec leurs descripteurs. Les tables *record* et *describes* de la base de données sont ensuite complétées avec les résultats expérimentaux extraits. Nous présentons dans la suite chacune des trois étapes du système.

4.3. *Reconnaissance des résultats quantitatifs*

Les valeurs numériques des expériences sont le point de départ de la reconnaissance des expérimentations. Il est nécessaire de pouvoir toutes les reconnaître, sans pour autant extraire toutes les valeurs numériques présentes dans l'article. La figure 4 présente différents types de résultats quantitatifs d'expériences. L'étude d'un corpus de développement [5] nous a montré que 94% des résultats quantitatifs sont donnés dans les parties *Results* et *Discussion*, les 6% restants étant des résultats provenant de figures et qui ne sont cités de façon textuelle que dans la section *Abstract*. Les informations contenues dans les champs de commentaires sont présentes majoritairement dans la section *Methods*. Nous pouvons donc limiter les parties de l'article à analyser pour repérer les valeurs numériques aux sections *Results* et *Discussion*. Pour les identifier, nous avons testé une méthode à base de règles et une méthode à base d'apprentissage. L'évaluation de ces deux méthodes est présentée dans la section 6.3.1.

5. Notre corpus de développement est composé de 5 articles, contenant 95 résultats expérimentaux enregistrés dans la base de données (cf. section 6.1).

The urinary Ca2+ concentration of the knockout mice reached values as high as 20 mM, compared with 6 mM for TRPV5+/+ littermates.

Apical membrane Pf averaged (in cm/s) 9.37 +- 0.77 e-4 (n = 5) at 20°C, and two values obtained at 37°C were 33.7 and 33.2 e-4 cm/s.

Parameter	Group 1 (n = 7)	Group 2 (n = 5)	Groupe 3 (n = 5)
Body weight (g)	24.8 +- 0.47	23.5 +- 0.48	26.1 +- 0.8

Figure 4. *Exemples de résultats quantitatifs à extraire*

4.3.1. *Reconnaissance par règles des valeurs numériques*

La première méthode est fondée sur des expressions régulières permettant de repérer dans le texte les valeurs numériques qui peuvent correspondre à des résultats d'expériences. Les patrons utilisés sont les suivants :

{nombre}% ? +- {nombre}% ?(e^{nombre}) ?
{nombre}% ? (e^{nombre}) ?

{nombre} décrit un nombre qui peut être décimal, négatif, ou contenant un séparateur de milliers. *e^{nombre}* indique que le nombre peut être suivi d'une puissance. Le premier patron permet de reconnaître les valeurs qui sont suivies d'une précision, et le deuxième annote toutes les autres valeurs numériques. Ce dernier a un rappel très élevé mais une précision très basse : en effet il repère les numéros des figures, les renvois bibliographiques, etc. Toutes les valeurs n'étant pas repérées avec le premier patron, il est nécessaire d'appliquer un patron peu sélectif pour ne pas omettre un résultat potentiel à ce stade de l'extraction. Pour améliorer la précision, nous appliquons ensuite un filtre qui permet de supprimer les valeurs numériques entre parenthèses, celles qui suivent les mots *figure* et *table*, et les numérotations de références. Après cette dernière étape, la précision reste encore faible ; une dernière sélection des valeurs numériques aura lieu lors de la mise en relation des descripteurs.

4.3.2. *Reconnaissance par apprentissage des valeurs numériques*

Nous avons également testé une méthode à base d'apprentissage considérant la reconnaissance des valeurs numériques des résultats expérimentaux comme un problème de classification binaire. Les valeurs numériques contenues dans les articles sont ou non des résultats d'expérience. Nous avons utilisé un classifieur à base de SVM : libSVM (Chang *et al.*, 2001). Nous avons repéré toutes les unités numériques de l'article avec un patron simple : {nombre}% ?(+- {nombre}% ?) ?, que nous avons appliqué dans les tableaux et dans le texte. Pour chaque valeur numérique annotée, nous avons construit un vecteur d'attributs, que nous avons fourni en entrée du classifieur. Les attributs suivants sont utilisés :

– le caractère décimal du nombre,

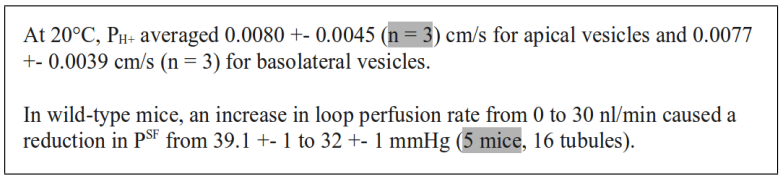

Figure 5. *Exemples du nombre d'animaux étudiés à extraire*

– le nombre de chiffres de la partie entière de la valeur numérique (s'il n'y a qu'un chiffre et que ce n'est pas un nombre décimal, il est probable que ce soit un numéro de figure et non pas un résultat),

– l'inclusion de la valeur dans un tableau,

– le lemme et la catégorie morpho-syntaxique des cinq mots avant et des cinq mots après la valeur numérique,

– la classe de VerbNet [6] (Kipper *et al.*, 2008) des verbes dans les cinq mots avant et après (les verbes sont importants car ils introduisent souvent les résultats, comme par exemple *average*),

– les unités dans les cinq mots avant et après (très peu de valeurs numériques de résultats sont indiquées sans unités),

– la distance entre la valeur numérique et les unités les plus proches,

– le type des descripteurs de la phrase (*Parameter*, *Solute*, etc.),

– le titre de la partie où est située la valeur, comme *Results*, *Discussion*, etc. (aucune valeur numérique de résultat ne sera donnée dans la partie *Methods* par exemple).

4.3.3. *Reconnaissance des autres attributs du résultat quantitatif*

Les autres attributs des résultats quantitatifs sont repérés par des patrons. Le nombre d'animaux étudiés est un nombre entier qui peut être précédé de *n* = (1er exemple de la figure 5), ou suivi du nom de l'espèce concernée ou de *males* ou *females* (2e exemple de la figure 5).

La précision est annotée avec le même patron que la valeur numérique associée : {nombre}% ? +- {nombre}% ?(e^{nombre}) ?.

Pour annoter les unités, nous utilisons un patron qui repère les chaînes de caractères composées de la combinaison d'unités de base (comme *g* ou *mol*), de préfixes (comme *k* ou μ), de suffixes ($^{-1}$) et de symboles de séparation (. ou /). Nous repérons des unités comme *cm/s*, *mmHg*, *pmol ATP/mm/h* ou encore *µmol/mg creatinine*. Pour repérer cette dernière unité, nous acceptons la présence d'un soluté juste après l'unité ou entre les composants de l'unité.

6. VerbNet est un lexique de verbes en anglais regroupés en classe (une extension des classes de Levin)http ://verbs.colorado.edu/∼mpalmer/projects/verbnet.html

4.4. *Reconnaissance des descripteurs*

La reconnaissance des descripteurs se heurte au problème de grande variabilité de ces termes. Les différents types de variations que nous avons observés dans le corpus de développement sont présentés dans le tableau 3.

	Exemple	
Type	**Dans la base**	**Dans le texte**
Dérivation/flexion	urine	urinary
Formule chimique	Na^+	sodium
Variation d'écriture	0.000937 ± 0.77	9.37 ± 0.77 e-4
Abréviation	permeability	P_f
Variation sémantique	flow rate	excretion

Tableau 3. *Types de variations observés*

Afin de reconnaître les termes associés aux descripteurs, ainsi que leurs variantes, nous avons utilisé les variantes présentes dans nos lexiques, ainsi que Fastr pour repérer des variantes supplémentaires. Fastr développé par (Jacquemin, 1996) est un analyseur de surface basé sur des métarègles, qui permet de détecter des variantes morphologiques, syntaxiques ou sémantiques.

Deux ressources de la langue générale sont utilisées : la base CELEX [7] pour détecter des variantes morphologiques et WordNet (Fellbaum, 1998) pour les synonymes. Par exemple, à partir de *inner diameter*, une règle de coordination reconnaîtra *inner and outer diameters* comme une variante.

Étant donné un texte et une liste de termes multi-mots, les données sont analysées par le TreeTagger, puis Fastr compile les métarègles, et applique les règles sur le texte étiqueté. De nombreuses variantes sont trouvées par Fastr : 3000 occurrences sont reconnues à partir de 194 termes multi-mots, dont 520 variantes flexionnelles (16,7%), 337 variantes syntaxiques (10,8%) et 84 variantes morpho-syntaxiques (2,7%) [8]. Par exemple, *glomerular cells and capillaries* est reconnu comme une variante du terme pour le tube segment *glomerular capillary*, et *fraction of filtered* comme une variante de *filtration fraction*. Dans la section 6.3.2, nous présentons une évaluation de l'apport de l'enrichissement du lexique.

4.5. *Mise en relation des attributs et descripteurs*

Une fois les résultats quantitatifs et les descripteurs annotés, il faut mettre en relation une valeur numérique avec ses attributs et descripteurs. Pour cela, il faut sélec-

7. www.ldc.upenn.edu/readme_files/celex.readme.html
8. 2059 occurrences ne sont pas des variantes mais le terme repéré tel quel.

tionner parmi tous les descripteurs annotés dans le texte ceux qui sont associés à la même expérimentation que la valeur numérique.

La difficulté vient de la présence d'ambiguïtés, c'est-à-dire que plusieurs termes correspondant à un même type de descripteur peuvent être exprimés dans la même phrase ou dans le même article. L'ambiguïté pour le champ espèce est très faible : en effet, dans le corpus de développement, une seule espèce est citée dans l'article pour 90% des expérimentations. Dans les autres cas, l'espèce est citée à proximité du résultat quantitatif. Il en est de même pour le champ organe. Pour les autres descripteurs, l'ambiguïté au sein d'une phrase peut être levée grâce à la ponctuation ou à des critères de proximité par rapport à la valeur numérique de l'expérience.

Par ailleurs, les descripteurs peuvent ne pas être proches de la valeur numérique du résultat dans l'article. Dans le corpus de développement, 90% des solutés sont dans la même phrase que la valeur numérique, ainsi que 65% des paramètres, mais seulement 35% des commentaires sont dans la même phrase que la valeur numérique. Extraire les descripteurs uniquement dans la phrase de la valeur numérique ne sera donc pas suffisant. Nous avons par conséquent mis en place des heuristiques de sélection des valeurs numériques, ainsi que des attributs et descripteurs. Le principe est d'utiliser la valeur numérique comme élément pivot, et d'associer les attributs ou descripteurs en fonction de leur proximité avec cette valeur dans l'article, ou de leur fréquence dans l'article complet.

Les valeurs numériques sont ainsi sélectionnées si au moins l'une de ces conditions est vérifiée :

– la valeur numérique est dans un tableau ;

– elle est suivie d'une précision ;

– une unité est à proximité.

Pour chaque valeur numérique sélectionnée, les attributs et descripteurs sont sélectionnés en fonction des critères suivants :

– pour les descripteurs *species* et *organ* : sélection du plus proche de la valeur numérique (dans la phrase) ou du plus fréquent dans l'article s'ils ne sont pas mentionnés dans la phrase ;

– si la valeur numérique est dans un tableau : sélection du descripteur dans l'entête de la colonne ou de la ligne ou dans la légende ;

– dans tous les autres cas : sélection de l'attribut ou du descripteur le plus proche dans la phrase et s'il n'y en a pas sélection du plus proche dans le paragraphe.

Ces critères de selection sont évalués dans les sections 6.3.3 et 6.3.4.

La figure 6 présente un exemple de mise en relation des attributs et descripteurs avec la valeur numérique *8.5 +- 0.5* (tous les descripteurs extraits sont surlignés dans l'exemple) et la figure 7 présente la mise en relation des attributs et descripteurs dans un tableau.

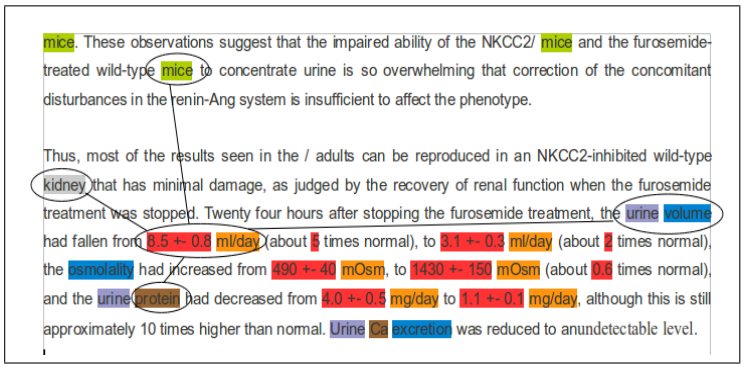

Figure 6. *Exemple de la mise en relation des descripteurs*

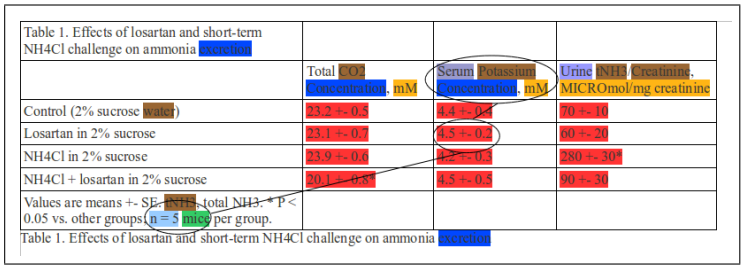

Figure 7. *Exemple de la mise en relation des descripteurs dans un tableau*

5. Assistant d'aide à l'annotation

Pour répondre au besoin d'assistance des experts lors du peuplement la base de données, nous avons développé un assistant d'aide à l'annotation d'article et au peuplement de la base.

5.1. *Spécification de l'outil*

L'outil d'extraction ne pouvant pas avoir une précision parfaite, les informations extraites par ce système doivent être vérifiées avant d'être insérées dans la base de données. Nous avons donc choisi d'intégrer le système d'extraction dans une interface d'aide à l'annotation afin de faciliter l'annotation des articles par les experts. L'objectif est de fournir un article pré-annoté aux experts, qui n'auront plus qu'à valider, modifier ou rejeter les propositions du système.

L'interface développée doit donc permettre de :

– ajouter un article dans la base d'articles ;

– sélectionner un article à annoter ;

– afficher les annotations extraites ou existantes de l'article ;

– ajouter ou modifier manuellement les annotations de l'article.

Cette interface étant destinée à être utilisée par des experts curateurs divers, elle est conçue comme une application Web, et mise à disposition de ces experts [9].

5.2. *Descriptif*

Les différents processus utiles à l'analyse d'un texte, allant de sa conversion au format XML requis jusqu'à la proposition des enregistrements à intégrer dans la base de données, sont intégrés au travers d'une interface Web, développée en PHP et JavaScript (figure 8).

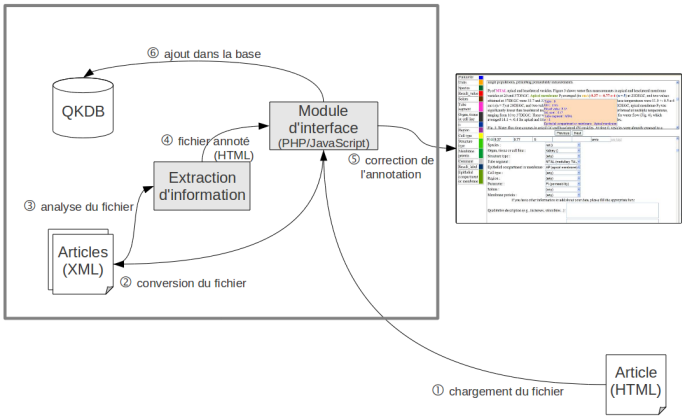

Figure 8. *Architecture de l'assistant d'aide au peuplement de la base de données*

L'utilisateur doit fournir à l'assistant un article HTML ainsi que des fichiers contenant les tables associées. Les articles sont transformés dans notre format XML. Pour le moment, seuls les articles HTML sont acceptés. Ensuite l'extraction est exécutée sur l'article XML. Le fichier de sortie est l'article XML dont les descripteurs ont été annotés.

Après la phase d'extraction, l'assistant propose la visualisation présentée dans la figure 9 [10].

L'interface permet à l'utilisateur de :

– visualiser, pour chaque résultat, les informations qui lui sont associées, surlignées en couleur (avec un code couleur par champ) (1) ;

9. http ://qkdb.limsi.fr/
10. Dans ce qui suit les numéros entre parenthèses renvoient aux numéros sur la figure.

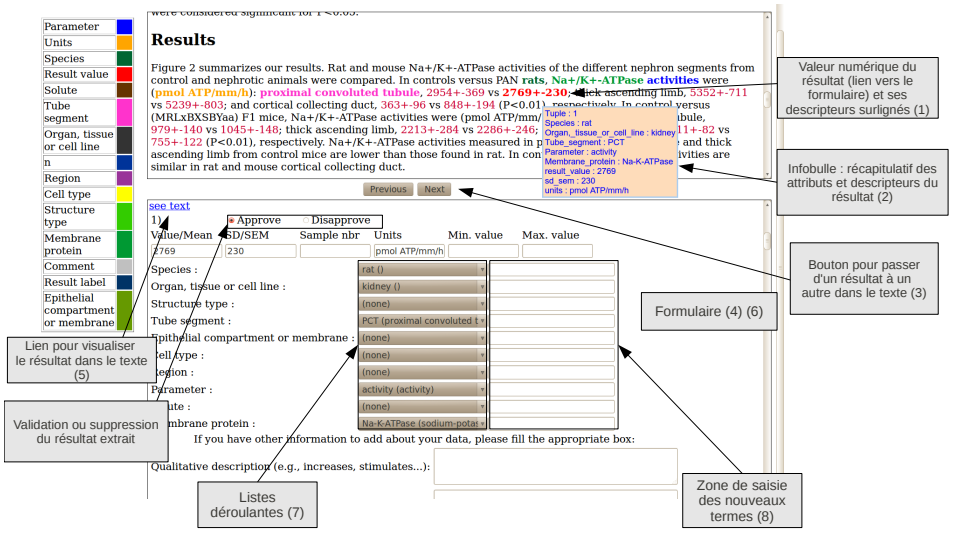

Figure 9. *Interface*

– avoir un récapitulatif des attributs caractérisant une expérimentation ; celui-ci est affiché dans une infobulle (2) ;

– parcourir l'article en passant d'un résultat à l'autre (3) ;

– visualiser les attributs de la base de données dans un formulaire modifiable affiché en bas d'écran, qui correspond au résultat affiché dans la partie texte (4) ;

– parcourir les schémas extraits, avec un affichage de la partie texte correspondante (5) ;

– modifier des descripteurs via le formulaire ; les modifications sont répercutées sur l'annotation du texte de manière à conserver la cohérence entre les deux points de vue sur le texte (6).

Le formulaire est composé de listes déroulantes pour les champs pour lesquels les listes des valeurs sont fermées (7). À droite de ces listes, des zones de saisies ont été ajoutées pour permettre à l'expert d'écrire le terme utilisé dans le texte pour décrire le concept, dans le cas où un mauvais concept aurait été extrait (8).

6. Évaluations

Nous avons évalué d'une part les performances du système d'extraction d'information et d'autre part l'apport de l'interface pour l'aide au peuplement de la base de donnée. Nous présentons dans cette partie le corpus sur lequel nous avons évalué notre système ainsi que le lexique utilisé, puis la façon dont nous avons mené ces évaluations et les résultats que nous avons obtenus.

6.1. *Corpus*

Pour développer et évaluer notre système, nous avons constitué un corpus avec des articles de physiologie rénale référencés dans QKDB afin de pouvoir les annoter automatiquement avec les données contenues dans la base.

Les articles contenus dans la base sont disponibles au format PDF et parfois en XHTML sur le Web. Pour pouvoir analyser les articles, il était nécessaire d'en disposer dans un format avec une structuration simple et facilement convertible en texte, comme le XML. De ce fait, nous n'avons gardé que les articles disponibles au format XHTML, soit 20 articles car la conversion du format PDF au format XML n'est possible que pour des articles récents (d'une quinzaine d'années) et pose des problèmes, principalement pour extraire les tableaux et certains caractères spéciaux (comme $\pm$). Les 20 documents conservés sont convertis dans un format XML structuré avec les balises suivantes : titre, auteurs, corps de l'article, paragraphes, tableaux (avec lignes et colonnes) et notes de bas de page. Les tableaux ont été conservés car ils contiennent de nombreux résultats numériques : dans le corpus, 73% des valeurs numériques des résultats sont dans des tableaux. Les balises XHTML `td`, `th`, `caption`, etc. ont été converties en balises `colonne`, `ligne`, `legende` et `tableau`. Un attribut `num` a été associé aux balises `colonne` et `ligne`, il a pour valeur le numéro de la ligne ou de la colonne du tableau. Grâce à cet attribut nous pouvons récupérer toute l'information contenue dans une ligne donnée ou dans une colonne. La figure 10 est un exemple d'un tableau dans un article en XHTML, visualisé avec Mozilla Firefox.

Table 1. Basal values for three groups[a]

Parameter	Group 1 (n = 7)	Group 2 (n = 5)	Group 3 (n = 5)
Body weight (g)	24.8 ± 0.47	23.5 ± 0.48	26.1 ± 0.8
Kidney weight (g)	0.30 ± 0.01	0.27 ± 0.01	0.33 ± 0.01
Hematocrit (%)	47 ± 1	39 ± 1[b]	41 ± 1[b]
Plasma sodium concentration (mEq/L)	149 ± 3	151 ± 2	144 ± 3
Plasma potassium concentration (mEq/L)	4.6 ± 0.8	3.4 ± 0.6	3.8 ± 0.8
Plasma osmolality (mosmol/kg)	292 ± 4	293 ± 3	290 ± 5

[a]Group 1 is the euvolemic group, group 2 is the volume-expanded group, and group 3 is the volume-expanded Ang II-infused group.
[b]$P < 0.05$ compared with group 1.

Figure 10. *Exemple d'un tableau provenant d'un article en XHTML*

Le corpus est composé d'environ 950 résultats expérimentaux (provenant des 20 articles). 95 résultats (issus de cinq articles) ont été utilisés pour le développement et 855 (issus des quinze autres) pour l'évaluation. Nous avons choisi cette répartition en considérant que la variété de 95 résultats expérimentaux était suffisante pour développer notre système.

Dans la figure 11 nous avons représenté la proportion de descripteurs par rapport au nombre de résultats expérimentaux, calculs effectués à partir de la base de données. Nous observons par exemple que pour tous les résultats expérimentaux, l'espèce et le paramètre sont renseignés ; en revanche le soluté l'est dans 60% des cas.

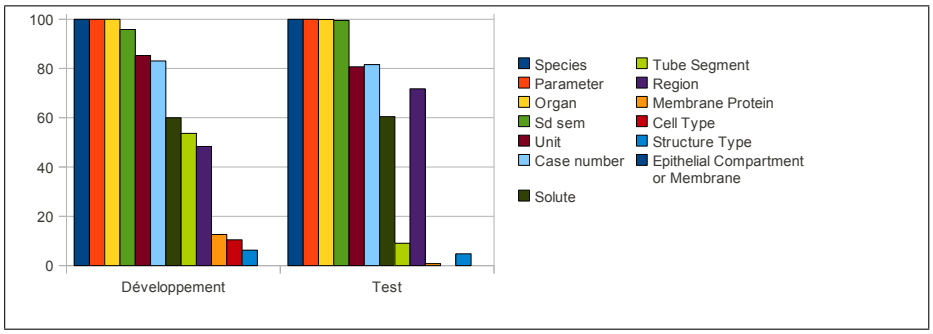

Figure 11. *Proportion de descripteurs par rapport au nombre de résultats expérimentaux dans chaque corpus*

Pour le développement et l'évaluation du système d'extraction d'information, nous avions besoin d'annoter les corpus. Pour cela nous avons projeté les descripteurs des résultats expérimentaux contenus dans la base dans les 20 articles de notre corpus. Nous avons donc réalisé la tâche inverse à l'extraction : nous avons utilisé les données de QKDB qui avaient été extraites manuellement des articles scientifiques et nous les avons recherchées et annotées dans les textes. Une vérification et une complétion manuelle ont ensuite été faites.

6.2. *Lexique*

Pour extraire les descripteurs, nous avions besoin d'un lexique des termes du domaine. Nous avons construit un lexique de base avec les termes associés aux concepts de la base de données, ainsi que leurs variantes, dont nous donnons le détail dans le tableau 4. L'identifiant associé à chaque concept dans la base est conservé dans le lexique pour relier les termes au concept désigné. Les entrées du lexique ont la forme suivante : *PAR_16 concentration*. *PAR* donne le type de l'entrée, en l'occurrence paramètre, *16* est l'identifiant du concept de la base, et *concentration* est le terme relié à ce concept. Si plusieurs termes sont reliés au même concept, il y aura plusieurs entrées avec le même type et le même identifiant, par exemple : *TUS_57 cortical CD* et *TUS_57 cortical collecting duct*.

Pour enrichir le lexique de base, nous avons ajouté des termes provenant de ressources externes.

Nous avons utilisé des listes provenant de sites spécialisés que nous avons sélectionnés pour l'espèce et le soluté [11]. Nous avons également voulu ajouter des données provenant d'une ressource du domaine médical : l'UMLS (Unified Medical Lan-

11. Exemple des sites utilisés : www.kterre.org/dossiers/atomes_liste.php ou en.wikipedia.org/wiki/Dictionary

	Lexique de base		Lexique enrichi
Descripteur	**# concepts**	**# termes**	**# termes**
Species	24	25 (1,0)	68 (2,8)
Organ	12	15 (1,2)	22 (1,8)
Structure type	7	10 (1,4)	10 (1,4)
Tube segment	46	82 (1,8)	94 (2,0)
Epithelial compartment or membrane	14	28 (2,0)	30 (2,1)
Cell type	16	24 (1,5)	24 (1,5)
Region	30	47 (1,6)	81 (2,7)
Parameter	85	153 (1,8)	228 (2,7)
Solute	66	92 (1,4)	119 (1,8)
Membrane Protein	11	22 (2,0)	22 (2,0)

Tableau 4. *Contenu du lexique : nombre de termes associés à chaque type de concept, et nombre moyen de termes par concept, dans le lexique de base et le lexique enrichi*

guage System). C'est un métathésaurus très riche, développé par la NLM (National Library of Medicine). Il rassemble plusieurs thésaurus en créant des relations entre les concepts. Il est très complet mais aussi très complexe. Pour évaluer la couverture de l'UMLS dans le domaine de la physiologie rénale, nous avons utilisé l'outil MetaMap de (Aronson, 2001) qui annote les concepts de l'UMLS dans des textes biomédicaux. Sur le corpus de développement, nous avons observé un grand nombre de termes annotés non pertinents, et ceux annotés par notre système ne l'étaient pas par MetaMap : seulement 6,5% des termes repérés par MetaMap étant dans notre lexique, le bruit est donc très important. Par exemple, dans la phrase « we performed a separate series of experiments in NHE3+/+ and NHE3-/- mice to directly evaluate P_sf during loop of Henle perfusion », MetaMap annote *directly* comme un concept qualitatif, qui n'est pas pertinent pour notre tâche, et il n'annote pas le descripteur *loop of Henle* de type *tube segment*. Aussi, nous n'avons pas retenu cette ressource.

Pour extraire les unités nous avions besoin d'une liste des unités de base. Nous avons pour cela utilisé l'ontologie units.obo (dans le format de Gene Ontology), dans laquelle nous avons extrait les unités de base (*cm*, *mol*, etc.) et les unités composées (*ml/kg*, etc.). Nous avons ajouté manuellement quelques préfixes, comme par exemple *k* pour kilo, ce qui nous permet d'extraire des unités complexes qui étaient absentes de notre lexique.

Le lexique de base n'était pas assez complet pour permettre une bonne identification des descripteurs et la complétion manuelle des listes de termes à partir de terminologies ou d'ontologies existantes ne permettait pas de disposer de tout le vocabulaire nécessaire. Aussi nous avons appliqué une méthode d'acquisition semi-automatique de termes par apprentissage de patrons à partir de corpus (voir (Minard *et al.*, 2010) pour plus de détails), ce qui nous a permis d'augmenter notre lexique de 5%. Le tableau 4 indique la taille du nouveau lexique après l'ajout de termes grâce à des ressources externes et à l'acquisition de nouveaux termes.

6.3. *Évaluations du système d'extraction d'information*

L'évaluation est faite sur trois critères différents. Premièrement, nous calculons le rappel, la précision et la F-mesure pour l'extraction des résultats expérimentaux (nous nommons cette étape *évaluation générale*) :

$$Rappel = \frac{\text{résultats correctement extraits}}{\text{nombre de résultats à extraire}}$$

$$Précision = \frac{\text{résultats correctement extraits}}{\text{nombre de résultats extraits}}$$

$$F - mesure = \frac{2 \cdot (\text{précision} \cdot \text{rappel})}{(\text{précision} + \text{rappel})}$$

Nous comptons un descripteur ou un attribut correctement extrait s'il est du bon type et associé au bon résultat expérimental. Si une valeur numérique est sélectionnée à tort, tous ses attributs et descripteurs sont comptabilisés comme faux. Nous évaluons également l'extraction et la sélection des valeurs numériques uniquement, ce qui correspond à évaluer l'identification des résultats quantitatifs (nous nommons cette étape *évaluation des résultats quantitatifs*). Le troisième type d'évaluation que nous faisons, concerne uniquement les résultats expérimentaux pertinents, c'est-à-dire que nous calculons le rappel, la précision et la F-mesure pour les valeurs numériques correctement sélectionnées (nous nommons cette étape *évaluation des résultats expérimentaux correctement extraits*). Ce dernier type d'évaluation permet de n'évaluer que l'extraction des descripteurs et des attributs, et leur mise en relation.

Rappelons que le corpus de test est composé de 855 résultats expérimentaux pertinents (issus de 15 articles).

L'évaluation du système final est présentée dans le tableau 5. Ce système utilise des règles pour extraire les valeurs numériques, le lexique enrichi pour la détection des descripteurs et les critères de proximité et fréquence pour la mise en relation. L'évaluation des résultats expérimentaux correctement extraits est faite sur une base de rappel de 0,63, c'est-à-dire que l'évaluation est faite sur les 63% de résultats expérimentaux correctement extraits. Nous obtenons une F-mesure de 0,78 pour les résultats pertinents, ce qui correspond à un bon résultat, au niveau de l'état de l'art si on se réfère à la tâche similaire de i2b2 2009 portant sur des termes médicaux. La F-mesure 0,61 reste bonne, ce qui signifie que même si trop de résultats sont proposés, ils ne constitueront pas un bruit trop lourd à gérer par les curateurs.

	Rappel	**Précision**	**F-mesure**
Évaluation générale	0,75	0,51	0,61
Évaluation des résultats quantitatifs	1,00	0,63	0,77
Évaluation des résultats expérimentaux correctement extraits	0,75	0,81	0,78

Tableau 5. *Évaluation du système final*

Dans la suite de cette partie, nous présentons l'évaluation de l'impact de certains choix que nous avons faits pour chaque étape de la méthode. Chaque évaluation est

effectuée en comparant les performances du système final et du système en faisant varier un paramètre à la fois.

6.3.1. *Évaluation de l'extraction des valeurs numériques*

Le tableau 6 présente les résultats de l'extraction des valeurs numériques par apprentissage et par règles. La précision (P) et la F-mesure (F) sont meilleures pour l'extraction par apprentissage, mais le rappel (R) est meilleur avec les règles. Ce qui nous intéresse pour cette tâche est d'avoir un très bon rappel : en effet il paraît plus aisé pour les experts qui vérifieront les données annotées d'en enlever que d'en ajouter. De plus le système d'apprentissage n'a pas assez de données pour apprendre la pertinence des valeurs numériques. En effet, seuls les résultats suivis d'une approximation ($\pm$) et ceux dans un tableau sont extraits, mais les résultats pertinents qui ne rentrent pas dans ces deux cas ne sont pas extraits. Nous avons par conséquent conservé les règles pour cette extraction.

Évaluation des résultats quantitatifs	**R**	**P**	**F**
Apprentissage	0,93	0,78	**0,85**
Règles	**1,00**	0,63	0,77

Tableau 6. *Évaluation de l'extraction des résultats numériques par apprentissage et par règles*

6.3.2. *Évaluation de la complétion du lexique*

Dans le tableau 7 nous comparons les performances du système avant l'enrichissement du lexique et du système final. Dans les deux cas, les résultats numériques sont extraits avec des règles et les descripteurs sont reliés selon des critères de proximité et de fréquence. Les précisions des deux systèmes restent très proches, alors que plus les lexiques sont complets plus le nombre de descripteurs annotés augmente, il y a donc plus de risques d'erreurs lors de la mise en relation. Comme on peut s'y attendre plus le lexique est complet et plus le rappel augmente. Nous pouvons observer que le système final a un rappel qui augmente de 12% par rapport au système de base.

Évaluation des résultats expérimentaux correctement extraits	**R**	**P**	**F**
Système avec le lexique de base (QKDB)	0,67	0,79	0,73
Système final, avec le lexique enrichi	0,75	0,81	0,78

Tableau 7. *Évaluation de la complétion du lexique*

6.3.3. *Évaluation de la mise en relation*

La sélection de l'espèce et de l'organe se fait selon un critère de proximité puis de fréquence. Si un organe ou une espèce sont cités dans la même phrase que le résultat quantitatif de l'expérience, alors cet organe ou cette espèce est sélectionné(e). Dans le cas contraire, nous sélectionnons l'organe ou l'espèce le (la) plus fréquent(e) dans

l'article. En effet la majorité des articles portent sur une seule espèce et sur un organe en particulier (dans notre cas, les articles portent sur le rein). Si ce n'est pas le cas, l'espèce et l'organe sont donnés au moment de la description du résultat et à proximité de celui-ci. Nous avons évalué le choix de ce critère de fréquence par rapport à un critère de proximité uniquement. Dans un premier temps, nous avons évalué l'extraction de l'espèce et de l'organe avec des critères de sélection de fréquence et de proximité, et dans un second temps uniquement avec un critère de proximité. Le système utilisé est le système final avec le lexique enrichi. Les résultats de cette évaluation sont donnés dans le tableau 8.

	Critère prox/freq			Critère prox		
	R	**P**	**F**	**R**	**P**	**F**
Évaluation générale	0,75	0,51	**0,61**	0,52	0,45	0,48
Évaluation des résultats expérimentaux correctement extraits	0,75	0,81	**0,78**	0,52	0,76	0,62
Espèces	0,97	0,98	**0,97**	0,21	1,00	0,35
Organes	0,96	0,98	**0,97**	0,07	0,83	0,13

Tableau 8. *Évaluation de l'extraction de l'espèce et de l'organe selon le critère utilisé*

Nous observons que le rappel et la précision pour l'extraction des descripteurs de type espèce et de type organe avec un critère de fréquence sont proches de 1,00. Cela montre que la majorité des descripteurs de ces deux types sont extraits. En revanche, en utilisant un critère de proximité très peu de descripteurs sont extraits. En effet, les descripteurs de type espèce et organe sont cités principalement au début du document, et ne sont généralement pas rappelés dans la partie résultat de l'article. Une des raisons pour lesquelles le rappel et la précision pour l'extraction des descripteurs de type espèce ne sont pas à 1,00 est une mauvaise utilisation de la structure des tableaux. En effet, certains tableaux ont des structures assez complexes, avec par exemple des légendes communes à plusieurs lignes, et les règles de mise en relation des descripteurs dans ces tableaux ne s'appliquent pas correctement.

6.3.4. *Évaluation de la reconnaissance des résultats expérimentaux dans des tableaux*

Dans le corpus de développement 35% des valeurs numériques des résultats expérimentaux étaient dans des tableaux, et dans le corpus de test 77% le sont. Les descripteurs des résultats expérimentaux dont les valeurs numériques sont dans les tableaux peuvent être dans les tableaux ou dans le texte. Dans les tableaux, les descripteurs sont dans les en-têtes des colonnes ou des lignes, ainsi que dans les légendes du tableau. Les descripteurs qui sont dans les en-têtes sont faciles à mettre en relation avec la valeur numérique ; pour cela nous prenons en compte leur position dans le tableau. Mais il est plus difficile de relier les informations de la légende aux valeurs numériques.

Les 33 résultats du corpus de développement qui sont dans des tableaux, sont décrits par 204 descripteurs, dont 180 sont dans les tableaux et seulement 24 dans le texte (dans le texte on retrouve souvent l'espèce et l'organe). Les problèmes posés par

l'extraction des informations données sous forme de tableau ou de texte étant différents, nous avons évalué séparément l'extraction des résultats expérimentaux dans des tableaux et dans le texte.

Le système utilisé pour cette évaluation est le système final avec le lexique enrichi et l'utilisation des critères de proximité et de fréquence pour la mise en relation. Les résultats sont dans le tableau 9. Nous observons que le rappel est identique pour l'extraction dans le texte et dans les tableaux, en revanche la précision est meilleure pour l'extraction dans les tableaux. Cela est dû au fait que l'information présente dans les tableaux est structurée ce qui permet de limiter les erreurs.

	Tableaux			Texte		
	R	**P**	**F**	**R**	**P**	**F**
Évaluation des résultats quantitatifs	1	0,86	0,92	1	0,33	0,50
Évaluation générale	0,75	0,65	0,70	0,75	0,39	0,51

Tableau 9. *Évaluation de l'extraction des résultats numériques dans les tableaux et dans le texte*

6.4. *Évaluation de l'interface*

Nous nous sommes inspirés de la méthode adoptée par (Alex *et al.*, 2008) pour l'évaluation de l'outil. Nous avons demandé à cinq experts [12] d'annoter trois articles chacun dans trois conditions différentes (au total cinq articles différents ont été annotés) : un à partir de l'annotation de référence (c'est-à-dire à partir de l'annotation faite par projection des données de la base), un avec les annotations de notre système d'extraction et le troisième *manuellement*, c'est-à-dire sans annotation préalable de l'article. Nous avons calculé le temps qu'ils passaient sur chaque article, le nombre d'expérimentations annotées, le nombre de descripteurs et le nombre de commentaires (les commentaires ne sont pas annotés par notre système d'extraction mais très important pour la base de données QKDB). À la suite de l'annotation, les experts ont répondu à un questionnaire.

Le tableau 10 présente le nombre total de résultats, de descripteurs et de commentaires annotés, ainsi que le temps moyen passé pour annoter un résultat dans deux conditions : *manuellement* (le formulaire était vide) et avec l'assistant. Le temps total passé par les 5 experts sur les 5 articles est de 170 mn avec l'assistant et 150 mn sans, mais il y a presque deux fois plus de résultats annotés avec l'assistant. Nous observons que huit fois plus de commentaires sont ajoutés lorsque l'expert utilise l'assistant, par rapport à l'annotation *manuelle*.

12. Un étudiant en bioinformatique, un post-doctorant, un ingénieur d'étude en bioinformatique, une *assistant professor* en ingénierie chimique et biologique et un directeur de recherche spécialisé en physiologie rénale (le concepteur de QKDB).

Condition	# résultats	# descripteurs	# commentaires	Temps passé par résultat
Manuellement	52	448	9	174s
Avec l'assistant	96	846	75	105s

Tableau 10. *Nombre de résultats expérimentaux insérés dans la base manuellement et avec l'assistant, et temps moyen passé pour annoter un résultat.*

Une partie des questions posées aux experts après l'annotation est présentée dans le tableau 11. Pour chaque affirmation, nous leur demandions de mettre un score de 1 à 5, de d'accord avec l'affirmation à pas d'accord. Ils étaient tous d'accord pour dire que l'assistant facilite et accélère la tâche d'annotation, et que le surlignage des descripteurs permettait une meilleure visibilité, ce qui simplifiait et accélérait la tâche.

Après chaque étape, nous demandions à l'expert s'il pensait avoir annoté tous les résultats pertinents de l'article. Ils répondaient tous oui quand ils utilisaient l'assistant, et seulement la moitié répondaient oui quand ils annotaient l'article à la main, et trois experts sur cinq ont dit que de tout annoter à la main prenait trop de temps. Un expert a lu les articles en entier avant de procéder à l'annotation, les autres n'ont lu que les parties *Results* et *Methods* (et *Abstract* pour l'un d'entre eux).

	Score
L'assistant facilite la tâche d'annotation.	1
L'assistant accélère le traitement de l'article.	1
Le surlignage des termes simplifie la tâche.	1
Le surlignage des termes accélère la tâche.	1
La visualisation des résultats un par un est utile.	1,4
Les infobulles sont utiles.	1,4
Les liens entre le formulaire et le texte facilitent la tâche.	1,2
L'interface est simple d'utilisation.	1,4

Tableau 11. *Moyenne des scores du questionnaire. Les scores vont de 1, si les experts sont d'accord avec l'affirmation, à 5, s'ils ne sont pas d'accord.*

7. Discussion et Conclusion

Nous avons présenté dans cet article un système complet permettant le peuplement d'une base de données. Cette base est formalisée par une ressource termino-ontologique qui définit un résultat d'expérimentation par une relation n-aire. Ce système constitue l'un des rares travaux qui permet de remplir ce type de base, plus complexe que les bases sur la génomique par exemple où l'on trouve plutôt des relations binaires à instancier. Il répond à un besoin dans le domaine, puisque la base conçue par des biologistes, et adaptée à leurs besoins, n'est pas alimentée à cause du

temps que cette opération demande. Les premières présentations faites ont vivement intéressé les chercheurs du domaine.

Du point de vue de l'extraction d'information, l'originalité de notre approche réside dans le fait que les données sont extraites des articles complets et nécessitent d'être retrouvées dans différentes sections de l'article, et que nous extrayons les informations qu'elles soient présentes dans le texte ou dans des tableaux. Les principales difficultés proviennent de la présence de nombreuses variations terminologiques et de la mise en relation des descripteurs avec un résultat. Le système d'extraction atteint un résultat de très bon niveau (F-mesure de 0,78), et permet de trouver tous les résultats des articles, sans pour autant fournir trop de bruit devenant gênant pour les experts curateurs (rappel de 1 et précision de 0,63).

L'assistant développé pour la curation des données extraites a été évalué par différents experts. Ses performances et son ergonomie entraînent les experts à extraire tous les résultats d'un article, là où leur extraction manuelle était moins exhaustive, notamment dans le cas des tableaux dont le traitement est allégé et rendu moins fastidieux. Les résultats obtenus sur ces cas particuliers sont d'ailleurs très bons, ce qui permet de se reposer sur l'assistant.

La méthodologie mise en œuvre ainsi que les choix de développement permettent de transposer ce travail à d'autres domaines où il est important pour leur étude de représenter et décrire des résultats expérimentaux. La formalisation sous forme d'ontologie permet de voir les concepts à définir et le niveau de généralité désiré. Le lexique initial contient les termes qui désignent ces concepts, il peut ensuite être étendu dès lors que l'on dispose d'articles. Si des ontologies sont créées, elles pourront être intégrées sans problème dans notre système, puisque nos lexiques séparent bien les concepts des termes qui les désignent.

Nous n'avons pas encore pu évaluer l'adaptation de notre système à un autre domaine, mais nous souhaiterions le faire prochainement. Il est envisagé d'incorporer notre assistant dans le package QxDB générique (distribué dans la communauté du Physiome/VPH Européen), afin de faciliter l'adoption de cette approche dans d'autres domaines.

Plusieurs améliorations du système sont envisageables ; dans un premier temps, il faudrait travailler sur la mise en relation des descripteurs et de la valeur numérique dans les tableaux complexes. Au préalable une amélioration de la transformation des tableaux du format XHTML en XML est nécessaire, car certaines structures ne sont pas conservées lors de la conversion. Dans un deuxième temps, lors de la mise en relation des descripteurs, il serait intéressant d'étudier et de prendre en compte les liens implicites existant entre les descripteurs. Par exemple, si le descripteur de type paramètre est *concentration* alors il y a forcément un soluté associé. Ou encore, si l'unité est *g* le paramètre sera *weight*. Le dernier point pouvant permettre une amélioration de la précision du système serait d'analyser plus précisément avec un expert en quoi un résultat d'expérimentation est pertinent ou non pour la base de données. Ces améliora-

tions nécessitent d'augmenter la taille du corpus ainsi que d'améliorer son annotation. La mise en ligne de notre outil sera un pas important dans ce sens.

8. Bibliographie

Ahn D., « The stages of event extraction », *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, ARTE '06, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 1-8, 2006.

Alex B., Grover C., Haddow B., Kabadjov M., Klein E., Matthews M., Roebuck S., Tobin R., Wang. X., « Assisted Curation : does Text Mining Really Help ? », *In Proceedings the Pacific Symposium on Biocomputing*, 2008.

Ananiadou S., « A methodology for automatic term recognition », *Proceedings of the 15th conference on Computational linguistics - Volume 2*, COLING '94, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 1034-1038, 1994.

Ananiadou S., Pyysalo S., Tsujii J., Kell D. B., « Event extraction for systems biology by text mining the literature », *Trends in Biotechnology*p. 381-390, 2010.

Aronson A. R., « Effective Mapping of Biomedical Text to the UMLS Metathesaurus : The MetaMap Program », *AMIA 2001 Symposium Proceedings*, 2001.

Aronson A. R., Mork J. G., Gay C. W., Humphrey S. M., Rogers W. J., « The NLM Indexing Initiative's Medical Text Indexer », *In Proceedings of the 11th World Congress on Medical Informatics Demner-Fushman and Lin Answering Clinical Questions (MEDINFO 2004*, p. 268-272, 2004.

Ashburner M., Ball C. A., Blake J. A., Botstein D., Butler H., Cherry J. M., Davis A. P., Dolinski K., Dwight S. S., Eppig J. T., Harris M. A., Hill D. P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J. C., Richardson J. E., Ringwald M., Rubin G. M., Sherlock G., « Gene ontology : tool for the unification of biology. The Gene Ontology Consortium. », *Nature genetics*, vol. 25, n° 1, p. 25-29, 2000.

Baumgartner W. A., Cohen K. B., Fox L. M., Acquaah-Mensah G., Hunter L., « Manual curation is not sufficient for annotation of genomic databases », *Bioinformatics*, vol. 23, n° 13, p. i41-i48, 2007.

Björne J., Ginter F., Pyysalo S., Tsujii J., Salakoski T., « Complex event extraction at PubMed scale », *Bioinformatics*, vol. 26, p. i382-i390, 2010.

Bodenreider O., Rindflesch T., Burgun A., « Unsupervised, corpus-based method for extending a biomedical terminology », *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3*, Association for Computational Linguistics, p. 53-60, 2002.

Buyko E., Faessler E., Wermter J., Hahn U., « Event extraction from trimmed dependency graphs », *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing : Shared Task*, BioNLP '09, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 19-27, 2009.

Chang C.-C., Lin C.-J., *LIBSVM : a library for support vector machines*. 2001.

Cohen A. M., Hersh W. R., « A survey of current work in biomedical text mining », *Briefings in Bioinformatics*, vol. 6, n° 1, p. 57-71, 2005.

Corney D., Buxton B., Langdon W., Jones D., « BioRAT : extracting biological information from full-length papers », *Bioinformatics*, vol. 20, n° 17, p. 3206, 2004.

Dinh D., Tamine L., « Biomedical concept extraction based on combining the content-based and word order similarities », *Proceedings of the 2011 ACM Symposium on Applied Computing*, SAC '11, ACM, New York, NY, USA, p. 1159-1163, 2011.

Donaldson I., Martin J., de Bruijn B., Wolting C., Lay V., Tuekam B., Zhang S., Baskin B., Bader G., Michalickova K., Pawson T., Hogue C., « PreBIND and Textomy - mining the biomedical literature for protein-protein interactions using a support vector machine », *BMC Bioinformatics*, vol. 4, n° 1, p. 11, 2003.

Dzodic V., Hervy S., Fritsch D., Khalfallah H., Thereau M., Thomas S. R., « Web-based tools for quantitative renal physiology », *Cellular and Molecular Biology*, vol. 50, n° 7, p. 795-800, 2004.

Fellbaum C. (ed.), *WordNet An Electronic Lexical Database*, The MIT Press, Cambridge, MA ; London, 1998.

Garten Y., Altman R., « Pharmspresso : a text mining tool for extraction of pharmacogenomic concepts and relationships from full text », *BMC bioinformatics*, vol. 10, n° Suppl 2, p. S6, 2009.

Gold S., Elhadad N., Zhu X., Cimino J. J., George H., « Extracting Structured Medication Event Information from Discharge Summaries », *AMIA 2008 Symposium Proceedings*, p. 237-241, 2008.

Grouin C., Deléger L., Zweigenbaum P., « Extracting medical information from narrative patient records : the case of medication-related information », *JAMIA*, vol. 17, p. 555-558, 2010.

Guarino N., *Formal ontology in information systems : proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy*, vol. 46, Ios Pr Inc, 1998.

Hunter P., Coveney P. V., Bono B. d., Diaz V., Fenner J., Frangi A. F., Harris P., Hose R., Kohl P., Lawford P., McCormack K., Mendes M., Omholt S., Quarteroni A., Skar J., Tegner J., Thomas S. R., Tollis I., Tsamardinos I., Beek J. H. G. M. v., Viceconti M., « A vision and strategy for the virtual physiological human in 2010 and beyond », *Phil. Trans. R. Soc. A*, vol. 368, p. 2595-2614, 2010.

Jacquemin C., « A symbolic and surgical acquisition of terms through variation », *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, Springer-Verlag, London, UK, UK, p. 425-438, 1996.

Karamanis N., Lewin I., Seal R., Drysdale R., Briscoe E., « Integrating Natural Language Processing with Flybase Curation », *In Proceedings of the Pacific Symposium on Biocomputing*, 2006.

Kipper K., Korhonen A., Ryant N., Palmer M., « A large-scale classification of English verbs », *Language Resources and Evaluation*, vol. 42, n° 1, p. 21-40, 2008.

Maynard D., Peters W., « Using lexicosyntactic ontology design patterns for ontology creation and population », *Proc. of the Workshop on Ontology Patterns*, Citeseer, 2009.

McDonald R., Pereira F., Kulick S., Winters S., Jin Y., White P., « Simple algorithms for complex relation extraction with applications to biomedical IE », *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 491-498, 2005.

Minard A., Ligozat A., Grau B., « Extraction de résultats expérimentaux d'articles scientifiques pour le peuplement d'une base de données », *10th International Conference on statistical analysis of textual data (JADT)*, 2010.

Nenadic G., Ananiadou S., McNaught J., « Enhancing automatic term recognition through recognition of variation », *Proceedings of the 20th international conference on Computational Linguistics*, Association for Computational Linguistics, p. 604-610, 2004.

Nguyen Q. L., Tikk D., Leser U., « Simple tricks for improving pattern-based information extraction from the biomedical literature », *Journal of Biomedical Semantics*, 2010.

Noy N., Rector A., Hayes P., Welty C., « Defining N-ary Relations on the Semantic Web », *W3C Working Group Note*, 2006.

Patrick J., Li M., « High accuracy information extraction of medication information from clinical notes : 2009 i2b2 medication extraction challenge », *JAMIA*, vol. 17, p. 524-527, 2010.

Ribba B., Tracqui P., Boix J. L., Boissel J. P., Thomas S. R., « QxDB : a generic database to support mathematical modelling in biology », *Philos Transact A Math Phys Eng Sci*, vol. 364, n° 1843, p. 1517-32, 2006.

Rosse C., Mejino J. L. J., « A reference ontology for biomedical informatics : the Foundational Model of Anatomy », *J Biomed Inform*, vol. 36, n° 6, p. 478-500, 2003.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Proceedings of the International Conference on New Methods in Language Processing*, p. 44-49, 1994.

Touhami R., Buche P., Dibie-Barthélemy J., Ibanescu L., « An Ontological and Terminological Resource for n-ary Relation Annotation in Web Data Tables », *10th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2011)*, Crete, Grèce, p. 662-679, October, 2011.

Uzuner O., Solti I., Cadag E., « Extracting medication information from clinical text », *JAMIA*, vol. 17, p. 514-518, 2010.

Van Auken K., Jaffery J., Chan J., Muller H.-M., Sternberg P., « Semi-automated curation of protein subcellular localization : a text mining-based approach to Gene Ontology (GO) Cellular Component curation », *BMC Bioinformatics*, vol. 10, n° 1, p. 228, 2009.

Wattarujeekrit T., Shah P. K., Collier N., « PASBio : predicate-argument structures for event extraction in molecular biology », *BMC Bioinformatics*, 2004.

Wimalasuriya D. C., Dou D., « Ontology-based information extraction : An introduction and a survey of current approaches », *Journal of Information Science*, vol. 36, n° 3, p. 306, 2010.

Yeh A., Hirschman L., Morgan A., « Evaluation of text data mining for database curation : lessons learned from the KDD Challenge Cup », *Bioinformatics*, vol. 19, n° Suppl 1, p. i331, 2003.

Annexe

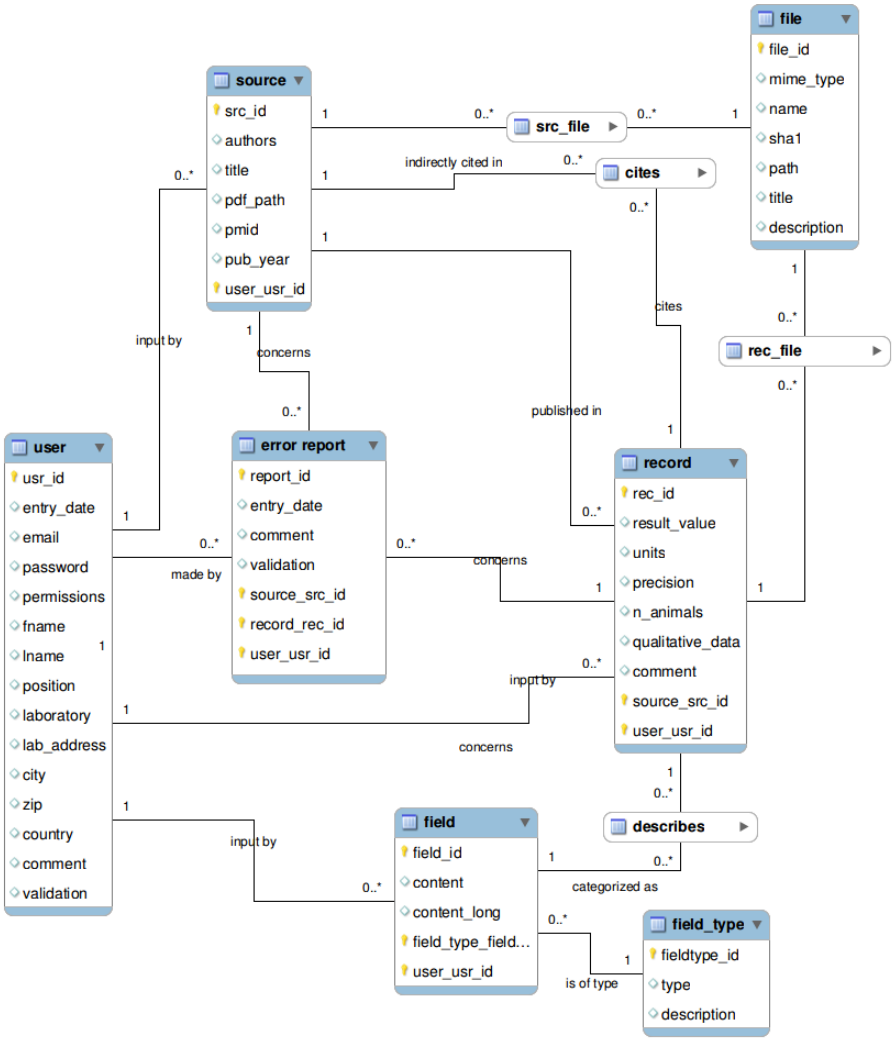

Figure 12. *Schéma complet de la base de données QKDB*

On peut noter que les tables *field* et *field_type* peuvent être étendues au besoin, et les changements sont immédiatement visibles dans l'interface Web. Dans la mesure où le caractère «rénal» de QKDB dépend seulement du contenu de ces deux tableaux, il suffit de modifier leurs contenus pour adapter la base à un domaine de recherche différent (d'où la version maquette générique, QxDB (Ribba *et al.*, 2006)).

Article reçu le 19/03/2012.
Version révisée le 20/12/2012.
Rédacteur responsable : GUILLAUME LAURENT

SERVICE ÉDITORIAL – HERMES-LAVOISIER
14 rue de Provigny, F-94236 Cachan cedex
Tél. : 01-47-40-67-67
E-mail : revues@lavoisier.fr
Serveur web : http://www.revuesonline.com

ANNEXE POUR LE SERVICE FABRICATION
A FOURNIR PAR LES AUTEURS AVEC UN EXEMPLAIRE PAPIER
DE LEUR ARTICLE ET LE COPYRIGHT SIGNE PAR COURRIER
LE FICHIER PDF CORRESPONDANT SERA ENVOYE PAR E-MAIL

1. ARTICLE POUR LA REVUE :

L'objet. Volume X – n°X/2011

2. AUTEURS :

Anne-Lyse Minard[*, **] — *Brigitte Grau*[*, ***] — *Anne-Laure Ligozat*[*, ***] — *Stephen Randall Thomas*[**, ****]

3. TITRE DE L'ARTICLE :

Extraction de relations complexes : application à des résultats expérimentaux en physiologie rénale

4. TITRE ABRÉGÉ POUR LE HAUT DE PAGE MOINS DE 40 SIGNES :

Extraction de relations complexes

5. DATE DE CETTE VERSION :

21 novembre 2012

6. COORDONNÉES DES AUTEURS :

– adresse postale :
[*] LIMSI-CNRS, rue John von Neumann, 91400 Orsay, France
[**] Université Paris-Sud, 91400 Orsay, France
[***] ENSIIE, 1 square de la résistance, 91000 Évry, France
[****] IR4M CNRS UMR 8081, 94805 Villejuif, France
prenom.nom@limsi.fr, stephen-randall.thomas@u-psud.fr
– téléphone : 01 69 85 80 10
– télécopie : 01 69 85 80 88
– e-mail : annlor@limsi.fr

7. LOGICIEL UTILISÉ POUR LA PRÉPARATION DE CET ARTICLE :

LaTeX, avec le fichier de style `article-hermes.cls`,
version 1.23 du 17/11/2005.

8. FORMULAIRE DE COPYRIGHT :

Retourner le formulaire de copyright signé par les auteurs, téléchargé sur :
`http://www.revuesonline.com`

SERVICE ÉDITORIAL – HERMES-LAVOISIER
14 rue de Provigny, F-94236 Cachan cedex
Tél. : 01-47-40-67-67
E-mail : revues@lavoisier.fr