Qualité des données de santé disponibles en France et de leurs modèles - Comment la garantir pour répondre aux enjeux de la gestion des connaissances médicales ?

Karine Steinberg

CONSERVATOIRE NATIONAL DES ARTS ET METIERS

Ecole Management et Société-Département CITS

INTD

MEMOIRE pour obtenir le Titre enregistré au RNCP

"Chef de projet en ingénierie documentaire"

Niveau I

Présenté et soutenu par

Karine Steinberg

Le 08/12/2016

Qualité des données de santé disponibles en France et de leurs modèles

Comment la garantir pour répondre aux enjeux de la gestion des connaissances médicales ?

Jury :
Monsieur Gérald Kembellec, Maitre de conférences, INTD-CNAM
Monsieur Jean Charlet, chargé de mission à l'Assistance Publique – Hôpitaux de Paris, chercheur INSERM au LIMICS (Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en eSanté – U1142)

Promotion 46

A mon fils, ma fierté, Jordan, qui me construit chaque jour.

Remerciements

Je tiens ici à remercier toutes les personnes proches qui m'ont soutenue moralement durant ce travail de recherche passionnant, pour mener à bien l'écriture de ce mémoire. En particulier, merci au père de mon fils qui a toujours respecté mes choix, et tenté de les accompagner au mieux, et à ce grand fils, pour sa patience, lorsque j'étais totalement absorbée dans mon étude. Je n'oublie pas mes parents, toujours présents pour moi, et prêts à me soutenir.

J'ai un grand plaisir à remercier les membres de mon jury : Monsieur Gérald Kembellec, mon professeur à l'INTD qui m'a enseigné durant cette année de formation, les bases du Web sémantique d'un point de vue technique, et Monsieur Jean Charlet, mon maître de stage, qui m'a donné l'opportunité de travailler dans son équipe au sein du LIMICS, et m'a permis de découvrir le domaine scientifique des ontologies, en me communiquant sa passion et sa maîtrise du sujet, avec une humilité et un professionnalisme remarquables.

Je remercie également particulièrement Monsieur Xavier Aimé, ontologue, qui m'a beaucoup guidée et conseillée sur les orientations à suivre durant mon projet (en l'occurrence le sien à l'origine), Madame Marie-Christine Jaulent pour son accueil bienveillant au sein du laboratoire qu'elle dirige, ainsi que Madame Isabelle Verdier-Kehren, chargée de gestion, pour sa gentillesse et son implication dans le suivi administratif, et la vie quotidienne de l'équipe.

En règle générale, l'équipe de chercheurs et doctorants que j'ai pu côtoyer durant ce stage, m'ont apporté un enrichissement personnel et humain, de par la diversité, et l'intérêt des sujets qu'ils traitent, ainsi que la variété de leurs parcours respectifs. Je leur en suis très reconnaissante, et n'aurais eu qu'un seul souhait, celui d'avoir pu prolonger ce court stage de 3 mois en leur compagnie.

Je tiens enfin à remercier Madame Claudine Milstein, architecte de données à l'ANSM, pour ses explications sur les données mises à disposition sur les plateformes gouvernementales, et la fourniture des fichiers non disponibles directement.

Ce stage a mis un terme à une année d'étude à l'INTD, qui m'a beaucoup apporté, que ce soit au niveau des nouvelles connaissances et compétences acquises, mais aussi et surtout, concernant la richesse des auditeurs, enseignants et intervenants rencontrés. J'en sors avec la satisfaction et la fierté d'avoir partagé tous ces moments avec des personnes d'univers si différents, que je n'oublierai pas.

Notice

STEINBERG Karine. Qualité des données de santé disponibles en France et de leurs modèles : comment la garantir pour répondre aux enjeux de la gestion des connaissances médicales ? Mémoire professionnel INTD, Titre I, Chef de projet en ingénierie documentaire. Conservatoire national des arts et métiers – Institut national des Sciences et Techniques de la Documentation, 2016, 117p. Promotion 46.

Résumé : Ce mémoire effectue une approche de l'ouverture des données publiques en France, axée sur l'évaluation de la qualité de ces données mises à disposition, ainsi que des modèles permettant de les véhiculer, les partager et les réutiliser. Les recherches et constats réalisés concernent en particulier le domaine médical et la gestion des connaissances en e-Santé, et s'appuient sur un cas pratique de création d'une ontologie française des médicaments. Il soulève ainsi les contraintes politiques, législatives, techniques et structurelles auquel un projet innovant de mise en œuvre d'un référentiel terminologique doit se confronter, pour exploiter et valoriser les données ouvertes. Les apports liés à la standardisation et l'interopérabilité offertes par les technologies du Web sémantique et des données liées, sont largement abordés.

Descripteurs : ouverture des données en France; ontologie des médicaments ; Web sémantique ; Web de données ; qualité des données ; modèles conceptuels de données ; réutilisation des données; ingénierie des connaissances ; référentiels terminologiques ; e-santé ; interopérabilité.

Abstract: This master thesis is an approach to Open Data in France, focused on evaluating the quality of the data provided, as well as models which carry, share and reuse them. The research and findings particularly concern the medical field and the management of e-health knowledge, and rely on a case study of creating a French drugs ontology. It thus raises the political, legislative, technical and structural which an innovative implementation project of a terminology repository must confront, in order to exploit and promote Open Data. The contributions related to standardization and interoperability offered by the technologies of the semantic Web and linked Open Data are widely discussed.

Keywords: Open Data ; drugs ontology ; semantic Web, linked Open Data ; data quality ; data model ; data reuse, knowledge management ; interoperability.

Table des matières

Table des matières

Liste des tableaux

Liste des figures

Introduction

Les données constituent une valeur précieuse et essentielle de notre civilisation, car elles permettent l'élaboration de la connaissance: elles définissent, décrivent et structurent l'information, qui elle-même conduit au savoir et à l'enrichissement de nos pensées, idées, et des progrès humains en général.

Ces données sont considérées comme un bien commun, c'est-à-dire une ressource partageable par la communauté, mais dans un cadre défini de conditions d'accès, afin de la préserver. Le nouveau modèle économique qui en résulte offre une vision qui bouscule les perspectives, car la diffusion, la circulation et l'utilisation des données permettent leur valorisation et non pas leur épuisement. La collaboration permet l'enrichissement de la connaissance, comme on peut le constater dans l'exemple de Wikipédia[1], considérée aujourd'hui comme étant l'encyclopédie de référence universelle et multilingue.

L'ouverture des données s'impose donc comme une condition à l'innovation, et constitue une réalité économique, et politique, qui n'a pas pu être niée par les Etats : le mouvement Open Data qui concerne les données publiques, a donc vu le jour en 2009, tel qu'on le connait aujourd'hui dans sa version moderne. Il est le résultat d'une lente gestation, et de certaines actions historiques concrètes, détaillées dans ce mémoire, mouvement qui s'est accéléré avec les progrès technologiques et les changements sociétaux.

Nous nous sommes intéressés à comprendre pourquoi les gouvernements français et internationaux ont été conduits à mettre en œuvre l'ouverture des données publiques (*Open Data*), et ainsi accepter de livrer de manière transparente les données issues des administrations. Les enjeux, les conditions politiques, législatives, et les moyens et outils de gouvernance détaillés tout au long du document, nous permettront de nous rendre compte de la difficulté et de l'ampleur de la tâche.

Un accent particulier est mis sur les résultats observés aujourd'hui concernant la qualité des données diffusées, sur leur capacité de réutilisation et jusqu'à quel point. Nous montrerons que les modèles porteurs de ces données, en constituent le fondement pour délivrer la qualité attendue, mais aussi la clé pour les comprendre et les exploiter à bon escient. L'évolution des modèles grâce aux technologies du web sémantique, et en l'occurrence des ontologies permet de donner une nouvelle dimension aux réseaux de connaissances. En effet, les relations entre toutes ces données sont ainsi rendues possibles, et des domaines complets peuvent être décrits, ce qui ouvre la voie aux raisonnements intelligents réalisés par la machine.

Un cas pratique, la création de l'ontologie française des médicaments, met en évidence les objectifs, attentes et contraintes liés à la réutilisation des données ouvertes, dans le domaine complexe de la e-Santé. Celui-ci est par conséquent étudié et analysé de ce point de vue, afin de situer ses enjeux particuliers, ainsi que ses problématiques. Les sources de données disponibles, les modèles existants qui ont servi à réaliser le projet d'ontologie illustrent les difficultés qu'observe chaque acteur de la chaine de distribution des données de santé, mais aussi les avancées et réalisations que la réutilisation et le partage permettent.

[1] https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Accueil_principal

L’Open Data est un formidable levier pour les citoyens mieux informés et collaborants, les entreprises qui développent chaque jour de nouvelles applications grâce aux données libérées, aux chercheurs qui bénéficient aussi de ces données et font avancer leurs résultats, mais aussi pour les administrations amenées à moderniser leurs processus internes.

Première partie : Open Data : quel contexte ?

1. Open Data à l'international

1.1. Historique et contexte international

Le mouvement pour la réutilisation des données publiques est un mouvement mondial né début 2009 et devenu un principe avec l'investiture de Barack Obama et le lancement en mars 2009 du projet Data.gov[2]. Aujourd'hui, dans le monde, une quinzaine d'Etats et une cinquantaine de communautés urbaines (sans compter les villes du Royaume-Uni, où une centaine de communautés urbaines se sont lancées dans la libération de données sous l'impulsion du gouvernement britannique) parmi les plus importantes villes américaines sont impliquées [1,GUILLAUD].

PublicData.eu[3] recense quelques 215 initiatives d'acteurs publics en Europe, que ce soit au niveau national (Belgique, Finlande, Suède, Norvège, Espagne, Grèce, France, Royaume-Uni…) ou régional.

Loin d'être né avec le numérique, le mouvement de libération des données s'inscrit dans la continuité d'une longue tradition qui a imposé un droit d'accès des citoyens à l'information produite par les administrations pour plus de transparence. Après la Seconde Guerre mondiale, le concept d'*Open Governement* a ainsi émergé pour désigner le droit des citoyens à accéder aux secrets de l'Etat. Aux Etats-Unis, la dénonciation de l'opacité de l'armée pendant la Guerre du Vietnam aboutit à l'adoption en 1966 du *Freedom of Information Act* qui oblige les agences fédérales des États-Unis à transmettre leurs documents non classifiés à tout citoyen qui en fait la demande.

En parallèle de ce mouvement, les milieux économiques se sont mobilisés en faveur de la libération des données, susceptible de créer de la valeur. Pour y parvenir la condition était d'imposer la gratuité des données, ce qu'est parvenu à faire le lobby de l'industrie de l'information dans les années soixante-dix.

À partir de 2007, on assiste à une surprenante convergence entre les enjeux techniques et politiques. Les acteurs du monde économique et les militants de la transparence vont inscrire l'ouverture des données comme une des priorités des politiques numériques. Libérer les données serait susceptible de créer de la richesse tout en développant la transparence de l'État. Ce tournant est incarné par la réunion dite « de Sébastopol », où un ensemble d'activistes du numérique vont demander la libération des données publiques dès leur production, dans leur intégralité et telles qu'elles sont collectées.

Les citoyens comme les entreprises pourraient alors refaire les calculs, en récupérant les données et les outils pour les interpréter, plutôt que de se contenter d'agrégats, et être ainsi en mesure de contester les décisions publiques, à partir des mêmes sources que l'administration.

Cela sous-entend de connaitre également comment sont développés les algorithmes [4,ERTZSCHEID], les hypothèses qu'on leur met en entrée et qui conditionne les résultats obtenus. En effet, *les algorithmes ne se trompent pas, les mathématiques ne mentent pas*[4], mais on peut leur faire tirer les conclusions que l'on souhaite, ils dépendent totalement de leurs créateurs et de ce qu'ils ont comme objectif. Le manque de transparence à ce niveau est avéré, et peut même être

[2] http://data.gov/
[3] http://publicdata.eu/
[4] http://affordance.typepad.com/mon_weblog/big-data/

considéré comme dangereux, puisqu'on ne peut vérifier par quel chemin un algorithme a donné un résultat, une « vérité ». En effet, le postulat de départ est peut-être faux, ou même le raisonnement peut comporter des règles de gestion qui « arrangent » le créateur.

Dans son blog Affordance, Olivier Ertzscheid donne sa vision de tous ces effets[5], et en arrive à l'une des conclusions suivantes : si l'action publique et les entreprises privées choisissent d'axer prises de décisions, ou analyses sur les algorithmes, des garanties doivent être données. Elles se doivent « d'être transparentes à l'inspection, prévisibles pour ceux qu'elles gouvernent, et robustes contre toute manipulation ».

En conclusion, le développement d'une politique de « gouvernement ouvert » implique de répondre à la fois à des impératifs politiques autour des questions de transparence et d'efficacité de l'action publique notamment, tout en tenant compte d'objectifs économiques, en termes de création de services et de valorisation des données.

1.2. Open Data en Europe

La Commission européenne lance en décembre 2012 son portail Open Data[6] [3], qui fournit aux européens les données publiques détenues par les institutions et organes de l'Union européenne.

Le portail «Données ouvertes» de l'Union européenne est un point d'accès unique à un éventail croissant de données produites par les institutions et organes de l'Union européenne (UE). Ces données peuvent être utilisées et réutilisées gratuitement à des fins commerciales ou non.

En proposant un accès simple et gratuit à ces données, le portail contribue à en promouvoir une utilisation innovante et à en exploiter le potentiel économique. L'objectif est également de renforcer la transparence et la responsabilité des institutions et organes de l'UE.

Le portail des données ouvertes de l'UE est géré par l'Office des publications de l'Union européenne. La mise en œuvre de la politique en matière de données ouvertes de l'UE incombe à la direction générale «Réseaux de communication, contenu et technologies» de la Commission européenne.

En 2013, l'Union européenne a adopté la directive 2003/98/CE concernant la réutilisation des informations du secteur public (directive PSI), actuellement transposée en France au travers du Projet de loi pour une République numérique.

1.3. Open Data en France

On peut considérer que l'Open Data trouve son origine dans la Déclaration des Droits de l'Homme et du Citoyen de 1789 et son article 15 qui stipule clairement que *La société a le droit de demander compte à tout agent public de son administration*. Il s'agit là d'un souhait de transparence de l'Etat, tel qu'il est aujourd'hui remis au goût du jour [2, MESZAROS *et al.*].

Il a ensuite fallu attendre les lois de 1978 pour poursuivre cette volonté d'ouverture. La première, le 6 janvier 1978 avec la promulgation de la loi Informatique

[5] http://affordance.typepad.com/mon_weblog/2016/06/france-culture-data-gafa-emploi.html
[6] https://data.europa.eu/

et libertés et la création simultanée de la Commission Nationale de l'Informatique et des Libertés (CNIL) chargée de la protection des données personnelles ; la seconde, le 17 juillet 1978 avec la mise en en place de la Commission d'Accès aux Documents Administratifs (CADA) dont l'objectif est de mettre à disposition du public, les données produites ou détenues par les administrations, dans le cadre de leurs missions de service public.

En 1998, le gouvernement a décidé de la publication en ligne et gratuite des *données publiques essentielles*, afin de prendre en compte les évolutions de l'internet et son importance pour la croissance du marché de l'information (programme d'action « Préparer l'entrée de la France dans la société de l'information » (16 janvier 1998)).

Quelques années plus tard, c'est-à-dire le 23 octobre 2000, le portail de l'administration française service-public.fr est mis en place par la Documentation Française (devenue Direction de l'Information Légale et Administrative DILA en 2010). Celle-ci lance également en 2002, le portail vie-publique.fr, entièrement refondu en novembre 2008, avec l'objectif d'aider les citoyens à comprendre les questions qui qui animent le débat public français.

C'est ensuite Kéolis et Rennes Métropole qui ont lancé la première initiative de portail « Open Data » en juin 2010 avec l'entrepôt de données de Rennes Métropole[7] [1,GUILLAUD].

D'autres initiatives régionales et territoriales l'ont rejoint, comme les plates-formes Open Data de Paris[8], de la Saône-et-Loire[9], de la Loire Atlantique[10], du Loir-et-Cher[11], de la Gironde[12], de la Communauté urbaine de Bordeaux[13], du Grand Toulouse[14], de Nantes[15], de Montpellier[16].

On assiste par la suite au lancement de la mission Etalab[17] le 21 février 2011, initiative portée par l'Etat pour mette en place des conditions nécessaires au développement des données ouvertes. Cela s'est traduit par la création de la Licence ouverte par Etalab, en octobre 2011, destinée entre autres, à régir les données du futur portail gouvernemental Datagouv.fr, ouvert, à son tour en décembre 2011 (en 2013, a eu lieu une refonte du site).

C'est également en 2011 que le principe de la gratuité du droit à réutiliser des documents et données publiques a été posé par le décret n° 2011-577.

Progressivement, les développeurs ont pu profiter des milliers de jeux de données « libérés » pour créer des applications mobiles, et améliorer ainsi l'efficacité des services publics, mais aussi la transparence de l'action publique.

Récemment une impulsion importante a été donnée à cette culture de l'Open Data avec l'adhésion de la France au mouvement international de l'*Open Gov*

---

[7] http://www.data.rennes-metropole.fr/
[8] http://opendata.paris.fr/
[9] http://www.opendata71.fr/
[10] http://data.loire-atlantique.fr/accueil/
[11] http://www.pilote41.fr/
[12] http://www.datalocale.fr/
[13] http://data.lacub.fr/
[14] http://data.grandtoulouse.fr/
[15] http://data.nantes.fr/
[16] http://opendata.montpelliernumerique.fr/
[17] La mission et les actions Etalab sont détaillées plus finement dans les chapitres qui suivent.

Partnership (OGP) en avril 2014 (dont elle devrait prendre la présidence en octobre 2016) et la création d'un poste d'administrateur général des données, occupé depuis le 16 septembre 2014 par Henri Verdier, responsable de la mission Etalab (aujourd'hui remplacé à ce poste par Laure Lucchesi en janvier 2016). Celui-ci, placé sous l'autorité du premier ministre est chargé de superviser et d'améliorer l'utilisation des données par l'administration et ses nombreux opérateurs [5].

Figure 1 : Evénements marquants Open Data en France [5].

L'OGP est une organisation internationale lancée en 2011 et composée de 65 pays chargés de valoriser les principes du gouvernement ouvert et de diffuser des « bonnes pratiques ». Ses membres s'engagent à élaborer, conjointement avec la société civile, un plan d'action sur les mesures qui seront développées pour accélérer l'ouverture de l'action publique. Il ambitionne notamment d'établir un cadre légal pour la publication des revenus des hauts fonctionnaires, de promouvoir une transparence budgétaire et fiscale et d'augmenter la participation des citoyens dans l'action publique.

A ce titre, chaque pays membre doit produire un Plan d'Action national visant à faire progresser la transparence et l'ouverture de l'action publique dans les domaines promus par le partenariat.

Aujourd'hui, les constats restent encore mitigés sur l'avancée de la libération des données, c'est un processus qui prend du temps, demande la participation active de toutes les parties prenantes, et des efforts à la fois politiques, économiques, réglementaires, organisationnels [6,CHIGNARD]. Un grand pas vient d'être effectué au travers de l'adoption du projet de loi pour le Numérique par le Sénat le 28 septembre 2016[18].

2. Fondamentaux et enjeux de l'Open Data

2.1. Open Access et Open Data : à ne pas confondre

Le mouvement du libre accès (*Open Access*) [7] désigne l'ensemble des initiatives prises pour une mise à disposition des résultats de la recherche au plus grand nombre, sans restriction d'accès, que ce soit par l'auto-archivage ou par des revues en libre accès. Cela implique la gratuité sur l'Internet public, permettant à tout un chacun de lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces articles, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale, sans barrière financière, légale ou technique autre que celles indissociables de l'accès et l'utilisation d'Internet. La seule contrainte sur la reproduction et la distribution, et le seul rôle du copyright dans ce domaine devrait être de garantir aux auteurs un

18 https://www.senat.fr/espace_presse/actualites/201603/projet_de_loi_pour_une_republique_numerique.html

contrôle sur l'intégrité de leurs travaux et le droit à être correctement reconnus et cités (Initiative de Budapest pour l'Accès Ouvert).

Le mouvement du Libre Accès est jalonné de déclarations, réactions à la prise de conscience quant à la difficulté d'accéder aux résultats scientifiques, qui proposent des solutions pour y remédier. Trois changements peuvent être repérés : (1) l'appropriation par la société d'un mouvement initié par des individus, (2) la prise en compte de toutes les données utiles à la recherche et pas seulement des publications, et (3) l'élargissement à tous les domaines de la science.

Le partage des résultats de la recherche publique donne l'avantage d'une gestion correcte de ces données, une meilleure efficacité dans leur production, une garantie de conservation pérenne, une facilité de citation, de partage et de réutilisation. Grâce à cette démarche structurante, on améliore ainsi l'efficience dans la création et le traitement de ces données. C'est-à-dire tout au long de leur recherche et non seulement en amont et/ou en aval [8].

La gestion des données de la recherche aide à la rédaction préliminaire d'un plan de gestion des données, à l'optimisation de la collecte de celle-ci en fonction des outils d'analyse, au traitement des données, à leur sécurisation et anonymisation, à la mise en place d'une préservation pérenne et de la « citabilité » qui l'accompagne.

Un monde entier sépare les données gouvernementales publiques, dont l'ouverture fait l'enjeu des politiques « *Open Data* », des données de la recherche proprement dites, adressée à un public essentiellement universitaire, et dont le potentiel économique reste, dans la plupart des cas, à démontrer [17,MULLER]. En matière d'éthique, rendre les données disponibles permet la validation des résultats de la recherche, ce qui pourrait éviter certains abus en termes de publications médicales ou de lobbying économiques des entreprises pharmaceutiques. Une des initiatives dans ce but, doit voir le jour vers mi-2017, et concerne la consultation des IPD (International Patient Data) anonymisées[19].

Nous nous intéresserons volontairement dans ce mémoire, uniquement à la politique *Open Data*, pour réduire le champ très vaste de l'ouverture des données en général, mais également parce que c'est l'aspect qui touche directement le sujet du stage au LIMICS (on présente le Laboratoire plus loin dans ce mémoire).

2.2. Les Fondamentaux

L'Open Data désigne donc l'effort que font les institutions, notamment gouvernementales, qui partagent les données dont elles disposent. Ce partage doit être gratuit, dans des formats ouverts, et permettre la réutilisation des données. Entre autres, le projet de loi pour une République numérique[20] énonce le principe selon lequel les informations publiques sont librement réutilisables par quiconque et à ses propres fins.

Pour être considérées comme « ouvertes », un certain nombre de principes doivent être respectés. Les données se doivent d'être :

[19] http://www.h2mw.eu/redactionmedicale/2016/01/licmje-groupe-de-vancouver-va-imposer-le-partage-des-données-et-dans-10-ans-cela-deviendra-commun.html

[20] http://www.lagazettedescommunes.com/462650/ce-quil-faut-retenir-du-projet-de-loi-pour-une-republique-numerique/

1. **Complètes**. Toutes les données sont mises à disposition. Elles ne doivent donc pas être concernées par des limitations sur la vie privée, la sécurité ou sur des privilèges d'accès.

2. **Primaires**. Les données sont telles que collectées à la source, avec la plus grande granularité possible, et ne se présentent pas sous des formes agrégées ou modifiées.

3. **Opportunes**. Elles sont mises à disposition aussi rapidement que nécessaire pour préserver leur valeur.

4. **Accessibles**. Les données sont accessibles au plus grand éventail d'utilisateurs possible et pour des usages aussi divers que possible.

5. **Lisibles par des machines**. Les données sont structurées pour permettre le traitement automatisé.

6. **Non discriminatoires**. Les données sont accessibles à quiconque, sans aucune obligation préalable ni inscription.

7. **Non propriétaires**. Les données sont accessibles dans un format sur lequel aucune entité ne dispose d'un contrôle exclusif.

8. **Sans permis**. Elles ne sont pas soumises au droit d'auteur, à brevet, au droit des marques ou au secret commercial. Des règles raisonnables de confidentialité, de sécurité et de priorité d'accès peuvent être admises.

2.3. Les enjeux

Produire et mettre à disposition des jeux de données est bien sûr le cœur de la politique *Open Data*, mais il s'agit également d'un enjeu majeur pour les états, car il ouvre la voie à la modernisation pour les acteurs publics. En effet, ce mouvement oblige à revoir l'ensemble des organisations publiques, et à lancer des chantiers pour recenser les données, rationaliser les systèmes d'information, aider au décloisonnement des services, améliorer la qualité des données [12,FAUVEL]. De plus, il sous-tend une démocratisation des outils et services avec les usagers, en créant le dialogue et l'interaction entre les administrations et leur public [15,PERES].

Les différentes parties sont donc plus impliquées dans le processus : au niveau de la production, les procédures de qualité se développent que ce soit pour réaliser les inventaires de données[21], les collecter, les créer ou les mettre à jour. Il s'agit ensuite de donner l'intérêt et les moyens pour que le public s'en empare, les utilise et trouve de nouvelles fonctionnalités. Et ce, de manière libre et gratuite, sous un format permettant leur traitement automatisé et leur réutilisation [14,VILLE]. Par exemple, les séries complètes, les données permettant de construire des référentiels, les données fréquemment actualisées, les données géo-localisées ou encore les données portant sur la transparence de l'action publique sont particulièrement utiles et recherchées.

Avec la révolution numérique, les données prennent par ailleurs une place centrale dans l'économie [11,FAUVEL]. Ouvrir et partager les données publiques, c'est organiser la mise en ligne de données essentielles, qui vont enrichir les analyses de nombreux décideurs, permettre de nombreuses économies de temps de travail ou

[21] Informations issues d'une autorité administrative dans le cadre de sa mission de service public

permettre, dans de nombreux secteurs, des prises de décisions mieux informées. C'est créer de grands référentiels partagés par tous les acteurs et encourager le développement de nombreux services à forte valeur ajoutée, par exemple dans le tourisme, le transport, la santé ou la maîtrise de la consommation d'énergie [13,CONTE].

Ainsi, la « valeur » des données ouvertes constitue une richesse, d'une part via les économies qui peuvent être réalisées grâce à la transparence de la gestion de l'action publique ou privée [14], et d'autre part, via la création de divers services pouvant se traduire par de nouvelles initiatives entrepreneuriales, disposant d'une valeur marchande et pouvant donc être chiffrée. Le bénéfice ne se mesure plus par la vente directe, mais au travers des services rendus, tels que l'amélioration de la fidélisation des usagers ou des clients d'une marque ou d'une société, l'évolution de la logique de communication et de marketing plus transparents des acteurs concernés, ou encore la participation de la communauté d'utilisateurs à l'enrichissement des services, à partir de besoins non identifiés initialement [2, MESZAROS *et al.*].

En effet, la notion des données ouvertes n'est pas réservée strictement aux données publiques mais c'est l'ensemble des secteurs et des acteurs qui produisent ou travaillent avec les données qui sont concernés [18,TRANGER]. De nombreuses sociétés publiques et privées se prêtent également au jeu de l'ouverture pour des raisons qui peuvent aller d'une simple volonté de transparence, comme par exemple le groupe ENEL[22] en Italie, jusqu'à intégrer cette logique de l'ouverture, en tant que partie intégrante du business plan, comme c'est le cas pour Keolis[23]. Comme on peut alors le voir, les données publiques n'ont pas le monopole de l'ouverture même si ce sont elles qui représentent pour le moment la majorité des données ouvertes mises à disposition.

Enfin, malgré la volonté de l'État de renforcer la gratuité de l'accès aux données ouvertes et même si la logique de la redevance semble être en déclin, cela reste une question à considérer. En effet, certains prévoient déjà, comme c'est le cas du Grand Lyon, de reconsidérer leur approche vis-à-vis du sujet en préparant, par exemple, plusieurs niveaux d'accès aux données qui vont du gratuit jusqu'au payant.

Afin d'encadrer et de soutenir les transformations nécessaires à l'ouverture des données, de faire en sorte que les fondamentaux soient respectés, que les conditions de réglementation pour la mise à disposition, et l'utilisation soient remplies, un organisme d'Etat a donc été mis en place, Etalab. On peut également noter la présence d'initiatives locales pour être force de proposition et nourrir le débat grâce à la collectivité, comme l'action du *Think Tank* « Renaissance Numérique », et ses 13 propositions d'axes d'amélioration des actions de l'Etat [16].

Le chapitre suivant décrit le rôle et les actions d'Etalab.

2.4. Mission Etalab[24]

La politique d'ouverture et de partage des données publiques (« *Open Data* ») est pilotée, sous l'autorité du Premier ministre, par la mission Etalab, dirigée par Mme Laure Lucchesi.

[22] http://www.usinenouvelle.com/article/les-entreprises-tentees-par-l-ouverture.N168879
[23] https://datamobilite.wordpress.com/2016/04/25/la-solution-open-data-mobilite-keolis/
[24] https://www.etalab.gouv.fr/qui-sommes-nous

Etalab coordonne l'action des services de l'Etat et de ses établissements publics pour faciliter la réutilisation la plus large possible de leurs informations publiques, c'est-à-dire la mise à disposition gratuite, libre et facile des données publiques, conformément au principe fixé par les circulaires du Premier ministre du 26 mai 2011 et du 13 septembre 2013 relatives à l'ouverture des données publiques en général, et particulièrement celles concernant la santé, l'éducation, et à fort potentiel d'innovation économique et sociale.

2.4.1. Portail datagouv.fr[25]

Etalab administre le portail interministériel data.gouv.fr destiné à rassembler et à mettre à disposition librement l'ensemble des informations publiques de l'Etat, de ses établissements publics et, si elles le souhaitent, des collectivités territoriales et des personnes de droit public ou de droit privé chargées d'une mission de service public. Data.gouv.fr contribue à rendre des comptes aux citoyens sur le fonctionnement de l'État et de ses administrations en permettant une plus grande transparence de leur fonctionnement. Etalab est chargé de coordonner le travail des administrations. Il fournit le site et son moteur de recherche, le cadre, mais les agents d'Etalab ne produisent pas les données. Etalab a une autre mission, qui est d'organiser le dialogue entre les producteurs de données et les réutilisateurs.

En juin 2016, 21 420 séries de données sont disponibles sur data.gouv.fr. La plateforme data.gouv.fr permet aux services publics de publier des données et à la société civile de les enrichir, modifier, interpréter en vue de coproduire des informations d'intérêt général. Il est le premier site au monde à proposer aux usagers d'enrichir et d'améliorer les données disponibles, et également d'en déposer de nouvelles.

Des exemples d'ouvertures de données depuis début 2014 : liste des maires au 17 juin 2014 (Ministère de l'intérieur), les condamnations (statistiques à partir du casier judiciaire nationale - Ministère de la justice), prix des carburants en France depuis 2007 (Ministère de l'économie, de l'industrie et du numérique), émissions de CO2 et de polluants des véhicules commercialisés en France (Ademe), base de données publiques du médicament (Ministère des affaires sociales et de la santé), l'aide publique au développement de la France (Ministère des affaires étrangères),…

2.4.2. Vademecum [26]

Etalab fournit un certain nombre de recommandations et principes de mise à disposition des données aux administrations.

Elle souhaite que les données publiques ouvertes correspondent à des données brutes dans des formats normalisés, permettant une réutilisation simplifiée dans des applications. La diffusion est permise au travers d'interfaces de programmation (API). Elle demande également dans la mesure du possible, que les données diffusées soient précises, exhaustives, et que les référentiels soient décrits. Ceci implique entre autres la qualification des métadonnées et l'indexation : Etalab propose ainsi des champs de description normalisée à tous les producteurs de données publiques afin de spécifier le contexte et le contenu des données. Il leur est notamment demandé de caractériser leurs données (titre, description, mots-clés…).

Étant donné que les données publiques sont produites ou reçues dans le cadre d'une mission de service public, elles sont généralement d'une qualité permettant le

[25] https://www.data.gouv.fr/fr/

[26] http://www.modernisation.gouv.fr/sites/default/files/fichiers-attaches/vademecum-ouverture.pdf

travail quotidien de l'administration. Il est donc suggéré que le document annexe présentant les jeux de données précise les méthodes de production et les limites intrinsèques des données proposées. De plus, la plateforme data.gouv.fr doit rester autant que possible la référence nationale des acteurs publics : dans le cas de portails existants pour certaines administrations, collectivités locales ou opérateurs, une fiche de référencement sur data.gouv.fr suffit, à la condition qu'elle soit bien documentée avec les métadonnées concernées.

La publication des données peut être réalisée manuellement sur l'espace d'administration de data.gouv.fr, ou de manière automatisée pour des volumes de données importants issus de systèmes d'informations ou fréquemment mis à jour, via une interface standardisée, documentée et gratuite.

Etalab est conscient que l'existence d'erreurs peut subvenir, mais insiste sur le fait que cela ne doit pas ralentir la démarche globale d'ouverture et de partage des données publiques, et qu'au contraire, leur signalement par les ré-utilisateurs favorise l'amélioration de la qualité.

Concernant le concept de gratuité, il est précisé que même si le droit n'interdit pas systématiquement la facturation du coût de mise à disposition des données publiques pour des questions d'équilibre budgétaire des producteurs de données, la redevance ne doit pas être un frein à l'innovation, et doit donc rester limitée.

2.4.3. Licence Etalab

La réutilisation des données publiques peut susciter le développement de nouveaux services comme les applications mobiles, des sites Internet, des visualisations de données ou « datavisualisation ». Elle doit être autorisée sans restriction autre que celles prévues par la loi CADA (qui demande que ces *informations ne soient pas altérées, que leur sens ne soit pas dénaturé et que leurs sources et la date de leur dernière mise à jour soient mentionnées*). Les données publiques peuvent être aussi réutilisées par les chercheurs, les enseignants, les étudiants, les responsables associatifs, les citoyens, pour construire de nouveaux états sur la société ou sur l'action publique.

Etalab a donc conçu la « Licence Ouverte / *Open Licence* ».

Cette licence garantit la plus grande liberté de réutilisation tout en apportant la plus forte sécurité juridique aux producteurs et aux ré-utilisateurs des données publiques. La reproduction, redistribution, adaptation et exploitation commerciale des données est autorisée, et ce de manière compatible avec les standards internationaux (ODC-BY, CC-BY 2.0), et notamment celles du gouvernement britannique (*Open Government Licence*)[27].

D'où une exigence forte de transparence de la donnée et de qualité des sources en rendant obligatoire la mention de la paternité : le standard réutilisable permet également de pouvoir mutualiser les données et se servir d'autres données existantes.

[27] Les licences Open Data sont détaillées dans un chapitre ci-après de ce mémoire

Figure 2 : Logos utilisables licence Etalab.

2.4.4. Plan d'action national[28]

En rejoignant le Partenariat pour un Gouvernement Ouvert (PGO) en avril 2015, la France s'est engagée à élaborer, en concertation avec la société civile, un plan d'action national.

Au préalable, le Premier ministre a instauré en septembre 2014 la fonction d'Administrateur général des données (« *Chief Data Officer* ») au niveau national. L'Administrateur général des données est chargé de mettre en place une gouvernance de la donnée, veiller à la qualité des données produites par l'État, et donner la volonté, les moyens et la capacité aux administrations de suivre le mouvement.

Le plan d'action national a été publié et transmis au PGO en juillet 2015 et couvre la période 2015-2017. La France prend, au total, 26 engagements. Ces derniers entrent dans le cadre des prochaines cibles à atteindre, les enjeux de l'ouverture des données publiques dépassant aujourd'hui les premières actions de création de la mission Etalab et du portail data.gouv.fr et de création de la fonction d'Administrateur général des données. Il s'agit à présent de construire un droit à la donnée publique et un droit des données publiques, alors que les économies mondiales et les pratiques gouvernementales évoluent de plus en plus rapidement sous l'effet de la révolution numérique.

Un rapport d'autoévaluation a été réalisé à mi-parcours et sera mis à jour en continu et en mode participatif, jusqu'en juillet 2017, date à laquelle la France devra remettre une version finale, et soumettre son deuxième plan d'action[29].

Côté budget, le projet de loi de finance 2017 prévoit d'allouer 2,5 millions d'euros à la mission Etalab pour la gestion du portail data.gouv.fr, la valorisation des projets de Data Sciences et les actions du PGO[30].

[28] https://www.etalab.gouv.fr/plan-daction-national
[29] https://suivi-gouvernement-ouvert.etalab.gouv.fr/fr/Introduction.html
[30] http://www.nextinpact.com/news/101755-open-data-25-millions-deuros-pour-mission-etalab.htm

2.5. Aspect juridique et licences Open Data

2.5.1. Cadre législatif

Les données publiques sont principalement régies par la loi CADA de 1978 et lois connexes, et les données issues du secteur privé dépendent du code de la propriété intellectuelle et de ses diverses lois connexes.

Dans le cadre de la loi CADA, trois conditions sont imposées : (1) indication de la source des données (paternité), (2) indication de leur date de mise à jour, (3) respect de l'intégrité des données (ne pas altérer et dénaturer les données).

Sur le plan de la propriété intellectuelle, deux conceptions majeures s'affrontent, celle basée sur le copyright et celle basée sur le droit d'auteur comme c'est le cas en France. Le cas des bases de données est complexe à ce titre, car les *données brutes* ne sont pas toujours considérées comme des créations et ne tombent pas directement sous le régime du droit d'auteur. Au final, la nécessité d'encadrer l'accès aux données pour protéger les investissements engagés pour les créer, les bases de données sont couvertes par la loi sur les droits d'auteur [2,MESZAROS].

Par ailleurs, au niveau législatif, le projet de loi pour une République numérique[31], adoptée définitivement le 28 septembre 2016 par le sénat, après son vote par l'assemblée nationale en juillet, garantit un meilleur accès à aux données, notamment avec une ouverture accrue des données publiques, en même temps qu'il légifère un certain nombre de points, tels que le droit à l'oubli numérique pour les mineurs, la portabilité des données (un des sujets préconisés par le GFII[32]), le renforcement des pouvoirs de sanction de la CNIL, la non-discrimination des contenus[33] (*neutralité du net*),…

2.5.2. Les licences

Les licences définissent à la fois les droits et obligations des ré-utilisateurs de données ouvertes (diffusion, utilisation, exploitation…).

La nécessité de faire appel à une licence concerne les organismes qui prévoient de mettre en place une redevance, ou ceux dont les données sont gérées selon un régime spécial (EPIC[34]-service public industriel et commercial), comme c'est le cas par exemple de la RATP ou de la SNCF, qui peuvent eux-mêmes décider des conditions de la libération des données (licence sur mesure), afin de valoriser économiquement les données ouvertes. Le nombre de licences s'est donc multiplié, ce qui a généré davantage de confusion que de clarté et d'interopérabilité. Sans compter que l'agrégation de données s'ajoute des contraintes, car elle implique l'existence de plusieurs sources, et donc très souvent de différentes licences, pas forcément compatibles entre elles.

---

[31] http://www.assemblee-nationale.fr/14/projets/pl3318-ei.asp
[32] http://www.gfii.fr/fr/groupe/diffusion-des-donnees-publiques
[33] http://www.journaldugeek.com/2016/01/22/arcep-sanction-neutralite-net/, http://www.nextinpact.com/news/101526-les-15-mesures-cles-loi-numerique.htm
[34] http://www.vie-publique.fr/decouverte-institutions/institutions/administration/organisation/structures-administratives/que-sont-etablissements-publics-administratif-epa-industriel-commercial-epic.html

La simplicité et la clarté des licences Open Data est une des conditions clés de succès. Le Royaume-Uni a réussi sur ce point à créer une licence simple d'utilisation et compatible avec les autres licences ouvertes, l'OGL[35].

2.5.2.1. La « licence ouverte » d'Etalab (LO)

En octobre 2011, Etalab a publié sa propre licence libre, baptisée « licence ouverte » (LO). D'autres gouvernements ont suivi la même démarche.

Elle est dite permissive car la seule contrainte est de mentionner la paternité et la date de la dernière mise à jour. Pour le reste, il est possible de reproduire, redistribuer, modifier, exploiter à titre commercial. Outre le portail data.gouv.fr, de nombreuses collectivités se sont ralliées à elle telles que Bordeaux, Montpellier, l'Auvergne, les Hauts-de-Seine [10,BLANC].

La création de la licence Etalab a permis de simplifier le choix de licence pour les administrations, les services de l'Etat se devant de l'utiliser.

2.5.2.2. Les licences Creative Commons

La licence CC0[36] est la seule licence développée par Creative Commons compatible avec le droit des bases de données. Elle est issue du projet *Sciences Commons* sur les données scientifiques. Créée en 2002 pour assouplir le droit d'auteur, elle est caractérisée par 4 clauses : paternité obligatoire (« by »), usage commercial interdit (« nc »), partage à l'identique (« share-alike »), non dérivative (« nd »), c'est-à-dire non modifiable. Les clauses nd et nc sont incompatibles avec l'Open Data. En France, nous n'avons pas trouvé de collectivité qui l'utilise.

Creative Commons propose un service en ligne[37] aux utilisateurs pour aider au choix de la licence appropriée à l'usage, sous la forme d'un mini-questionnaire. La licence adéquate est automatiquement indiquée, en fonction des réponses saisies.

2.5.2.3. Les licences de l'Open Knowledge Foundation (OKF)

L'Open Knowledge Foundation[38] (OKF) est une association de promotion du savoir libre, qui a développé trois jeux de licences orientés bases de données, basés sur le droit anglo-saxon : *Open Data Commons* (ODC-by), *Open Database License* (ODbL), et *Public Domain Dedication and License* (PPDL).

La licence *Open Data Commons* (ODC-by) est permissive puisqu'elle autorise toutes les utilisations à condition que la paternité (by) soit indiquée. Elle est donc proche de la licence ouverte d'Etalab.

L'Open Database License (ODbL) est une licence qui permet de copier, modifier, de faire un usage commercial, sous trois conditions : citer la source, redistribuer sous des conditions de partage identiques les modifications, maintenir ouverte techniquement la base de données redistribuée. Il existe une dérogation au « *share alike* » (partage à l'identique) moyennant contrepartie. Elle fonctionne sur le mode pot commun (si j'utilise, je dois recontribuer).

---

[35] http://www.nationalarchives.gov.uk/information-management/re-using-public-sector-information/licensing-for-re-use/what-ogl-covers/
[36] https://creativecommons.org/share-your-work/
[37] https://creativecommons.org/choose/
[38] https://fr.wikipedia.org/wiki/Open_Knowledge_Foundation

La licence ODbL a été transposée en français par *Veni, Vedi, Libri*[39], une association dédiée aux licences libres, en partenariat avec Paris, dans le cadre du lancement de son portail. C'est la seconde licence standard la plus courante en France (après Etalab): la communauté urbaine de Toulouse Métropole, des offices du tourisme en Paca ont fait ce choix.

La licence *Public Domain Dedication and License*[40] (PDDL) revient à renoncer à tout droit puisque la base de données est placée dans le domaine public. L'auteur abandonne son droit moral. Elle se rapproche de la licence CC0 [19,GRUTTEMEIER *et al.*].

<table>
<tr><td rowspan="2"></td><td colspan="3">Libertaires</td><td rowspan="2">Participatives</td></tr>
<tr><td>Totalement</td><td colspan="2">Partiellement</td></tr>
<tr><td>Données ouvertes</td><td>CC-0</td><td>CC-BY</td><td rowspan="2">Licence Ouverte</td><td>CC-BY-SA</td></tr>
<tr><td>Spécifiques à des bases de données</td><td>PDDL</td><td>ODC-BY</td><td>ODbl</td></tr>
</table>

Tableau 1 : Licences majeures du monde de l'Open Data [2, MESZAROS et al.*].*

2.6. La question des formats et de l'interopérabilité

Après la question des licences, L'enjeu suivant sur lequel se concentrer concerne les formats. Pour pouvoir manipuler facilement et librement les données publiques, il est nécessaire de se désengager des formats propriétaires (type Excel). En effet, l'utilisation de tels formats fait encourir un important risque juridique aux ré-utilisateurs qui deviennent dépendants du propriétaire. Les formats ouverts, et reconnus sont par exemple CSV, ODS, CSV, ODS, XML, JSON, RDF, TXT, SQL,..

La question du format de distribution a une importance à la fois du point de vue de l'accès, de l'interopérabilité et de la pérennité. Il faut s'assurer que le format distribué peut être facilement accessible, et ne nécessite pas de logiciels spéciaux pour être ouverts, avec une contrainte financière, par exemple. Il doit également être transformable vers un autre format, et être durable. Autrement dit, les formats choisis permettent d'accéder à tout moment aux contenus des fichiers d'une manière *non-surveillée* et pérenne pour en extraire des données pertinentes [2,MESZAROS, *et al.*].

Etant donné que cette lisibilité doit être réalisable par une machine, et pas seulement par l'humain, la Directive 2013/37/UE du Parlement Européen et du Conseil indique :

« Un document devrait être considéré comme présenté sous un format lisible par une machine s'il se présente dans un format de fichier structuré de telle manière que des applications logicielles puissent facilement identifier et reconnaître des données spécifiques qu'il contient et les en extraire. Les données encodées présentes dans des fichiers qui sont structurés dans un format lisible par une machine sont des données lisibles par la machine. Les formats lisibles par la machine peuvent être ouverts ou propriétaires ; il peut s'agir de normes formelles ou non. Les

[39] http://wlibri.org/
[40] http://opendatacommons.org/licenses/pddl/

documents encodés dans un format de fichier qui limite le traitement automatique, en raison du fait que les données ne peuvent pas, ou ne peuvent pas facilement, être extraites de ces documents, ne devraient pas être considérés comme des documents dans des formats lisibles par la machine. Les États membres devraient, le cas échéant, encourager l'utilisation de formats ouverts, lisibles par la machine. »

Par exemple le format PDF, tout à fait adapté à une utilisation manuelle et visuelle, du fait de son indépendance vis-à-vis des différentes plateformes, ne fait en revanche pas partie des formats permettant l'extraction ou l'analyse de son contenu de manière native et simple.

La structuration de ces formats est un point essentiel pour le rendre « *machine readable* » (lisible par la machine) : elle permet d'enrichir les données en les décrivant, de créer des métadonnées, et donc de donner un sens à ces données, de savoir comment les traiter de manière automatisée, à quoi elles correspondent,…

Un autre aspect technique à prendre en compte concernant l'*Open Data* concerne l'hétérogénéité et la dissémination des données sur le Web, ce qui rend leur intégration difficile au sein d'un entrepôt de données.

Comme on le verra par la suite dans ce mémoire, ces problématiques sont prépondérantes dans les domaines du Web sémantique et de la santé, et seront développées dans ce cadre de manière spécifique. Au-delà des formats de fichiers mis à disposition, s'ouvre la question au cœur de notre sujet, et qui pose le problème de la qualité des modèles de données. Il s'agit d'un élément central, permettant la cohérence du système, qui va prendre en charge des données pour les rendre utilisables à d'autres.

2.7. Qualité des modèles de données

La donnée générée par la révolution numérique et à notre disposition aujourd'hui est inutile si on ne sait pas s'en servir, l'exploiter [20,DELAYAT].

L'enjeu pour l'entreprise ou l'administration est de gérer cette chaîne de valeur des données autour desquelles se greffent des contraintes fortes de protection, de sécurité, de qualité et d'éthique. Les données collectées peuvent être structurées ou non, internes ou externes, en provenance de sources institutionnelles, commerciales, techniques, financières ou diverses. Elles peuvent prendre des formes aussi variées que des bases de données, du texte, des images et du son, des vidéos.

La gestion des métadonnées est donc une activité centrale dans la gestion des données car elle contribue à la valorisation des données et en facilite les usages. Pour les données structurées, gérer des métadonnées suppose de pouvoir s'appuyer au préalable sur une démarche de modélisation des données : en effet, les métadonnées sont attachées aux données modélisées, quel que soit le niveau de modélisation (global au niveau de l'Entreprise, ou bien local au niveau d'un métier voire d'une application).

Concernant le modèle de données, l'enjeu principal est de se doter de la capacité à réutiliser et partager plus facilement les données, associé à une plus grande efficacité dans l'assurance de leur qualité. En effet, si les données sont contrôlées à la source, et si la qualité n'est pas altérée dans les processus de transmission de celles-ci, alors la qualité peut être plus facilement garantie dans les processus qui se situent au bout de la chaine de traitement.

Pour évaluer la qualité d'un modèle de données, les éléments suivants peuvent être étudiés :

- La donnée est-elle partagée ? Peut-on y accéder ?
- La donnée est-elle fiable / pérenne ? Peut-on lui faire confiance ? Est-elle à jour ? Le *versionning* des changements est-il à jour ?
- La donnée est-elle compréhensible ? Peut-on l'utiliser ? Est-elle bien documentée ?
- La donnée est-elle sécurisée ? Dans quelles conditions et qui est autorisé à lire, réutiliser (Licences, gratuité) ?

C'est la réponse à l'ensemble de ces questions qui permet d'établir si un modèle de données répond aux attentes.

La qualité des données est directement liée à la nécessité d'avoir un vocabulaire commun, des référentiels, des modèles. Les propriétaires et producteurs de données sont nécessairement responsables de la qualité de leurs données. Ils définissent les contrôles à opérer lors de l'acquisition ou de la transformation.

Le site Opquast[41] a recensé 72 règles, autant de clés pour offrir un modèle de données « idéal » : un bon modèle est un modèle simple d'utilisation, ce qui implique en arrière-plan une réflexion poussée sur des critères tels que :

- l'*animation* : expliquer le fonctionnement général,
- les *API :* donner la possibilité d'exploiter ;
- les *applications* : montrer des cas d'usage ;
- le *catalogage :* décrire les jeux de données ;
- le *format* : donner la possibilité de réutiliser ;
- l'*historique* : indiquer les versions ;
- l'*identification* : donner les sources des données ;
- la *licence* : indiquer les conditions d'utilisation ;
- le *linkedata* : respecter les standards W3C du Web de données ;
- le *nommage* : indiquer la charte le cas échéant ;
- la *transparence* : donner toutes les informations à disposition ;
- l'*utilisabilité* : prévoir des fonctionnalités permettant d'effectuer des actions simples ;
- la *vie privée* : anonymiser si besoin.

2.8. Production, distribution, standardisation et accès: La libération des jeux de données publiques

2.8.1. Chaine de production

Il convient tout d'abord de décrire le rôle des différents intervenants dans la fabrication, mise à disposition et exploitation des données ouvertes. Chaque acteur peut être classé dans un des trois groupes suivants : les producteurs, les re-distributeurs (passerelles) et les ré-utilisateurs [2,MESZAROS *et al.*].

Ils sont réciproquement interdépendants. Ainsi, les re-distributeurs et surtout les ré-utilisateurs sont largement soumis à la nature et avant tout à la qualité des données mises à disposition et des services associés. De l'autre côté, les producteurs ne peuvent pas ignorer les besoins et les possibilités des acteurs se trouvant dans les étapes suivantes de la chaîne de distribution des données ouvertes. Ils doivent

[41] https://checklists.opquast.com/fr/opendata/

être toujours à l'écoute des besoins et savoir adapter leurs manières de faire pour une ouverture efficace de ces données. En résumé, chaque partie a son rôle et ses obligations afin que le jeu de construction soit solide, et que chaque pièce occupe correctement sa place.

Alors, quels sont ces caractéristiques propres à chacun ?

Les producteurs représentent l'ensemble des acteurs qui mettent à disposition des autres, les données dont ils sont propriétaires. Ils peuvent correspondre à des instances de l'Etat, institutions publiques, mais aussi à des sociétés privées et publiques, comme JCDecaux, la RATP ou Keolis. Ils se doivent de suivre un certain nombre d'étapes constituant le processus de traitement des données :

- **Sélection** : étape cruciale de collecte, les domaines prioritaires et la nature des données à ouvrir étant déterminants et impactants pour la suite des opérations ;
- **Extraction** : la récupération des données depuis leur emplacement d'origine passant parfois par leur numérisation ;
- **Nettoyage** : il s'agit de faire en sorte que les données soient valides et exhaustives ;
- **Transformation** : le(s) format(s) de publication sont déterminés à ce niveau, en s'assurant de leur exploitabilité ;
- **Publication** : elle s'accompagne de la documentation correcte des données (métadonnées les plus complètes possible), du choix de licence d'utilisation, et des procédures de mises à jour ;
- **Réutilisation** : l'accès aux données doit être simple et sans contrainte particulière.

Les distributeurs ou les plateformes passerelles représentent le médiateur entre les producteurs et les ré-utilisateurs. Il s'agit souvent de portails tels que data.gouv.fr. Ainsi, les données ouvertes issues d'environnements différents, sont regroupées en un seul endroit. Il s'agit donc de maîtriser l'ensemble des solutions développées et utilisées par d'autres acteurs. Le re-distributeur doit savoir gérer l'ensemble des formats utilisés, sans oublier les différentes normes de présentation de ces données et leur encodage. Il se doit également d'être en veille permanente, pour que les liens vers les producteurs qu'il recense, fonctionnent toujours, et être à jour en continu sur d'éventuelles modifications.

Les ré-utilisateurs sont les utilisateurs finaux, c'est-à-dire un simple citoyen qui souhaite avoir un droit de regard sur certains aspects de la sphère publique, ou bien une personne, une entreprise qui cherche à développer de la valeur autour des données, en créant des applications. Ce type d'utilisateur va alors devoir faire face à un certain nombre d'obstacles, en particulier s'il a besoin d'agréger des données de sources différentes : on y retrouve les problématiques du re-distributeur, auxquels d'ajoutent la contrainte des licences à respecter.

2.8.2. Conditions de production des données ouvertes

Pour que les données soient véritablement utilisables, nous avons vu qu'un cycle de traitements doit être suivi rigoureusement, afin d'assurer leur pérennité, leur consistance, leur cohérence, et leur identité.

Tout d'abord, on procède au « *casting* » des données, afin d'identifier celles qui seront candidates à l'ouverture.

Il s'agit ensuite d'y accéder et de les récupérer, opérations loin d'être aisées, étant donné que les données sont stockées dans des bases de données ou des outils qui permettaient jusqu'alors leur simple consultation. L'organisation physique des données qui, jusque-là, importait peu, devient un des aspects à explorer, pour comprendre les modalités de stockage et d'organisation. C'est le passage obligé pour les extraire en conservant leur sens, et leurs caractéristiques propres. Ce sont donc généralement des outils ad hoc qui sont développés pour l'extraction, avec la difficulté de se trouver devant une accumulation de ces bases de données, des logiciels qui y donnent accès et de leurs différentes versions. Ce travail de véritable fouille parmi une multitude de dispositifs attachés aux données représente un coût important.

Le côté positif des chantiers mis en œuvre pour l'extraction dans les institutions, est la reprise de contrôle sur les données de celles-ci, puisqu'on passe par une déconstruction et une désarticulation de l'agencement technique, et parfois, à un désengagement commercial et juridique envers les entreprises privées, jusqu'alors maitres du jeu.

Un des aspects central et particulier de l'*Open Data* concerne le caractère brut des données qui doivent être diffusées. On parle de données « non modifiées », « inaltérées », ou encore « primaires. ». En réalité, elles sont reconfigurées. Ce n'est pas qu'elles soient de mauvaise qualité, puisqu'elles répondent le plus souvent parfaitement à leur usage interne et collent à leur organisation, leur métier et leur environnement de production. Elles ne peuvent cependant être publiées telles quelles. On passe donc par des transformations afin de rendre les données intelligibles. *« Nettoyées », les données sont également adaptées pour être compréhensibles par le plus grand nombre, mais aussi formatées pour devenir « machine readable »* [21, DENIS *et al.*]. On peut donc considérer que les données sont à la fois « brutes », puisqu'issues de leur source mère, mais également « travaillées » pour devenir exploitables. D'autant plus que l'on ne connait pas l'ensemble des usages potentiels, et qu'il s'avère donc nécessaire de rendre les données génériques et universelles pour répondre à la majorité des besoins.

Cela se traduit par un nettoyage des données, afin de les corriger (erreurs de saisie, absence de valeur,…), et par une amélioration de leur clarté, en les harmonisant (deux données identifiées différemment peuvent correspondre à la même information), en les classant,… Ces transformations permettent donc de construire une cohérence à travers les différences et les redondances de jeux de données hétérogènes. Elles peuvent être réalisées au travers des métadonnées qui constituent une première approche, mais doivent également comporter des manipulations et des modifications à même les données.

Par ailleurs, étant donné que les données ouvertes doivent être aussi techniquement intelligibles, l'adoption de standards, et de formats partagés et une nécessité. *L'un de ces formats, basique et partagé par la plupart des acteurs du domaine, notamment parce qu'il est considéré comme ouvert et lisible par le plus grand nombre d'instruments : le CSV (comma-separated values)*.

Toutes ces étapes constituent le travail de *brutification* des données primaires, qui conditionne leur réutilisation hors de leur contexte de production.

2.9. Constats d'avancement en France

Nous sommes actuellement, et dans de nombreux pays, à un stade très artisanal de la diffusion des données produites par les administrations [12,FAUVEL].

En France, certaines administrations telles que la DILA, l'IGN ou l'INSEE ont une certaine avance, car elles disposent de moyens natifs pour diffuser des données. Ce n'est pas la majorité, et en général, les institutions publiques montrent des difficultés à suivre le processus de production décrit précédemment. Face à la volonté d'automatiser le plus possible l'ouverture de leurs données, et de donner de la qualité, on trouve encore beaucoup de traitements manuels, réalisés parfois laborieusement, et donc couteux. L'enjeu est de taille, car il faut s'attaquer à des volumes de données de l'ordre du milliard, de provenances diverses, donc techniquement, humainement, politiquement, juridiquement, organisationnellement et sémantiquement hétérogènes.

Il existe donc une lourde dette technologique ainsi que des manques structurels évidents. Ainsi, Henri Verdier, Administrateur général des données, rappelle que *les choix d'architecture sont antérieurs à la révolution de la donnée*. De plus, l'État ne maîtrise pas totalement son système d'information, du fait de la place historique qu'ont pris différentes sociétés prestataires dans la mise en œuvre, le suivi et la maintenance des usages et outils existants. Cette méconnaissance a des conséquences sur la volonté d'ouverture, et oblige une réorganisation pour mettre en place *de nouvelles règles d'audit des projets informatiques de l'État* et assurer l'extractibilité des données et leur accès, de manière sécurisée et pérenne [13, CONTE].

Par ailleurs, une impulsion juridique est indispensable pour favoriser la libération des données et leur inclusion dans les procédures administratives. Le projet de loi pour une République Numérique, lancé en septembre 2015 (cf $1.6.1 Cadre législatif), en est une étape essentielle, d'autant plus qu'il émane d'un processus de co-création, et donc de transparence participative. En effet, 1389 modifications au projet de loi, apportées par plus de 21000 citoyens ont pu être apportées. Ce projet inédit est l'exemple même de l'esprit d'ouverture. Il a même provoqué l'adoption par les députés d'un amendement, en vertu duquel le gouvernement devra remettre un rapport portant sur *la nécessité de créer une consultation publique en ligne pour tout projet de loi ou projet de loi avant son inscription à l'ordre du jour au Parlement* [16].

Dans le même ordre d'idées, Etalab a publié en novembre 2016, le résultat d'une consultation des citoyens, tenue du 29 septembre au 20 octobre 2016[42], à laquelle entreprises, associations, administrations publiques et particuliers ont pu contribuer. Les contributions ont pu montrer quelles sont les données de références les plus attendues, par quels moyens et avec quel niveau de qualité les répondants souhaitent leur mise à disposition.

[42] https://www.etalab.gouv.fr/consultation-spd

Deuxième partie : Open Data et Web sémantique

Avancées en quelques illustrations

3. *Open Data* et Web sémantique

Celui qui contrôle les métadonnées contrôle le web[43].

3.1. L'ingénierie des connaissances et ses enjeux

L'Ingénierie des Connaissances (IC) propose des concepts, méthodes et techniques permettant de modéliser, de formaliser et d'acquérir des connaissances dans les organisations dans un but d'opérationnalisation, de structuration ou de gestion au sens large [22, AUSSENAC-GILLES *et al.*].

Dans la lignée des travaux en Intelligence Artificielle (IA) sur la formalisation des connaissances et du langage, l'IC a pour objectif de livrer des systèmes capables de traiter les données et contenus symboliques à partir de leur sémantique. Il s'agit donc de modéliser les connaissances, les diffuser et les rendre utilisables dans des applications « intelligentes » (avec une capacité de raisonnement).

Avec l'expansion du Web et des Web technologies, L'IC s'est adossée au Web sémantique lié à la donnée, et non plus au simple document, pour construire des architectures faisant circuler le savoir. Il s'agit de représenter les connaissances hétérogènes de manière formelle, pour pouvoir ensuite les externaliser et les diffuser dans un contexte d'interactivité. Le modèle conceptuel est le support de la représentation, et les ontologies, ou autres langages de représentation des connaissances, en sont les méthodes de mise en œuvre. Le modèle décrit l'objectif à atteindre, avec toutes les étapes nécessaires (son cycle de vie), en séparant les taches des moyens d'y parvenir. On obtient ainsi un système de base de connaissances (SBC).

La modélisation peut prendre forme à partir de l'analyse des besoins exprimés au départ par les utilisateurs ou demandeurs, c'est ce qu'on nomme la démarche ascendante ou « *bottom-up* ». Le procédé descendant ou « *top-down* » part de l'existant afin de l'améliorer, le transformer, le corriger en fonction des remontées utilisateur. Les deux méthodes ne sont et ne doivent pas être exclusives, pour obtenir un système alliant l'usage observé à la théorie initiale visée.

Les sources de connaissances correspondent, soit à des savoir-faire de haut niveau automatisés dans des systèmes experts, soit à des procédures partagées à disposition des utilisateurs dans les systèmes à base de connaissances. Pour exploiter les sources de connaissances, on cherche à les représenter et donc à les transcrire dans un modèle : c'est la tâche que se donne l'IC.

Au préalable, il s'agit d'identifier ces sources de connaissances, et comment y accéder pour les exploiter ensuite au sein du processus. Des données déjà structurées sont préférables, même si cela suppose des adaptations ou alignements par rapport au modèle cible : la réutilisation est, au final, moins couteuse puisqu'une part d'analyse est déjà réalisée. Il s'agit également de réfléchir d'ores et déjà à la maintenance du cycle de vie, vis-à-vis des futures évolutions techniques, fonctionnelles et des données en elles-mêmes.

Les sources peuvent être issues d'enquêtes, de questionnaires et donc soumises à l'interprétation ou mauvaise retranscription ; elles peuvent également provenir de documents textuels, qu'il est alors nécessaire de traiter en fonction de

[43] fabien, gandon, @fabien_gandon, http://fabien.info

leur type (langage naturel et traitement automatique du langage TAL, documents structurés), pour leur donner ensuite un sens.

La gestion des documents produits par les organisations, leur cycle de vie, leur type et format, mode d'utilisation (partage, circulation) relève donc de l'IC, ou en tout cas y prend sa part.

En ce qui concerne les outils utilisés par l'IC, sous l'impulsion du Web sémantique, les ontologies utilisées initialement comme source de métadonnées pour indexer des documents, sont très adaptées. Elles touchent également aujourd'hui la Recherche d'Information (RI), certaines étant même spécialisées pour cette application, avec une composante linguistique forte (ressource termino-ontologique ou RTO). Les ontologies jouent un rôle clé en intégration de sources multiples et hétérogènes dans un même domaine et en temps réel, ce qui permet ensuite de pouvoir établir des relations entre les concepts définis [31, KEMBELLEC].

Pour être efficace et viable, une ontologie se veut relative à un domaine réduit, et se rapporte à des expériences d'usage : c'est en abandonnant tout objectif d'universalité que l'on peut obtenir des représentations sans ambiguïté. On peut cependant faire appel à des « niveaux » d'ontologies différents, avec les ontologies de référence, qui ont des *visées de représentations larges*, et les ontologies d'interface développées pour des applications spécifiques[44]. On réutilise ainsi ce qui nous est nécessaire dans l'ontologie de référence, et on procède à des alignements entre les 2 types d'ontologies, grâce à des services ou standards tel que *Common Terminology Services* CTS2[45].

3.2. Technologies du Web sémantique

Le terme "Web 2.0" a été proposé dans le cadre d'une conférence tenue en août 2004 pour désigner le Web en tant que plateforme partagée par tous les usagers, et une architecture permettant la contribution à la création des contenus, via les réseaux sociaux (blogs, wikis,…), le développement collaboratif, mais aussi le mixage ou « *mashup* » d'applications [46]. C'est cet aspect qui nous intéresse le plus ici, appliqué à l'origine aux documents, et aujourd'hui aux données, chaque élément de la chaine d'information pouvant servir de source, et en même temps se nourrir de toute autre entité du Web. On parle fréquemment de révolution, au même titre que l'apparition de l'écriture.

Le Web sémantique s'inscrit dans ce cadre et constitue une sorte d'extension au Web, permettant de donner du sens au contenu des pages, en les structurant et les rendant ainsi interprétables par les machines (« *machine readable* »). Le terme « *linked data* » (« données liées » ou encore « Web de données ») est utilisé pour décrire les meilleurs pratiques pour mettre à disposition, partager et interconnecter les données, informations et connaissances sur le Web sémantique, en utilisant les URIs et RDF [23, BIZER *et al.*]. On constitue de cette manière *une base de données à l'échelle du Web* [22, AUSSENAC-GILLES *et al.*], dans laquelle les données sont reliées entre elles. L'effet obtenu est de minimiser la duplication des données (meilleure cohérence), et d'y accéder à tout moment et de partout. La question de la fréquence de mise à jour se pose d'ailleurs à ce niveau, pour être certain d'avoir à

[44] ROSENBLOOM, MILLER, JOHNSON, ELKIN, BROWN. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc. 2006 May-Jun;13(3):277–88. Disponible sur : <http://www.ncbi.nlm.nih.gov/pubmed/16501181>

[45] http://www.3mtcs.com/resources/hl7cts

[46] http://www.ladocumentationfrancaise.fr/dossiers/internet-monde/web2.0.shtml

disposition la donnée la plus « fraiche » possible. Il est nécessaire de réfléchir en amont s'il s'agit de données statiques (publication d'une statistique, par exemple), dynamiques (diffusion du prix du carburant) ou fluides, c'est-à-dire dans un *flux ininterrompu* de données (état de la circulation routière) [2, MESZAROS *et al.*].

Pour bien comprendre le fonctionnement du Web de données, quelques explications sont nécessaires.

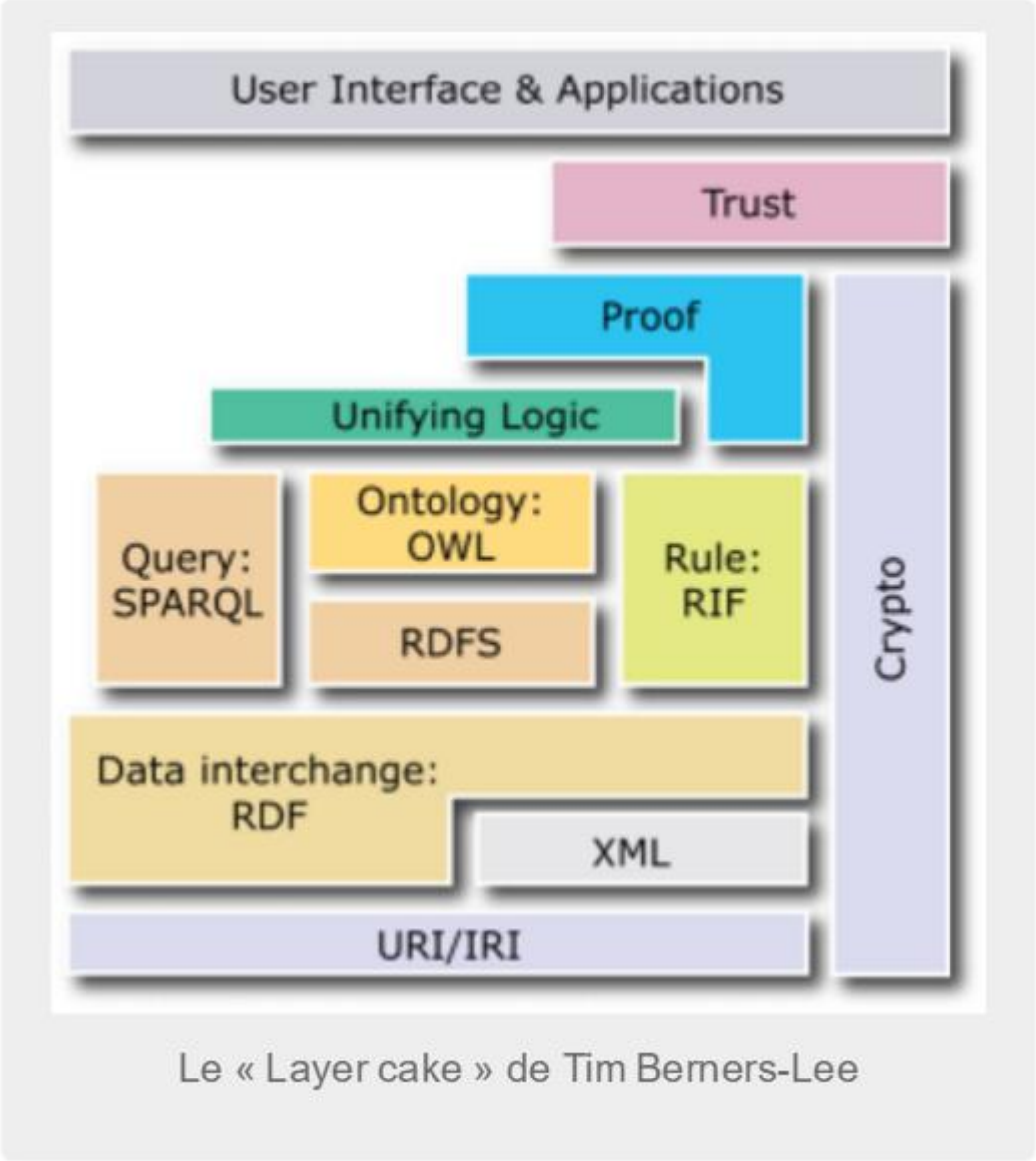


Figure 3 : Le « Layer cake » du Web de données.

Le schéma du *Layer cake* (Figure 3) illustre l'architecture des *linked data*, qui se présente en couches successives de langages, chacun participant à l'ensemble des recommandations du W3C (*World Wide Web Consortium*) [24].

Chaque ressource du Web est identifiée par son URI/IRI (*Uniform Resource Identifier)/ Internationalized Resource Identifier*), c'est-à-dire une adresse URL unique sur internet[47].

Le modèle de description RDF (*Resource Description Framework*)[48] de ces ressources, est un standard du W3C, qui correspond à un triplet (ressource, propriété, valeur). Ainsi la ressource décrite est le sujet, la propriété renseignée le prédicat et la valeur de cette propriété l'objet. Le sujet et le prédicat sont identifiés par des URIs/IRIs, tout comme l'objet, qui peut en plus être déterminé par une valeur littérale (nombre, chaîne de caractères). Une même ressource peut être sujet, prédicat ou objet dans plusieurs triplets.

Les formats utilisés pour RDF peuvent être XML[49], Turtle[50], N-Triples et N-Quads [51], JSON[52], en fonction du besoin de simplicité, de lisibilité ou d'utilisation.

Les triplets reliés par les URIs forment un graphe (ensemble de nœuds et d'arcs), où les sujets et les objets sont des nœuds et les prédicats des arcs. Les graphes offrent un aspect visuel, et permettent également d'être manipulés par des outils informatiques, comme les ontologies. Celles-ci indiquent les classes et propriétés qui servent de support à l'expression des descriptions RDF, ce qui ouvre le champ d'exploitation ultérieur des données. Les ontologies permettent donc de spécifier les ressources via un vocabulaire exprimé en RDF Schema (RDFS)[53] et OWL (*Web Ontology Language*)[54].

En partant des classes et leur hiérarchie, RDFS donne la possibilité d'organiser les concepts, d'en indiquer les propriétés et les notions de domaine/portée. OWL complète la description de ces propriétés, des relations, et introduit la notion d'équivalence entre concepts, pour aboutir à la documentation formelle du domaine.

Le langage permettant de faire des requêtes sur des triplets RDF est SPARQL[55]. Ce dernier est supporté par les *triplestores*[56], des bases de données permettant de stocker nativement des triplets RDF. Il est à RDF ce que SQL est aux bases de données relationnelles. Un *triplestore* est interrogeable en SPARQL sur le Web via un SPARQL *endpoint*[57] (point d'accès, protocole http), et donne un résultat au format XML (ou JSON et bientôt JSONLD).

En conclusion, Bruno Menon [27, MENON] donne cette illustration très claire en parlant du Web de données : *Les URIs/IRIs en sont les mots et les ontologies conçues en RDF Schema ou en OWL sont les dictionnaires qui permettent d'en élucider la sémantique. RDF, qui régit la construction des assertions, en est la syntaxe. Et XML est l'un des systèmes d'écriture avec lesquels cette langue peut être notée. Les jeux de données RDF sont les textes ou les ouvrages de référence rédigés dans cette langue, que SPARQL permet de consulter et de compiler*.

4747 https://www.w3.org/TR/uri-clarification/
48 https://www.w3.org/RDF/
49 https://www.w3.org/XML/
50 https://www.w3.org/TR/turtle/
51 https://www.w3.org/TR/n-triples/ , https://www.w3.org/TR/n-quads/
52 http://www.w3schools.com/js/js_json_intro.asp
53 https://www.w3.org/TR/rdf-schema/
54 https://www.w3.org/2001/sw/wiki/OWL
55 https://www.w3.org/TR/sparql11-overview/
56 https://fr.wikipedia.org/wiki/Triplestore
57 https://www.w3.org/wiki/SparqlEndpoints

3.3. Aller plus loin avec les ontologies de domaine

L'ontologie de domaine sert le plus souvent à hiérarchiser et classer les éléments composant le domaine (classes d'entités du domaine, concepts), ainsi qu'à décrire leurs relations. Elle constitue *le vocabulaire logique qui permet d'exprimer des faits et des connaissances du domaine sur lesquelles raisonner* [22, AUSSENAC-GILLES *et al.*]. C'est également grâce au lien qu'elle fournit entre le format, la structure exploitable par la machine, et la connaissance humaine, qu'elle est considérée comme référentiel terminologique le plus abouti.

Une ontologie peut présenter plusieurs niveaux de complexité, en fonction du besoin, et aller d'une représentation simple des concepts d'un domaine, à une formalisation complète de ce même domaine, en y intégrant des raisonnements. Elle sera alors qualifiée de légère dans le premier cas, et lourde dans le second.

Nous avons évoqué précédemment le caractère spécifique d'une ontologie de domaine, qui se veut la moins ambiguë possible pour le domaine qu'elle modélise, et doit donc se restreindre aux cas d'usage pour lesquels elle est mise en œuvre. Cela va à l'encontre de la pensée initiale de disposer de représentations universelles ou génériques. On trouve cependant cette « catégorisation » sur des niveaux supérieurs à celui l'ontologie de domaine. En effet, la top-ontologie, le degré « père », est par essence commun et donc générique, par sa vision philosophique des connaissances (Sowa[58], SUMO[59], DOLCE[60], …). Le second niveau d'ontologie correspond à la core-ontologie, qui concerne les concepts généraux d'un domaine dans sa globalité, comme par exemple *LKIF-Core dans le droit*. L'ontologie de domaine prend sa place ensuite, et caractérise une des facettes du domaine pour un besoin défini, à partir de sources existantes (textuelles ou structurées).

Au moment de la construction d'une ontologie, plusieurs problématiques doivent être pensées en termes de contenu informationnel et de sa qualité, de méthodes de traitement des sources de données (TAL, réutilisation,..), de moyens humains nécessaires et outils à disposition pour la conception, de gestion de projet, et d'environnement technique. On trouve ainsi des plateformes de gestion globales intégrant un éditeur d'ontologie, tel que Protégé (Stanford), sur lequel nous reviendrons plus après dans ce mémoire, de manière théorique et pratique. Concernant les méthodes de construction d'ontologies, on peut citer à titre d'exemples, OntoClean[61] et ARCHONTE[62], spécialisées sur la qualité du contenu de l'ontologie.

La validation de l'ontologie développée est une étape primordiale dans le processus global, et permet de vérifier l'adéquation entre les résultats attendus et observés. Plusieurs méthodes et techniques sont employées, allant des tests automatiques aux interrogations approfondies des experts du domaine concerné [59, LE PICARD].

---

[58] http://www.jfsowa.com/ontology/index.htm
[59] http://www.adampease.org/OP/
[60] http://www.loa.istc.cnr.it/old/DOLCE.html
[61] https://www.researchgate.net/publication/226934944_An_Overview_of_OntoClean
[62] DHOMBRES Ferdinand, JOUANNIC Jean-Marie, JAULENT Marie-Christine, CHARLET Jean. Choix méthodologiques pour la construction d'une ontologie de domaine en médecine prénatale. In DESPRES Sylvie. 21èmes Journées Francophones d'Ingénierie des Connaissances, Nîmes, France. Ecole des Mines d'Alès, pp.171-182, 2010. <hal-00487736>.

3.4. Etat de l'art dans le domaine de la santé

Prévention, prédiction, participation, personnalisation : tels sont les enjeux de la médecine, qui tire parti des progrès de la science pour mieux comprendre la complexité du corps humain, prévenir les maladies mais aussi améliorer l'accompagnement des patients par une personnalisation des soins et un meilleur partage entre médecins et patients.

En Juillet 2016, la ministre de la Santé a ainsi dévoilé la stratégie nationale e-santé 2020, avec des mesures destinées à développer la médecine connectée [32].

L'un des chantiers marque la volonté de structurer l'action des éditeurs de logiciels en santé autour de règles et de terminologies partagées. Marisol Touraine a également annoncé la création d'un administrateur des données de santé au ministère, dont l'une des missions sera d'accélérer l'Open Data en santé. Un plan *Big Data* en santé serait par ailleurs lancé en automne 2016, pour mieux valoriser les données, et améliorer par exemple l'interprétation de données médicales dans l'aide au diagnostic.

L'informatique, via l'intelligence artificielle et les systèmes experts, constitue un outil essentiel d'aide pour les médecins, sans remettre en cause leur valeur ajoutée ni les *remplacer* [63] : ce sont les professionnels de santé qui posent toujours le diagnostic, le logiciel participe au choix de *stratégie thérapeutique*. On trouve par exemple des bases de données de références de *bonnes pratiques établies par des spécialistes et des organismes comme la Haute Autorité de santé (HAS)*, qu'il s'agit d'étendre et de développer aujourd'hui en France.

Ainsi, pour aider à imposer l'appropriation de ces systèmes et outils, la loi de modernisation de notre système de santé de janvier 2016[64] a introduit l'Open Data des données de santé dans le code de la santé publique. Il s'agit d'améliorer, simplifier et de mieux encadrer l'accès aux données, comme celles du Système National d'Information InterRégimes de l'Assurance Maladie (SNIIRAM), la base de données nationale.

3.4.1. Organismes clé, référentiels

3.4.2.1. Le SNIIRAM

Créé en 1999 par la loi de financement de la Sécurité sociale[65], le SNIIRAM (Système National d'Information InterRégimes de l'Assurance Maladie) est une base de données nationale dont les objectifs sont de contribuer à une meilleure gestion de l'Assurance Maladie et des politiques de santé pour une meilleure qualité des soins, et de mettre à disposition des professionnels de santé les informations pertinentes sur leur activité (finalités définies par l'article L161-28-1 du code de la sécurité sociale).

---

[63] VANLERBERGHE Cyrille. Quand l'ordinateur inquiète les médecins. sante.lefigaro.fr, [En ligne]. 05 juillet 2016.Disponible sur : <http://sante.lefigaro.fr/actualite/2016/07/05/25174-quand-lordinateur-inquiete-medecins?__scoop_post=e0641f50-42e9-11e6-9f1a-00221934899c&__scoop_topic=5960664#__scoop_post=e0641f50-42e9-11e6-9f1a-00221934899c&__scoop_topic=5960664>

[64] https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000031912641&categorieLien=id

[65] http://www.ameli.fr/l-assurance-maladie/statistiques-et-publications/sniiram/finalites-du-sniiram.php

Son périmètre, ses finalités, son alimentation et l'accès aux données sont définis dans un arrêté du ministère des affaires sociales et de la santé transmis pour avis à la Commission Nationale de l'Informatique et des Libertés (CNIL), puis publié au Journal officiel.

Le SNIIRAM prend la forme d'une base de données complète et détaillée sur le parcours des patients et l'organisation du système de soins : les données sont anonymes, et proviennent des remboursements effectués par l'ensemble des régimes d'assurance maladie, pour les soins du secteur libéral (1,2 milliard de feuilles de soins pour l'ensemble de la population vivant en France). La structure est en pratique découpée en 15 bases de données thématiques de données agrégées (« *datamarts* ») ayant une finalité particulière : suivi des dépenses (Damir), analyse de l'offre de soins libérale, biologie, pharmacie, dispositifs médicaux, établissements privés.

On y trouve des informations sur le prestataire de soin (spécialité, mode d'exercice, sexe, âge, département d'implantation), le codage détaillé (médicaments, actes techniques des médecins, dispositifs médicaux, prélèvement biologiques), ainsi que la date des soins et les montants remboursés par l'Assurance maladie et payés par les patients. La qualité des soins fait l'objet d'un suivi par la comparaison des pratiques aux référentiels et les accords de bons usages. Un dictionnaire référence l'ensemble des variables et documente les données et leurs règles de gestion : le « wiki-sniiram ».

Les données recueillies ne se prêtent pas à la base à un objectif d'études puisqu'elles servent à verser des prestations aux assurés, et donc avec des contraintes de productions et un cadre législatif imposé. Cependant, en 2013, une cinquantaine de chercheurs a interrogé de manière régulière cette base de données, et a réalisé plus de 17 000 requêtes, en augmentation de 30 % par rapport à l'année précédente. Ainsi, à la demande des autorités sanitaires, des études visant à évaluer la sécurité des médicaments ont été menées, avec la publication de travaux sur le risque cardiaque associé à la consommation de Benfluorex [66] ou sur la sécurité de la Pioglitazone[67].

En termes d'accès, les données médicales permettant une identification indirecte des bénéficiaires (mois et année de naissance, date des soins, commune de résidence et date de décès), sont consultables par un nombre très restreint d'utilisateurs, sous la responsabilité d'une autorité médicale (services ministériels, agences sanitaires, organismes publics de recherche, et exclusion de tout organisme poursuivant un but lucratif - arrêté du 19 juillet 2013) et avec des profils de droits d'accès spécifiques. La liste des organismes et des accès autorisés est définie par arrêté : le 6 octobre 2016, un nouvel arrêté[68] a cependant fait disparaitre la notion de caractère lucratif, remplacé par une approbation du bureau de l'Institut des données de santé. Reste à savoir si de nouvelles autorisations vont se concrétiser dans ce sens.

---

[66] Weill A, Païta M, Tuppin P, Fagot JP, Neumann A, Simon D, Ricordeau P, Montastruc JL, Allemand H. Benfluorex and valvular heart disease: a cohort study of a million people with diabetes mellitus. Pharmacoepidemiol Drug Saf. 2010 Dec;19(12):1256-62.

[67] Neumann A, Weill A, Ricordeau P, Fagot JP, Alla F, Allemand H. Pioglitazone and risk of bladder cancer among diabetic patients in France: a population-based cohort study. Diabetologia. 2012 Jul;55(7):1953-62. Epub 2012 Mar 31.

[68] www.entreprise.news/lusage-donnees-de-sante-tres-limite/

3.4.2.2. La haute Autorité de Santé (HAS)

La Haute Autorité de Santé[69] est une entité indépendante avec 2 activités principales : évaluation et recommandation des pratiques cliniques, accréditation et évaluation des établissements de santé. Elle met également à disposition un ensemble de guides et méthodes pour mettre en œuvre des projets dans le respect de la qualité et de la sécurité du patient. La HAS a été dotée de plus de responsabilités depuis 2005, par de nombreuses modifications législatives, ces missions sont définies aux articles 161-37 et suivants du code de la sécurité sociale.

La HAS donne ainsi des avis sur les médicaments, les dispositifs médicaux, les technologies de santé, les bonnes pratiques professionnelles,… Les certifications sont issues de procédures d'évaluation d'établissements de santé publics et privés, concernant le niveau de prestations et soins, la qualité et sécurité de ces soins. Les accréditations des médecins concernent la gestion des risques médicaux en établissement de santé.

La HAS est associée au Programme Santé Connectée[70], dont les enjeux touchent l'interopérabilité sémantique, socle technique de la santé publique par le développement de la valeur d'usage de la donnée clinique. L'objectif est de développer la capacité des logiciels à utiliser les données recueillies au cours des soins, en évitant les saisies multiples d'une même information, et en optimisant la structuration du système. Ce programme est soutenu en partenariat avec l'ASIP Santé, qui constitue le socle technique pour la mise en œuvre de l'interopérabilité des systèmes d'information de santé.

3.4.2.3. L'ASIP Santé

L'ASIP Santé (Agence des Systèmes d'Informations Partagées de Santé) existe depuis 2009 pour accompagner l'émergence des nouvelles technologies en santé, et mettre en place une gouvernance des systèmes d'information de ce secteur. Elle a donc un rôle de maitrise d'ouvrage pour des projets tels que le Dossier Médical Personnel (DMP), et l'une de ses missions concerne le processus de mise en œuvre et de suivi des référentiels et standards entrant dans le cadre de l'interopérabilité des Systèmes d'Information de Santé (SIS).

Ainsi, L'ASIP Santé définit, assure la maintenance et publie des référentiels nationaux sur lesquels s'appuient les SIS. Ces référentiels recouvrent les domaines de l'identification, de l'interopérabilité et de la sécurité.

Le Répertoire National des Référentiels (RNR)[71] est l'espace de publication des référentiels suivants :

- Modèle des Objets de Santé (MOS) : c'est une bibliothèque de composants sémantiques qui centralise, dans une documentation de référence unique, les mêmes définitions, nommages, structures et codages de l'information. Le MOS concerne les nomenclatures déjà publiées ou à venir.
- Spécifications de référence, telles que : le Cadre d'Interopérabilité des Systèmes d'Information de Santé (CI-SIS), l'Identifiant National de Santé

[69] http://www.has-sante.fr/portail/

[70] http://fr.slideshare.net/esante_gouv_fr/2014-0206-asipsanterirprogrammesanteconnectee?ref=http://esante.gouv.fr/actus/regions/retour-sur-les-rencontres-inter-regionales-du-6-fevrier-2014-a-paris

[71] http://esante.gouv.fr/services/referentiels/presentation-du-repertoire-national-des-referentiels-rnr/presentation-du

(INS), la Politique Générale de Sécurité des Systèmes d'Information de Santé (PGSSI-S). Les nomenclatures métier du CI-SIS peuvent être existantes et gérées par L'ASIP Santé (SNOMED 3.5, Code Identifiant de Spécialité CIS, Code Identifiant de Présentation CIP)[72], ou créées par l'ASIP Santé pour de nouveaux besoins.

3.4.2. Cartographie des bases de données publiques de santé françaises

Une cartographie de plus de 260 bases ou jeux de données ayant été recensés, dans le cadre du *débat thématique sur l'ouverture des données publiques de santé* lancé par le Ministère des Affaires Sociales et de la Santé en novembre 2013, a été publiée par la mission Etalab[73].

Chaque base de données identifiée a fait l'objet d'une évaluation de son "niveau d'ouverture" actuel, selon de 4 critères:

- la liberté d'accès (qui a accès aux données ?) ;
- le coût d'accès (les données sont-elles disponibles gratuitement ?) ;
- le format de mise à disposition (les données sont-elles proposées dans des formats facilitant la réutilisation ?) ;
- les conditions juridiques de la réutilisation (la réutilisation des données est-elle explicitement autorisée ?).

Par ailleurs, deux niveaux de précision sont également identifiés pour chaque base ou jeux de données :

- le niveau granulaire correspondant aux données de niveau le plus fin possible en fonction de leur origine et du système de collecte ;
- le niveau agrégé résultant du regroupement des données granulaires selon une ou plusieurs caractéristiques communes.

Comme l'illustrent les résultats obtenus dans le tableau 2 ci-dessous, les restrictions d'accès se concentrent principalement sur les données granulaires. Ce sont en effet celles-ci qui concentrent les difficultés d'anonymisation, essentielles pour éviter la ré-identification des patients. Ce sont aussi ces données difficilement accessibles qui intéressent le plus les entreprises, les militants et les chercheurs à l'origine de l'Initiative Transparence Santé, un regroupement d'acteurs hétéroclites, lancée le 25 janvier 2013, qui s'interroge sur les restrictions légales et techniques qui limitent la réutilisation des données ouvertes[74].

[72] Cf chapitre de ce mémoire sur les médicaments
[73] https://www.data.gouv.fr/fr/datasets/cartographie-des-bases-de-donnees-publiques-en-sante/
[74] http://publications-sfds.fr/index.php/stat_soc/article/view/313/294

Critère	Données agrégées	Données granulaires
Liberté d'accès	Accessible à tous : 95 En accès restreint : 2 Non précisé : 189	En accès restreint : 154 Accessible à tous : 47 En accès fermé : 19 Non précisé : 66
Cout d'accès	Gratuit : 97 Non précisé : 189	Gratuit : 138 Payant : 40 Payant ou gratuit : 10 Non précisé : 98
Format d'accès	Non exploitable : 73 Exploitable : 22 Non précisé : 191	Exploitable : 163 Non exploitable : 25 Non précisé : 98
Condition de réutilisation	Explicité avec restriction : 57 Explicité sans restriction : 22 Non explicité : 9 Non précisé : 198	Explicité avec restriction : 107 Explicité sans restriction : 10 Non explicité : 14 Non précisé : 155

Tableau 2 : Synthèse de la cartographie des données de santé, réalisé par la Société de Statistiques Française, données Etalab avril 2014 [75].

3.4.3. Mise en œuvre de terminologies médicales de référence en France

Les terminologies de référence ont été identifiées comme un sujet prioritaire dans un objectif d'interopérabilité sémantique. Dans ce cadre, En mars 2014, L'ASIP Santé a été chargée par la DSSIS (Délégation à la Stratégie des Systèmes d'Information de Santé du Ministère des Affaires Sociales, de la Santé, et du Droit des Femmes) de la réalisation d'une *étude sur la mise en œuvre de terminologies de référence pour le secteur santé-social en France* [34]. Les axes de l'étude concernent l'organisation des contributions françaises à la normalisation, ainsi que les règles de mise en œuvre et d'utilisation.

Une explication sur les éléments essentiels à l'interopérabilité sémantique est donnée dans un premier temps.

On trouve tout d'abord les modèles de structures d'information interopérables permettant les échanges, comme la e-prescription, le volet de synthèse médicale, la feuille de soins électronique,…

Il est ensuite indispensable de disposer de concepts codés qui associent un code porteur du sens pour les systèmes informatiques, un terme exprimant le concept sous forme verbale pour les utilisateurs, et la référence de la terminologie associée (par exemple : E11 | Diabète sucré de type 2 | CIM-10-FR V2015).

Puis, les jeux de valeurs permettent d'énumérer les concepts codés permis dans un contenu conforme à un modèle de structure d'information interopérable. Ils peuvent prendre la forme d'une liste de choix possible lors de la saisie d'un champ de formulaire.

Les terminologies de référence fournissent quant à elles, les définitions formelles et univoques des concepts codés d'un domaine spécifique, avec leur

[75] http://publications-sfds.fr/index.php/stat_soc/article/view/313/294

structure et leurs interrelations éventuelles. Ce sont par exemple CIM-10[76], CISP-2[77], SNOMED CT[78], LOINC[79], CCAM[80], MedDRA[81],…

Les alignements sémantiques définissent les relations entre les concepts codés de deux terminologies de référence, afin d'aider au transcodage de l'une à l'autre. Par exemple, la classification internationale des soins primaires CISP-2 dispose d'un alignement sémantique de ses diagnostics et maladies avec la CIM-10, entre le concept codé T90 (diabète de type 2) de la CISP-2 et les concepts codés E11, E12, E13, E14 et leurs spécialisations de la CIM-10.

Enfin, les services normalisés d'accès aux ressources terminologiques mises en ligne (jeux de valeurs, terminologies, alignements sémantiques) sont utilisés par les logiciels métier des acteurs de santé.

La qualité des ressources terminologiques produites est primordiale. Les processus doivent être organisés et outillés pour satisfaire aux exigences fortes de qualité : univocité et non ambiguïté des concepts codés, cohérence des relations sémantiques entre concepts, pertinence des traductions, justesse des alignements sémantiques, adéquation des jeux de valeurs aux cas d'usage, pérennité dans le temps des différentes versions (l'historique d'un patient se doit d'être complet et sans altération de sens).

Au niveau économique, la question des licences est également essentielle, cette dernière concernant également la protection de la propriété intellectuelle. Ainsi, le financement peut être indirect et réparti sur la communauté, avec une mise à disposition gratuite, sans licence (CCAM, CIM-10 de l'ATIH[82]), ou via une licence nominative gratuite (SNOMED 3.5 VF portée par l'ASIP Santé), ou encore subventionnée par des organisations (LOINC utilisable sous licence gratuite partout dans le monde, est financée par de nombreuses organisations américaines dont la *National Library of Medicine* NLM[83]). Le financement peut également être direct avec par exemple, une licence « par produit » lorsqu'on intègre la ressource dans un produit commercial (MedDRA), ou bien une licence « par usage » si une organisation souhaite utiliser la ressource à l'intérieur de son périmètre (SNOMED CT).

Une autre problématique majeure vient de la diversité des terminologies de référence à maîtriser. Par exemple pour les médicaments, on trouve :

- IDMP : *Identification of Medicinal Products* (ISO 11615, 11616, 11238, 11239, 11240) ;
- CIS : Spécialités ;
- CIP : Présentations ;
- UCD : Unités Communes de Dispensation ;
- CIOsp : Référentiel interopérable des spécialités, agrégeant plusieurs terminologies ;
- ATC : *Anatomical Therapeutic Chemical* ;
- DCI (INN) : Dénomination Commune Internationale des principes actifs Standard Terms ;

[76] http://ec.europa.eu/health/indicators/international_classification/index_fr.htm
[77] http://www.refcisp.info/index.php5?page=listeCodes&rubrique=consultation
[78] http://www.snomed.org/snomed-ct/what-is-snomed-ct
[79] http://www.health.belgium.be/fr/terminologie-et-systemes-de-codes-loinc
[80] http://www.ccam.sante.fr/
[81] Cf chapitre Benchmark des modèles de ce mémoire
[82] http://www.atih.sante.fr/
[83] Cf chapitre Benchmark des modèles de ce mémoire

- EDQM - pharmacopée européenne : formes galéniques, voies d'administration ;
- DC : Dénomination Commune des principes actifs, à utiliser pour la prescription en France ;
- MedDRA, WHO-ART : Evénements indésirables et pharmacovigilance.

Cette variété s'explique par les cas d'usage auxquelles chaque terminologie se réfère, afin de satisfaire au mieux le besoin. En effet, lorsqu'on s'écarte des usages prévus à la conception, la terminologie perd en pertinence et en efficacité de codage (voir tableaux 3, 4, et 5).

Cas d'usage	Acteur producteur	Terminologie
Codage des causes de mortalité	CépiDc	CIM-10
Codage des causes de morbidité. Codage médico-administratif des diagnostics pour le PMSI	Unité de soins et/ou DIM	CIM-10
Codage des ALD	Médecin déclarant	CIM-10
Documentation des problèmes de santé par la médecine de premier recours	Médecin généraliste	CISP-2 ou DRC
Documentation des problèmes de santé dans le dossier patient informatisé	Médecin spécialiste en établissement ou en ville	SNOMED CT ou SNOMED 3.5 VF
Documentation d'une maladie rare	Médecin spécialiste à l'hôpital	ORPHANET, OMIM, HPO, SNOMED CT

Tableau 3 : Cas d'usage pour documenter les problèmes de santé [34].

Cas d'usages pour l'utilisation de LOINC
Typage d'un document de santé dématérialisé (CR d'imagerie, fiche RCP ...)
Typage d'une section d'un document de santé dématérialisé (problèmes en cours, facteurs de risques, allergies, résultats de microbiologie ...)
Typage d'une observation (poids pré-dialyse, type histologique d'une lésion tissulaire, sodium sanguin, tension systolique, identification d'une bactérie dans une hémoculture ...)
Identification d'un examen demandé (biologie, imagerie, anatomie pathologique et cytologie)[2]

[2] Sous réserve de validation de la couverture sémantique de LOINC par des groupes d'experts métier.

Tableau 4 : Cas d'usage pour l'utilisation de la terminologie LOINC [34].

Cas d'usage	Terminologie
Résultat d'une observation qualitative	SNOMED CT
Unité de mesure accompagnant le résultat d'une observation numérique (mmol/L)	UCUM
Unité de mesure pour posologies journalières, volumes des médicaments en solution, masses pondérales de base active des médicaments[4]	UCUM

[4] Source : Interview de la HAS

Tableau 5 : Les terminologies utiles pour les résultats d'observation et de mesure [34][84].

Cas d'usage	Besoin identifié	Commentaire
Consolidation des résultats de biologie médicale issus de différents laboratoires pour un même patient	Codage des techniques analytiques, afin de sécuriser la comparabilité inter-laboratoires des résultats	Besoin formulé par Interop'Santé en interview. A instruire avec biologistes et industriels du diagnostic in vitro
Prescription de médicament en dénomination commune	Besoin d'un référentiel de médicament virtuel	Besoin formulé par la HAS et par l'ANSM en interview. L'ANSM envisage d'instruire le sujet.
Vigilances sur les DM	Besoin de codage des dysfonctionnements des DM, de leurs causes et de leurs effets sur les patients	Besoin formulé par l'ANSM en interview. A instruire.
Dématérialisation du carnet de vaccination, aide automatisée à la gestion des vaccinations d'une personne	Codage des valences vaccinales	Besoin formulé par la HAS en interview. Un point de départ possible : les abréviations de valences vaccinales publiées par le CDC aux Etats Unis[5]

[5] Abréviations de valences vaccinales du CDC : http://www.cdc.gov/vaccines/about/terms/vacc-abbrev.htm

Tableau 6 : Besoins de codage non couverts par les terminologies de référence existantes [34].

Certains cas d'usage font apparaître des besoins de codage de l'information qui ne semblent couverts par aucune terminologie de référence connue à ce jour (Tableau 6) : le référentiel de médicaments en fait partie, d'où l'intérêt de la création d'une ontologie française des médicaments.

D'un autre côté, on constate des chevauchements de domaines sémantiques entre plusieurs terminologies, ce qui oblige à devoir choisir entre deux langages différents pour coder la même information. Par conséquent, la réutilisation des données d'un système à l'autre, nécessite de passer par plusieurs étapes intermédiaires de transcodages, et donc augmente les risques d'erreur ou de dégradation de sens. Certains arbitrages ont été ou vont être rendus au niveau européen (Tableau 7).

[84] La terminologie de référence des unités de mesure est UCUM. De plus, UCUM est aussi nécessaire pour préciser les mesures liées aux caractéristiques des médicaments prescrits

Cas d'usage	Terminologies concurrentes	Arbitrage européen[6]
Documentation des problèmes de santé par la médecine de premier recours	CISP-2 ; DRC	n.a., le DRC n'étant pas un standard international
Documentation des problèmes de santé par les spécialistes, observations qualitatives, infections, ...	SNOMED 3.5 VF ; SNOMED CT	n.a., SNOMED 3.5 VF n'étant pas un standard international
Codage des diagnostics, interventions et résultats de la pratique des soins infirmiers	NNN (NANDA-I, NIC, NOC) ; ICNP	Pas d'arbitrage connu à ce jour
Typage des dispositifs médicaux pour un usage autre que l'usage médico-administratif assuré par la LPPR (art L-165-1 du CSS)	ClaDiMed ; GMDN ; UMDNS	Le règlement européen en préparation sur les DM semble s'orienter vers GMDN. A suivre en 2015
Pharmacovigilance, signalement à l'ANSM des effets indésirables de médicaments	WHO-ART ; MedDRA	MedDRA obligatoire en Europe

[6] n.a. dans cette colonne du tableau signifie « non applicable »

Tableau 7 : Terminologies en situation de concurrence sur un même cas d'usage [34].

On peut citer d'autre part, le cas particulier des deux terminologies SNOMED.

On trouve d'abord la « SNOMED International 3.5 VF », traduction française d'une terminologie créée au Canada et qui a été achetée pour la France par l'ASIP Santé. Cette version a été essentielle pour bâtir les modèles de contenus du CI-SIS et permettre ainsi le partage de contenus médicaux structurés. Elle est utilisable sous licence gratuite par toute organisation qui le souhaite.

On distingue ensuite la SNOMED CT qui est l'évolution sous forme d'ontologie et avec de nombreux enrichissement de la SNOMED International 3.5 et est utilisable sous licence nationale (en général gratuite pour les utilisateurs finaux) dans les 27 pays membres d'IHTSDO (organisme à but non lucratif) dont 15 sont en Europe (Belgique, Danemark, Espagne, Estonie, Island, Lituanie, Malte, Pays Bas, Pologne, Portugal, République Tchèque, Royaume Uni, Slovaquie, Slovénie, Suède), ainsi que dans les pays dits émergents. En dehors de ces pays et notamment en France, l'utilisation de SNOMED CT dans un contexte de production de soins requiert l'acquisition auprès d'IHTSDO d'une licence « affilié » par l'organisation de soins demandeuse. Le coût de cette licence est fonction du nombre de systèmes présents dans l'organisation exploitant SNOMED CT en production ou en analyse de données. Le coût est d'environ 1 700 USD annuel par système.

Comme on peut le constater dans le tableau 8 qui résume leurs caractéristiques distinctives, 18 ans d'expertise médicale et de conception sémantique séparent aujourd'hui les deux versions.

	SNOMED 3.5 VF	SNOMED CT
Standard international	Jusqu'en 1998	Oui
Pérennité et évolutivité	Figée depuis 1998[10] SNOMED 3.5 est non maintenue par IHTSDO et sera déclarée obsolète en avril 2017.	Deux versions par an produites avec le concours d'une communauté internationale d'experts et de professionnels de santé utilisateurs, par les « *editors* » et les consultants terminologistes d'IHTSDO.
Gestion des relations sémantiques	Hiérarchies exprimées à travers la valeur des codes. Relations limitées, sous la forme de références, par exemple une morphologie rappelle entre parenthèses le code topographie auquel elle s'applique.	Codes non signifiants. Ontologie sous-jacente représentée par le « Concept model » de SNOMED CT, et exprimée à travers 1,5 million de relations sémantiques explicites entre les concepts, dont les relations hiérarchiques multiaxiales.
Gestion des termes synonymes	Limitée	Illimitée
Traduction en français (origine canadienne dans les deux cas)	90 % des 150 000 concepts codés	10 % des 350 000 concepts codés (pas de termes synonymes traduits en français à ce jour)
Droits d'usage en France	Sans limitation, sous licence gratuite	- Projets de recherche - En production de soins sous licence « affilié » confinant les concepts SNOMED CT dans les systèmes déclarés pour la licence

[10] Source : Rapport Fieschi 2009 "La gouvernance de l'interopérabilité sémantique est au cœur du développement des systèmes d'information en santé" http://www.sante.gouv.fr/IMG/pdf/RapportFieschi.pdf

Tableau 8 : Comparaison entre SNOMED 3.5 VF et la version courante SNOMED CT [34].

De nombreuses sociétés industrielles interviewés par l'ASIP Santé (AGFA, MAINCARE, VIDAL, MONDECA et SANOFI) souhaitent une ouverture de droits d'usage en France de SNOMED CT, car ce standard international permettrait de garantir l'évolutivité, la pérennité, l'échange facilité et sécurisé des données de santé des patients.

De manière générale, l'ASIP Santé dégage deux axes d'amélioration possibles : l'harmonisation des formats techniques de diffusion des informations, et le développement de services standardisés d'accès aux ressources terminologiques.

3.5. Quid de l'ouverture des données dans d'autres domaines ou pays

Il est intéressant de faire un rapide tour d'horizon sur des domaines autres que la santé, pour comparer l'état d'avancement et ce qui peut être dégagé au niveau retours d'expérience.

3.5.1. Enseignement et Culture

Dans le secteur de l'enseignement, le ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche (M.E.N.E.S.R.) propose aujourd'hui la première plateforme ministérielle de mise à disposition de données sur l'enseignement supérieur et la recherche[85]. Par la mise à disposition d'un vaste ensemble de données réutilisables, le M.E.N.E.S.R. s'engage dans le développement d'une expertise partagée sur l'enseignement supérieur et de la recherche. En plaçant

[85] http://www.enseignementsup-recherche.gouv.fr/cid76441/open-data-notre-demarche.htm

l'*Open Data* au cœur de la stratégie de modernisation de son action, le ministère souhaite également encourager l'innovation et le développement économique. Le site a été rénové en octobre 2016.

55 jeux de données sont donc disponibles, sous licence ouverte Etalab, sur le site, pour l'exploration, le téléchargement ou la réutilisation vie des APIs. On trouve par exemple la « liste des écoles doctorales accréditées »[86], avec une description générale, les résultats sous forme de tableau, de carte géolocalisée, la possibilité d'exporter les fichiers aux formats CSV, JSON, EXCEL et les cartes aux formats GeoJSON[87], Shapefile[88] ou KML[89], et l'utilisation d'une API.

Dans le secteur de la culture, l'exemple de la Bibliothèque nationale de France (BnF) est emblématique de l'ouverture des données, et des innovations qu'elle a mises en place liées au Web sémantique. En 2011, la BnF a été la première institution culturelle à déposer sur le site data.gouv.fr un premier jeu de données mis à disposition du public sur son site data.bnf.fr[90]. Aujourd'hui, elle offre un véritable service, consistant à regrouper les informations des différents catalogues et bases de la BnF (catalogue général, catalogue archives et manuscrits, reliures.bnf.fr,...) et de sa bibliothèque numérique Gallica [35, DUHAMEL].

L'ensemble de ses données est librement et gratuitement réutilisable sous réserve de la mention d'attribution de la source (Licence ouverte Etalab). Cette interopérabilité favorise ainsi les échanges avec d'autres bibliothèques, y compris celles de l'enseignement supérieur et de la recherche, voire d'autres communautés présentes sur le Web, la Bnf servant de fédérateur ou pivot.

Les données des différentes ressources sont, de ce fait, valorisées, intégrées et liées entre elles, grâce aux standards du Web sémantique et du Web de données. En effet, les URIs identifient les ressources, le formalisme RDF est utilisé, ainsi que les ontologies, pour les décrire, et les données sont interrogeables depuis le mois d'août 2014 avec un service SPARQL endpoint. Le modèle de données (voir figure 4) offre également des alignements vers des référentiels extérieurs, tels que Geonames[91] pour les lieux, VIAF[92] pour les données d'autorité,…

[86] https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-ecoles_doctorales_annuaire/
[87] http://geojson.org/
[88] https://fr.wikipedia.org/wiki/Shapefile
[89] https://fr.wikipedia.org/wiki/Keyhole_Markup_Language
[90] http://www.bnf.fr/fr/professionnels/anx_recuperation_donnees/a.ouverture_donnees_bnf.html
[91] http://www.geonames.org/
[92] http://www.bnf.fr/fr/professionnels/donnees_autorites/a.viaf.html

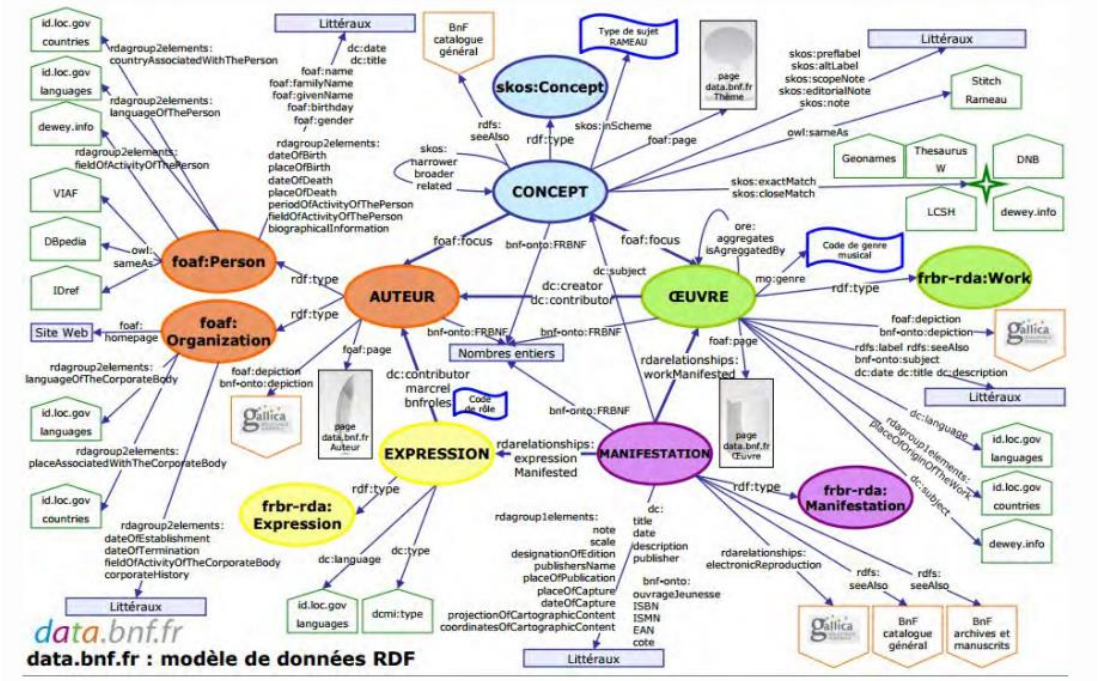


Figure 4 : Fédération des informations des différents catalogues de la BnF autour d'entités œuvre, auteur [93].

3.5.2. Royaume-Uni

Avec les États-Unis avec data.gov en 2009, le Royaume-Uni avec data.gov.uk en janvier 2010, fait partie des précurseurs de l'Open Data. Auparavant, c'est déjà une association née en 2004 au Royaume-Uni, l'*Open Knowledge Foundation* (OKFn, Fondation pour des connaissances ouvertes), qui a défini l'ouverture des données : « *Une donnée est ouverte si n'importe qui peut l'utiliser, la réutiliser, et la redistribuer. Les seules conditions étant l'attribution de l'auteur, et la conservation du statut de donnée ouverte pour les réutilisations suivantes*. ». Une autre phase célèbre, *Free our data* (« Libérons nos données »), a accompagné la création et le développement du mouvement Open Data au Royaume-Uni, où il y a pris une réelle ampleur (avant la France) [36, DHRZANOWSKI].

Au Royaume-Uni, chaque administration gouvernementale ou locale est un producteur ou un détenteur de données publiques. On trouve les ministères eux-mêmes, les administrations comme le NHS (*National Health Service*, service national de santé), les autorités locales (régions, comtés, districts et villes), et les « *Trading Funds* », administrations financièrement indépendantes (Met Office, par exemple, Service de météorologie du Royaume-Uni).

Le portail data.gov.uk, constitue le pivot central d'accès aux données publiques du Royaume-Uni, et se présente comme un catalogue de jeux de données détenues par les administrations. Il permet de consulter les données et de les télécharger (formats XLS, XML, PDF, ZIP, HTML, CSV, RDF,…), ainsi que de connaitre les applications utilisatrices.

Techniquement, data.gov.uk repose sur le gestionnaire de contenus Drupal[94], et sur l'application CKAN[95] développée par l'OKFn, permettant la création de

[93] http://www.bnf.fr/documents/3_afnor2016_open_data_usages.pdf
[94] http://www.drupal.fr/
[95] http://ckan.org/

catalogues de jeux de données. Les producteurs de données peuvent directement s'identifier sur le portail, pour ensuite répertorier leurs jeux de données, sous forme d'URL dans un module prévu à cet effet. Les données sont ensuite téléchargeables depuis data.gov.uk, et exploitables par les applications via les APIs CKAN.

En termes de copyright, les producteurs de données obéissent au « *Crown Copyright* » spécifique aux données publiques appartenant à l'État (*sous la couronne*). Concernant les licences de réutilisation, c'est en majorité *l'Open Government Licence* (OGL) qui est utilisé: sur les 293 autorités locales qui avaient publié leurs données fin janvier 2011, 141 avaient choisi l'OGL.

La question centrale des formats, à laquelle se confronte la France, est également un enjeu poursuivi par le Royaume-Uni, car actuellement, un tiers des données sont publiées dans le format XLS non adapté au Web, et à la réutilisabilité. Le gouvernement cherche à convaincre les administrations de publier des données dans les formats préconisés, XML, RDF ou JSON mais également d'y associer des métadonnées. L'objectif étant toujours de standardiser, et de rendre les données lisibles par les ordinateurs pour construire le Web de données. Ce changement reste lent et difficile pour les administrations, autant techniquement que du point de vue organisationnel.

Les applications répertoriées (plus de 135 en 2015) ne répondent pas à un vrai besoin pour la plupart ; on trouve tout de même dans le domaine de la santé des applications iPhone comme « *NHS local* » ou « *UK Dentists* » permettent de trouver l'hôpital ou le dentiste le plus proche, ou encore « *HealthyApps* » donne la possibilité de comparer les services chirurgicaux en fonction du nombre d'infections nosocomiales répertoriées.

3.6. Open Data : quelle réalité ?

L'ouverture des données publiques n'a pas encore répondu aux attentes jugées parfois idéalistes des militants Open Data. La gestion de l'information publique est un territoire nouveau pour les entrepreneurs et les citoyens et les données, sans les analyses et les outils pour les comprendre, restent pour beaucoup encore des chiffres obscurs.

En effet, l'enjeu n'est pas simplement l'ouverture, mais aussi et surtout, leur mise à disposition correcte et cohérente, leur exploitation et leur interprétation. Que les données soient issues du secteur public, de la recherche scientifique, des entreprises ou des citoyens, il s'agit d'inventer et de mettre en place les technologies et mode de gouvernance nécessaires pour les traiter, les analyser et en faire ressortir le sens et le potentiel. Les défis concernent à la fois les formats de données et de métadonnées utilisées, les portails qui en assurent l'accès, l'architecture du Web qui en permet la diffusion et enfin les applications qui permettent de les traiter et les visualiser.

Le cheminement est en cours pour tirer le bénéfice des nouvelles technologies du Web sémantique et du *linked data*, accroitre la capacité à échanger et traiter des informations homogènes, et interconnectées. Dans le domaine de la santé, les données issues des études de terrain seront à la base des nouveaux traitements fabriqués dans les laboratoires. En partie, grâce à l'établissement de formats standards et à la création de base de métadonnées communes, avec la dimension de représentation formelle des connaissances, les ontologies.

Troisième partie : Open Data en pratique

L'ontologie française des médicaments

Ou comment contribuer à l'interopérabilité technique, économique et politique de l'ouverture des données dans le domaine de la santé.

4. L'ontologie des médicaments : contexte

L'e-science vise la production de nouvelles connaissances par l'accès et le traitement informatique de jeux de données à grande échelle et permet l'exploration et l'émergence d'hypothèses scientifiques jusque-là impossibles à vérifier [28, PRIME-CLAVERIE].

4.1. Le médicament : *Un produit pas comme les autres*

[37 SAFON *et al.*] Selon l'article Art. L. 5111-1 du code de la santé publique, le médicament est définit comme *toute substance ou composition présentée comme possédant des propriétés curatives ou préventives à l'égard des maladies humaines ou animales, ainsi que toute substance ou composition pouvant être utilisée chez l'homme ou chez l'animal ou pouvant leur être administrée, en vue d'établir un diagnostic médical ou de restaurer, corriger ou modifier leurs fonctions physiologiques en exerçant une action pharmacologique, immunologique ou métabolique….*

Un certain nombre de concepts concernant le médicament nécessite d'être abordés.

Le principe actif est la substance responsable de l'action pharmacologique, par exemple, le paracétamol. Il constitue l'un des composants de la spécialité. Celle-ci correspond à la définition suivante de l'article Art. L. 5111-2 du code de la santé publique : *On entend par spécialité pharmaceutique, tout médicament préparé à l'avance, présenté sous un conditionnement particulier et caractérisé par une dénomination spéciale*. Par exemple " Voltarène comprimés 50mg " et " Voltarène suppositoire 100 mg " et " Voltarène comprimés 25 mg " sont des spécialités différentes contenant un même médicament ou principe actif, le diclofénac.

Pour être commercialisées en France, les spécialités présentent une Autorisation de Mise sur le Marché (AMM). Elle est donnée suite à une procédure, définie par l'article Art.5121-8 du code de la Santé Publique (et suivant), selon trois critères principaux de qualité, sécurité et efficacité. Elle est délivrée par le Directeur de l'ANSM (Agence Nationale de Sécurité du Médicament et des Produits de Santé)[96] ou son homologue européen, le directeur de l'European Medicines Agency (EMEA), puis publiée au Journal Officiel.

Un numéro d'enregistrement d'AMM est attribué à la spécialité pharmaceutique (reporté sur le conditionnement, sous le libellé "Médicament autorisé n° ...").

Concernant la question de la prise en charge du médicament par la Sécurité Sociale (liste des médicaments remboursables), la Commission de la Transparence intégrée à la HAS (Haute Autorité de Santé)[97] donne un avis, et une appréciation sur le service médical rendu (SMR). Un autre critère, l'ASMR (relatif), permet de spécifier une valeur ajoutée au médicament déjà présent sur le marché.

Le Comité Economique des Produits de Santé (CEPS), réunissant des représentants de différents ministères (Economie et Finances, Sécurité sociale, Industrie et Santé), ainsi que des représentants de la l'UNCAM, sont chargés de la fixation des prix du médicament après avis de la Commission de Transparence.

[96] Cf. chapitre sources de données dans ce mémoire
[97] Cf. chapitre domaine de la santé

Lors de la publication au Journal Officiel, un arrêté mentionne le nom commercial de la spécialité, sa dénomination commune (nom du principe actif), son taux de remboursement ainsi que les indications thérapeutiques remboursables. La figure 5 illustre l'ensemble de la procédure.

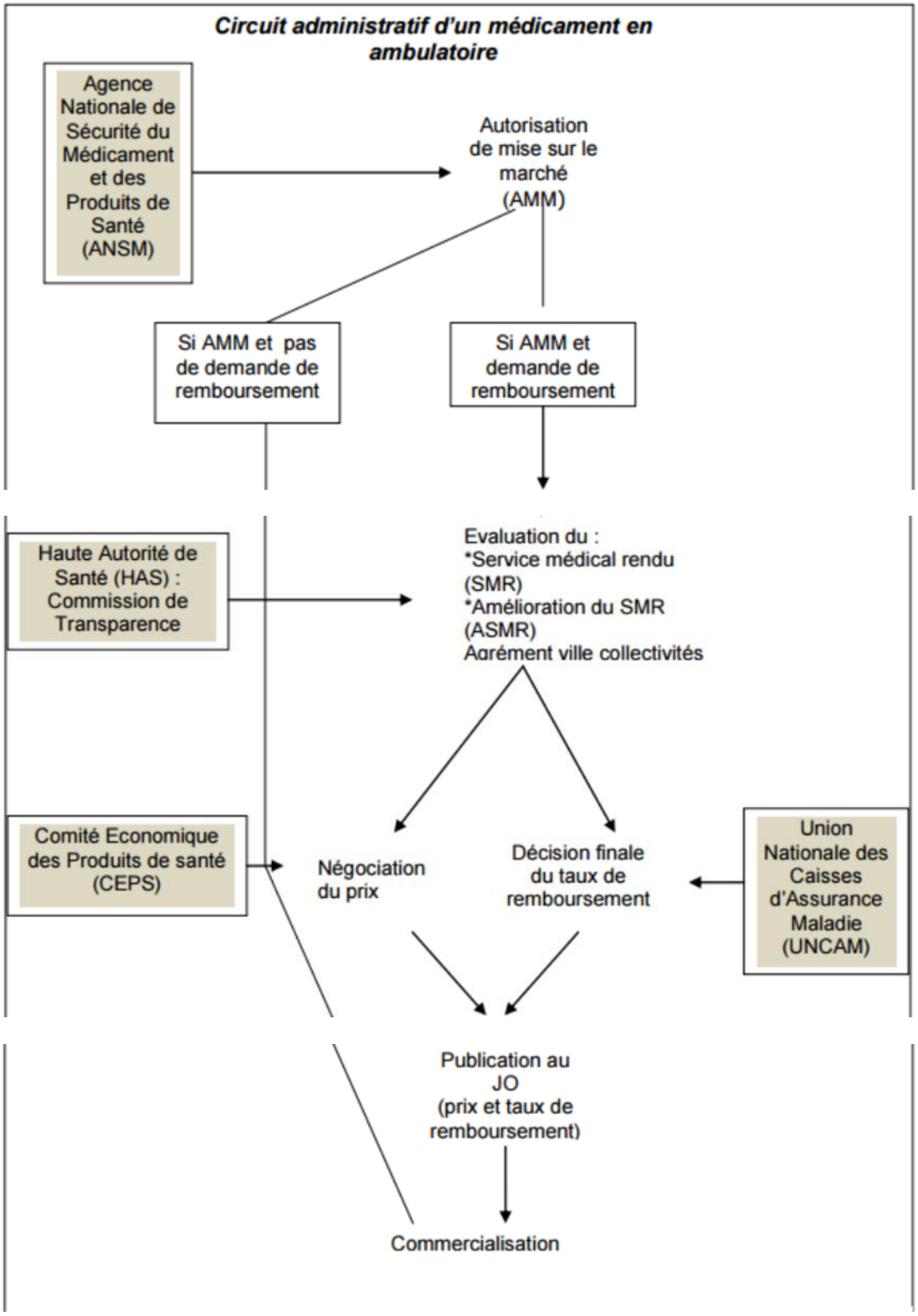


Figure 5 : Procédure d'Autorisation de Mise sur le Marché (AMM) en France [37 SAFON et al.*].*

La rigueur de cette procédure s'explique par la forte croissance de la consommation pharmaceutique, et donc du coût imputable à l'assurance maladie, que l'État souhaite donc réguler. Par ailleurs, la question de la pharmacovigilance est également prégnante, afin de minimiser les risques d'effets indésirables des médicaments. Sur ce point, l'ANSM dispose d'un fort pouvoir de sanction financière, en cas de non transparence sur d'éventuels effets indésirables.

4.2. Pourquoi une ontologie française des médicaments?

Les constats actuels, et retours suite à l'enquête menée par l'ASIP Santé [34] relèvent les points suivants :

- La disponibilité de données de santé structurées de bonne qualité est nécessaire pour améliorer la coordination entre les professionnels de santé impliqués dans le parcours de soins du patient. Ce besoin n'est pas uniquement national, et les terminologies de référence se doivent donc d'être standardisées au niveau international. Cela constitue également un point d'entrée dans un cercle vertueux de bonnes pratiques.
  Dans ce cadre, et comme on le verra plus loin dans ce mémoire, l'ANSM donne ainsi accès à ses tables de référence, mais dans des formats complexes à traiter, et sous la forme de nombreux fichiers. Notre valeur ajoutée, dans le cadre du stage, revient alors à délivrer nos propres sources respectant les standards du Web sémantique, à partir de celles de l'ANSM, éventuellement complétées par d'autres.
- *La donnée de santé est à la fois source de connaissance et la clé d'accès à celle-ci*, c'est pourquoi elle doit faire l'objet d'attention au niveau de son extraction puis de sa structuration sémantique.
- Nous avons vu précédemment que l'ensemble des cas d'usage n'était pas couvert par les terminologies actuellement disponibles, ou de manière partielle. Ainsi, la prescription en dénomination commune a besoin d'un référentiel de médicaments virtuels français. *La médecine de premier recours utilise aujourd'hui illégalement la terminologie de référence internationale CISP-2 qui répond à son besoin*. SNOMED CT qui permettrait de produire des données de qualité à l'échelle française, reste payante faute de licence.
- La nécessité d'homogénéiser également la mise à disposition des ressources terminologiques est également un critère d'amélioration pour aboutir à une standardisation globale.

Comme l'explique Jean Charlet dans son article [30, CHARLET], le monde médical a des particularités historiques, en termes de référentiel terminologique, mais aussi de sa nature même, qui se doit d'être très précise dans ses descriptions de cas. Il se voit aujourd'hui doté de nombreux outils informatiques de classement, dont il a eu besoin très tôt pour gérer et utiliser toute sa gamme de savoirs, concepts et entités liées à la santé. Cet existant oblige à des rénovations et adaptations, souvent plus difficiles à mettre en œuvre que des créations.

Les ontologies répondent bien aux contraintes de la médecine, puisqu'elles constituent des modèles de représentation formelle, capables d'intégrer des relations sémantiques riches ; le point d'attention étant de considérer qu'une ontologie ne peut être générique, car intimement liée au domaine qu'elle décrit, et donc forcément spécialisée. Afin d'obtenir une couverture plus large, les ontologies peuvent « dialoguer » entre elles, grâce à des *alignements entre leurs concepts* et un fédérateur qui orchestre l'ensemble en évitant ainsi la cacophonie.

Concernant les sources de données, un mélange entre la réutilisation d'ontologies existantes et les techniques de fouille de texte (*80 % de l'information*

médicale est non structurée) peut être réalisé : ce sont les méthodes « *top down* » « *bottom up* » abordées précédemment dans la partie sur l'ingénierie des connaissances. La question de l'*Open Access* est ici soulevée, concernant les différentes politiques et droits d'accès aux corpus scientifiques.

Pour en revenir à ce qui nous concerne, l'élaboration d'une ontologie française des médicaments permettra d'offrir les avantages suivants :

- L'existence d'un outil commun pour le référencement, la recherche d'information ou l'export vers d'autres modèles favorisera la collaboration entre les chercheurs, médecins et pharmaciens français.
- En tant que ressource ouverte gratuite des données publiques hétérogènes, elle est attendue comme standard français, et pourrait être par la suite reconnue en tant que tel au niveau mondial, si elle se révèle pertinente.

Il s'agit donc de développer une terminologie de référence qui s'insère comme une brique connectée à la fois aux ressources qu'elle utilise (*d'autres référentiels pivots reconnus dans leur domaine ou les domaines proches* [29, LE PICARD]), et aux applications qu'elle nourrit, et ce de manière la plus normalisée et standardisée possible. Autrement dit, on va chercher à réutiliser au maximum toutes les ressources pérennes mises à disposition, en gommant leur hétérogénéité lors de leur intégration, et offrir ainsi une cohérence au final. Le tout grâce à des tâches les plus automatiques possible au sein du processus de création ou de mise à jour.

L'ontologie finale doit rester évolutive et permettre l'accroissement des connaissances, de par sa structuration et hiérarchisation, liée à l'univers du médicament, mais aussi des relations sémantiques qui lui donnent vie. La documentation des annotations et propriétés offertes par les ontologies est la clé d'enrichissement du domaine, donnant la possibilité d'établir des raisonnements et d'interroger le système obtenu de manière intelligente.

Pour devenir ce langage commun et pérenne, avec des liens et un vocabulaire appropriés, afin de donner une information la plus complète, et une sérendipité (« don de faire des trouvailles »), elle nécessite un certain nombre de moyens de mise en œuvre et une gouvernance adaptée. C'est l'objet des chapitres suivants.

4.3. De l'utilité d'un modèle

Comme évoqué précédemment, une des spécificités du monde médical est la richesse de la connaissance à manipuler. Par exemple, UMLS[98] contient 2 millions de concepts avec plus de 7 millions de termes de 140 terminologies biomédicales.

Il n'est pas rare d'avoir plusieurs termes pour définir le même concept et vice versa. Ainsi, la même maladie peut être désignée par des noms ou des expressions différentes (synonymie), le même terme peut avoir un sens différent suivant le contexte ou le locuteur (polysémie). Cette situation rend la numérisation de l'information médicale difficile, c'est la raison pour laquelle la discipline s'est intéressée aux systèmes de codage et de structuration de l'information issue de l'ingénierie des connaissances [38, CHOQUET].

La connaissance médicale évolue constamment. Dans le but de garantir une pérennité des systèmes d'informations biomédicaux, l'information médicale se subdivise en plusieurs ensembles : les données, les terminologies et la connaissance. Les données correspondent à une réalité observée ou mesurée dans un domaine,

[98] Cf. chapitre Benchmark des modèles et RxNorm de ce mémoire

les terminologies formalisent les termes (et leurs caractéristiques) utilisés pour stocker ces données, et les ontologies représentent les connaissances associées au domaine considéré.

Les modèles structurant l'information et les modèles structurant la connaissance offrent une formalisation nécessaire au partage de l'information. Le contenu de l'information partagée et véhiculée dans ces modèles d'information doit également faire preuve de qualité pour maintenir la cohérence de l'ensemble.

Déterminer la source de données et comment y accéder constitue la première étape. La question se pose ensuite de la façon d'extraire les données dans la structure dans laquelle elle est stockée. Les structures sont les garantes de la performance d'un système et de sa capacité à être interrogé, donc interopérable. En effet, plus un modèle de données stocke les données de manière implicite (l'approche EAV[99] par exemple), plus la connaissance humaine sera nécessaire pour construire des requêtes et plus celles-ci seront difficiles à mettre en œuvre. A l'inverse, plus la structure de stockage sera explicite, plus lourde seront les mises à jour de ces structures.

Il s'agit donc de trouver le bon compromis, et faire les choix corrects, en termes de modèle. Les attentes dans le domaine de la santé sont fortes, afin de mieux soigner tout en gagnant du temps : l'amélioration de l'accès à la connaissance, malgré la complexité de l'information médicale, impose l'utilisation de modèles sémantiques en lieu et place de systèmes *classiques*.

Mike Loukides [39, LOUKIDES] explique très justement que sans modèle de données, les données sont en quelque sorte orphelines, sans signification ou domaine de couverture, dénuées de contexte : *Il n'y pas que les données qui doivent être ouvertes : il y a aussi les modèles ! (…) Vous pouvez avoir toutes les données sur la criminalité que vous voulez, toutes les données de l'immobilier que vous voulez, toutes les données sur les performances des élèves que vous voulez, toutes les données médicales que vous voulez, mais si vous ne savez pas quels modèles sont utilisés pour générer des résultats, vous n'aurez pas beaucoup de réponses*.

Le modèle va donc permettre de documenter la donnée, encore une fois aider à lui retirer toute ambiguïté de sens, et donc faciliter son partage et sa reproduction ou réutilisation. Le rôle des métadonnées est là essentiel au niveau du modèle et de ses données.

Par ailleurs, un modèle ne doit pas être figé, et il doit permettre d'interagir avec lui, offrir des paramètres modifiables, et ce dans l'objectif d'une amélioration constante : Pour Cathy O'Neil [40, O'NEIL], il est nécessaire que nous puissions *jouer avec* les modèles, et donc qu'ils soient ouverts.

4.4. Benchmark des modèles français et américains

Nous avons recensé et balayé un certain nombre de modèles de données sur les médicaments ou la chimie, pour avoir un aperçu de l'état de l'art dans ce domaine, et ce qui pouvait être appliqué ou adapté à l'ontologie française des médicaments. Nous en avons retenu 3, que nous allons détailler ci-après, les plus pertinentes par rapport à notre besoin, et aux rapprochements que nous pouvions effectuer avec.

[9999] https://en.wikipedia.org/wiki/Entity%E2%80%93attribute%E2%80%93value_model

4.4.1. MedDRA (*Medical Dictionary for Regulatory Activities*)

En 1994, une terminologie médicale standardisée est adoptée pour faciliter, sur le plan international, le partage d'informations réglementaires concernant les produits médicaux à usage humain. Il s'agit de MedDRA [41], outil ICH (Conférence Internationale sur l'Harmonisation), disponible librement lors de l'enregistrement, de la documentation de la surveillance des produits médicaux, à la fois avant et après qu'un produit ait été autorisé sur le marché. Les produits couverts par le domaine d'application de MedDRA comprennent les produits pharmaceutiques, les produits biologiques, les vaccins et les produits associant un dispositif et un médicament.

Le travail de création de ce dictionnaire international a été effectué conjointement entre la FDA (Etats-Unis), l'EMA[100] (Europe) et le Ministère de la Santé, du travail et du Bien- être (Japon), à partir de vocabulaires validés au niveau national, puis agrégées. Il est révisé tous les six mois en prenant en compte les suggestions de changement proposées par les utilisateurs.

L'abonnement à MedDRA est disponible gratuitement pour toutes les autorités gouvernementales à travers le monde, mais payant pour les entreprises, suivant une échelle ascendante liée au chiffre d'affaires. Les universitaires et professionnels de santé peuvent également avoir accès à MedDRA gratuitement auprès de la MSSO (*Maintenance and Support Services Organization*)[101].

MedDRA est construit selon une hiérarchie de 26 classes de haut niveau permettant de définir et traduire les renseignements médicaux selon 5 niveaux de précision :

1. Classe Organes (*System Organ Class* SOC): il s'agit du plus haut niveau de la hiérarchie qui offre le plus large concept de regroupement par :
   - étiologie (*Infections and infestations*) ;
   - site d'atteinte (*Gastrointesinal disorders*) ;
   - action (*surgical and medical procedures*).

2. Termes de haut niveau (*High Level Term* HLT): ils regroupent :
   - le terme préféré (*Preferred Terms* PT) décrivant un concept médical unique. Il doit être le moins ambigu et le plus spécifique et auto-descriptif possible. Un PT doit être relié à au moins un SOC ;
   - les groupes de termes de haut niveau (*High Level Group Term* HLGT) rassemblent plusieurs HLT ayant un lien anatomique, physiopathologique, étiologique ou fonctionnel ;
   - le terme de bas niveau (*Low Level Terms* LLT), niveau préférentiel de codage, il couvre en effet le plus grand nombre d'entrées possibles. Chaque LLT est relié à un seul PT.

La figure 6 illustre par des exemples chacun de ces types de niveaux.

[100] https://eudravigilance.ema.europa.eu/human/evMpd01.asp
[101] http://www.meddra.org/about-meddra/organisation/msso

Type de terme	Exemple de terme	Nombre de termes dans MedDRA
System Organ Class (SOC)	Troubles du foie et des voies biliaires	26
High Level Group Term (HLGT)	Maladies hépatobiliaires	332
High Level Term (HLT)	Hépatite	1 682
Preferred Term (PT)	Adipose douloureuse de Dercum	17 867
Low Level Terms (LLT)	Syndrome abdominal aigu	56 580

Figure 6 : Exemples et nombres de termes MedDRA selon le type de terme [41].

Le modèle MedDRA étant très spécifique et difficilement adaptable à ce qui était souhaité pour l'ontologie française des médicaments, nous nous sommes simplement demandé si nous pouvions nous inspirer de certaines annotations ou propriétés pour enrichir notre propre modèle.

4.4.2. RxNorm[102] et le méta-thésaurus UMLS

RxNorm est un catalogue permettant de standardiser les noms de médicaments cliniques et leurs dispositifs d'administration aux États-Unis, afin de rendre les systèmes interopérables, indépendamment de la compatibilité logicielle et matérielle. RxNorm est intégré au sein du méta-thésaurus UMLS, permettant l'alignement avec d'autres terminologies, et d'élargir ainsi cette notion d'interopérabilité.

UMLS est un système puissant pour lier des informations sur la santé, les termes médicaux, les noms de médicaments, et les codes de facturation à travers différents systèmes informatiques. Considéré comme la plus grande base de données terminologique, le méta-thésaurus constitue entre autres, la base unifiée des concepts médicaux. Il comprend des synonymes, des variations lexicales et des concepts associés : 2 millions de concepts avec plus de 7 millions de termes de 140 terminologies biomédicales.

Le contrat de licence[103] requise permet d'obtenir les différentes versions des fichiers RxNorm, nécessite de disposer d'un compte *UMLS Terminology Services* (UTS), et donc de se conformer aux conditions établies.

Les règles du méta-thésaurus UMLS consistent à :

1. regrouper sous un même concept les différents termes qui l'expriment. On y trouve quatre niveaux :
   - *Concept Unique Identifiers* (CUI) : il regroupe tous les termes qui partagent le même sens. Par exemple, les termes « Froid (Cold) » issu de MeSH[104] et « température froide (cold temperature) » issu de CISP[105] doivent être regroupés dans un même concept UMLS. A chaque concept correspond un identifiant unique CUI, le sens d'un concept étant par essence conservé d'un système à un autre, des concepts de systèmes différents peuvent donc partager le même CUI.
   - *Lexical Unique Identifiers* (LUI) : il rassemble toutes les variations lexicales pour un terme donné. Cependant, ce regroupement est appliqué seulement pour les termes en anglais.

[102] https://www.nlm.nih.gov/research/umls/rxnorm/
[103] https://uts.nlm.nih.gov/license.html
[104] http://mesh.inserm.fr/mesh/
[105] https://fr.wikipedia.org/wiki/Classification_internationale_des_soins_primaires_(CISP)

- *String Unique Identifiers* (SUI) : chaque nom de concept ou terme dans chaque langue est associé à un identifiant unique SUI. De plus, chaque variation dans le nombre de caractères, la ponctuation est considéré comme des termes différents ce qui implique des SUI différents. Par exemple, les deux termes « Adrenal Gland Diseases » (maladies de la glande surrénale) et « Disease of adrenal gland » ont des SUI différents. Alors que, les termes « Cold » du MeSH et « Cold » de la SNOMED ont un même SUI.
- *Atom Unique Identifiers* (AUI) : chaque occurrence d'un terme dans chaque terminologie est associée à un unique identifiant AUI. Par exemple, les deux mêmes termes « Cold » du MeSH et « Cold » de la SNOMED ont des AUI différents.

2. inclure toutes les hiérarchies dans le méta-thésaurus si les mêmes concepts appartiennent à différents contextes hiérarchiques ;
3. inclure également les différentes relations entre concepts de différentes terminologies.

À chaque concept correspond : une définition, un terme préférentiel, éventuellement des synonymes, des variantes lexicales, un ou plusieurs types sémantiques et un identifiant unique « Concept Unique Identifier » (CUI).
Plusieurs relations existent entre différents concepts. Ce sont des relations qui proviennent des terminologies d'origine.

L'UMLS repose sur une base de données relationnelle répartie en 11 entités pour les SOC et 13 entités pour les informations du méta-thésaurus. Il existe également un certain nombre d'index pour chaque langue.

Concernant le modèle en lui-même, il se découpe de la manière suivante (Tableau 9) :

Catégorie	**Abréviation**	**Instance**
Ingredient	IN	*Cetirizine*
Brand Name	BN	*Zyrtec*
Clinical drug component	SCDC	*Cetirizine 5 MG*
Branded drug component	SBDC	*Cetirizine 5 MG [Zyrtec]*
Clinical drug name	SCD	*Cetirizine 5 MG Oral Tablet*
Branded drug	SBD	*Zyrtec 5 MG Oral Tablet*
Clinical drug form	SCDF	*Cetirizine Oral Tablet*
Branded drug form	SBDF	*Cetirizine Oral Tablet [Zyrtec]*

Tableau 9 : Principales catégories RxNorm.

Pour chaque concept « *brand name* », il existe 1 ou plusieurs « *branded drug components* », « *branded drugs* » et « *branded drug forms* ».

Chaque « *ingredient* » est associé à 1 ou plusieurs « *clinical drug components* », « *clinical drugs name* » et « *clinical drug forms* ».

Concernant les relations entre concepts, il s'agit des suivantes :

- *tradename_of / has_tradename* (entre ingrédient et marque) ;
- *form_of / has_form ;*
- *precise_ingredient_of / has_precise_ingredient;*
- *ingredient_of / has_ingredient ;*
- *consists_of / constitutes ;*
- *dose_form_of / has_dose_form;*
- *isa / inverse_isa.*

Toutes les entités RxNorm sont liées entre elles grâce à ces relations : par exemple, le concept « *brand name* » est relié au concept « *branded drug component* » par la relation *ingredient_of* et *has_ingredient*, la relation inverse. Toutes les instances de relation sont bidirectionnelles.

Voici deux exemples d'instances de relations permettant de les illustrer :

branded drug → branded drug component
Zyrtec 5 MG Oral Tablet **consists_of** *Cetirizine 5 MG [Zyrtec]*
clinical drug component → clinical drug
Cetirizine 5 MG **constitutes** *Cetirizine 5 MG Oral Tablet*

Par ailleurs, pour chaque médicament à 1 seul ingrédient, il y a une correspondance stricte : à chaque « *brand name* » correspond un « *ingredient* ».

Pour un médicament à plusieurs ingrédients, cela fonctionne de la manière suivante : chaque multi ingrédient « *branded drug* », « *branded drug component* » and « *branded drug form* » est lié à un seul « *brand name* », en revanche, les multi-ingrédients « *clinical drugs* » et « *clinical drug forms* » sont liés à plusieurs « *ingredients* » et « *clinical drug components* ». De plus, un multi-ingrédients « *brand name* » est lié à plusieurs « *ingredients* », et un multi-ingredient « *branded drug component* » ou « *branded drug* » sont liés à plusieurs « *clinical drug components* ».

La figure 7 illustre ces considérations relationnelles.

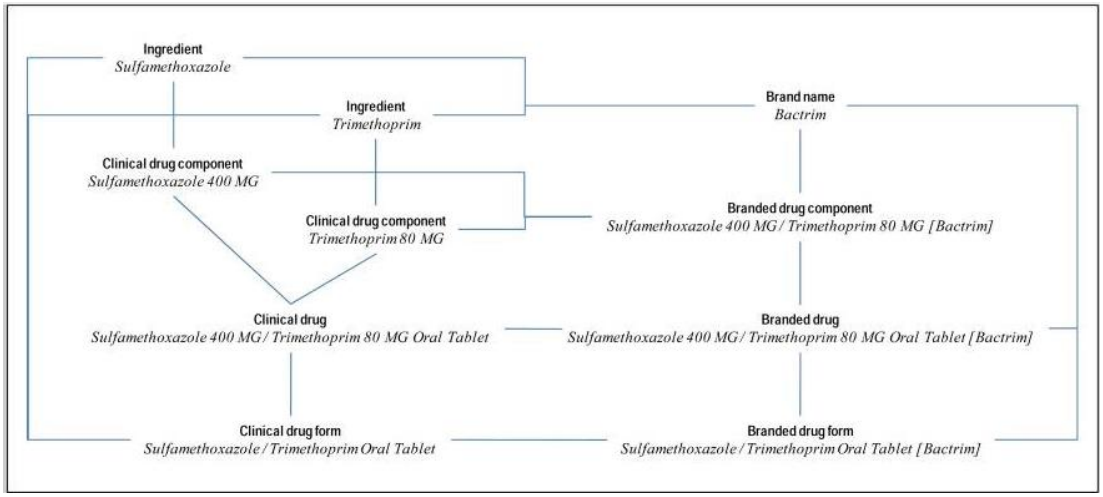


Figure 7 : Représentation multi-ingrédient dans RxNorm [42].

Nous reviendrons sur RxNorm plus loin dans ce mémoire, lors de la question du choix de modèle à suivre ou pas pour l'ontologie des médicaments, car un travail de recherche plus poussé a été réalisé à cette occasion.

4.4.3. La classification ATC

La classification ATC (Anatomique Thérapeutique et Chimique)[106] est utilisée pour classer les médicaments. Elle présente une classification multi-niveaux ou le même concept (médicament) peut être classé à différents endroits (doublons). De plus, l'ATC recense des substances actives qui sont parfois des noms de médicaments, et parfois contenues en combinaison dans un médicament avec une autre substance active. Parfois en combinaison avec un autre produit, parfois non.

Les médicaments sont divisés en plusieurs groupes selon l'organe ou le système sur lequel ils agissent et/ou leurs caractéristiques thérapeutiques et chimiques. Le code ATC a la forme générale *LCCLLCC* (où *L* représente une lettre et *C* un chiffre).

Dans ce système, les médicaments sont classés en groupes à cinq niveaux différents (voir figure 8) :

- groupe anatomique (un caractère alphabétique) ;
- groupe thérapeutique principal (deux caractères numériques) ;
- sous-groupe thérapeutique/ pharmacologique (un caractère alphabétique) ;
- sous-groupe chimique/ thérapeutique/ pharmacologique (un caractère alphabétique) ;
- sous-groupe pour la substance chimique : le principe actif individuel ou l'association de principes actifs (deux caractères numériques) ;
  Puis DCI (Dénominations Communes Internationales) et la substance, quand elle existe.

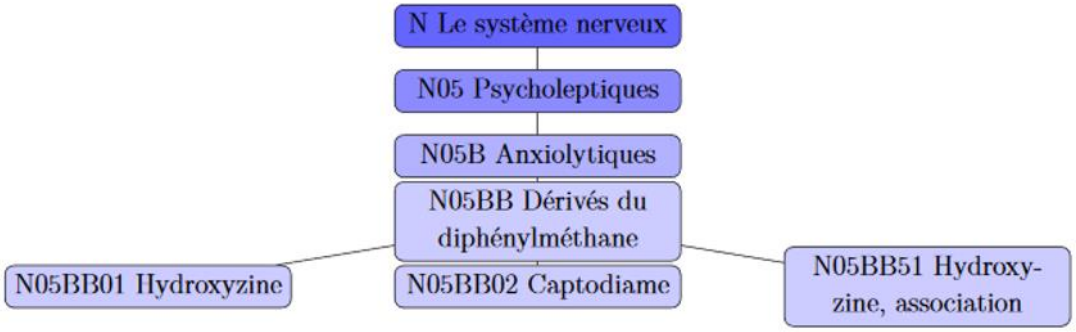


Figure 8 : Les 5 niveaux de l'ATC [38, CHOQUET].

Le choix a été fait pour l'ontologie française des médicaments de suivre la structure de l'ATC pour regrouper les spécialités et présentations de médicaments autorisés sur le marché français. On y revient plus loin dans ce mémoire, dans la partie réalisation du stage.

4.5. Analyse des sources de données françaises autour du médicament

La donnée n'est qu'un matériau brut de base. Pourtant sa collecte et son traitement peuvent conduire au savoir, correspondant à une *vérité provisoire scientifiquement acceptée* [15, PERES] et à la connaissance c'est-à-dire la manière dont chacun en tire parti. Une donnée est, par définition, une information numérique

[106] https://www.vidal.fr/classifications/atc/

ou alphanumérique codée, lisible par une machine, figée et transmissible avec un cycle de vie qui lui est propre (création ou récupération, traitement et communication, conservation).

En termes de volumes, on ne peut que constater la sur-mulitiplication des données, liée à « l'infobésité » ou surcharge informationnelle [107]: depuis les origines de l'humanité et jusqu'en 2003, l'humanité avait produit 5 exaoctets de données numériques, soit 5 milliards de milliards d'octets. En 2010, il suffisait de 2 jours pour produire le même volume.

Les métadonnées participent à l'amélioration de la mise à disposition de ces quantités d'informations, en les structurant : elles les décrivent, les expliquent, les localisent et en facilitent donc la découverte et l'utilisation. Ainsi, une donnée n'est pas quelque chose de naturel, mais de construit. Comme nous l'avons vu précédemment, le processus de fabrication est aussi important que la donnée en elle-même.

Pour qu'une source de données remplisse correctement sa fonction d'accessibilité et de disponibilité des données dont elle dispose, plusieurs choix sont offerts. La mise à disposition peut s'effectuer, par exemple, via un lien qui mène directement vers le jeu de données, quel que soit son format, ou vers le fichier qui contient l'ensemble des liens présents dans le catalogue. Un lien unitaire sur un portail, vers chaque format de données téléchargeable, est également réalisable ; ou encore des liens vers des catalogues entiers. On trouve aussi des types de liens indirects, qui peuvent être difficiles à repérer dans une page internet (exemple de insee.fr), ou mis à disposition sur des sites externes (exemple de Data.gouv.fr), via un formulaire. Enfin, les API (*Application Programming Interface*) deviennent de plus en plus répandues, même dans le monde de l'*Open Data*. Une interface permet d'accéder aux données, en interrogeant les bases de données via différentes requêtes.

L'*Open Data* est un mouvement visant à généraliser la mise à disposition des internautes des informations détenues par le secteur public ; un mouvement qui considère que l'ouverture et le partage des données publiques est un bien commun. Nous avons donc cherché à nous servir de ces ressources française ouvertes et disponibles.

4.5.1. L'ANSM et Data.gouv.fr

Depuis avril 2012, l'ANSM[108] se substitue à l'AFSSAPS dont elle reprend les missions, les droits et les obligations. Elle évalue la sécurité d'emploi, l'efficacité et la qualité des produits de santé, et assure également la surveillance des effets indésirables liés à leur utilisation. Elle exerce des activités de contrôle en laboratoire et sur les sites de fabrication et de recherche, avec des moyens renforcés pour assurer la surveillance et l'évaluation des produits de santé. Enfin, elle mène des actions d'information auprès des professionnels de santé et du public pour améliorer le bon usage des produits de santé.

Le Répertoire des spécialités pharmaceutiques avec Autorisation de Mise sur le Marché (AMM) [46], réalisé par l'ANSM, permet d'accéder à une partie de l'information officielle sur les spécialités pharmaceutiques ayant obtenu une AMM, qu'elles soient commercialisées ou non. La recherche se fait par dénomination, substance active ou laboratoire/titulaire. Il est possible de limiter la recherche sur une

[107] https://fr.wikipedia.org/wiki/Surcharge_informationnelle
[108] http://ansm.sante.fr/

période d'AMM, sur l'état de l'AMM (valide, abrogation, suspension, retrait), ou sur les médicaments uniquement commercialisés. On obtient une fiche de synthèse sur le médicament recherché, contenant un ensemble d'informations pharmaceutiques et réglementaires, et comprenant la dénomination, la composition en substances actives, le nom du ou des titulaires de l'AMM, la date de l'AMM, son statut, ses différentes présentations, la date de déclaration de leur commercialisation et la notice patient.

L'ANSM dispose d'une base de données pour collecter, évaluer et analyser les cas de pharmacovigilance [47, BOIDIN]. Cet outil est développé et exploité en interne. Cette base de données rencontre plusieurs limites :

- Au niveau du contenu de l'information, on remarque un manque d'harmonisation des pratiques de codage et l'absence d'un référentiel universel des médicaments.
- D'un point de vue technologique, cette base de données rencontre des problèmes de connexion et le système est lent. Pour des raisons techniques, la France ne communique plus ses données de pharmacovigilance au centre de surveillance internationale des médicaments d'UPPSALA (UMC) depuis 2007.
- Enfin, l'ANSM n'utilise pas de dispositif systématique de détection de signaux alors que la plupart des autres pays le font.

Au niveau du portail gouvernemental porté par Etalab, Data.gouv.fr, il semble que les données qui s'y trouvent ne sont pas toujours à jour et qu'elles sont parfois manifestement erronées [2, MESZAROS *et al.*]. Par ailleurs, la question sur la quantité réelle des données présentes sur le portail se pose, et impose une méfiance vis-à-vis des valeurs fournies. Il faut se rendre alors à chaque fois directement sur le site des producteurs des données pour en savoir plus.

En termes de contenus, le portail est loin de donner le meilleur exemple, certains jeux de données souffrant de problèmes de qualité, *comme des budgets publiés en PDF, des agrégations rendant l'interprétation impossible, des liens html présentés dans la liste de données ouvertes, des tableaux et trombinoscope en .doc, des données en vrac dans les fichiers* [109]. Même si ce constat date de 2012, il est toujours d'actualité à certains égards, sur l'utilisation du format PDF, par exemple, comme nous le verrons plus loin dans la partie du mémoire concernant le stage.

Le portail Data.gouv.fr, en tant que fédérateur de données issues de divers producteurs, doit en effet constamment affronter des difficultés, telles que la veille constante des sources primaires, le suivi de changement de dénominations des sites sources, la gestion des doublons, etc. Quoi qu'il en soit, et malgré certaines défaillances, ce genre de portails est une nécessité. Avec l'augmentation du nombre des acteurs et du volume des données, il sera de plus en plus difficile d'aller chercher les données à chaque fois chez les fournisseurs primaires.

4.5.2. La qualité au rendez-vous de l'*Open Data* ?

Ce qui ressort de l'ensemble des retours d'expérience sur le sujet (y compris le mien), divers problèmes existent ou sont à prendre en compte pour toute personne

[109] http://www.c-radar.com/blog/2012/11/26/lopen-data-est-tres-mal-estime-interview-de-claire-gallon-de-lassociation-libertic/

ou organisation souhaitant réutiliser des données ouvertes, qu'elles soient publiques ou privées.

Tout d'abord, la diversité des formats de distribution est un souci majeur au niveau de l'interopérabilité des données, puisqu'elles sont censées être non propriétaires et « *machine readable* » : on se retrouve encore trop souvent avec des fichiers XLS ou PDF, qui sont respectivement contraires à ces deux principes. D'autant plus que cela oblige le réutilisateur à transformer lui-même ces données « brutes » afin de les intégrer dans ses traitements. Le format CSV doit encore s'étendre pour répondre à cette problématique.

Ensuite, concernant la qualité intrinsèque des données, on a évoqué précédemment le fait qu'un travail important en amont de la diffusion était indispensable pour livrer des données complètes, sans erreurs, bien documentées et avec les métadonnées les plus exhaustives possibles et s'appuyant sur certains des référentiels existants. Beaucoup d'efforts en temps et en investissements, ainsi qu'une conduite du changement dans les organisations doivent encore être opérés pour parvenir à cet objectif.

Les moyens de distribution des données sont, quant à eux, loin d'être uniformes. Un réel effort est fait pour mettre en place, par exemple, des API, mais l'accès à des données reste assez fastidieux et dépendant des différentes interfaces. On manque encore de procédures automatiques permettant l'accès direct par la machine pour traiter les données, comme le monde de l'Open Data le requiert.

La pérennité des données n'est pas non plus au niveau attendu. On souhaite en effet des données « fraiches », à jour, ce qui n'est toujours le cas, car cela demande également un investissement au long cours de la part des producteurs de données, ou des passerelles de distribution. On rencontre encore des cas ou les données disparaissent tout simplement ou deviennent inaccessibles pour une raison quelconque. Cette insécurité quant à la pérennité des données est donc très préjudiciable pour l'ensemble de l'écosystème que les données ouvertes constituent.

Enfin, le manque de standards propres au monde des données ouvertes est un axe d'amélioration également pour permettre un meilleur encadrement et des normes appropriées.

Ce que l'on peut en conclure, et malgré les progrès accomplis par l'administration en termes de transparence, *la mission commune d'information du Sénat sur l'accès aux documents administratifs constate « l'inertie persistante » des administrations sur l'ouverture des données publiques et la qualité inégale des informations publiées* [2, MESZAROS *et al.*]. Les pratiques actuelles des principaux acteurs de l'*Open Data* doivent évoluer pour la mise à disposition des données et de leur documentation, car des intermédiaires sont encore trop souvent nécessaires pour donner l'accès direct attendu par le citoyen ou le réutilisateur. Ce qui altère également le caractère gratuit de l'*Open Data*.

5. Gouvernance et processus à mettre en œuvre

5.1. La gestion de projet

Tout au long d'un projet de réutilisation de données ouvertes, la question suivante doit rester présente à l'esprit : comment contribuer à l'interopérabilité technique, économique et politique de l'ouverture des données, tout en atteignant ses

propres objectifs fonctionnels et techniques ? Dans notre cas, s'ajoute le respect et la couverture du domaine qui nous intéresse, celui de la e-Santé.

Comme pour tout autre projet, le choix de la méthode mise en œuvre pour gérer chaque étape (budgétisation, spécifications fonctionnelles et techniques, planification, réalisation, tests / recettes, documentation, mise en place, maintenance) est primordiale.

Issues de l'informatique, les méthodes « agiles » sont de plus en plus adoptées car elles préconisent de créer des prototypes rapidement et de fonctionner par itérations, en commençant par de petites entités. Par exemple, pour le portail data.gov.uk [36, DHRZANOWSKI], les britanniques ont commencé par convaincre tous les ministères, et ont ensuite avancé administration par administration, avec des résultats visibles dès le début. Là où des réticences existent, on essaye d'expliquer et de convaincre plutôt que d'imposer. Cette méthode fonctionne, car elle implique les différentes parties prenantes, et tient compte des remontées éventuelles avant de poursuivre la démarche ; ce qui évite de passer à côté des attentes et des besoins réels.

Concernant le processus de mise en place de projets sur les données ouvertes et liées, la Revue Electronique Suisse de Science de l'Information (RESSI) a publié une étude décrivant les étapes de leur mise en œuvre : celles-ci ont été reprises et résumées dans le tableau 10 par Sébastien Chambonnet dans son mémoire [48, CHAMBONNET].

Étape 1 : Revue de la littérature	• Analyse des applications déjà existantes et des pratiques d'autres institutions en matière de web sémantique • Se tenir au courant des évolutions techniques du domaine ainsi que des nouveaux standards et modèles
Étape 2 : Analyse des données	1. Quelles données publier ? Au niveau des jeux de données entiers ou au niveau plus fin des champs de données au sein d'une base de ressources. 2. Critères : pertinence qualité, quantité, normalisation
Étape 3 : Modélisation	1. Choisir un modèle de données 2. L'appliquer à ses propres données en identifiant des équivalences : quelles données correspondent à quelles entités du modèle ? 3. Définir la forme des URI de chaque ressource.
Étape 4 : Mapping	1. Choix des vocabulaires. Ex : Dublin Core, FOAF, SKOS… 2. Établir des règles de conversion pour chaque élément de donnée (chaque champ).
Étape 5 : Liens externes	1. Génération de liens avec des référentiels du web sémantique. Ex : VIAF pour les personnes, DBPedia, etc.
Étape 6 : Transformation	2. Formulation en langage informatique des règles établies et application sur les données.
Étape 7 : Contrôle qualité	3. Mesure de la qualité des données sur des échantillons représentatifs des données.
Étape 8 : Publication	4. Publication sur un serveur ou un entrepôt de données, avec licence ouverte.

Tableau 10 : Etapes de mise en place d'un projet d'ouverture de données [48, CHAMBONNET].

En parallèle, concernant les ontologies qui nous intéressent plus particulièrement, il est indispensable de considérer chaque étape de son cycle de vie pour la rendre pérenne (et pas uniquement la phase de conception, petite part de l'ensemble au final) : voir figure 9.

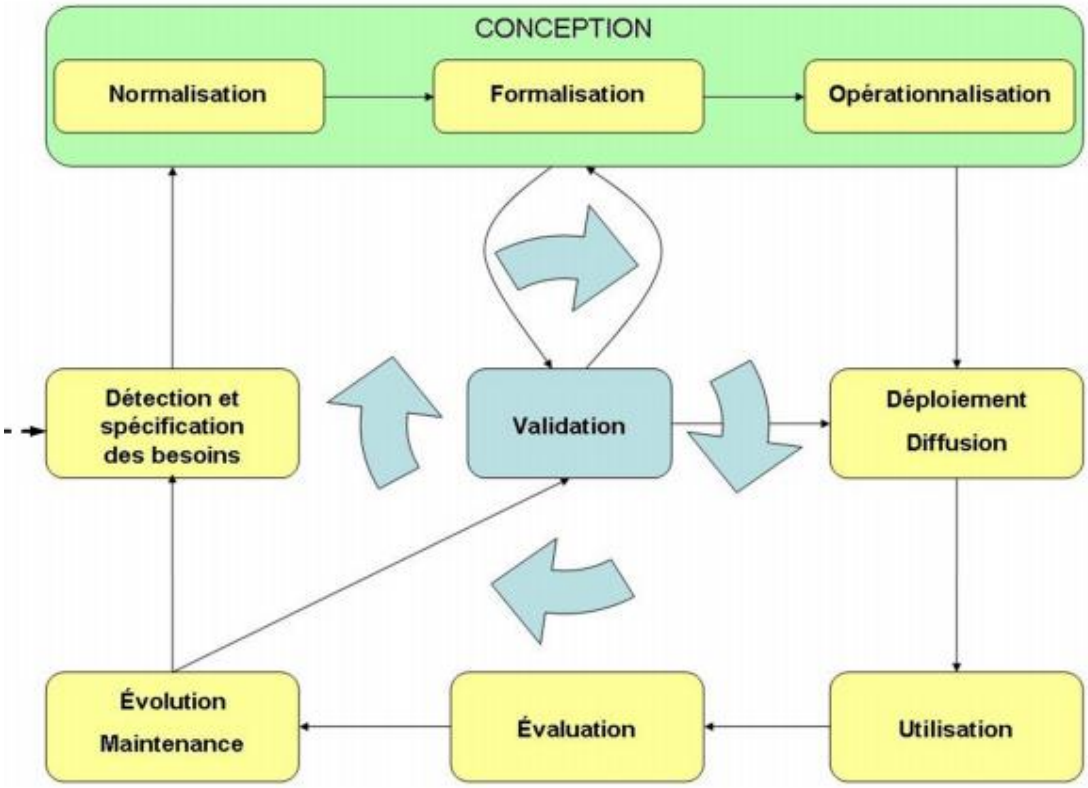


Figure 9 : Le cycle de vie d'une ontologie [49, BANEYX et al.*].*

5.2. Les standards et normes

La standardisation intervient à deux niveaux, tout aussi importants l'un que l'autre dans le processus de mise à disposition d'applications ouvertes : l'intégration de données et l'interopérabilité. Lorsqu'on intègre des données, on cherche à agréger des informations de sources hétérogènes, que l'on souhaite ensuite rendre intelligibles par d'autres systèmes avec qui on communique. Il s'agit donc standardiser les méthodes de structuration des données (modèles d'information), voir standardiser les données elles-mêmes (vocabulaires), mais aussi les méthodes de transport et d'accès aux données (XML, HTTP) [38, CHOQUET].

Dans le domaine de la santé, un enjeu économique de l'interopérabilité est, par exemple, de diminuer le nombre de tests diagnostic réalisés pour un citoyen malade en déplacement, ce qui est de plus en plus courant.

Le standard permet d'aider à cette interopérabilité, et donc à la qualité des données, mais il peut avoir tendance à appauvrir le contenu informationnel qu'il code. En effet, son essence étant d'uniformiser, il doit s'adapter à tout le monde, et dans le cas de l'information biomédicale, c'est chose quasiment impossible. Il faut également être conscient qu'au travers des différentes versions du même standard, l'interopérabilité avec lui-même peut se perdre (par exemple, ICD-9 et ICD-10).

L'objectif est de résoudre les obstacles suivants qui freinent l'exploitation des données et en l'occurrence en e-Santé :

- manque d'interopérabilité technique: l'intégration de bases de données propriétaires et hétérogènes reste une étape conséquente ;
- manque d'interopérabilité sémantique: les différentes données de même sens doivent pouvoir être analysées conjointement ;

- manque de transparence concernant la provenance des données: cette dernière doit accompagner toute analyse de données ;
- pauvre qualité des données primaires: les données réelles cliniques sont, de manière intrinsèque, de mauvaise qualité (manquantes, erreurs et bruit) ;
- multiplicité des formats et des types de données: les données biomédicales en sont un exemple significatif ;
- la confidentialité des données: la capacité d'agrégation depuis différentes sources de données ne garantit pas la vie privée des patients.

L'apport de standards dans chacune de ses problématiques est un moyen d'y apporter des améliorations, et est actuellement surtout attendu au niveau de l'interprétation du sens des données par la machine. Ainsi, elle pourra contrôler une mauvaise saisie d'un utilisateur, savoir qu'une donnée est mal codée, que deux codes différents sont identiques dans certains contextes,… En résumé, faire en sorte qu'elle puisse raisonner.

Le Web sémantique apporte des réponses grâce, notamment à l'apport de la gestion des métadonnées mais surtout de la sémantique des données. Cependant, un travail de standardisation des ressources existantes structurées doit se poursuivre pour utiliser des références sémantiques partagées et fiables, telles qu'on peut les retrouver dans les terminologies ou les systèmes de classification.

En termes de normes, on peut noter que l'ISO distingue les terminologies (listes de termes), les thésaurus (index et synonymes), les classifications (avec des relations génériques) ou les vocabulaires (avec des définitions) et les ontologies (ISO TS17117[110]) utilisées dans le domaine de la santé pour représenter une formalisation du domaine (concepts) et des termes d'un système d'information clinique.

L'AFNOR a mené une enquête qualitative [50, HUOT *et al.*] pour cerner les impacts et besoins fonctionnels en termes de normalisation, et les résultats qui en ressortent (voir figure 10) ont été regroupés par activités liées aux processus de collecte, transformation et analyse (associée à la restitution, la représentation, et la visualisation de l'information après traitement).

110 http://www.iso.org/iso/fr/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32883

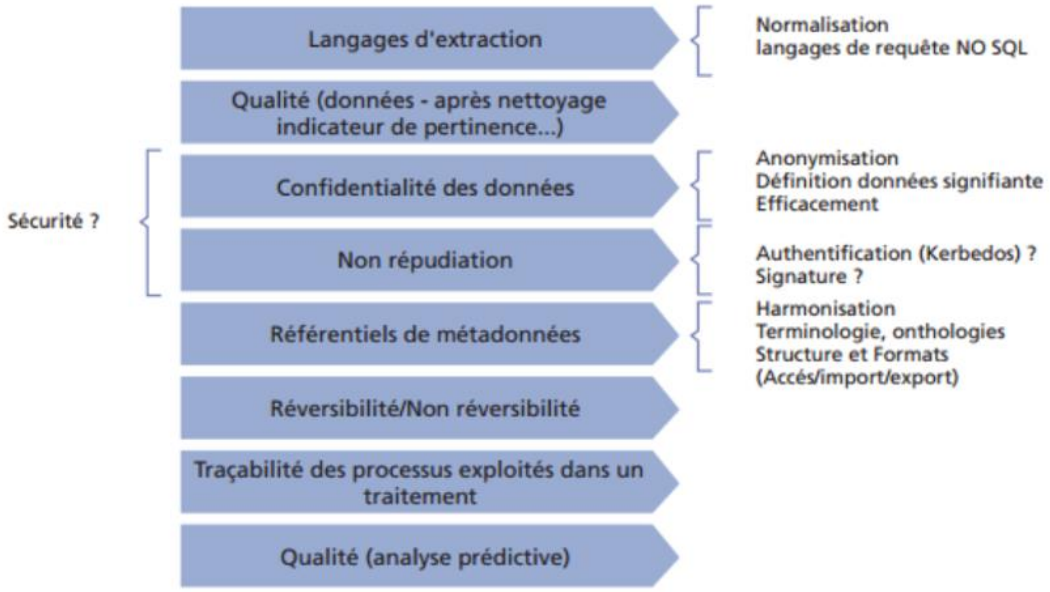


Figure 10 : Besoins fonctionnels associés aux activités de collecte, traitement et analyse de contenus [50, HUOT et al*.].*

Pour être en capacité à traiter l'ensemble de l'information se posent des enjeux en matière de qualité des données. Comme nous l'avons évoqué précédemment, dès la collecte des données hétérogènes, on se trouve face à des problèmes récurrents, tels que des données manquantes qu'il va falloir interpréter, des affichages décimaux différents, des erreurs diverses,… Par la suite, lors du traitement, il va être question de définir les référentiels à modéliser pour rapprocher les sources de données, les agréger, les nettoyer (*data cleaning*) et les catégoriser. C'est également à ce moment-là que doit intervenir la traçabilité des opérations et son historisation, par la documentation de ce qui a été accompli.

Cet ensemble participe à la capacité de maintenance des systèmes mis en place qui doit pouvoir s'appuyer sur des standards et normes permettant d'homogénéiser à la fois les données et les métadonnées. Par exemple, le fait que la donnée devienne un *objet d'échange*, impose de connaitre sa provenance, son propriétaire, son circuit de création, transformation, diffusion, mais aussi les droits qui y sont associés, d'autant plus si elle a acquis de la valeur ajoutée dans son parcours. Or, trop souvent, la question de l'origine ne se pose pas suffisamment pour ceux qui exploitent des ressources partagées, ce qui finira par mener à des situations instables.

Il apparaît donc important de définir des méthodologies pour qualifier la qualité de la source d'une part, et la confiance dans l'information d'autre part. La normalisation peut apporter des solutions de référence, et faciliter les relations entre les différents acteurs, à l'image des normes de systèmes d'enregistrement que développe le comité technique ISO TC 46[111] sur la documentation (ISAN, ISBN, etc.).

Les conditions d'utilisation des données et des licences à gérer sont aussi vouées à être encadrées de manière plus normative, pour éviter d'être face à des règles différentes en la matière (voire opposées) de la part des fournisseurs des

[111] http://www.iso.org/iso/fr/iso_technical_committee.html?commid=48750

données que l'on souhaite intégrer. Mais aussi de savoir à son tour quel droit d'accès accorder pour les données obtenues en sortie.

Si l'on s'intéresse maintenant de plus près aux ontologies, et à la volonté d'harmonisation et de référencement que leur développement suscite, on peut citer l'initiative du projet LOV (*Linked Open Vocabulary*)[112] lancé en 2011. Il s'agit d'un dispositif *Open Source*, actif depuis 2012 et reconnu au niveau international, permettant de recenser les ontologies de référence[113] (plus de 450 en 2015). Il se définit donc comme un catalogue avec des fonctions supplémentaires d'évaluation, de documentation, et de *versionning* de chaque ressource référencée, doté également d'un moteur de recherche. Les données issues du LOV *sont elles-mêmes sémantisées,* et mises à jour quotidiennement, ce qui le rend donc garant de l'interopérabilité des données liées référencées. Il lui manque aujourd'hui de ne pas être reconnu dans le cadre de l'ISO, ne faisant l'objet d'aucune norme.

5.3. Evaluation de la qualité

Une fois réalisé suivant le respect de son cahier des charges, et des standards en application, un modèle se doit d'être évalué, pour vérifier qu'il répond aux attentes et besoins définis.

Un modèle doit en effet offrir une qualité syntaxique par rapport au langage utilisé, une qualité sémantique par rapport au domaine représenté (termes employés exhaustifs, justes, à jour et évolutifs), et une qualité pragmatique par rapport à son utilisation en contexte réel (être compréhensible grâce à la documentation et au nommage, et répondre au cas d'usage) [51, DUPUY-CHESSA *et al.*].

Évaluer des modèles n'est pas facile à mettre en œuvre, car tout n'est pas mesurable directement, et requiert la participation de différentes parties prenantes, tels que les experts du domaine métier, les clients finaux, le chef de projet, les développeurs. Leur implication doit avoir lieu dès le début et tout au long de la chaine de fabrication pour prendre toute son ampleur au moment de l'évaluation puis de son interprétation avec des conséquences pour la prise de décision.

L'évaluation de la qualité des modèles peut s'appuyer sur les retours d'expérience de groupes d'utilisateurs pilotes: les expérimentations peuvent servir à évaluer, mais aussi à co-construire avec les utilisateurs ou à explorer de nouvelles solutions. D'autre part, l'insuffisance d'outils pour automatiser ou systématiser le processus d'évaluation, combinant si possible un ensemble de méthodes à ce sujet, conduit à un processus d'évaluation souvent fastidieux et couteux. L'ingénierie des modèles mérite d'être encore développée à ce niveau pour élaborer des techniques d'évaluation plus avancées et plus nombreuses, permettant d'obtenir les indicateurs nécessaires.

Au niveau des référentiels sémantiques, *la validation sert à s'assurer que l'ontologie modélise vraiment le monde réel (domaine) […]* et plus précisément les *connaissances que nous avons sur le monde […] pour lequel le modèle a été créé. Le but est de prouver la conformité du modèle du monde (s'il existe et est connu) à partir du monde modélisé formellement* [49, BANEYX *et al.*]. L'ontologie doit faire la preuve qu'elle répond au cas d'usage attendu, de la manière la plus proche possible.

La vérification périodique de la hiérarchie des classes par les utilisateurs est essentielle, car elle porte l'ontologie et en est le cœur ; elle doit prouver sa pérennité,

[112] http://lov.okfn.org/dataset/lov/
[113] https://liris.cnrs.fr/QetR2012/site/wp-content/uploads/2012/02/qetr2012_1.pdf

fiabilité et stabilité, tout en offrant une évolutivité. Particulièrement dans le domaine de la médecine qui offre constamment de nouvelles avancées, les mises à jour sont indispensables pour conserver l'intérêt de l'ontologie. L'automatisation des tâches de fabrication permet de limiter les coûts induits par le *versionning*, mais elle ne peut être généralisée, et des actions manuelles restent nécessaires, en particulier pour le nettoyage des données sources. Par conséquent, l'amélioration de la qualité des informations d'origine ne peut être que bénéfique, le gain serait directement mesurable, rien qu'en terme d'économie de temps et de charge.

Parmi les outils disponibles pour la validation d'ontologies, on trouve la méthode LOVMI (Les Ontologies Validées par Méthode Interactive) [52, RICHARD *et al.*], *qui s'appuie sur des outils déjà existants, et une collaboration entre les ontologues et les acteurs du domaine modélisé*.

5.4. Cadre de mise en application du stage : le LIMICS

Le LIMICS[114] est un laboratoire pluridisciplinaire en santé et en informatique, appartenant à l'INSERM, premier organisme européen de recherche biomédicale, créé en 1964, et représenté sur tout le territoire à travers 13 délégations régionales. Il est également rattaché au Pôle Modélisation et ingénierie de l'UPMC. Sa directrice est Marie-Christine JAULENT.

Il est constitué d'une cinquantaine de membres informaticiens, médecins et pharmaciens et a été créé le 1er janvier 2014, à partir de la fusion de deux équipes :

- l'ICS (Ingénierie des Connaissances en Santé) ;
- le LIM&BIO (Laboratoire d'Informatique Médicale et Bio-informatique).

Les membres du laboratoire sont majoritairement répartis sur deux sites principaux, le campus des cordeliers à Paris (aux côtés du Centre de Recherche des Cordeliers - CRC, et du Laboratoire d'imagerie biomédicale - LIB), et celui de l'UFR Santé, Médecine et Biologie humaine sur le campus de l'université Paris 13 à Bobigny. Quelques-uns se trouvent sur St-Etienne et Rouen.

Le LIMICS s'inscrit dans une démarche de traitement de l'information de santé, avec une approche sur 2 axes :

- les systèmes décisionnels pour la recherche médicale et la prise en charge des patients : collecte de données structurées normalisées (dossier patient informatisé ou constitution de cohortes et d'essais cliniques);
- l'ingénierie de l'information en santé correspondant aux méthodes et outils d'acquisition, à la modélisation et à la formalisation des connaissances pour la e-santé : ressources terminologiques ou ontologiques en santé, permettant le traitement de masses de données volumineuses (*Big Data*) ou encore de répondre aux nouveaux enjeux dans le domaine du Web Sémantique.

Mon stage est suivi par Jean Charlet, Chargé de missions APHP et Xavier Aimé, post-doctorant Ontologue. Tous deux travaillent sur la conception des méthodes et outils pour l'élaboration et l'évaluation de ressources terminologiques ou ontologiques en santé, afin de participer à l'élaboration des nouveaux référentiels internationaux d'informatique de santé (modèles d'information, terminologies et ontologies en santé).

[114] http://www.limics.fr/

En effet, la sémantique associée à l'information doit être élaborée, normalisée, standardisée, car l'intégration sémantique de ces données est un enjeu majeur pour la compréhension des mécanismes qui régulent notre santé. La sémantique associée à l'information (modèles de données et vocabulaires) doit pouvoir être publiée et alignée sur les standards pour pouvoir être partagée et mise en relation pour en extraire de nouvelles connaissances modulaires, interopérables et adaptables. En particulier, leurs recherches dans ce domaine se concentrent sur la construction d'ontologies.

Les outils informatiques utilisés au LIMICS sont :

- *Talend open source Data Integration* qui offre de puissantes fonctionnalités d'intégration de données, dans une architecture ouverte et évolutive. C'est le logiciel qui permet d'agréger diverses sources de données collectées en entrée, après une action curative (les nettoyer et les structurer). On obtient en sortie un fichier au format OWL, correspondant à l'ontologie finale. En entrée, on peut trouver soit des morceaux d'ontologies, soit des fichiers CSV ayant été préparés ;
- outils de fouille de texte (Biotex, SYNTEX, YATEA) ;
- *Protégé*, logiciel java créé à l'université Stanford, qui permet l'intégration de l'ontologie OWL, et est très populaire dans le domaine du Web Sémantique.

5.5. Projet de recherche du stage

Le projet auquel j'ai participé durant mon stage au LIMICS a consisté à créer une ontologie du médicament, en partant du travail déjà réalisé par Xavier Aimé depuis 2013.

Les sources de données concernées sont par exemple data.gouv.fr et l'ANSM, et les schémas pouvant servir de modèles comme RXNorm ou autres bases de données ouvertes, principalement dans le domaine de la chimie.

L'ontologie obtenue à mon arrivée dans le laboratoire par Xavier Aimé comportait plus de 50 000 concepts, mais présentait des imperfections dues aux sources souvent brutes, mal structurées, et qu'il s'agissait donc de formater et normaliser de manière pérenne, dans le processus (voir figure 11).

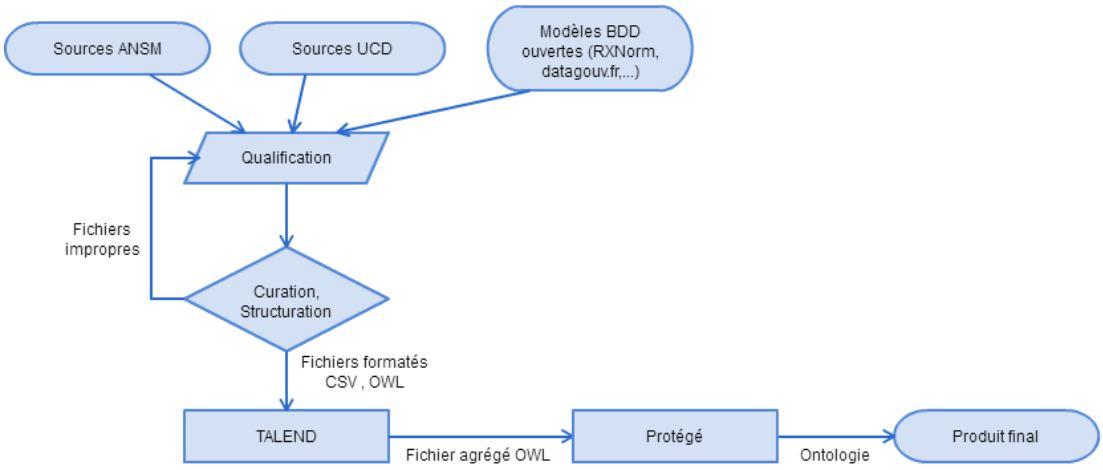


Figure 11 : Processus de création / mise à jour de l'ontologie des médicaments à mon arrivée au LIMICS.

Ma mission a consisté à poursuivre le travail de réalisation de l'ontologie française des médicaments à partir de ressources structurées en utilisant un ETL-*Extract-Transform-Load*[115] (outil Talend Open Studio – Data Integration), avec une étape préalable et indispensable d'analyse de la qualité des modèles et des données mises à disposition en *Open Data*, principalement l'ATC de l'OMS, RXNorms du NIH et les AMM (liste des autorisations de mise sur le marché) accessibles via la base du répertoire des spécialités pharmaceutiques de l'ANSM.

Ces données, une fois recueillies, ont dû être nettoyées, homogénéisées et intégrées pour obtenir la forme finale de l'ontologie au format OWL. L'idée générale est ainsi de générer une chaine de traitement le plus automatique possible afin de combiner un ensemble de données pouvant être de nature hétérogène dans une ontologie de domaine.

Les objectifs se sont donc découpés de la manière suivante :

- Passer en revue les concepts à revoir, au niveau de l'étape de curation :
  - concept Voies d'administration à normaliser ;
  - entêtes ANSM à intégrer dans le fichier de données source (elles se trouvent dans un fichier PDF !) ;
  - concept Conditionnements à nettoyer et structurer ;
  - concept Médicaments à améliorer sémantiquement en se basant sur la hiérarchie des codes ATC ;
  - concepts Conditions et Formes pharmaceutiques à améliorer ou compléter.
- Étudier les possibilités des modèles de données des bases de données ouvertes et liées (*Linked Open Data*), et repérer si elles sont interrogeables en SPARQL, ou si elles offrent des APIs.
- Réfléchir sur la problématique de l'amélioration de l'ensemble du processus, en particulier en ce qui concerne la qualité des sources de données, outre l'aspect technique de mise en forme des fichiers d'entrée CSV (on souhaite automatiser le plus possible, et donc intervenir par programmation sur la constitution des fichiers).
- Mettre l'ontologie obtenue à disposition dans un triplestore pour permettre son exploitation et sa réutilisation.

Pour débuter le stage, j'ai commencé par le travail de recherche concernant les différentes ressources exploitables, afin de repérer des modèles sur lesquels s'inspirer et peut-être compléter nos données. J'avais conscience que la partie curation serait certainement ardue, et me demanderait peut-être des développements ou l'utilisation de composants techniques spécifiques, voire des heures de nettoyage manuel. La variété des tâches à réaliser ou à étudier ont rendu la mission passionnante, et les résultats concrets auxquels j'ai pu aboutir m'ont donné une grande satisfaction personnelle.

6. Réalisations et constats concrets

L'élaboration d'une ontologie passe par plusieurs phases. Dans notre cas, elle s'est articulée autour des actions suivantes (les différents outils et notions sont développés par la suite) :

- créer ou mettre au point le vocabulaire constitué par l'ontologie : c'est le rôle de Protégé ;

[115] https://fr.wikipedia.org/wiki/Extract-transform-load

- agréger, intégrer et mettre les données sources au format RDF : l'outil Talend va permettre cette phase, même si on considère que celle-ci n'est pas encore optimisée, comme on le verra dans le détail du chapitre ;
- stocker les données dans un triplestore RDF : nous avons utilisé la base de données relationnelle paramétrée Virtuoso pour indexer des données en RDF (on peut citer également à titre d'exemple ARC[116], Oracle 11g[117], Sesame[118], 3store[119]).

L'exploitation des données (interrogations, exports,…) fera l'objet d'investigations ultérieures, très certainement via des requêtes SPARQL directement, ou pourquoi pas via des langages tels que Java (Jena[120]), PHP (RAP[121]), C (Redland[122]) ou encore python (RDFlib[123]).

6.1. Point sur les aspects techniques et logiciels (Protégé, Talend, Virtuoso)

Quelques explications et rappels du chapitre sur les technologies du Web sémantique, sont nécessaires pour bien comprendre les buts visés et les infrastructures utilisées.

Développé par le W3C [24], le *Resource Description Framework* (RDF) est un modèle de graphe destiné à décrire de façon formelle les ressources Web et leurs métadonnées, afin de permettre le traitement automatique de telles descriptions par raisonnement sémantique. Un document structuré en RDF est un ensemble de triplets (sujet, prédicat, objet) : Le sujet représente la ressource à décrire, le prédicat représente un type de propriété applicable à cette ressource, l'objet représente une donnée ou une autre ressource.

Le sujet, et l'objet dans le cas où c'est une ressource, peuvent être identifiés par une URI ou être des nœuds anonymes. Le prédicat est nécessairement identifié par une URI.

RDF est donc simplement une structure de données constituée de nœuds et organisée en graphe, c'est-à-dire un ensemble de triplets reliés les uns aux autres par les URI qu'ils ont en commun[124].

6.1.1. Protégé[125]

Protégé est un éditeur d'ontologies gratuit et *Open Source* qui permet de structurer une base de connaissances.

Concernant l'ontologie des médicaments, l'utilisation de Protégé a été effectuée à deux niveaux.

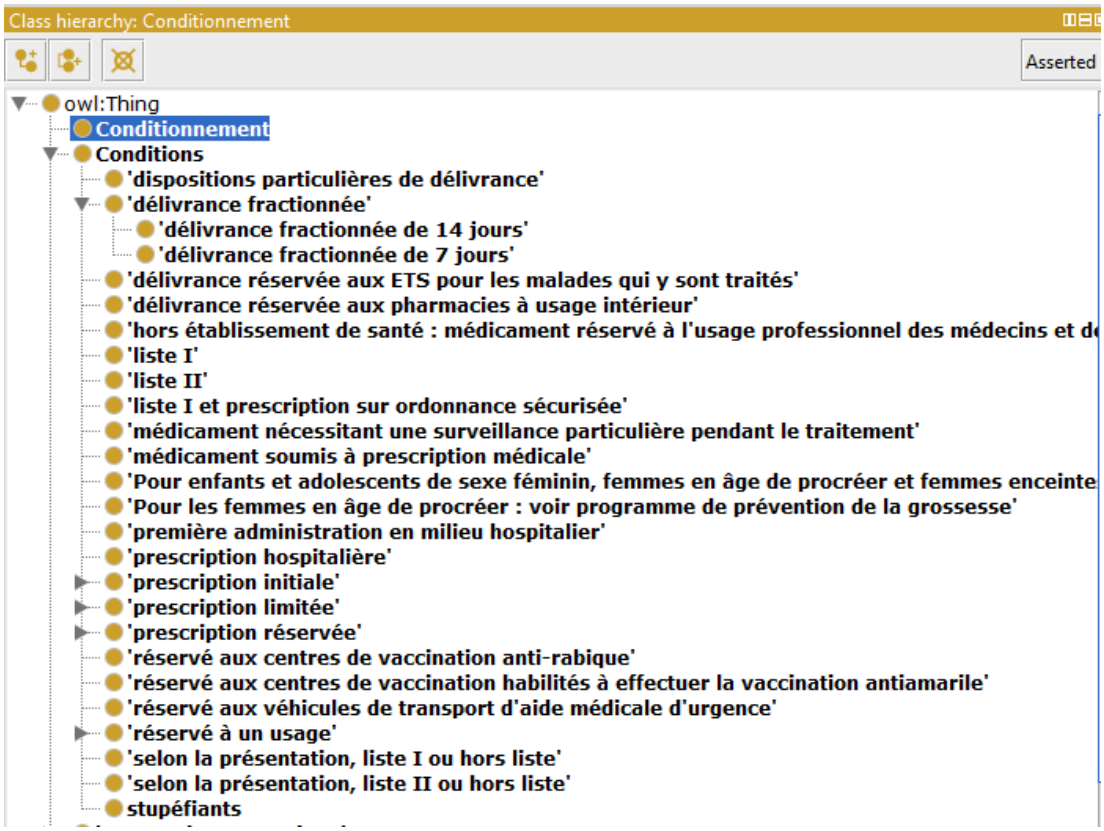
Nous avons tout d'abord créé ce que nous avons nommé l'ontologie « racine », c'est-à-dire, le squelette de notre terminologie. Le fichier OWL obtenu constitue

116 https://www.w3.org/2001/sw/wiki/ARC
117 https://www.cs.utexas.edu/~schrum2/DBTripleStore.pdf
118 https://www.w3.org/2001/sw/wiki/Sesame
119 http://semanticweb.org/wiki/3store.html
120 https://jena.apache.org/tutorials/rdf_api.html
121 http://wifo5-03.informatik.uni-mannheim.de/bizer/rdfapi/tests.html
122 http://librdf.org/
123 https://rdflib.readthedocs.io/en/stable/
124 https://www.w3.org/2013/data/
125 http://protege.stanford.edu/

ensuite une des sources en entrée du processus Talend, au même titre que les fichiers de données mis à disposition par l'ANSM et datagouv.fr, en l'occurrence dans notre cas.

Cette première phase a consisté à :

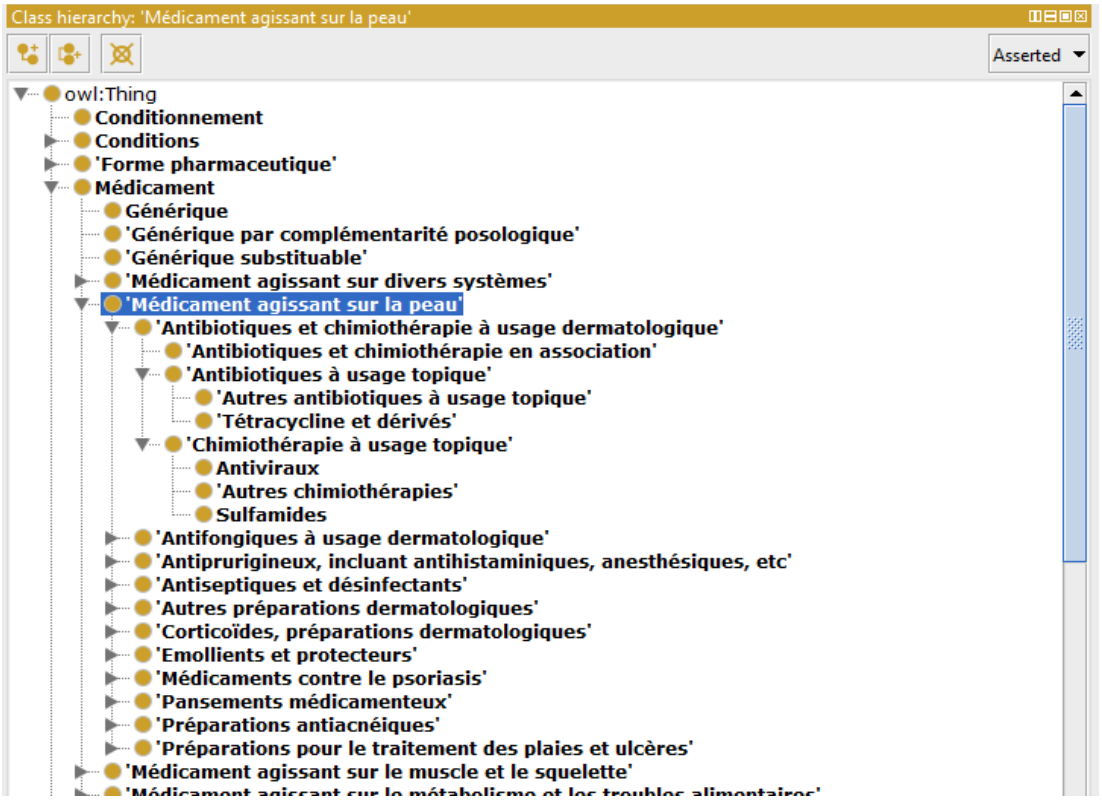
- Un nettoyage manuel de certaines données, sachant que cette tâche serait réalisée une fois pour toutes : correction du transcodage UTF8[126] de la classe des Conditions de délivrance des médicaments (voir Ecran 1) afin d'avoir les accentuations correctes, par exemple.



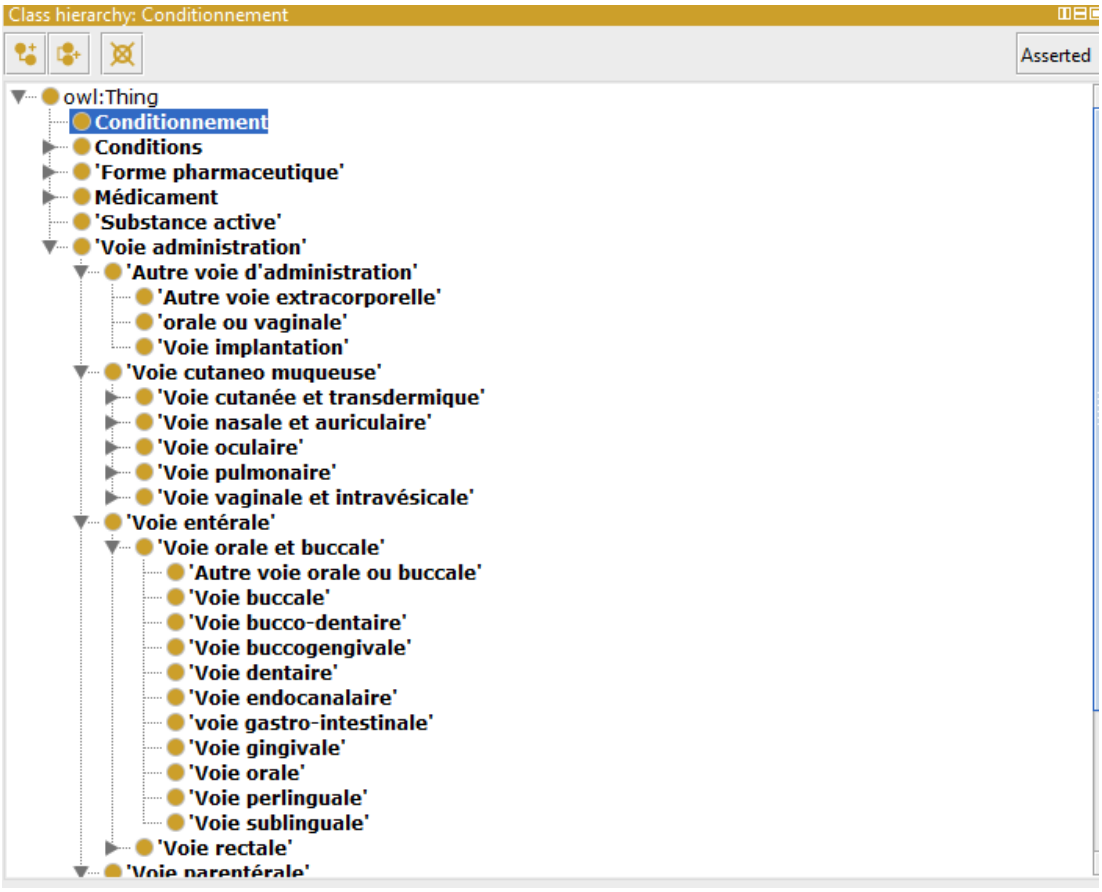
Ecran 1 : Extrait de la classe Conditions sous Protégé.

- La structuration des classes, en particulier l'alignement avec la nomenclature ATC (voir Ecran 2), ou bien les voies d'administration (voir Ecran 3).

[126] Cf §6.2.2 Automatiser la curation et traiter l'information



Ecran 2 : Extrait de la classe Médicaments sous Protégé selon la structure de l'ATC.



Ecran 3 : Extrait de la classe Voies d'administration sous Protégé.

- La documentation des concepts, c'est-à-dire leur affecter dans l'annotation « skos:preflabel [langage:fr] » un nom en français de manière normée (1er caractère en majuscule, le reste en minuscule et accentué), dans l'annotation « skos:preflabel [langage:en] » la traduction en anglais, dans l'annotation « locale » 'code ATC' la valeur du code ATC ou encore dans l'annotation « skos :altlabel » un ou plusieurs synonymes (pour les voies d'administration par exemple), ou enfin dans l'annotation « rdfs:comment » des explications sur le concept.
- L'amélioration de la complétude des propriétés (object properties, data properties, annotation properties) que l'on ne va pas détailler ici, car cela apporte peu d'intérêt à le faire.

Cette étape est consommatrice en termes de temps et nécessite également de faire des choix concernant l'organisation des données. De plus amples détails seront donnés dans la partie concernant le travail sur les données.

L'ontologie « racine » une fois générée, nous sommes passés au traitement et à l'intégration dans Talend de l'ensemble des sources de données. Nous avons ainsi obtenu en sortie, un nouveau fichier OWL correspondant à l'ontologie finale des médicaments. Protégé intervient à nouveau à ce moment-là, pour visualiser la base de connaissance obtenue, ou pour interroger celle-ci via des requêtes SPARQL.

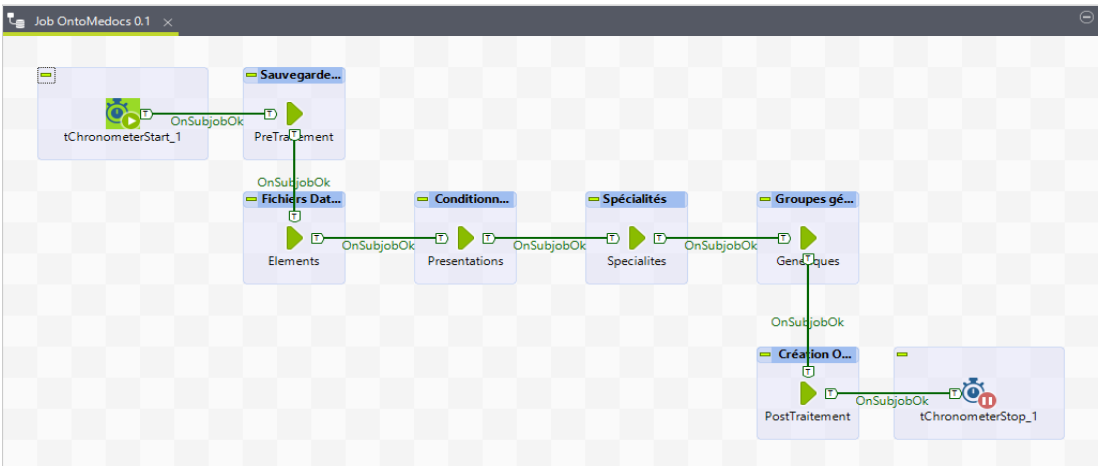
6.1.2. Talend Open Studio – Data integration[127]

Nous avons utilisé un des produits *Open Source* de Talend permettant l'intégration de données, *Data Integration*. Talend permet le chargement, l'extraction, la transformation et la manipulation de larges volumes de données disparates dans un environnement graphique, et avec une architecture ouverte et évolutive.

La première étape dans Talend a consisté à paramétrer l'environnement de travail et à recenser l'ensemble des sources de données des médicaments, avec le format ad hoc, le transcodage adapté et les métadonnées associées. Dans notre cas, il s'agit de données textuelles au format TXT ou CSV en majorité.

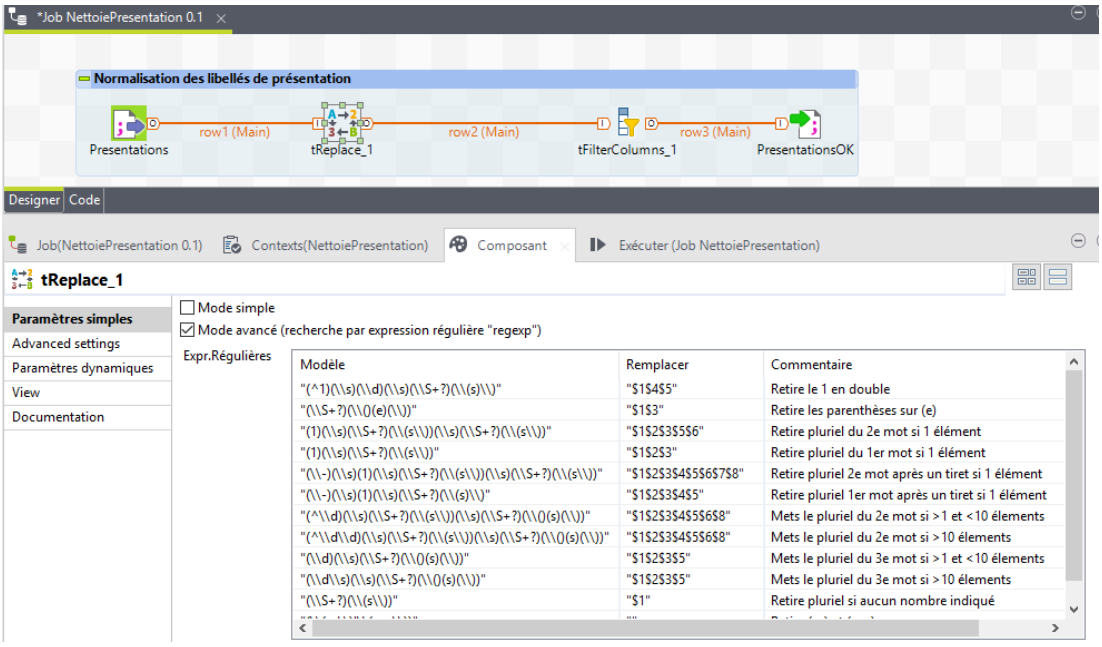
Talend fonctionne sous la forme de « jobs », sorte de modules graphiques correspondant à des fonctions, qu'il s'agit de manipuler les uns à la suite des autres, voire de les emboiter, afin de réaliser l'action souhaitée : on peut voir dans l'écran 4 la chaine complète de création de l'ontologie (elle-même constituée de sous-jobs, afin de découper le plus possible chaque phase du traitement).

[127] https://fr.talend.com/products/talend-open-studio



Ecran 4 : Chaine principale de création de l'ontologie sous Talend.

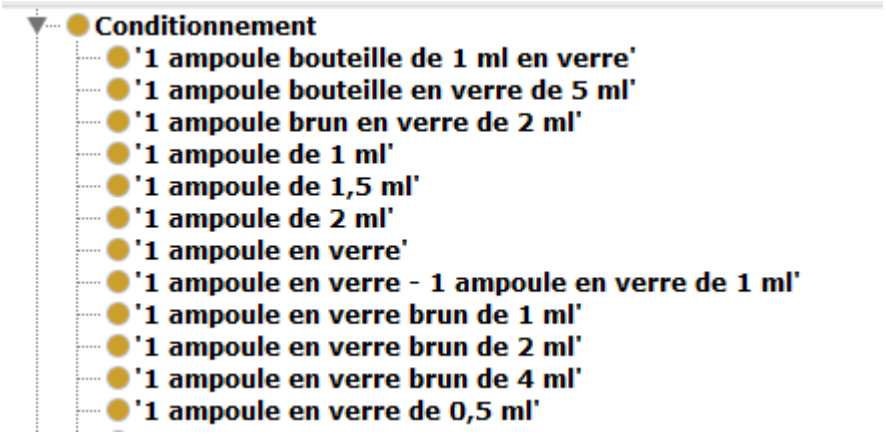
Concernant le nettoyage de données, un nouveau job (job « PreTraitement ») a été conçu et développé. En effet, le fichier issu de l'ANSM concernant le conditionnement ou présentation des médicaments, était inutilisable tel quel. On y trouvait des doublons, ou des formulations telles que « 1 ampoule(s) en verre - 1 ampoule(s) en verre de 1 ml », ou « 1 cartouche(s) polystyrène polypropylène de 60 dose(s) », « 20 plaquette(s) thermoformée(s) PVC-aluminium de 20 comprimé(s) ». Parmi les nombreux composants ou connecteurs disponibles dans Talend (plus de 900), l'un d'entre eux (tReplace) permet d'utiliser les expressions régulières[128] pour transformer des chaines de caractères et traiter ainsi en masse des expressions offrant le même modèle. L'écran 5 illustre comment ces expressions régulières ont été mises en œuvre dans Talend pour « nettoyer » le fichier des présentations.



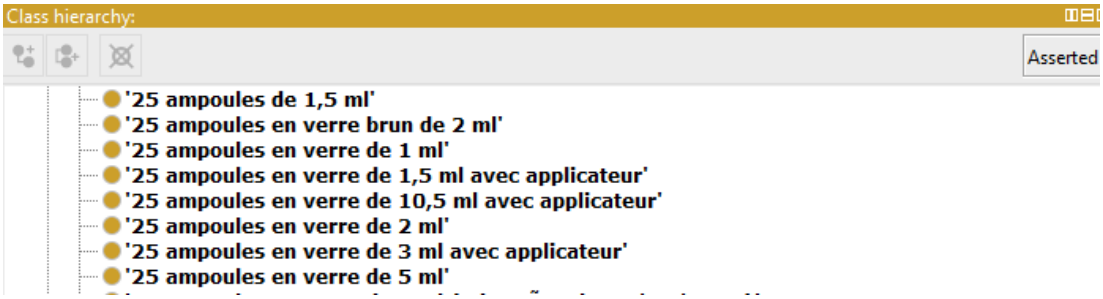
Ecran 5 : Expressions régulières créées dans le connecteur tReplace de Talend.

[128] Une expression régulière est une chaine de caractère qui prend la forme d'un motif (*Pattern*), permettant la recherche et le remplacement automatisés de ce motif dans un corpus.

Après exécution sur les fichiers d'entrée impropres, on obtient en sortie des résultats satisfaisants, comme on le voir sur les écrans 6 et 7.



Ecran 6 : Visualisation sous Protégé de la correction du format singulier par Talend.



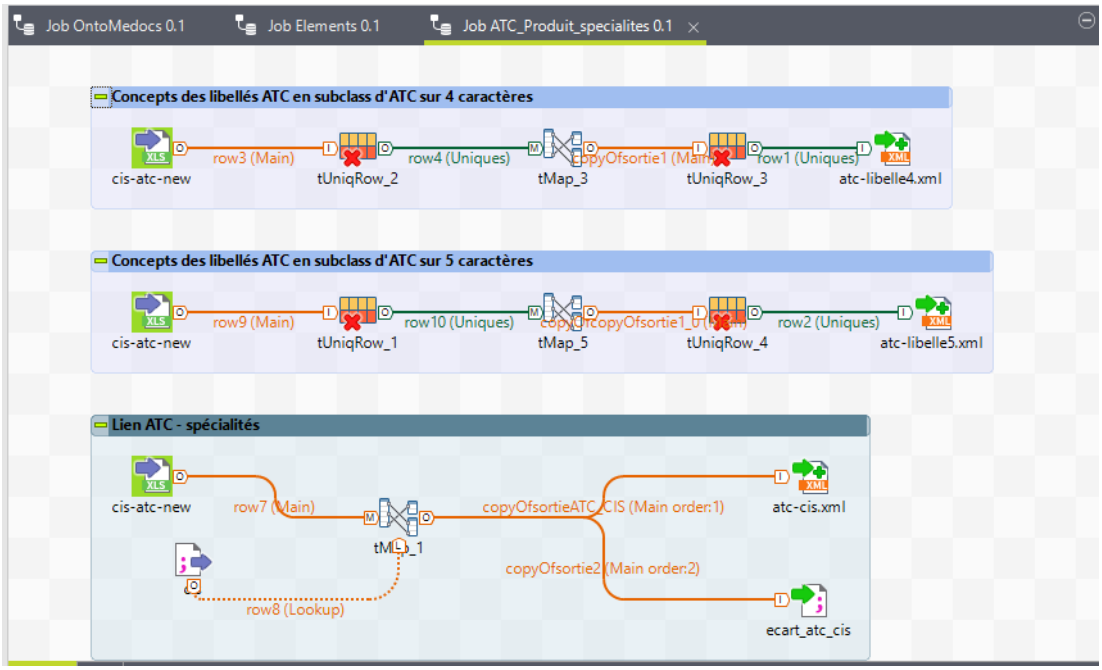
Ecran 7 : Visualisation sous Protégé de la correction du format pluriel par Talend.

C'est également via Talend que nous pouvons construire à la volée la structure hiérarchique des médicaments par rapport au modèle de l'ATC.

Comme nous l'avons évoqué précédemment dans la partie sur Protégé, les concepts de niveau le plus élevé de l'ATC ont été créés semi-manuellement (un processus Talend d'initialisation avait permis de formater les URI de ces classes à partir du code ATC), c'est-à-dire :

- le groupe anatomique (dans notre exemple de l'écran 2, « Médicament agissant sur la peau » / code ATC « D ») ;
- le groupe thérapeutique (dans notre exemple de l'écran 2, « Antibiotiques et chimiothérapie à usage dermatologique » / code ATC « D06 ») ;
- le sous-groupe thérapeutique (dans notre exemple de l'écran 2, « Antibiotiques à usage topique » / code ATC « D06A ») ;
- le sous-groupe chimique (dans notre exemple de l'écran 2, « Tétracycline et dérivés » / code ATC « D06AA »).

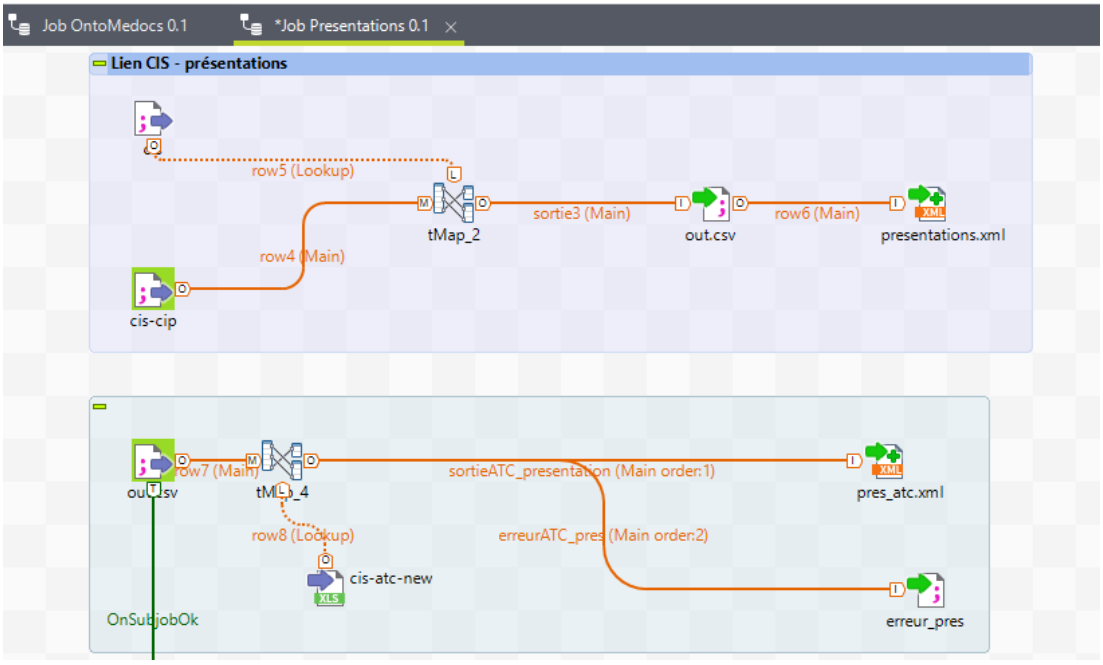
L'automatisation a donc consisté à traiter la partie variable, c'est-à-dire celle qui fera l'objet de mises à jour dans le futur, en fonction de l'évolution des fichiers de données sources. C'est l'objet du job Talend que l'on peut voir dans l'écran 8.

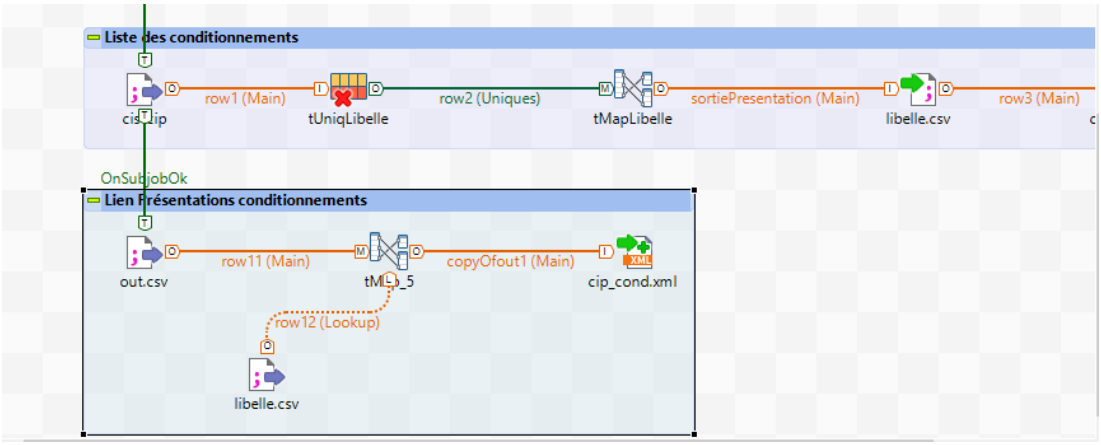


Ecran 8 : Job Talend pour la création automatique de la sous-classe « Spécialité » par rapport à l'ATC.

Dans son fonctionnement, ce job permet de créer pour chaque spécialité de médicament existante dans le fichier source issu de l'ANSM, l'enregistrement correspondant avec son URI propre, ainsi que son URI de rattachement, le tout au format XML. Un fichier d'erreur correspondant aux écarts constatés entre les fichiers d'entrée est généré, grâce à l'existence d'une clé d'identification unique des médicaments : le code CIS.

Le même type d'opération est réalisé avec un autre traitement (voir Ecran 9), concernant les présentations et conditionnements de médicament qui correspondent à un autre sous-niveau de classement dans l'ATC.





Ecran 9 : Job Talend pour la création automatique des sous-classes « Présentation » et conditionnement par rapport à l'ATC.

Là encore, les formats de sortie sont des fichiers XML, qui seront agrégés en fin de chaine de traitement pour former l'ontologie finale au format OWL, grâce à un programme Java inclus dans un composant Talend (tGroovy).

Les autres sous-jobs ne sont pas décrits dans ce mémoire, car ils n'ont pas été modifiés pendant le stage.

Enfin, pour permettre le suivi et la maintenance du processus Talend de génération de l'ontologie des médicaments, un guide de paramétrage initial et d'utilisation a été rédigé et mis à disposition du LIMICS.

6.1.3. Openlink Virtuoso[129]

Openlink Virtuoso est un triplestore[130], c'est-à-dire une base de données spécialement conçue pour le stockage et la récupération de données RDF. Il offre également une interface de requêtage ou SPARQL *endpoint*, déjà évoqué précédemment dans ce mémoire. Tout comme une base de données relationnelle, un triplestore stocke des données et il les récupère via un langage de requête :

- SQL pour une base de données relationnelle ;
- SPARQL pour un triplestore : il s'agit d'un protocole (W3C du 15 janvier 2008) et d'un langage de requêtes qui permet d'exploiter l'approche sémantique des données RDF. Il est doté : d'un langage de requêtes avec syntaxe basée sur des triplets, d'un protocole d'accès comme un service Web (SOAP), d'un langage de présentation des résultats (XML, JSON, HTML, RDF, N3). Grâce à cette technologie d'interrogation, les utilisateurs peuvent se concentrer sur leur recherche plutôt que sur la technologie de base de données ou le format sur lesquels repose le stockage des données. SPARQL cible donc l'interrogation de métadonnées RDF, structure de base du Web sémantique.

L'avantage d'un triplestore vient directement du type de données qu'il stocke, le triplet, car il ne nécessite pas la création de tables au préalable. Les données peuvent être directement enregistrées sans phase d'initialisation. Les volumétries peuvent être très importantes, les capacités de stockage pouvant atteindre des téras de triplets RDF (1000 milliards).

---

129 http://virtuoso.openlinksw.com/

130 https://fr.wikipedia.org/wiki/Triplestore

L'ensemble des triplets reliés les uns aux autres par des URIs (eux-mêmes étant des ressources représentés par des URIs) constitue un graphe, qui peut être créé ou interrogé via des requêtes SPARQL (méthode CONSTRUCT).

Dbpedia[131] est par exemple un des projets qui utilise Virtuoso.

Notre objectif est de stocker l'ontologie des médicaments dans Virtuoso. Celle-ci est au format OWL, en sortie de Talend et Protégé. Comme nous l'avons évoqué précédemment dans ce mémoire, OWL est un langage basé sur RDF. Il enrichit le modèle des RDF Schemas en définissant un vocabulaire riche pour la description d'ontologies complexes. Par conséquent, une ontologie peut constituer un graphe dans Virtuoso.

L'installation et le paramétrage de Virtuoso n'est pas forcément quelque chose de simple et les différents forums sur le Web ont été d'une grande utilité. Pour éviter à mes successeurs d'avoir à refaire ce travail de recherche, j'ai rédigé un petit guide d'installation et de chargement de l'ontologie des médicaments dans Virtuoso. On y trouve les étapes à suivre avec les liens sources pour l'installation en elle-même, puis pour l'utilisation, quelques requêtes SPARQL permettant de vérifier, par exemple, que l'ontologie a bien été chargée dans le graphe Virtuoso, créé à cet effet.

6.1.4. Conclusion : une belle ontologie mais une petite déception.

A la fin du stage, grâce à tous les outils et technologies que nous venons d'aborder, une ontologie des médicaments assez bien classée et structurée a donc pu être générée : elle contient 184 000 classes !

Cela n'a pas été sans un travail conséquent sur la qualité des données en entrée, en l'occurrence lors du développement du job de curation des présentations (conditionnements) de médicaments. L'ensemble des problèmes rencontrés en ce qui concerne la qualité des données, et leurs éventuelles solutions font l'objet de la partie suivante.

Pour finir sur l'aspect technique, notre regret est de ne pas avoir pu mettre en place une procédure automatique Talend de chargement de l'ontologie directement en triplets RDF dans Virtuoso, c'est-à-dire sans passer par une transformation intermédiaire en fichiers XML, ensuite concaténés par programmation pour en faire le fichier OWL final.

En effet, des composants spécifiques pour interfacer Virtuoso ont été développés par des membres de la communauté de développeurs Talend[132] , et nous auraient permis d'effectuer cela, sauf qu'il nous a été impossible de les faire fonctionner. En effet, leurs auteurs n'ont laissé aucune documentation aboutie, et aucune utilisation par d'autres équipes n'a été trouvée sur le Web. Nous avons donc abandonné cette idée au bout d'une bonne semaine de recherches infructueuses.

Cependant, l'idée de réaliser le chargement direct dans Virtuoso des morceaux d'ontologie au fur et à mesure de sa construction (au lieu de morceaux en XML) est toujours envisagée pour de futures évolutions, avec des développements spécifiques en Java, par exemple.

131 http://dbpedia.org/
132 http://fbelleau.github.io/talend4sw/

6.2. Travail sur les données en entrée

6.2.1. Sources de données, constats et traitements « one shot »

6.2.2.1. Données de l'ANSM

Les fichiers téléchargeables sur le site de l'ANSM[133] contiennent toutes les Autorisations de Mise sur le Marché (AMM) **existantes ou ayant existées depuis 2001**. On trouve :

- le fichier des spécialités : cis.txt ;
- le fichier des présentations : cis_cip.txt ;
- le fichier des compositions : compo.txt.

Ce sont les codes CIS (identifiant unique) des bases de l'ANSM qui sont les références des processus avant complémentation par les autres fichiers (datagouv,…).

6.2.2.2. Données issues de la base de données publique des médicaments[134]

Les fichiers téléchargeables sur le site du Ministère des Affaires sociales et de la Santé [53] contiennent les AMM **avec 3 ans de recul** sont :

- le fichier des avis SMR de la HAS : CIS_HAS_SMR_bdpm.txt ;
- le fichier des avis ASMR de la HAS : CIS_HAS_ASMR_bdpm.txt ;
- le fichier des liens vers les avis de la commission de la transparence (CT) de la HAS : HAS_LiensPageCT_bdpm.txt ;
- le fichier des groupes génériques : CIS_GENER_bdpm.txt ;
- le fichier des conditions de prescription et de délivrance : CIS_CPD_bdpm.txt ;
- le fichier des informations importantes : CIS_InfoImportantes_AAAAMMJJhhmiss_bdpm.txt.

On y trouve également la version correspondante des fichiers des spécialités, des présentations et des compositions, **sans qu'il ne soit précisé nulle part l'existence de ce même jeu sur le site de l'ANSM, dans une version des règles de gestion différente (depuis 2001 pour l'ANSM vs. 3 ans pour le site du gouvernement)**. Ainsi, nous avons utilisé dans un premier temps les fichiers du gouvernement pour nous rendre compte par la suite de l'existence de gros écarts de données dans nos traitements : c'est en nous renseignant auprès de l'ANSM directement que nous avons eu l'explication sur les différences de contenu de fichiers.

6.2.2.3. Les problèmes de qualité traités sur les fichiers ANSM et Datagouv

Le point qui vient d'être évoqué est le premier problème de qualité qui nous a fait perdre du temps, du fait de l'utilisation des mauvais fichiers et des explications recherchées pour comprendre les écarts constatés. Sans l'existence de notre traitement de gestion des erreurs, nous aurions mis à disposition dans l'ontologie des médicaments, des données incomplètes et donc fausses !

Un autre élément à considérer concerne les métadonnées et la manière dont elles sont fournies :

- Sur le site de l'ANSM, un texte libre accompagne les fichiers téléchargeables et indique les noms des colonnes du fichier correspondant.

---

[133] http://agence-prd.ansm.sante.fr/php/ecodex/telecharger/telecharger.php
[134] http://base-donnees-publique.medicaments.gouv.fr/telechargement.php

- Sur Datagouv, il est nécessaire d'ouvrir un fichier PDF fourni dans la page de téléchargement pour avoir l'explication des métadonnées pour chacun des fichiers disponibles.

Le fait de disposer d'un entête de colonnes avec la métadonnée directement dans chaque fichier éviterait aux réutilisateurs des données, à avoir à le faire lui-même, avec les erreurs que cela peut engendrer.

Un autre problème a concerné la qualité intrinsèque du fichier des présentations (ou conditionnement), comme nous l'avons déjà abordé dans la partie logicielle sur Talend et Protégé. Comme nous avons pu le voir, nous avons pu gérer la question du format textuel avec des « faux » pluriels (« 1 ampoule(s)… »), grâce à l'utilisation des expressions régulières, processus désormais automatisé, donc résolvant le problème pour les futures mises à jour.

Cependant, nous souhaitions également structurer ces données sur plusieurs niveaux de détail, et donc les catégoriser, par exemple : « ampoule en verre brun de 2 ml » peut se décomposer en « ampoule/ampoule en verre/ampoule en verre brun/ ampoule en verre brun de 2 ml ». Cela aurait donné une classification hiérarchique intéressante, idée que nous avons dû abandonner en tentant de le faire, car il s'est avéré impossible de réaliser ce découpage proprement jusqu'au bout.

En effet, pour de nombreux cas, des tournures de phrases différentes signifient en fait la même chose, et il devient impossible de prévoir un traitement automatique pour gérer ce fait. Cela est dû, très certainement, à des saisies provenant de sources différentes, au niveau humain ou logiciel.

Voici un exemple de ce que l'on peut trouver : pas moins de 11 manières sur ce cas pour signifier la même chose, avec plus ou moins de précision :

- bouteille acier aluminium de 15 l munie d'un robinet en laiton avec manodétendeur et prises normalisées ;
- bouteille acier de 15 l munie d'un robinet en laiton avec raccord de sortie normalisé ;
- bouteille acier aluminium de 15 l munie d'un robinet en laiton avec manodétendeur RM200light® et prises normalisées ;
- bouteille acier de 15 l avec manodétendeur-débitmètre Combistar(R) de 0,5 à 5 l/min et prise normalisée ;
- bouteille acier de 15 l avec manodétendeur-débitmètre Ministar(R) de 1 à 15 l/min et prises normalisées ;
- bouteille acier ou aluminium de 15 l avec un robinet en laiton avec manodétendeur et prise normalisée ;
- bouteille aluminium ou acier de 15 l munie d'un robinet en laiton avec raccord de sortie normalisé ou d'un robinet en laiton à pression résiduelle avec raccord de sortie normalisé ;
- bouteille aluminium ou acier de 15 l- robinet laiton avec raccord de sortie normalisé ;
- bouteille de 15 l en acier, en aluminium ou en aluminium frettée, muni d'un robinet en laiton avec raccord de sortie normalisé ;
- bouteille en acier, en aluminium ou en aluminium frettée de 15 l munie d'un robinet en laiton avec raccord de sortie normalisé ;
- bouteille en aluminium ou en acier de 15 l munie d'un robinet en laiton avec raccord de sortie normalisé ou d'un robinet en laiton à pression résiduelle avec raccord de sortie normalisé.

6.2.2.4. Fichier ATC

Il s'agit d'un fichier au format XLS, qui nous est fourni directement et manuellement par l'ANSM. Une des questions est de savoir si ce fichier sera mis à disposition de manière ouverte sur Data.gouv.fr par exemple.

Nous nous servons du fichier ATC pour faire le lien entre le code ATC et les spécialités et présentations de médicaments, grâce au code CIS, l'identifiant unique déjà évoqué précédemment, servant de clé pour établir les relations entre chaque fichier. Cela permet ainsi de classer un médicament dans sa classe thérapeutique, et donc groupe anatomique.

Pour la création de l'ontologie, quelques axes d'amélioration du fichier seraient les bienvenus :

- Avoir tous les niveaux ATC renseignés, car certains codes intermédiaires sont absents, ce qui a pour effet, de déstructurer l'ontologie en bout de chaine. Pour corriger le problème, le niveau manquant a été créé manuellement dans l'ontologie racine, avec le libellé trouvé dans Vidal[135]. Exemple des codes ATC S01AE01, S01AE02 et S01AE03, pour lesquels la classe mère S01AE (Fluoroquinolones) absente du fichier a dû être ajoutée manuellement.
- Ne pas conserver de codes ATC sans libellé : nous avons su par l'ANSM, que ce sont des codes inactivés et remplacés par l'OMS en 2005. Exemple des codes ATC J01DA01, J01DA04, J01DA08, J01DA14, J01DA31, J01DA37 et J01DA38, devenus respectivement J01DB01, J01DB04, J01DC04, J01DC05, J01DB09, J01DE02 et J01DC08.

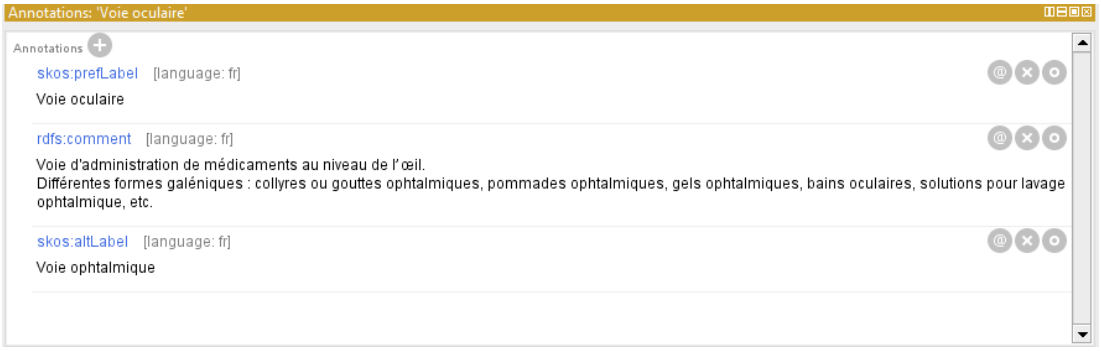
6.2.2.5. Voies d'administration

Les classes créées dans l'ontologie pour les voies d'administration sont issues des fichiers de l'ANSM et extraites via un job Talend.

La liste obtenue présente de « faux » doublons (exemple « intra-artérielle » et « intraartérielle ») et surtout n'offre pas de classification qui la rendrait plus structurée pour regrouper les termes.

En combinant et synthétisant plusieurs sources internet généralistes[136] ou spécialisées[137] en médecine/pharmacie [54], une taxinomie a été définie pour l'ontologie des médicaments, et créée une fois pour toute dans l'ontologie racine. Le nommage des URIs de dernier niveau devant correspondre à celui affecté dans le job Talend lors de l'exploitation du fichier source ANSM, le « skos:prefLabel » contenant le terme servant à l'affichage. Une documentation des termes et concepts a également été ajoutée : certains synonymes dans « skos:altLabel » ou description dans « rdfs:comment ». L'écran 10 illustre le résultat pour le concept « Voie oculaire ».

[135] https://www.vidal.fr/classifications/atc/
[136] https://fr.wikipedia.org/wiki/Voie_d%27administration
[137] http://www.decitre.fr/media/pdf/feuilletage/9/7/8/2/2/9/4/7/9782294738265.pdf



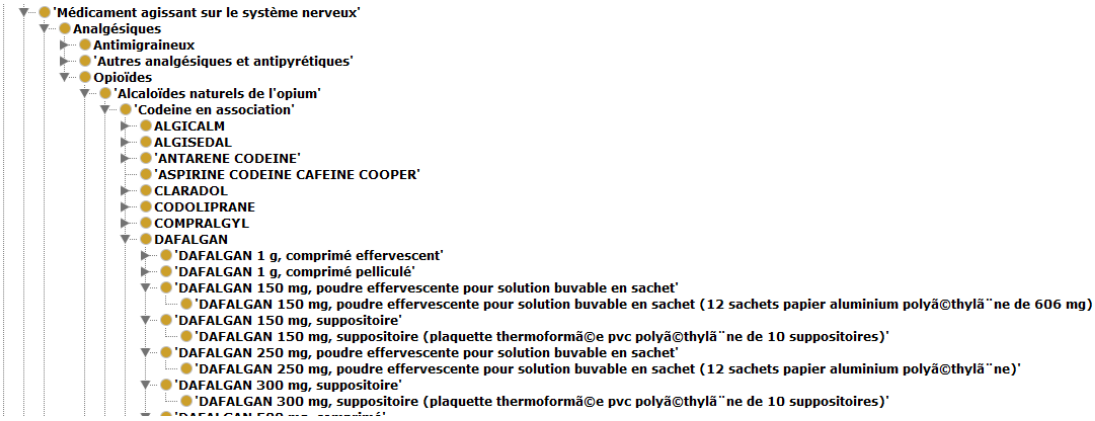
Ecran 10 : Annotations Protégé caractérisant le concept « Voie oculaire » dans l'ontologie des médicaments.

Le classement adopté pour les voies d'administration se trouve en annexes du mémoire.

6.2.2. Automatiser la curation et traiter l'information

Nous avons vu qu'un certain nombre traitements ou procédures manuels ont pu être mis en place pour semi-automatiser la mise à jour de l'ontologie.

Nous pouvons aussi y associer un élément important qui conditionne un affichage correct des chaines de caractère : le transcodage et l'utilisation appropriée d'UTF8[138]. Ce code UNICODE permet de représenter tous les caractères spécifiques aux différentes langues. De nouveaux codes sont régulièrement attribués pour de nouveaux caractères: caractères latins (accentués ou non), grecs, cyrilliques, arméniens, hébreux, thaï, hiragana, katakana... L'Unicode définit une correspondance entre symboles et nombres, par lesquels passe forcément le codage informatique. Il faut donc jouer entre UTF8 et ISO/CEI 8859[139], une norme commune de l'ISO et de la CEI pour le codage de caractères sur 8 bits et le traitement informatique du texte simple, permettant d'obtenir en sortie un affichage correct des caractères spéciaux ou accentués. Un transcodage incorrect génère des erreurs d'affichage comme on peut le voir dans l'écran 11 (« 12 sachets papier aluminium polyã©thylã¨ne de 606 mg » au lieu de « 12 sachets papier aluminium polyéthylène de 606 mg »).



Ecran 11 : Caractères spéciaux obtenu sous Protégé avec un transcodage incorrect.

138 https://fr.wikipedia.org/wiki/UTF-8
139 https://fr.wikipedia.org/wiki/ISO/CEI_8859

C'est au niveau de Talend et de l'intégration des fichiers sources que l'on effectue l'action de transcodage des fichiers sources.

6.2.3. Sensibiliser les organismes fournisseurs pour une amélioration des données à la source

L'ensemble des problèmes rencontrés doit faire l'objet d'une remontée vers les producteurs si on souhaite une amélioration du modèle de données dans sa globalité.

En effet, le fait de trouver des solutions de « bricolage » n'est pas pérenne, même si elles permettent dans un premier temps d'obtenir un résultat assez satisfaisant. Dans la durée, cela ne peut être que problématique et lourd à gérer. On ne fait que reporter plus en aval le problème initial, et obtenir de ce fait d'autres impacts.

De plus, c'est bien le propriétaire de la donnée qui en est responsable, et doit de ce fait s'acquitter des garanties de disponibilité, utilisabilité, sécurité, pérennité, authenticité de cette donnée.

6.3. Choix du modèle de données

Par rapport aux différents modèles de données existants, il faut tout d'abord noter qu'aucun d'entre eux ne correspond à une ontologie au sens strict, c'est-à-dire dans un format OWL.

Des ontologies basées sur RxNorm ont vu le jour, telles que DrOn (*Drug Ontology*) [43, HANNA *et al.*] pour pallier le manque de ressources de ce type dans le Web sémantique. DrOn correspond à un travail de fouille de RxNorm couplé à un « *mapping* » avec les classes ChEBI (*Chemical Entities of Biological Interest*). Cela a abouti, d'après les auteurs, à une ontologie des médicaments modulaire, extensible, décrivant les ingrédients, les effets biologiques et autres éléments répondant aux besoins de leurs cas d'utilisation. Des APIs sont utilisées pour transformer les différents modules en objets OWL.

Concernant l'ontologie française des médicaments, le modèle de données DrOn nous a semblé trop spécifique : son étude et alignement éventuel aurait demandé plus de temps que nous ne pouvions y consacrer, c'est-à-dire la durée de mon stage. D'autant plus que la structure initiale mise au point par Xavier Aimé constituait un point de départ solide, sur lequel j'ai choisi de partir pour l'améliorer, plutôt que de tout recommencer.

Un site très intéressant, AberOWL *ontology repository and semantic search engine* [44], donne accès à une description du vocabulaire RxNorm et permet de consulter l'ensemble des relations, propriétés, classes, instances, d'interroger la base via des requêtes SPARQL, et de télécharger celle-ci au format .ONT. Nous n'avons pas pu tirer parti de ce fichier non standard, et trop volumineux.

Pour en revenir à RxNorm, modèle qui a suscité le plus d'intérêt pour nous, il a pour objectif de permettre à des systèmes utilisant différentes nomenclatures de médicaments de partager et échanger des données efficacement. Il donne un vocabulaire standard et un modèle de référence, avec une structure hiérarchique et des relations entre les concepts : hiérarchie des classes de médicaments au niveau structure chimique, mécanismes d'action cellulaires, effets physiologiques, relations médicament/maladie pour décrire l'effet thérapeutique, et mécanismes d'absorption d'un médicament par le corps. RxNorm inclut maintenant le fichier national des drogues - terminologie de référence (NDF-RT) de l'administration de la santé des

anciens combattants. NDF-RT catégorise les médicaments via une hiérarchie d'héritage simple de 550 classes de médicaments environ.

RxNorm est également présente sur le site The National Center for Biomedical Ontology [45].

Pour étudier ce qu'il était possible de retirer de RxNorm, ou s'en inspirer, j'ai réalisé la représentation du modèle (Figure 12), grâce aux différentes descriptions et schémas publiés (les concepts et relations sont décrites précédemment dans ce mémoire).

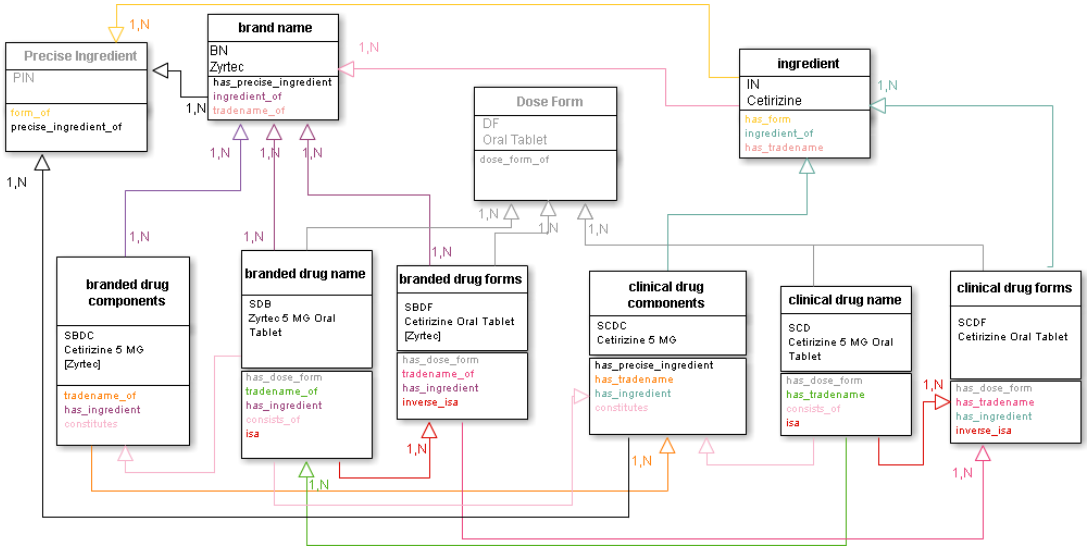


Figure 12 : Interprétation du modèle de données RxNorm.

J'ai ensuite procédé à une traduction du modèle en français, de ses concepts et relations (voir tableaux 11 et 12), pour étudier si un alignement simple pouvait être réalisé concernant l'ontologie française des médicaments, ce qui est illustré dans la figure 13.

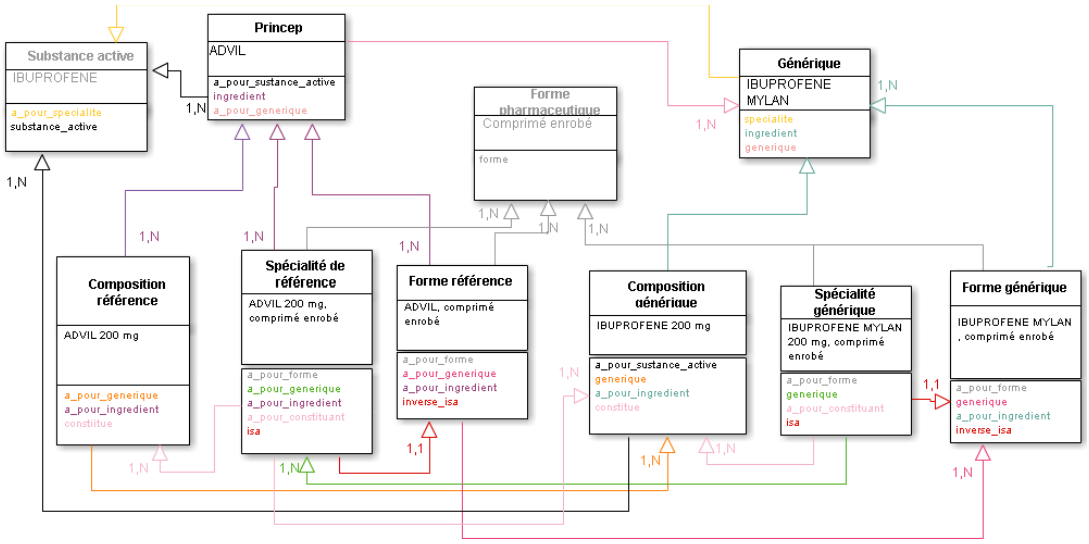


Figure 13 : Traduction en français du modèle de données RxNorm.

Concept RxNorm	Concept traduit en français	Signification
Ingredient (IN)	Générique	
Clinical Drug Component (SCDC)	Composition générique	Générique + dosage
Clinical Drug Form (SCDF)	Forme générique	Générique + forme
Clinical Drug name (SCD)	Spécialité générique	Générique + dosage + forme
Brand Name (BN)	Princeps	
Brand Drug Name (SDB)	Spécialité référence	Princep + dosage + forme
Branded Drug Component (SBDC)	Composition référence	Princep + dosage
Branded Drug Form (SBDF)	Forme référence	Princep + forme
Precise Ingredient (PIN)	Substance active	
Dose form	Forme	Forme pharmaceutique

Tableau 11 : Traduction des concepts RxNorm en français.

Relations RxNorm	Relations traduites en français	Signification / exemple
tradename_of / has_tradename	a_pour_generique / generique	Un princeps ou une spécialité de référence « a pour générique » un (ou plusieurs) générique(s) ou une (ou plusieurs) spécialité générique(s). *Ex : Advil a pour générique Ibuprofène Mylan*
form_of / has_form	A_pour_specialite / specialite	Une substance active a pour spécialité un (ou plusieurs) générique(s). *Ex : Ibuprofène a pour spécialité Ibuprofène Mylan*
precise_ingredient_of / has_precise_ingredient	a_pour_substance_active / substance_active	Un princep ou une composition générique « a pour substance active » une (ou plusieurs) substance(s) actives(s). *Ex : Advil a pour substance active Ibuprofène*
ingredient_of / has_ingredient	Ingredient / a_pour_ingredient	Une (ou plusieurs) composition(s), forme(s) et spécialité(s) de référence a pour ingrédient un princep. Une (ou plusieurs) composition(s), forme(s) générique(s) a pour ingrédient un générique (pas une spécialité générique ?). *Ex : Advil 200 mg a pour ingrédient Advil*

Relations RxNorm	Relations traduites en français	Signification / exemple
consists_of / constitutes	a_pour_constituant / constitue	Une spécialité de référence a pour constituant une (ou plusieurs) composition(s) référence, ou composition(s) générique(s). Une spécialité générique a pour constituant une (ou plusieurs) composition(s) générique(s). Ex : *Advil 200 mg, comprimé enrobé a pour constituant Advil 200 mg*
dose_form_of / has_dose_form	forme / a_pour_forme	Une spécialité de référence, forme référence, spécialité générique, forme générique a pour forme pharmaceutique. Ex : *Advil 200 mg, comprimé enrobé a pour forme comprimé enrobé*
isa / inverse_isa	est	Une spécialité de référence ou spécialité générique est une forme de référence ou une forme générique. Ex : *Advil 200 mg, comprimé enrobé est un sous concept de Advil, comprimé enrobé*

Tableau 12 : Traduction des relations RxNorm en français.

J'ai ensuite tenté de réaliser une correspondance entre la structure des données ANSM et RxNorm détaillée dans le tableau 13. On peut noter qu'un certain nombre de points sont restés sans réponse, illustrés par des points d'interrogation.

Concept ANSM	Concept RxNorm	Exemple médicament ANSM	Exemple médicament Rxnorm
Dénomination du médicament (+Titulaire pour la marque ?)	Branded Drug Form (SBDF) ou Clinical Drug Form (SCDF)	A 313 200 000 UI POUR CENT, pommade	Cetirizine Oral Tablet [Zyrtec] ou Cetirizine Oral Tablet
Forme pharmaceutique + Voies d'administration	Dose Form (DF)	pommade + cutanée	Oral Tablet
Libellé de la présentation	?	1 tube(s) aluminium verni de 50 g	?
Elément pharmaceutique (=forme pharmaceutique ?)	Dose Form (DF) ?	pommade	Oral Tablet ?
Dénomination de la substance	Ingrédient (IN)	CONCENTRAT DE VITAMINE A SYNTHÉTIQUE, FORME HUILEUSE	Cetirizine
Dosage de la substance	Clinical Drug Component (SCDC)	200 000 UI	Cetirizine 5 MG
Référence de ce dosage	?	100 g de pommade	?

Tableau 13 : Ebauche de correspondance entre les concepts ANSM et RxNorm.

Ce travail de traduction / alignement a conduit aux conclusions suivantes :

- concernant les relations, toutes ne seraient pas à conserver car elles semblent redondantes ;
- il manque la voie d'administration (oral, buccal,…) ;
- concernant le modèle, il peut y avoir des problèmes au niveau de la granularité de l'information, car le découpage de celle-ci n'est pas le même au niveau des sources de données.

Il n'a donc pas été possible de réaliser une correspondance immédiate entre la structure des données françaises (ANSM) et celles issues de RxNorm.

Au final, nous avons donc acté le fait de conserver les bases de notre modèle de données initial, en y incluant des améliorations notables en termes de hiérarchie entre les concepts, son principal défaut étant d'être « plate ». Le fait d'utiliser le modèle de l'ATC a permis de classer les médicaments par groupe anatomique, puis groupe thérapeutique, pour aller jusqu'à la substance chimique et Dénomination Commune Internationale (DCI) et donc d'être très logique et parlant pour les publics concernés.

6.4. Bénéfices attendus et exigences intrinsèques

De manière générale, l'usage d'une terminologie de référence apporte des améliorations dans les systèmes d'échanges de données, tels que :

- la fiabilisation de la saisie et l'aide à la décision pour le producteur ;
- l'interopérabilité sémantique vers d'autres systèmes ;
- la réutilisabilité des données dans d'autres contextes (pour la santé : recherche, veille sanitaire, épidémiologie, pilotage) ;
- l'amélioration de la coordination entre les acteurs du parcours ;
- l'amélioration du pilotage de l'activité.

En ce qui nous concerne, l'accent est mis sur les questions d'interopérabilité, d'exploitabilité grâce à la modularité par exemple, et le travail de documentation (liée intimement à la structuration).

6.4.1. Répondre à la nécessité d'interopérabilité

Depuis une vingtaine d'années, l'accès et l'utilisation des données médicales sont devenus des enjeux majeurs pour les professionnels de santé comme pour le grand public. Dans ce contexte, plusieurs terminologies médicales spécialisées ont été créées. Ces terminologies ont pour la plupart des formats de représentation et visées différentes : la nomenclature SNOMED 3.5 pour le codage d'informations cliniques, les classifications CIM10 et CCAM pour le codage épidémiologique puis médico-économique, le thésaurus MeSH pour la bibliographie [55, MERABTI].

Devant ce constat et la nécessité grandissante de permettre la coopération de différents acteurs de la santé et des systèmes d'information associés, il apparait nécessaire de rendre les terminologies interopérables.

En effet, le partage de données est un élément important dans l'amélioration des soins. Ainsi, si un médecin peut accéder aux données de son patient qui sont enregistrées sur un site externe au sien, où est passé ce patient, il pourra éviter de lui prescrire des examens déjà effectués. Le bénéfice se place donc au niveau du diagnostic, mais également du coût financier. Ce « dialogue » entre deux (ou plusieurs) systèmes permet de partager les connaissances, et donc d'aider

également à la prise de décision clinique, ainsi qu'à la recherche, dont les études peuvent avoir lieu à grande échelle, et basées sur des résultats statistiques plus fiables [56, AUBAIN].

Il existe plusieurs niveaux d'interopérabilité. Le niveau politique permet de définir la vision stratégique et partagée qui favorise les échanges. Il nécessite et s'appuie sur le niveau juridique pour le respect du cadre légal et des accords contractuels entre parties prenantes. Le niveau organisationnel va ensuite donner les moyens humains et structurels pour mettre en œuvre ces échanges, secondé par le niveau technique donnant les outils matériels pour véhiculer l'information. Pour que l'ensemble du dispositif fonctionne, il s'agit enfin de donner du sens à cette information, grâce aux niveaux syntaxique, et sémantique, ce dernier ôtant toute ambiguïté d'interprétation, via un vocabulaire commun.

6.4.2. Viser la modularité pour être mieux exploitable

Les terminologies de référence, et en l'occurrence les ontologies de domaine peuvent devenir très rapidement très volumineuses, et contenir un nombre trop important de classes et concepts. L'ontologie française des médicaments obtenue à la fin de mon stage avait atteint, par exemple, 184 000 classes, ce qui est pose des problèmes pour l'exploiter, et simplement rien que pour l'éditer dans Protégé, si l'ordinateur utilisé n'est pas assez puissant. Cependant, même si l'on peut procéder à une optimisation en améliorant la structuration, il n'est pas envisageable de supprimer des notions, puisque l'on souhaite couvrir l'ensemble du domaine. Une manière de simplifier et rendre moins complexe une ontologie peut consister à la découper en modules, ce qui évite d'avoir à la charger dans sa totalité.

Il serait possible de suivre l'exemple d'Orphanet, et du projet OntoOrpha[140] mené par Ferdinand Dhombres, Xavier Aimé et Jean Charlet de 2010 à 2013 [57, AIME].

ORPHANET est un portail d'information de référence sur les maladies rares et les médicaments orphelins, qui s'adresse à la fois aux professionnels de la santé et au grand public.

Le projet OntoOrpha s'est traduit par l'élaboration d'une ontologie couvrant le domaine du portail Orphanet et la valorisation des connaissances qu'il porte, en répondant aux standards du Web sémantique, et donc aux exigences d'interopérabilité. Ce qui nous intéresse particulièrement ici concerne le fait d'avoir également démontré que *les ontologies sont conçues également pour gérer la génération de plusieurs classifications de connaissances complexes*. La modularisation permet l'organisation des connaissances en plusieurs éléments indépendants, autonomes et réutilisables, chacun répondant à un besoin ou une fonction, par exemple. Ainsi, la gestion de l'ensemble est simplifiée, et chaque composant peut être réutilisé seul, ce qui est un avantage supplémentaire.

Cette ontologie trouve une application dans le projet ACCORDYS[141] pour représenter, et partager l'ensemble des connaissances connues et reconnues par les spécialistes, concernant les malformations du fœtus. Hormis les comptes rendus textuels qui y sont associés, les données sources collectées sont multimédias (*photos d'autopsie de fœtus, images d'échographie, images de différentes techniques de radiologie, résultats d'examens biologiques, etc.*).

[140] http://bioportal.bioontology.org/ontologies/HRDO
[141] https://bioportal.bioontology.org/projects/accordys

6.4.3. Documenter et annoter

L'ontologie obtenue en fin de stage, n'est pas aboutie et nécessite également des retours arrière au niveau de sa conception. En effet, la recherche nécessite d'être pragmatique, et c'est en appliquant certaines théories que l'on se rend compte que ce ne sont pas les bonnes, ou que l'on comprend certains fonctionnements.
Ainsi, lors du découpage hiérarchique suivant la nomenclature de l'ATC, nous avons fait le choix de rattacher directement les spécialités et les présentations sous le dernier niveau de l'ATC, sans y inclure la descendance avec les substances. Nous n'avions pas décelé à ce moment-là l'importance de ce lien, et ce n'est qu'à l'observation du résultat obtenu que cela s'est avéré nécessaire, et donc à modifier dans une prochaine version.

Par ailleurs, des évolutions sont à prévoir dans l'ontologie, au niveau de sa structuration et des annotations à y ajouter. Actuellement, en plus du découpage des médicaments en classes de l'ATC, les classes suivantes ont été également définies : Princeps, Générique, Génériques par complémentarité posologique, Générique substituable, Spécialité, Spécialité de référence, Spécialité générique, Spécialité substituable.
On y retrouve en doublons les médicaments déjà présents dans l'ATC : il serait plus judicieux de faire figurer cette information en tant que propriété, par exemple dans une annotation. Cela demande une réflexion et une analyse pour savoir comment implémenter cela au plus juste et valoriser le contenu de l'ontologie.

Une évolution potentielle aussi est de dériver des ontologies plus simples et plus petites à partir de la première pour donner accès, aux utilisateurs, à une ressource paramétrable et plus ou moins complexe à utiliser en fonction des besoins.

Enfin, un autre sujet a été soulevé durant le stage et concerne les équivalences entre les molécules associées au dosage et au mode d'administration. Il s'agit d'une information très utile aux médecins, et qu'il serait intéressant de faire figurer, par exemple grâce à une nouvelle propriété à créer « Est substituable ». Une autre stagiaire en pharmacie a travaillé sur ce sujet, en l'occurrence sur des équivalences pour les antidépresseurs, et il existe donc une base de travail pour étudier la question et pousser plus loin la manière de la traiter.

6.5. Passer du projet au produit

6.5.1. Quels critères pour être opérationnel ?

On peut retenir les critères relevés par l'ASIP Santé dans ses interviews [34], permettant à un futur usager de valider le choix d'une terminologie de référence par rapport à ses besoins.

La pertinence sémantique du contenu par rapport aux cas d'usage à satisfaire constitue le cœur du service rendu, avec en plus l'assurance que le codage de ces besoins pourra être satisfait dans sa totalité à partir de la terminologie (*jeux de valeurs*). On attend également une pérennité, et une évolutivité de la solution, en disposant d'un « objet vivant », avec les mises à jour régulières, demandées par les nouvelles connaissances ou pratiques du domaine.

La reconnaissance en tant que standard international conditionne fortement l'éligibilité d'une terminologie de référence, car elle implique un partage étendu au sein de la communauté d'experts, de manière libre et ouverte. La collaboration et les échanges sont améliorés, ce qui permet plus de mobilité transnationale des données ainsi qu'une meilleure qualité.

En conséquence de ce gommage des frontières, le multilinguisme est à prendre en compte pour permettre une adaptation au langage et au vocabulaire de l'utilisateur. Concernant la langue natale, la gestion des synonymes a son importance, puisqu'elle laisse la possibilité aux professionnels d'utiliser leur vocabulaire habituel d'une spécialité ou d'un environnement à l'autre.

Au niveau économique, la question des coûts induits pour le respect de l'ensemble de ces critères, se traduit par exemple par le règlement des licences d'usage, les charges de codage d'information, d'intégration de données puis de leur diffusion, de mise en œuvre et mise à jour d'alignements, de traduction,…

6.5.2. Respect de la norme internationale ISO 11238[142]

La norme ISO 11238 (*Informatique de santé — Identification des médicaments — Éléments de données et structures pour l'identification unique et l'échange d'informations réglementées sur les substances*) a été élaborée par le Comité technique ISO/TC 215, Informatique de santé.

Cette norme internationale répond à une demande mondiale de spécifications sur les médicaments, afin de les harmoniser au niveau international. Il fait partie d'un groupe de cinq normes qui constituent, ensemble, la base de l'identification unique des médicaments.

Pour atteindre les objectifs principaux de la réglementation des médicaments et de la pharmacovigilance, il est nécessaire d'échanger des informations sur les médicaments d'une manière robuste et fiable. La norme ISO 11238 fournit une structure qui permet d'établir et d'assurer la maintenance d'identifiants uniques pour toutes les substances entrant dans la composition des médicaments ou des matériaux de conditionnement contenant les médicaments.

6.5.3. Quelles applications et cas d'usage ?

De manière générale, les réutilisations de données ouvertes à des fins économiques sont, à ce jour, marginales. *Le modèle économique reste à construire, reconnaît Jean-Pierre Bailly, directeur des ressources numériques de Nantes métropole. L'Open Data sert surtout à certains développeurs ou sociétés, pour promouvoir leur savoir-faire* [14, VILLE].

229 applications ont été réalisées en France[143] à partir des données publiques, sans que l'on sache combien sont rentables aujourd'hui. La réutilisation économique reste cloisonnée par territoire et entravée par des difficultés de comparaison, la standardisation des données (licences, formats, contenus) devant encore progresser. Leur qualité reste inégale et en quantité insuffisante : les coûts engagés au niveau de la production, associés à une gratuité de diffusion sont-ils compatibles pour avoir des résultats satisfaisants ?

La question peut se poser pour l'ontologie française des médicaments ; un retour sur investissement doit être dégagé à un moment donné pour faire sa place et perdurer. En termes de clients potentiels, l'ANSM est envisagée, car cela lui permettrait de mettre à disposition un fichier cohérent, satisfaisant les demandes de data.gouv.fr. D'autres acteurs de la e-Santé peuvent également être ciblés, tels que le SNIIRAM afin de les aider à analyser des données hospitalières ou plus largement de santé : un référentiel abouti sur les médicaments serait un élément d'amélioration dans leur processus d'étude et structuration de données. Les acteurs de la e-Santé

[142] https://www.iso.org/obp/ui#iso:std:iso:11238:ed-1:v1:fr
[143] http://www.opendatafrance.net/wp-content/uploads/2015/06/ic15.pdf

au sens large pourraient trouver beaucoup d'intérêt à exploiter une ressource normalisée et disponible dans les formats du Web sémantique.

La force de l'ontologie réside dans son utilisation des *principes du Web de données*, à l'instar des *récentes évolutions de l'algorithme du leader des moteurs de recherche* [26, CHARLET]. De plus, s'il s'avère que le modèle que nous avons choisi de mettre en œuvre est suffisamment formel pour donner vie aux connaissances qu'il gère, que celles-ci puissent circuler, s'échanger, se compléter avec d'autres ; des applications multiples sont donc possibles.

Pour terminer ce mémoire, un cas d'usage est décrit afin d'illustrer par une réalisation existante, l'utilisation d'une ontologie dans le domaine de la santé, à partir d'un référentiel sur les médicaments. Cette application figure sur le site du W3C[144], au sein d'une liste approuvée et mise à jour par le consortium.

Ce cas concerne le domaine de la pharmacovigilance aux Etats Unis : il s'agit presque d'une obsession des américains, qui se rendent sur les sites internet de santé plusieurs fois par mois, pour connaitre les effets secondaires des médicaments ou avoir des informations sur les interactions médicamenteuses. En effet, plus de 200 000 morts chaque année résultent d'accidents médicamenteux aux Etats Unis. Les trois-quarts de ces morts sont supposés évitables grâce à la connaissance des effets et interactions des médicaments entre eux.

Un critère simple est insuffisant pour guider le choix d'un régime plus sûr; par exemple, si un médicament présente une incidence légère d'inflammation de la vessie, il peut engendrer un risque plus important de saignement du tube digestif; un autre médicament réduit le risque du saignement abdominal, mais présente un risque plus grand de l'inflammation de la vessie. D'autres facteurs doivent être pris en compte, tels que l'efficacité d'une thérapie pour un cas donné, les caractéristiques favorables à un patient, comme un dosage peu fréquent ou la petite taille de pilule, ou encore le fait que la plupart des personnes de 40 ans et plus absorbent plusieurs médicaments quotidiennement.

L'application PharmaSURVEYOR[145] permet la navigation dans cet espace complexe et identifie le meilleur compromis pour chaque patient. Le PharmaSURVEY fournit en sortie 2 types de profils optimisés, avec les meilleures options pour le patient. Les deux profils améliorent significativement la sécurité du régime de médicament actuel, avec une élimination du risque d'un effet indésirable, ou au moins la minimisation du risque. Cependant le patient et son médecin font face à un choix : quel est le plus important entre éliminer le risque de palpitations modérées du cœur et éliminer le risque de la transpiration excessive mineur, c'est un compromis inévitable dans le traitement.

Cette application a été développée grâce à une ontologie pour spécifier des conditions médicales : les langages RDF et OWL fournissent un formulaire normalisé, interopérable pour cette ontologie, permettant son développement collaboratif, sa maintenance et l'accès à des applications diverses. Elle permet le partage de données au sein d'une communauté de patients collaborant, de médecins, de pharmaciens, de chercheurs. Il est donc crucial que les données disponibles soient accessibles et significatives pour chacune de ces sous-communautés diverses. Les technologies de Web sémantique satisfont à ces deux exigences.

144 https://www.w3.org/2001/sw/sweo/public/UseCases/
145 https://www.w3.org/2001/sw/sweo/public/UseCases/PharmaSurveyor/

On retrouve les avantages évoqués dans les chapitres précédents pour l'utilisation d'une ontologie :

- accès Web aux ressources permises par le nommage à base d'URI ;
- non-ambiguïté grâce au langage formalisé OWL ;
- format de triplet permettant l'évolution dynamique des schémas ;
- échange de connaissance facilité par RDF et OWL.

Les liens à SNOMED CT et d'autres vocabulaires médicaux existants sont planifiés, ce qui permettra d'aller au-delà d'une simple taxonomie, en fournissant des relations latérales et hiérarchiques, et un accès aux conditions granulaires via des classifications multiples.

6.5.4. Quelles conclusions sur les applications des ontologies médicales

En réalisant les recherches nécessaires à la rédaction de ce mémoire, nous avons pu constater que les idées d'applications mettant en jeu les ontologies, fleurissent dans le domaine de la santé. Des chercheurs comme Jean Charlet œuvrent continuellement sur de nouvelles méthodes et modes d'utilisation des ontologies, souvent en rapport direct et nécessaire avec les thésaurus et terminologies déjà existantes dans le domaine médical, mais aussi des outils de traitement automatique du langage (TAL) pour tenir compte du *caractère linguistique de la médecine*. On peut citer par exemple les projets Ontopneumo[146] ou LERUDI[147], ce dernier utilisant OntolUrgences[148], dans lesquelles on cherche à être informatiquement opérationnel, et donc à impliquer les experts du domaine dans une étroite collaboration pour y parvenir.

L'apport des ontologies est avéré pour développer des applications permettant de gérer de manière fiable et efficace la masse d'informations constituant un domaine, et en retirer le contenu pertinent, avec tous ses liens, grâce aux alignements et réseaux sémantiques. Elle devient un outil d'aide et d'analyse pour le professionnel qui a ainsi à sa disposition l'ensemble des données utiles à son diagnostic.

---

146 https://bioportal.bioontology.org/ontologies/ONTOPNEUMO

147 http://esante.gouv.fr/actus/services/le-projet-lerudi-fiche-signaletique

148 CHARLET Jean, DECLERCK Gunnar, DHOMBRES Ferdinand, GAYET Pierre, MIROUX Patrick, *et al.*. Construire une ontologie médicale pour la recherche d'information : problématiques terminologiques et de modélisation. 23es journées francophones d'Ingénierie des connaissances, Jun 2012, Paris, France. pp.33-48, 2012, IC 2012. <hal-00717807>

Conclusion

Aujourd'hui l'innovation d'une entreprise ou d'une organisation au sens large ne peut être appréhendée de manière auto centralisée, en ne comptant que sur ses propres ressources. L'ouverture des données publiques ou privées constitue une rupture du modèle classique et oblige à revoir l'ensemble des problématiques politiques, économiques, sociétales, organisationnelles, techniques, juridiques,…

L'État français, poussé par les initiatives lancées dans d'autres pays européens ou internationaux, s'est engagé dans cette voie. Il a aujourd'hui pris des engagements de transparence et mis en place un certain nombre d'actions concrètes (création d'Etalab par exemple) pour suivre le mouvement de mise à disposition des données publiques.

Ce chantier de longue haleine avance, soutenu par de nouvelles lois (projet de loi pour une République numérique), et a permis de délivrer des jeux de données circulant librement, sous licence gratuite, avec la possibilité d'être réutilisés et ainsi valorisés. Cependant, le processus est loin d'être global et reste très perfectible : la production de ces données nécessite beaucoup de transformations pour les rendre brutes (et non plus primaires), de qualité et dans des formats de diffusion interopérables. Il s'agit d'un réel bouleversement au sein des administrations qui doivent analyser et restructurer leurs systèmes d'information, souvent lourds et complexes, mais également leurs habitudes et organisation interne. Sans oublier le coût induit par tous ces changements, et un retour sur investissement pas forcément immédiat ou comptable.

Les données de santé représentent un des domaines les plus complexes de mise en œuvre de l'*Open Data*, avec de nombreux enjeux (santé publique, information des citoyens, gestion administrative), un nombre important d'acteurs et de professionnels de santé, organismes institutionnels ou privés, chercheurs (etc) qui n'ont pas les mêmes besoins fonctionnels ou attentes économiques. Les horizons sont donc différents, et une gouvernance concernant l'accès à l'information doit se développer pour donner plus de cohérence à l'ensemble : il s'agit d'harmoniser les échanges, d'augmenter et d'améliorer la collaboration, avec l'objectif principal d'un meilleur suivi du patient.

La problématique posée en début de chapitre 5, consistant à analyser comment contribuer localement à l'avancement global du travail d'ouverture des données dans le domaine particulier de la santé, a abouti à un certain nombre de conclusions, illustrées à travers le cas pratique de l'ontologie française des médicaments. Il a fallu dans un premier temps se rendre compte des besoins attendus en termes de ressources médicales françaises de référence, au sein de la communauté internationale. Puis, étudier les sources de données françaises disponibles dans le cadre de l'*Open Data*, les évaluer et en déduire comment nous pouvions les utiliser (récupération, nettoyage, intégration). De même, les modèles conceptuels ont fait l'objet de recherches et d'investigations, sachant qu'ils sont au cœur de la problématique et la clé de voute du système. Sans modèle associé, les données n'ont aucun sens. Les technologies du Web sémantique et du Web de données correspondent à l'autre pilier, sur lequel on s'appuie aujourd'hui, pour mettre en œuvre de manière universelle et standardisée, les moyens d'accès, de partage et de collaboration aux données ouvertes.

L'appui d'une démarche projet est indispensable pour accompagner le cycle de vie des processus mis en jeu, de leur identification à leur pilotage, en passant par leur formalisation. Les démarches de co-construction et d'avancement par lots permettent d'ajuster au fur et à mesure la réponse au juste besoin, de corriger certaines interprétations grâce aux retours d'utilisateurs, et ainsi de valider le projet par étape. Cela se traduit pour les ontologies par la méthode ascendante ou *bottom-up*, par exemple, qui se nourrit de l'usage pour l'implémentation de nouvelles règles. L'amélioration en continu œuvre à pérenniser et à augmenter l'utilisabilité, et donc les retours potentiels, et ainsi de suite… Un aspect peu abordé ici concernant la communication et le fait de se rendre visible, est un sujet à part entière, et déterminant pour le passage de projet à produit.

Ce qui peut être retiré de notre questionnement sur l'*Open Data* en général, est que l'ouverture des données constitue la partie visible de l'iceberg, et que derrière, l'ouverture doit se situer à tous les niveaux pour assister à un réel développement de l'intelligence économique : ouverture des modèles, ouverture de l'innovation (voir l'entreprise comme élément de l'écosystème et non plus comme entité totalement indépendante), ouverture des systèmes physiques et techniques, et … ouverture des « esprits » ! Mais aussi que sans les avancées offertes par le Web sémantique et les données liées, par le biais des ontologies par exemple, l'exploitation de ces quantités d'informations ne pourrait atteindre les capacités de traitement et les résultats observés aujourd'hui [58, NOYER].

Bibliographie

Historique, fondamentaux, enjeux, qualité, production

[1] GUILLAUD Hubert. Open Data (1/4) : Où en est-on ? [En ligne]. 30 mai 2012. Disponible sur : < http://www.internetactu.net/2012/05/30/open-data-14-ou-en-est-on/>

[2] MESZAROS Branislav, SAMATH Sitthida, GUERIN-HAMDI Sonia, FAURE Céline. Livre blanc sur les données ouvertes. [En ligne]. 2015. Disponible sur : <https://halshs.archives-ouvertes.fr/halshs-01162692/document>

Ce livre blanc assez récent, propose un état des lieux et une analyse complète sur l'ouverture des données.

Il en donne un historique, définit les fondamentaux et les enjeux, indique le rôle important des choix politiques et de la règlementation, les licences permettant la réutilisation. La question des formats et standards est également détaillée, au niveau production et distribution des jeux de données, ainsi que les technologies et concepts associés au niveau du Web sémantique et des linked data. La question de la qualité de données est par conséquent récurrente tout au long de ce document, source précieuse d'information pour notre sujet.

[3] Le portail «Données ouvertes» de l'Union européenne. Office des publications de l'Union européenne, [en ligne] < http://data.europa.eu/euodp/fr/data/>

Le portail «Données ouvertes» de l'Union européenne est un point d'accès unique aux données produites par les institutions et organes de l'Union européenne (UE). Ces données peuvent être utilisées et réutilisées gratuitement à des fins commerciales ou non. Il est géré par l'Office des publications de l'Union européenne et la mise en œuvre de la politique en matière de données ouvertes de l'UE incombe à la direction générale «Réseaux de communication, contenu et technologies» de la Commission européenne.

Le portail propose un catalogue de métadonnées qui donne accès à des données des institutions et organes de l'UE. Pour faciliter la réutilisation, ces métadonnées s'appuient sur des règles d'encodage communes et des vocabulaires normalisés[149]. Les données sont disponibles dans des formats interprétables par l'utilisateur ou par l'ordinateur, pour une utilisation immédiate.

Il est également possible de participer de plusieurs manières, en suggérant des jeux de données, en faisant part de commentaires et suggestions ou en partageant des applications, via le formulaire de contact.

[4] ERTZSCHEID Olivier. Des datas, des GAFAs, et (peut-être) de l'emploi (France Culture). Affordance.info, ISSN 2260-1856, [En ligne]. 27 juin 2016. Disponible sur : <http://affordance.typepad.com//mon_Weblog/2016/06/france-culture-data-gafa-emploi.html>

[149] http://data.europa.eu/fr/linked-data

Le blog Affordance livre de nombreuses publications et articles d'analyse, de réflexion et d'avis sur les sciences de l'information, leur contexte, évolution,... Les sujets sont abordés dans toutes leurs dimensions, et l'auteur communique son savoir, ses connaissances et opinions de manière pragmatique et illustrée, non sans un certain humour. Il est question dans cet article de la place des algorithmes dans l'utilisation des données ouvertes.

[5] Projet du gouvernement porté par Jean-Vincent Placé, Axelle Lemaire. L'ouverture des données publiques. gouvernement.fr, [en ligne] <http://www.gouvernement.fr/action/l-ouverture-des-donnees-publiques>

[6] CHIGNARD Simon. 5 ans d'Open Data: qu'avons-nous appris ?. donneesouvertes.info, [en ligne]. 15 juin 2016. Disponible sur : <https://donneesouvertes.info/2016/06/15/5-ans-dopen-data-quavons-nous-appris/>

Données ouvertes est le blog de Simon Chignard, qui a participé dès 2010 à l'animation de l'ouverture des données publiques de Rennes Métropole, territoire pionnier en France. Conférencier, il intervient régulièrement sur le sujet et a publié plusieurs articles sur l'approche politique, sociale et économique de l'Open Data.

[7] Portail du Libre accès à l'information scientifique et technique. INIST, [en ligne] <http://www.inist.fr/>

Site internet du mouvement « Libre accès à l'information scientifique et technique » de l'INIST[150] (Institut de l'information scientifique et technique) du CNRS, on y trouve des informations d'actualité sur le mouvement libre accès (brèves, réalisations, débats, interrogations, études et prises de position de la part des acteurs impliqués dans la communication scientifique), ainsi que l'historique du mouvement[151], les textes de référence[152], les problématiques, manifestations sur le sujet.

[8] Portail L'École nationale supérieure des sciences de l'information. ENSSIB, [en ligne] < http://www.enssib.fr/>

L'École nationale supérieure des sciences de l'information et des bibliothèques (ENSSIB) s'appuie sur des enseignants-chercheurs et des doctorants mais également sur des conservateurs des bibliothèques, des étudiants et des professionnels, afin que ses activités scientifiques et de recherche constituent un espace de dialogue entre les communautés académiques et professionnelles.

[9] La Gazette des communes. Réutilisation des données publiques : des promesses vertigineuses. lagazettedescommunes.com, [En ligne]. 25 avril 2016. Disponible sur : <http://www.lagazettedescommunes.com/dossiers/reutilisation-des-donnees-publiques-des-promesses-vertigineuses/>

Lagazette.fr[153] est le site internet des fonctionnaires territoriaux, qui donne accès à toute l'actualité de la Fonction Publique Territoriale et des collectivités locales.

Il met à disposition des informations touchant à la fois la carrière dans la fonction publique, des textes officiels et dossiers juridiques régissant le droit des

[150] http://www.inist.fr/?Presentation&lang=fr
[151] http://openaccess.inist.fr/?Initiative-de-Budapest-pour-l
[152] http://openaccess.inist.fr/?-Textes-de-references
[153] *Editeur Infopro Digital (http://www.infopro-digital.com/)*

collectivités, mais aussi sur des sujets au cœur des débats actuels tels que l'Open Data. On y trouve aussi des dossiers complets sur le sujet sur la réutilisation des données publiques, ses implications et constats observés.

Ce dossier contient les articles suivants référencés dans le mémoire :

[10] BLANC Sabine. Le fouillis des licences Open Data s'éclaircit. lagazettedescommunes.com, [En ligne]. 26 novembre 2013. Disponible sur : <http://www.lagazettedescommunes.com/208893/le-fouilli-des-licences-open-data-seclaircit-fiche-pratique/>

[11] FAUVEL Virginie. Open Data : comment les collectivités s'y mettent. lagazettedescommunes.com, [En ligne]. 28 mai 2013. Disponible sur : <http://www.lagazettedescommunes.com/151408/open-data-comment-les-collectivites-sy-mettent/>

[12] FAUVEL Virginie. « L'Open Data est un révélateur violent de l'inadaptation de l'administration » – Denis Berthault, expert Open Data. lagazettedescommunes.com, [En ligne]. 10 octobre 2013. Disponible sur : <http://www.lagazettedescommunes.com/198393/lopen-data-est-un-revelateur-violent-de-linadaptation-de-ladministration-denis-berthault-expert-open-data/>

[13] CONTE Pierre-Alexandre. Gouvernance par la donnée, année zéro. lagazettedescommunes.com, [En ligne]. 21 janvier 2016. Disponible sur : <http://www.lagazettedescommunes.com/426499/gouvernance-par-la-donnee-annee-zero/>

[14] VILLE Frédéric. Open Data : le service public augmenté – 1. Cap sur la réutilisation. lagazettedescommunes.com, [En ligne]. 08 mars 2013. Disponible sur : <http://www.lagazettedescommunes.com/158021/open-data-le-service-public-augmente-%E2%80%93-1-cap-sur-la-reutilisation/>

[15] PERES Eric. Les données numériques : un enjeu d'éducation et de citoyenneté: Avis du Conseil économique, social et environnemental. lecese.fr, [En ligne]. Adopté le 13 janvier 2015. Disponible sur : <http://www.lecese.fr/travaux-publies/les-donnees-numeriques-un-enjeu-deducation-et-de-citoyennete >

Troisième assemblée constitutionnelle de la République après l'Assemblée nationale et le Sénat, le Conseil économique, social et environnemental (CESE) favorise le dialogue entre les différentes composantes de la société civile et les organes politiques. Le CESE regroupe des membres désignés par les principales associations de France, les syndicats de salariés, les organisations patronales et de nombreux acteurs de la société civile. Patrick Bernasconi est le président du CESE depuis le 1er décembre 2015.

Une saisine désigne le sujet et constitue le point de départ du travail des membres du CESE. Le Conseil peut être saisi par le Gouvernement, le Parlement, par voie de pétition citoyenne ou bien s'autosaisir d'une thématique afin de rendre un avis, une étude ou une résolution sur le sujet. Le CESE produit ainsi entre 25 et 30 avis par an.

Cet avis « dont le Conseil économique, social et environnemental a été saisi par décision de son bureau en date du 13 mai 2014 en application de l'article 3 de l'ordonnance no 58-1360 du 29 décembre 1958 modifié portant loi organique relative au Conseil économique, social et environnemental. Le bureau a confié à la section de l'éducation, de la culture et de la communication la préparation d'un avis intitulé : Les données numériques : un enjeu d'éducation et de citoyenneté. La section de

l'éducation, de la culture et de la communication, présidée par M. Philippe Da Costa, a désigné M. Eric Peres comme rapporteur. » Adopté le : 13/01/2015 | Mandature : 2010-2015

[16] RenaissanceNumérique. DEMOCRATIE : Mise à jour. Renaissancenumerique.org, [En ligne]. 18 avril 2016. Disponible sur : <http://www.renaissancenumerique.org/publications/rn/792-2016-04-18-08-25-24>

Renaissance Numérique réunit les grandes entreprises de l'Internet, françaises et multinationales, les entrepreneurs, les universitaires ainsi que les représentants de la société civile, pour participer à la définition d'un nouveau modèle économique, social et politique issu de la révolution numérique.

Paru le 18 avril 2016, ce rapport présente les opportunités qu'offre le numérique pour faire évoluer l'Etat et la démocratie vers davantage de transparence et de représentativité. Il se veut opérationnel, participatif, avec la volonté de faire bouger les inerties, et présente 13 propositions issues d'une réflexion collaborative, illustrées par des initiatives locales innovantes.

[17] MULLER Catherine. Bibliothèque, open science, Open Data et données de la recherche au Canada : quels enjeux ? - Par Alexandre Tur. enssib.fr, [En ligne]. 21 août 2015. Disponible sur : < http://www.enssib.fr/recherche/enssiblab/les-billets-denssiblab/bibliotheque-de-recherche-open-access-open-data-gestion>

Billet tiré d'ENSSILAB[154], service de l'Enssib depuis 2013, qui regroupe l'ensemble des projets numériques innovants, ayant en commun une dimension recherche et développement. Il explique clairement les différents concepts Open Science et Open Data, leur essence et politiques associées, afin de clarifier les enjeux des débats autour de ces sujets.

[18] TRANGER Hervé. L'Open Data : enjeux & opportunités. bva.fr, [En ligne]. 21 novembre 2011. Disponible sur : <http://www.bva.fr/fr/news/parole_d_expert/l_open_data_enjeux_opportunites.html>

Cette étude menée par BVA Healthcare en collaboration avec Bluenove auprès de décideurs au sein de grandes entreprises (responsables innovation, directeur marketing & communication…) sur tous les secteurs d'activité, montre que même si près de la moitié de l'échantillon interrogé (43%) se déclare ouvert à l'Open Data, l'autre moitié semble encore réticente et n'en perçoit pas l'intérêt. Ce résultat conduit à penser que, soutenu par une communication plus importante, l'Open Data aurait les capacités à gagner fortement en notoriété, et donc en potentiel.

[19] GRUTTEMEIER Herbert, HAMEAU Thérèse. Accès aux données scientifiques et contraintes juridiques – une question d'équilibre. », I2D – Information, données & documents 2016/2 (Volume 53), p. 20-22.

Article concernant l'aspect droit de l'information pour les données ouvertes, qui relève bien l'importance du domaine public pour y placer les données, et les licences associées.

[154] http://www.enssib.fr/enssibLab/presentation

[20] DELAYAT Régis. Comment gérer les données de l'entreprise pour créer de la valeur ? cigref.fr, [En ligne]. 30 octobre 2014. Disponible sur : <http://www.cigref.fr/rapport-cigref-enjeux-business-des-donnees>

Le CIGREF[155] est un réseau de Grandes Entreprises qui regroupe plus de 140 grandes entreprises et organismes français[156]. C'est « un carrefour d'informations, de réflexions, d'échanges et d'orientations sur l'entreprise au cœur du monde numérique », dont la mission pour 2020 est de « développer la capacité des grandes entreprises à intégrer et maîtriser le numérique ».

Ce rapport CIGREF propose une méthodologie de gestion des données, *est complété par un outil d'auto-évaluation[157] de la maturité des entreprises en matière de gestion des données. Il aborde donc largement la problématique de la qualité d'un modèle de données, et comment prévenir certaines failles dans sa constitution ou son audit.*

[21] DENIS Jérôme, GOETA Samuel. "Brutification" et instauration des données. La fabrique attentionnée de l'Open Data. i3WP_16-CSI-01.pdf. i3WP_16-CSI-01.pdf. 2016. Disponible sur : <https://hal.archives-ouvertes.fr/hal-01347301>

Résultat d'une enquête de deux ans menée dans plusieurs institutions françaises, cet article décrit tout le travail de préparation des données publiques avant leur publication et réutilisation de manière ouverte. Il parcourt l'ensemble des étapes du cycle de vie de ces données, c'est à dire leur identification, leur extraction des systèmes d'information puis leur transformation pour être enfin utilisables.

On comprend ainsi les difficultés rencontrées face à l'hétérogénéité, et au lien historique profond avec les métiers d'où ces données sont issues. Le terme de « brutification » sous-entend bien que les données brutes attendues dans la définition de l'Open Data ne correspondent pas aux données primaires, comme on peut l'imaginer, et que cela engendre un certain nombre d'investissements humains, et donc un coût d'analyse, d'enquête et de mise en œuvre.

Web sémantique

[22] AUSSENAC-GILLES Nathalie, CHARLET Jean, REYNAUD Chantal. Les enjeux de l'Ingénierie des connaissances. F. Sedes, J.M. Ogier, P. Marquis. Information, Interaction, Intelligence - Le point sur le i(3), Cepadues, 2012, 97823649300094. Disponible sur : <https://hal.archives-ouvertes.fr/hal-00787425>

Ce chapitre d'ouvrage aborde de manière pointue l'univers de l'Ingénierie des Connaissances et ses apports fondamentaux dans les systèmes informationnels.

Les auteurs insistent sur l'importance de la conception et modélisation formelle, préalable nécessaire pour avoir une représentation cohérente, voire philosophique des choses, et donc pour pouvoir en tirer profit, et gérer ainsi le savoir dans des buts opérationnels.

---

155 http://www.cigref.fr/qui-sommes-nous

156 http://www.cigref.fr/qui-sommes-nous/entreprises-membres

157 http://www.cigref.fr/wp/wp-content/uploads/2014/10/CIGREF-Maturite-Gestion-donnees-outil-auto-evaluation-2014.xlsx

Ils exposent ensuite la façon dont le Web sémantique et le Web de données contribuent fortement à définir des modèles de connaissances de plus en plus aboutis et permettant des raisonnements, avec une focalisation sur les ontologies.

[23] BIZER, HEATH C., BERNERS-LEE T. Linked Data – The Story so far. Special Issue on Linked Data in International Journal on Semantic Web and Information Systems, vol. 5, n° 3, 2009, pp. 1-22 doi:10.4018/jswis.2009081901. Disponible sur : <http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf>

Tim Berners-Lee, directeur du W3C, a inventé et défini le terme Linked Data et son synonyme Web of Data.

Les Principes du Web de Données[158] (par Tim Berners-Lee) :

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs. so that they can discover more things.

Référence incontournable du Web sémantique et Web de données pour en comprendre les concepts et principes techniques.

[24] The World Wide Web Consortium (W3C), w3.org. [En ligne]. Disponible sur : <https://www.w3.org/Consortium/>

The World Wide Web Consortium (W3C) est une communauté internationale qui a développé les standards ouverts pour assurer le développement du Web. Fondé et mené par l'inventeur du Web Tim Berners-Lee en octobre 1994, dirigé par le CEO Jeffrey Jaffe, il est chargé de promouvoir la compatibilité des technologies du World Wide Web telles que HTML5, HTML, XHTML, XML, RDF, SPARQL, CSS, XSL, PNG, SVG et SOAP.

Parmi tous les services proposés par le W3C (tutoriels, outils,…), on trouve aussi des exemples d'applications, les « Semantic Web Case Studies and Use Cases », comme l'exemple « Composing a Safer Drug Regimen for each Patient with Semantic Web Technologies » [159] qui a servi d'illustration dans ce mémoire.

[25] PRIME-CLAVERIE Camille, KEMBELLEC Gérald. Web de données et création de valeurs : le champ des possibles. Revue I2D – Information, données & documents, 2016/2 (Volume 53), p. 28-29, [En ligne]. 30 juin 2016. Disponible sur : <http://www.adbs.fr/b-dossier-Web-de-donnees-et-creation-de-valeurs-le-champ-des-possibles-b--156691.htm?RH=1426693578415 >

Ce numéro aborde l'ensemble des facettes du Web de données, de l'emploi et des métiers qu'il engendre, aux méthodes, techniques et outils qui le façonnent, en passant par le droit de l'information qui lui est propre. Il met ensuite en lumière les enjeux et la création de valeur dans différents domaines de l'information, tels que les bibliothèques, les secteurs de la culture et de la recherche, et se focalise enfin sur les usages métiers dans différents contexte, ainsi que sur les ouvertures rendues possibles en recherche d'information.

---

158 http://www.w3.org/DesignIssues/LinkedData.html
159 https://www.w3.org/2001/sw/sweo/public/UseCases/PharmaSurveyor/

L'ensemble du dossier est très riche, et la courte sélection de ressources suivantes s'explique par son ciblage par rapport au sujet du mémoire.

[26] CHARLET Jean, KEMBELLEC Gérald. Du Web sémantique au Web des données, quels enjeux professionnels ?. Revue I2D – Information, données & documents 2016/2 (Volume 53), p. 54-55. 30 juin 2016.

Article qui introduit le questionnement et les réponses apportées par l'ensemble des articles du dossier, sur le positionnement du Web de données au niveau de ses valorisations concrètes dans le monde industriel, les implications observées pour les moteurs de recherches et autres applications utilisant les schémas de données.

[27] MENON Bruno. Comprendre les standards du Web de données. Revue I2D – Information, données & documents 2016/2 (Volume 53), p. 32-34. 30 juin 2016.

Article très didactique sur l'ensemble des technologies, formats et composants du Web de données, car il pose bien les bases et concepts mis en jeu, et permet donc d'en saisir le fonctionnement et la manière d'en tirer parti.

[28] PRIME-CLAVERIE Camille. Linked science et Web de données. Revue I2D – Information, données & documents 2016/2 (Volume 53), p. 42-43. 30 juin 2016.

Article qui pose les enjeux de l'e-science et comment l'apport du Web de données enrichit les contenus scientifiques, grâce à l'ouverture et à la réutilisation des données, mais également à leur mise en relation (données liées), et introduit les perspectives que l'on peut en attendre dans le monde scientifique.

[29] LE PICARD Anne-Claire. OntoToxNuc : recherche d'information et enjeux pour la collaboration et la valorisation. Revue I2D – Information, données & documents 2016/2 (Volume 53), p. 59-59. 30 juin 2016.

Article qui, à partir d'un cas concret de mise en œuvre d'une ontologie sur la toxicité nucléaire, éclaire sur les enjeux d'une sémantique commune entre divers référentiels métiers d'un domaine, et les apports en termes de choix techniques, d'usage. Les perspectives de valorisation offertes grâce au Linked Open Data pour cette application, sont encore pleines de promesses.

[30] CHARLET Jean. Données liées et documentation médicale. Revue I2D – Information, données & documents 2016/2 (Volume 53), p. 56-58. 30 juin 2016.

Article qui met en lumière comment un domaine tel que celui de la médecine, très riche sémantiquement et relationnellement parlant, peut tirer profit du Web de données et des ontologies, en tant que Systèmes Organisés de Connaissances (SOC). Il indique quels sont les freins, liés directement aux forces du domaine de la santé, qui sont la multitude des modèles existants et le fait qu'une ontologie doive coller à un cas d'usage pour être utilisable (elle ne peut être universelle). Pour un système global, il s'agit donc d'aligner et de lier les modèles entre eux afin d'obtenir un réseau, et donc de l'interopérabilité entre les données.

[31] KEMBELLEC Gérald. Bibliographies scientifiques: de la recherche d'informations à la production de documents normés. Sciences de l'information et de la communication. Université Paris VIII Vincennes-Saint Denis, p.271-273. 2012. Disponible sur : <https://tel.archives-ouvertes.fr/tel-00771553>

Cette thèse concerne plus particulier l'univers des bibliothèques, mais m'a semblée intéressante comme référence au niveau du chapitre 9, concernant l'urbanisation des systèmes d'information. En effet, soutenue en 2012, les constats et explications abordés sont toujours d'actualité, tels que les technologies de l'information, l'interopérabilité des systèmes, les formats normés, et la question du Web sémantique. Le point de vue côté bibliothèques donne une vision parallèle du Web de données dans un autre domaine d'application.

Secteur E-Santé

[32] Le LEEM. Stratégie nationale E-santé 2020 : il faut désormais passer à la vitesse supérieure selon l'Alliance E-Health France. leem.org, [En ligne]. 07 juillet 2016. Disponible sur : <http://www.leem.org/strategie-nationale-e-sante-2020-il-faut-desormais-passer-vitesse-superieure-selon-l-alliance-e-heal>

Le LEEM - Les entreprises du médicament, fédère les entreprises du médicament opérant en France, avec une mission de respect de la déontologie des pratiques professionnelles de l'industrie pharmaceutique (règles d'éthique de la profession, conformité aux codes de bonne conduite). Créé il y a plus de 130 ans, le LEEM compte aujourd'hui près de 270 entreprises adhérentes[160], qui réalisent près de 98 % du chiffre d'affaires total du médicament en France.

Le site du LEEM propose une rubrique complète réservée au médicament, dans laquelle sont détaillés l'économie du médicament, la description des maladies, les caractéristiques des médicaments, l'avancée de la recherche sur le médicament et la situation des français face aux médicaments.

[33] Portail esante.gouv.fr, ASIP Santé. Esante.gouv.fr, [En ligne].Disponible sur : < http://esante.gouv.fr/ >

Le portail esante.gouv.fr est porté par l'ASIP santé (Agence des Systèmes d'Informations Partagées de Santé).

Il s'articule autour de 6 thématiques d'information : les actualités de la e-santé, la tribune permettant aux acteurs de la e-santé de prendre la parole, un espace (MAG) pour prendre connaissance de la e-santé, les initiatives locales/régionales de la e-santé, les travaux / projets de l'ASIP Santé complétés par des ressources externes, les ressources sur les services proposés par l'ASIP Santé et la présentation générale de l'ASIP Santé (missions, recrutements et marchés publics).

Au niveau des services proposés, on trouve :

- *Les référentiels[161],*
- *Les hébergeurs de données de santé[162],*
- *L'espace des produits de certification[163],*
- *Les repères juridiques[164],*
- *L'annuaire des projets e-santé[165].*

160 http://www.leem.org/annuaire/carte
161 http://esante.gouv.fr/services/referentiels
162 http://esante.gouv.fr/services/referentiels/securite/le-referentiel-de-constitution-des-dossiers-de-demande-d-agrement-des
163 http://esante.gouv.fr/services/espace-cps/qu-est-ce-que-la-carte-cps
164 http://esante.gouv.fr/services/reperes-juridiques
165 http://esante.gouv.fr/bdd

[34] Esante.gouv.fr. Publication du rapport de la phase 2 – "Diagnostic" dans la cadre de l'étude sur la mise en œuvre de terminologies de référence pour le secteur santé-social. esante.gouv.fr, [En ligne]. 05 mai 2015.Disponible sur: <http://esante.gouv.fr/actus/interoperabilite/publication-du-rapport-de-la-phase-2-diagnostic-dans-la-cadre-de-l-etude-sur>

La Délégation à la Stratégie des Systèmes d'Information de Santé (DSSIS), Ministère des Affaires Sociales, de la Santé, et du Droit des Femmes, Travaille depuis fin 2013 sur la contribution française aux normes et standards internationaux en informatique pour la santé.

Les terminologies de référence ont été identifiées comme un sujet prioritaire dans un objectif d'interopérabilité sémantique, et la DSSIS a confié en mars 2014 à l'ASIP Santé la réalisation d'une « étude sur la mise en œuvre de terminologies de référence pour le secteur santé-social en France ».

Cette étude a été découpée en trois phases :

• La phase 1 – « Fondamentaux », réalisée de mars à septembre 2014.
• La phase 2 – « Diagnostic », réalisée d'octobre 2014 à avril 2015.[166] *, résultat d'une campagne d'interviews d'industriels, institutions et organisations du secteur de la santé.*
• La phase 3 – « Propositions » se déroulera sur le second semestre 2015 (document non disponible à ce jour-demande d'information à l'ASIP le 12/10/16).

Le document issu de la phase 2 a été très utile dans la rédaction de ce mémoire, car il aborde de nombreuses questions abordées, telles que la diversité des terminologies françaises de référence qu'il s'agit de maitriser, leur regroupement par cas d'usage, comment mieux contribuer à la standardisation et interopérabilité sémantique, leur protection et sécurité, ainsi que leur mode de financement. Les tendances d'autres pays sont également évoquées, tels que le Portugal, Royaume Uni, les Etats Unis, l'Allemagne et la Belgique.

Secteur de la culture et Royaume-Uni

[35] DUHAMEL Benjamin. Les technologies du Web sémantique et du record linkage au service de data.bnf.fr et du Linked Open Data culturel : étude sur les nouveaux paradigmes informationnels. Documentation. 2014. Disponible sur : <http://memsic.ccsd.cnrs.fr/mem_01081739/document>

[36] DHRZANOWSKI Pierre. Data.gov.uk, l'ouverture des données publiques au Royaume-Uni. ambafrance-uk.org, [En ligne]. 21 Août 2015. Disponible sur : <http://www.ambafrance-uk.org/data-gov-uk-l-ouverture-des>

Les missions du Service pour la Science et la Technologie (SST)[167] *de l'Ambassade de France à Londres se déclinent selon trois grands axes : veille scientifique et technologique (suivi et productions de documents), promotion des échanges et de la coopération scientifique et technologique bilatérale et multilatérale (organisation de colloques, ateliers, mise en place de partenariats institutionnels,…), promotion des réalisations scientifiques et technologiques françaises.*

166 http://esante.gouv.fr/sites/default/files/asset/document/20150504_etude_terminos_phase2_diagnostic.pdf

167 http://www.ambafrance-uk.org/-Le-service-

L'équipe est constituée de 7 personnes placées sous la responsabilité d'un conseiller pour la science et la technologie.

Contexte ontologie du médicament

[37] SAFON Marie-Odile, SUHARD Véronique. Historique de la politique du médicament en France Synthèse thématique. Synthèses et dossiers bibliographiques, irdes.fr, [En ligne]. Mars 2016. Disponible sur : <http://www.irdes.fr/documentation/syntheses-et-dossiers-bibliographiques.html>

L'IRDES, Institut de Recherche et Documentation en Economie de la Santé, observe et analyse l'évolution des comportements des consommateurs et des producteurs de soins à la fois sous l'angle médical, économique, géographique... La mise à disposition de l'information ainsi que la formation font également partie de ses missions.

L'IRDES s'adresse à la fois à un public large[168], par la diffusion de résultats d'études, de recherche ou d'enquêtes en matière d'économie de la santé, ainsi qu'à la communauté scientifique, par des travaux de recherche (Documents de travail de l'Irdes[169]) ayant pour ambition la publication dans des revues à comités de lecture et alimenter le débat scientifique. Les rapports de l'Irdes[170] publient les résultats complets d'études, de recherche ou d'enquêtes dans une collection de référence.

Outre le service de documentation[171], L'Irdes produit également les logiciels Eco-Santé[172], bases de données qui rassemblent des séries statistiques dans le domaine sanitaire et social.

Le dossier Historique de la politique du médicament en France donne la définition du médicament, son circuit de mise sur le marché en France et brosse un historique de l'évolution juridique de 2012 à 2016.

Modèles conceptuels

[38] CHOQUET Rémy. Partage de données biomédicales : modèles, sémantique et qualité. BioInformatique, Biologie Systémique [q-bio.QM]. Université Pierre et Marie Curie - Paris VI, 2011. Français. Disponible sur : <https://tel.archives-ouvertes.fr/tel-00824931/document>

Cette thèse étudie le processus visant à l'échange d'informations biomédicales, depuis l'extraction de données depuis leur lieu de stockage (modèle relationnel, graphe RDF) par requêtage plus ou moins complexe, la gestion de la qualité des données et de leur fiabilité, pour pouvoir ensuite les relier entre elles, une fois formalisées de manière sémantique.

La question des standards permettant l'interopérabilité dans le domaine de la santé est largement analysé (initiative HL7 v3 - Modèle Health Level Seven), tout en pointant le fait qu'un modèle totalement générique ne puisse pas répondre à la complexité du domaine médical.

168 http://www.irdes.fr/recherche/questions-d-economie-de-la-sante.html
169 http://www.irdes.fr/recherche/documents-de-travail.html
170 http://www.irdes.fr/recherche/rapports-et-articles.html
171 http://www.irdes.fr/documentation/actualites.html
172 http://www.irdes.fr/recherche/eco-sante/eco-sante.html

L'auteur détaille ensuite l'expérimentation effectuée dans le cadre du projet DebugIT pour développer une plateforme d'interopérabilité entre les langages de représentation sémantique, sans oublier les dimensions techniques et syntaxiques. Le but étant de gérer les différentes visions du monde et tirer parti de chacune d'entre elles sans altérer leur essence, spécialité et usage propres.

[39] LOUKIDES Mike. We need open models, not just Open Data. radar.oreilly.com, [En ligne]. 11 novembre 2014. Disponible sur : <http://radar.oreilly.com/2014/11/we-need-open-models-not-just-open-data.html>

O'Reilly Media diffuse sur son site connaissances et innovations autour de la révolution Internet, sous différentes formes, telles que ouvrages, services en ligne, magazines, conférences.

[40] O'NEIL Cathy. Let them game the model. mathbabe.org, [En ligne]. 03 février 2012. Disponible sur : <https://mathbabe.org/2012/02/03/let-them-game-the-model/>

Mathbabe est le Blog de Cathy O'Neil , data scientist depuis 2011 à New York, qui a construit des modèles prédictifs sur les achats et clics internet, et écrit des ouvrages tels que « Doing Data Science » en 2013 ou actuellement en cours « Weapons of Math Destruction » sur le côté « noir » des big datas.

[41] MedDRA. Support Documentation, MedDRA Version 19.1 Français. meddra.org, [En ligne]. septembre 2016. Disponible sur : <http://www.meddra.org/how-to-use/support-documentation>

[42] BODENREIDER O, PETERS LB. J Biomed Inform. A graph-based approach to auditing RxNorm. 2009 Jun;42(3):558-70. doi: 10.1016/j.jbi.2009.04.004. Epub 2009 Apr 24.PMID:19394440. Disponible sur : < http://www.ncbi.nlm.nih.gov/pubmed/19394440>

[43] HANNA, JOSH *et al.* Building a Drug Ontology Based on RxNorm and Other Sources. Journal of Biomedical Semantics 4 (2013): 44. PMC. Web. 12 Septembre 2016. Disponible sur : <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3931349/>

[44] AberOWL ontology repository and semantic search engine. RXNORM – RxNORM. aber-owl.net, [En ligne]. Disponible sur : <http://aber-owl.net/ontology/RXNORM>

[45] The National Center for Biomedical Ontology. RxNORM. bioportal.bioontology.org, [En ligne]. Mis à jour le 09 juin 2016. Disponible sur : <https://bioportal.bioontology.org/ontologies/RXNORM>

Sources de données de santé françaises

[46] Agence Nationale de Sécurité du Médicament et des produits de santé (ANSM). Répertoire des médicaments. ansm.sante.fr, [En ligne]. 2016. Disponible sur : <http://ansm.sante.fr/Services/Repertoire-des-medicaments >

Le répertoire représente une partie de l'information officielle sur les spécialités pharmaceutiques ayant obtenu une autorisation de mise sur le marché (AMM), qu'elles soient commercialisées ou non.

[47] BOIDIN Célestine. Création d'une base de données des médicaments européens comme support de l'activité de pharmacovigilance [Texte imprimé] : l'XEVMPD. sous

la direction de Bernard Gressier - Mémoire ou thèse (version d'origine).Français. 2013. Disponible sur : <http://www.sudoc.fr/174404212>

[48] CHAMBONNET Sébastien. Définir une architecture de l'information pour la sauvegarde du patrimoine scientifique et technique contemporain (Patstec), à l'heure du Web de données. Documentation. 2015. Disponible sur : <http://memsic.ccsd.cnrs.fr/mem_01309411>

[49] BANEYX A., CHARLET J. Évaluation, évolution et maintenance d'une ontologie en médecine: état des lieux et expérimentation. Revue I3 – Information, Interaction, Intelligence, numéro spécial « Corpus et ontologies ». 2007.

Dans cet article, les auteurs partent de l'exemple de leur création d'une ontologie de la pneumologie, pour analyser et évaluer son cycle de vie : répond-elle aujourd'hui aux besoins et attentes du domaine ? Pourra t'elle être aisément suivre les évolutions inhérentes à ce domaine et de quelle manière ?

Dans une première partie, ils définissent et contextualisent ce que sont les ressources terminologiques ou ontologiques (RTO), puis donnent les clés d'évaluation qu'ils ont définies et expérimentées sur leur ontologie. Ils indiquent des méthodes de vérification pratiques, des tests en situation d'utilisation, et le degré de réutilisabilité.

La seconde partie de l'article détaille le mode d'évaluation opérée sur l'ontologie de la pneumologie, son application et ses résultats par rapport aux différents critères définis.

Un certain nombre de conclusions et questionnements sont identifiées pour donner des axes de réflexion au lecteur.

[50] HUOT Charles, LEGENDRE Jean-François (rapporteur). Données massives - Big Data Impact et attentes pour la normalisation. Livre blanc. wikipe.has-sante.fr. Juin 2015. Disponible sur : <http://wikipe.has-sante.fr/WikiPE/PHP/Synthese.php?Param=Semantique>

WikiPE[173] est le wiki de la Prescription Electronique de la Haute Autorité de Santé, administré par le Service Evaluation de la Pertinence des soins et amélioration des Pratiques et des Parcours. Il y est bien précisé cependant que « les données qui s'y trouvent ne sont pas les références de la HAS ». On y trouve donc de nombreux documents en discussion, regroupés par thèmes, des dossiers et synthèses thématiques.

Le livre blanc[174] sur les données massives fait donc partie des ressources mises à disposition sur le site, et pose la problématique de contrôle, fiabilité, propriété intellectuelle des données publiques, ainsi que de l'interopérabilité pour la collecte, extraction et restitution de l'information. La normalisation volontaire est un moyen efficace d'y répondre.

Elaboré au nom du Comité stratégique AFNOR information et communication numérique, Il s'appuie notamment sur une étude qualitative des besoins de normalisation volontaire au moyen d'un questionnaire, diffusé en 2014 au sein de la communauté française du big data.

[173] http://wikipe.has-sante.fr/WikiPE/PHP/Plan.php
[174] http://wikipe.has-sante.fr/WikiPE/PHP/Multimedia.php?Concept=112779&langue=FRA

Un de ses éléments de réflexion concerne l'apport déterminant des technologies sémantiques, du fait des informations hétéroclites issues des référentiels métiers (banque, édition, santé, etc.) et des relations sémantiques (ontologies de domaine). Une absence de normes volontaires est constatée, et il serait utile de définir un langage de base unifié pour la requête.

Il existe certes le LOV (Linked Open Vocabulary), une sorte de catalogue d'ontologies de plus de 450 références (ontologies du tourisme, de la météo, de la santé, etc.), qui a l'atout d'être dynamique en vérifiant la disponibilité des vocabulaires, mais il ne fait l'objet d'aucune norme volontaire et n'est pas encore reconnu dans le cadre de l'ISO.

Une norme internationale est fortement souhaitée afin de cadrer l'architecture de référence et le vocabulaire du big data, avec six axes de développement identifiés : la gouvernance de la donnée, la qualité et l'identification, les données ouvertes (Open Data), les opérateurs d'infrastructures, les opérateurs de service et la normalisation technique.

[51] DUPUY-CHESSA Sophie, MARCAL DE OLIVEIRA Kathia, SI-SAID CHERFI Samira. Qualité des modèles : retours d'expériences. 32ème congrès Inforsid'2014, May 2014, Lyon, France. pp.363-378, 2014. Disponible sur : <https://hal.archives-ouvertes.fr/hal-01002995/document>

Cet article s'intéresse au modèle conceptuel de données du SI, en tant que formalisation des besoins et conformité du SI par rapport au domaine : le modèle sert à comprendre et représenter les systèmes de plus en plus complexes. Les auteurs nous livrent donc leur étude et analyse concernant la qualité des modèles, et les méthodes pragmatiques d'évaluation qu'ils ont pu élaborer ou relever dans des travaux existants.

Après une définition des niveaux de qualité dans un modèle, ils abordent les outils et approches disponibles. Ils expliquent également qu'ils ont procédé à des enquêtes pour cerner les attentes des différents publics en termes de qualité de modèle, et quels critères mesurer. Ils décrivent enfin les différentes expérimentations qu'ils ont réalisées ou explorées sur des problématiques concrètes (système expert dans le domaine de la cardiologie, modèles de navigation pour la conception de sites Web, modèles de processus métiers,…).

Les conclusions livrées montrent qu'une évaluation correcte de la qualité passe par l'implication des différents acteurs concernés (co-construction), l'utilisation d'outils automatisés et le recours à un langage de modélisation adapté.

[52] RICHARD Marion, AIME Xavier, KREBS Marie-Odile, CHARLET Jean. LOVMI : vers une méthode interactive pour la validation d'ontologies. 26es journées francophones d'Ingénierie des Connaissances (IC), Jul 2015, Rennes, France. 2015. Disponible sur : <https://hal.archives-ouvertes.fr/hal-01166359>

Dans cet article, les auteurs présentent la méthode LOVMI (Les Ontologies Validées par Méthode Interactive) pour la validation d'ontologies.

Ils montrent tout d'abord qu'une vérification automatique des ontologies est aujourd'hui nécessaire, celles-ci n'étant plus forcément créées par les spécialistes métiers, mais devant cependant conserver leur qualité de réutilisabilité et de maintenabilité.

Après avoir listé les critères participant à la validation d'ontologies, un état de l'art de différents outils de test ou de pratiques collaboratives est effectué, au niveau structurel puis sémantique.

Enfin, les auteurs s'attachent à expliquer le cas de leur ontologie OntoPsychia, et le processus de validation qu'ils ont adopté. Ils détaillent leurs choix de moyens et critères de vérification, à la fois basés sur l'humain et la technique, constituant la méthode LOVMI.

[53] Ministère des Affaires sociales et de la Santé. Médicaments. medicaments.gouv.fr, [En ligne]. Dernière mise à jour le 27 juillet 2016. Disponible sur : <http://social-sante.gouv.fr/soins-et-maladies/medicaments>

La base de données publique sur le médicament est un référentiel sur les médicaments actuellement commercialisés ou en arrêt de commercialisation depuis moins de trois ans en France. Les informations émanent des différentes institutions en charge du médicament, la Haute autorité de santé (HAS), la Caisse nationale d'assurance maladie des travailleurs salariés (CNAMTS) et l'Agence nationale de sécurité des médicaments et des produits de santé (ANSM).

[54] Université de Strasbourg. Académie Nationale de Pharmacie. dictionnaire.acadpharm.org, [En ligne]. Disponible sur : <http://dictionnaire.acadpharm.org/w/Acadpharm:Accueil>

Le dictionnaire rassemble l'ensemble des connaissances touchant le médicament et autres produits de santé, les sciences physico-chimiques en rapport avec ces domaines, la biologie, la santé publique, l'environnement et la santé, la diététique, la nutrition, la cosmétologie. Sont également décrits les symptômes majeurs des principales pathologies.

Interopérabilité

[55] MERABTI Tayeb. Méthodes pour la mise en relations des terminologies médicales : contribution à l'interopérabilité sémantique Inter et Intra terminologique. Ecole doctorale sciences physiques, mathématiques et de l'information pour l'ingénieur, U.F.R. des sciences et techniques. Université de Rouen, 2010. Français. Disponible sur : <http://www.sudoc.fr/151801991>

Cette thèse pose la problématique de la multiplication des terminologies existantes dans le domaine médical, et des impacts que cela engendre au niveau interopérabilité technique, sémantique, syntaxique.

Dans ce cadre, le projet de « Serveur Multi-Terminologique de Santé » (SMTS) est étudié et analysé, avec un focus sur les méthodes d'alignement entre les différentes terminologies médicales francophones, à intégrer et mettre en relation (Ophanet, ATC, CCAM), vers les terminologies francophones de l'UMLS (F_UMLS), considérée comme la plus grande base de données terminologique avec plus de 140 terminologies.

Pour enrichir les terminologies, une projection des relations SNOMED CT entre trois terminologies francophones (CIM10, SNMI et MeSH) a également été opérée.

[56] AUBAIN Diane. Les référentiels sémantiques dans l'interopérabilité des systèmes d'information de santé. researchgate.net, [En ligne]. Mars 2016. Disponible sur :

<https://www.researchgate.net/publication/305806042_Les_referentiels_semantiques_dans_l'interoperabilite_des_systemes_d'information_de_sante>

Cet article décrit l'étude menée sur les référentiels sémantiques des systèmes d'information de santé qui permettent aux différentes parties prenantes d'avoir un langage commun.

Leur intégration au sein de serveurs de terminologie permet d'aligner les différentes terminologies qu'ils hébergent.

Après avoir défini la notion de référentiel sémantique, son rôle et pourquoi les SI en santé nécessitent une interopérabilité à différents niveaux, l'auteure présente Mediboard, progiciel libre de Gestion Intégrée des établissements de santé, puis le fonctionnement des serveurs de terminologie. Elle met en relief la nécessité de coller au contexte du besoin pour la mise en place de ces serveurs, afin de limiter les erreurs (mauvaise traduction de concepts, par exemple) ou écarts avec les attentes fonctionnelles des usagers.

[57] AIME Xavier. Orphanet le portail des maladies rares et des médicaments orphelins. xavier-aime.com, [En ligne]. Projet OntoOrpha (2010-2013). Disponible sur : <http://xavier-aime.com/ontology/projets-de-recherche/projet-ontoorpha/>

[58] NOYER Jean-Marc, CARMES Maryse. Le mouvement " Open Data " dans la grande transformation des intelligences collectives et face à la question des écritures, du Web sémantique et des ontologies. archivesic.ccsd.cnrs.fr, [En ligne]. 2012. Disponible sur : <https://archivesic.ccsd.cnrs.fr/sic_00759618>

Les auteurs analysent le mouvement Open Data au niveau des organisations publiques et domaines d'application, dans lesquels il prend sa forme et se déploie, et font le lien avec le Web sémantique, et en particulier les ontologies. Sont évoqués des aspects de mise en œuvre opérationnels tels que le « data mining», mais également des questions stratégiques et d'organisation pour gérer la gouvernance de la donnée ouverte.

[59] LE PICARD Anne-Claire. Le cycle de vie d'une ontologie : évaluation de l'ontologie du domaine de la Toxicologie Nucléaire. domain shs.info.docu.2014. Disponible sur : <https://memsic.ccsd.cnrs.fr/mem_01128938/document>

Ce mémoire aide à bien comprendre en quoi consiste une ontologie, il pose les bases de son essence propre, de la place qu'elle occupe par rapport à d'autres outils de classification (thésaurus, taxinomie,…), et de ses conditions de mise en œuvre au niveau technique et économique. Le cas pratique de l'ontologie ToxNuc étudié permet d'illustrer les propos de manière très complète pour servir de référence à un projet de création, évaluation, valorisation et maintenance d'ontologie.

Annexes

Classement adopté pour les voies d'administration des médicaments :

'Voie cutaneo muqueuse'
- **'Voie cutanée et transdermique'**
  - 'Voie cutanée'
  - 'Autre voie cutanée ou transdermique'
  - 'Voie épilésionnelle'
  - 'Voie péri-aréolaire'
  - 'Voie sous-muqueuse'
  - 'Voie transcutanée'
  - 'Voie transdermique'
- **'Voie vaginale et intravésicale'**
  - 'Voie endocervicale'
  - 'Voie intracervicale'
  - 'Voie intra-murale'
  - 'Voie intra-utérine'
  - 'Voie intravésicale'
  - 'Voie urétrale'
  - 'Voie vaginale'
  - 'Autre voie vaginale ou intravésicale'
- **'Voie oculaire'**
  - 'Voie intra-camérulaire'
  - 'Voie intra-oculaire'
  - 'Voie intravitréenne'
  - 'Autre voie oculaire'
  - 'Voie ophtalmique'
  - 'Voie péribulbaire'
  - 'Voie péri-oculaire'
  - 'Voie rétrobulbaire'
  - 'Voie sous-conjonctivale'
- **'Voie pulmonaire'**
  - 'Voie endotrachéobronchique'
  - 'Voie inhalée'
  - 'Autre voie pulmonaire'
- **'Voie nasale et auriculaire'**
  - 'Voie auriculaire'
  - 'Voie endosinusale'
  - 'Voie nasale'
  - 'Autre voie nasale ou auriculaire'
  - 'Voie oropharyngée'

'Voie entérale'
- **'Voie orale et buccale'**
  - 'Voie buccale'
  - 'Autre voie orale ou buccale'
  - 'Voie bucco-dentaire'
  - 'Voie buccogengivale'
  - 'Voie buccogingivale'
  - 'Voie bucco-pharyngée buccale'
  - 'Voie dentaire'
  - 'Voie endocanalaire'
  - 'voie gastro-intestinale'
  - 'Voie gingivale'
  - 'Voie intragingivale'
  - 'Voie orale'
  - 'Voie perlinguale'
  - 'Voie sublinguale'

'Voie rectale'
'Voie gastrique'
'Voie gastro-entérale'
'Voie intestinale'
'Voie intra cholangio-pancréatique'
'Voie rectale'
'Autre voie rectale'

'Voie parentérale'

'Voie osseuse'
'Voie intracrânienne'
'Voie intra-osseuse'
'Autre voie osseuse'
'Voie péri-articulaire'
'Voie périneurale'
'Voie péri-osseuse'

'Voie sous-cutanée'
'Voie injectable sous cutanée'
'Voie intradermique'
'Voie intralymphatique'
'Voie sous cutanée'
'Autre voie sous-cutanée'

'Voie épidurale'
'Autre voie épidurale'
'Voie intradiscale'
'Voie intradurale'
'Voie péridurale'

'Voie intramusculaire et intracardiaque'
'Voie intracoronaire'
'Voie intramusculaire'
'Voie intrapéricardiaque'
'Voie intraventriculaire cérébrale'
'Autre voie intramusculaire ou intracardiaque'

'Voie intraveineuse et artérielle'
'Voie intra-artérielle'
'Voie intraveineuse'
'Autre voie intraveineuse ou artérielle'
'Voie par perfusion'
'Voie intra veineuse en perfusion'

'Voie intra-articulaire, intra-séreuse, espaces virtuels'
'Voie hémodialyse'
'Voie hémofiltration'
'Voie infiltration'
'Voie intra-amniotique'
'Voie intraarticulaire'
'Autre voie intra-articulaire, intra-séreuse, espaces virtuels'
'Voie intrabursale'
'Voie intracaverneuse'
'Voie intracisternale'
'Voie intralésionnelle'
'Voie intrapéritonéale'
'Voie intrapleurale'
'Voie intrarachidienne'
'Voie intraséreuse'
'Voie intrathécale'
'Voie intratumorale'
'Voie paracervicale'
'Voie paravertébrale'
'Voie péritumorale'

'Autre voie parentérale'

'Autre voie d'administration'

'Autre voie extracorporelle'
'Voie implantation'
'orale ou vaginale'

Glossaire

API	*Application Programming Interface*. Interface de programmation permettant d'accéder à une application ou à un programme. Autre alternative au téléchargement pour de gros volumes de données, ou des mises à jour fréquentes.
BIG DATA	Pour des volumes de données atteignant une taille critique, on parle de *Big Data* : de nouvelles approches technologiques deviennent nécessaires pour gérer les bases de données, afin de maintenir le stockage (« volume »), la rapidité d'exécution et de mise à jour (« vitesse »), et les différents types de contenu (« variété ») – les 3 V du *Big Data*.
CRAWL	Méthode automatisée de collecte d'information sur le Web : les pages trouvées sont copiées et stockées dans une archive. Google procède ainsi dans son processus d'indexation du Web.
CROWDSOURCING	Procédé collaboratif pour créer des contenus et mutualiser des ressources, connaissances ou compétences au sein de communautés d'internautes. (Open Street Map par exemple).
DATA JOURNALISME **(ou Journalisme de données)**	Nouveau procédé journalistique basé sur des traitements automatiques de fouille de données complexes et volumineuses (*Datamining*), de traitement ou d'analyse de ces données, dans le but d'en extraire les informations pertinentes et de les présenter.
DATAVISUALISATION ***(ou « Dataviz »)***	Méthodes et outils de visualisation des données sous la forme de graphiques, camemberts, diagrammes, cartographies, chronologies, infographies,... La *Datavisualisation* a l'avantage de rendre les données directement explicites et compréhensibles.
DONNEES PUBLIQUES	Données publiées ou tenues à disposition du public, et produites ou collectées par un État, une collectivité territoriale ou un organisme public dans le cadre de leur mission de service public.
DONNEES PUBLIQUES PAYANTES	Certaines données publiques sont soumises à une redevance pour être réutilisables[175]. On peut citer par exemple des organismes tels que l'Institut National de la Statistique et des Etudes Economiques (INSEE), ou l'Institut National de la Propriété Industrielle (INPI).
HACKATHON	Référence au terme « Marathon » pour nommer un évènement organisé autour de la création d'applications innovantes dans une durée limitée et mené par des

[175] https://www.data.gouv.fr/fr/Redevances

	équipes pluridisciplinaires (développeurs, designers, graphistes,…).
JEU DE DONNEES **(ou *Dataset*)**	Liste de données composée d'attributs et de leurs valeurs. Dans un fichier au format CSV par exemple, les colonnes représentent les attributs et les lignes les valeurs. Les données publiques sont très souvent mises à disposition sous forme de jeux de données.
NOSQL	Bases de données qui ne sont pas fondées sur l'architecture classique des bases de données relationnelles, elles ont été conçues pour résoudre les problèmes de traitements de données volumineuses, multi-sources et multi-formats, dans des environnements *Big Data*. Les différents types de bases *NoSql* sont classées en 4 catégories : les bases de données orientées document, les bases clé/valeur, les bases en colonnes et les bases orientées graphes.
NOTATION CINQ ETOILES	Notation proposée par Tim Berners Lee pour mesurer le degré qualitatif des données ouvertes selon le modèle sémantique (de 1 à 5 étoiles) : ★ Données accessibles sur le web (sans conditions de formats) ★★ Données accessibles structurées (exemple: Excel au lieu de l'image d'un tableau) ★★★ Formats non-propriétaires (exemple: CSV au lieu d'Excel) ★★★★ Usage d'URL pour identifier les données ★★★★★ Données liées sémantiquement