Élaboration d'un lexique de l'eau stratifié en fonction de l'auditoire : du concept aux lexèmes

Jean-Louis Janin

Université Bordeaux Montaigne

École Doctorale Montaigne Humanités (ED 480)

THÈSE DE DOCTORAT EN « LINGUISTIQUE »

Élaboration d'un lexique de l'eau stratifié en fonction de l'auditoire : du concept aux lexèmes

Présentée et soutenue publiquement le 25 janvier 2019 par

Jean-Louis JANIN

Sous la direction de Frédéric Lambert

Membres du jury

Henri PORTNE, Président du jury, Professeur émérite, Université Bordeaux Montaigne.

Jean-Louis DUCHET, Professeur émérite, Université de Poitiers.

Frédéric LAMBERT, Professeur des Universités, Université Bordeaux Montaigne.

Jean-Louis OLIVER, Expert invité, Académie de l'Eau.

Julie TROTTIER, Directrice de Recherche, Université Montpellier 3 Paul Valéry.

Introduction

Historique

Le projet de lexique bilingue des textes et des données sur l'eau est né d'un sujet de master de linguistique choisi en 2010, à l'issue d'une licence d'anglais, en reprise d'études universitaires après une carrière d'ingénieur dans la fonction publique. D'abord axée sur la technique au service de l'agriculture, en hydraulique, machinisme et statistique agricoles, cette carrière s'est poursuivie, pour le compte de l'Ifen (Institut français de l'environnement), sur le projet d'une statistique publique des usages de l'eau à partir des données de la police de l'eau (loi sur l'eau de 1992). Ce projet a été mené en lien avec le projet « Apol'Eau » du ministère chargé de l'environnement. Les deux projets ont été abandonnés au début des années 2000. La réflexion sur ces abandons a mis en lumière les difficultés de compréhension d'une approche conceptuelle du domaine de l'eau dans la sphère technico-administrative, faute de lien explicite entre le lexique des modèles conceptuels et celui des textes réglementaires. Cette remarque valait *a fortiori* pour les difficultés des journalistes et du grand public aux prises avec un vocabulaire scientifique et technique incompréhensible. Cette situation a fait naître l'idée d'une recherche linguistique sur l'élaboration en français et en anglais d'un lexique de l'eau en fonction de l'auditoire, afin de remédier aux incompréhensions et malentendus, compte tenu des enjeux sociétaux du domaine.

Le caractère très technique du domaine de l'eau, en dehors de considérations générales du type « l'eau c'est la vie », résulte de l'ingénierie sophistiquée du développement de ses usages, dans une problématique technique axée sur les « services de l'eau », alimentation en eau potable et assainissement, et sur d'autres usages en grande quantité comme l'irrigation, l'hydroélectricité et le refroidissement industriel dans les centrales nucléaires. Les services de l'eau et ces usages quantitatifs sont apparus au début comme relativement peu coûteux, sans véritable réflexion socio-économique au niveau de l'État, sinon en affirmant que « l'eau paie l'eau », ce qui ramenait la question au niveau des collectivités locales en charge de l'eau potable et de l'assainissement et de l'économie interne des usages industriels et des usages collectifs en agriculture irriguée. Les questions de « qualité » de l'eau et de « pollution » des cours d'eau comme entrave à la vie piscicole ont mobilisé les scientifiques et les techniciens de cette époque dans une indifférence polie du monde politique.

Le développement des usages quantitatifs dans les années 1970 et 1980 avec des résultats alarmants sur l'étiage de certains cours d'eau dans les années de « sécheresse » (1978, 1988, 1989) a fait prendre conscience à l'opinion publique et aux élus d'une problématique de la *ressource*. Cette prise de conscience a débouché sur les autorisations préalables ou les décla-

rations préalables de prélèvement temporaire pour l'irrigation de la loi sur l'eau de 1992 et sur un pouvoir d'intervention des Préfets dans l'urgence, avec les arrêtés « sécheresse ». Le déficit structurel de certains cours d'eau en période d'étiage a conduit à l'instauration, avec la loi sur l'eau de 2006, de la possibilité d'une gestion collective des prélèvements d'eau pour l'irrigation par un organisme unique (OUGC). Ce mode de gestion de l'eau d'irrigation connaissait un début d'application en 2016 qui s'est étendu en 2017.

La question de la pollution, dont la prise de conscience était beaucoup plus ancienne, à travers la pratique de la pêche, était parfois ramenée au prix à payer pour conserver des emplois dans l'industrie. Elle était et reste traitée par un auto-contrôle dans les établissements industriels qui relèvent des textes sur les établissements classés. Cette question a pris une nouvelle dimension avec la persistance de la pollution diffuse d'origine agricole. L'arrivée d'une nouvelle génération de citoyens urbanisés a débouché sur le procès d'une production agricole industrielle, tributaire des engrais et des traitements phytosanitaires, en regard des promesses de l'agriculture biologique. Cette situation, combinée avec la surexploitation structurelle de la ressource déjà évoquée, continue d'alimenter l'incompréhension entre les tenants de l'écologie des milieux aquatiques et ceux de l'irrigation à partir de lacs collinaires. L'apparition des micro-polluants dans la sphère domestique a encore élargi le champ de la problématique environnementale de l'eau.

En 2018, la question de l'eau n'est plus une obscure question technique laissée aux ingénieurs. Elle est devenue, à travers un début de gestion pluri-acteurs préconisée par la directive cadre sur l'eau de 2000, et dans le contexte du réchauffement climatique, une source d'inquiétudes (inondations, sécheresse) et de conflits sociétaux. La question du lexique de l'eau s'est concrétisée en 2014 en tant que projet de « Dictionnaire de l'Académie de l'Eau », ou projet LEXEAU®, soutitré « Un lexique bilingue des textes et des données sur l'eau » (*A bilingual lexicon of water related texts and data*), avec le dépôt de la marque « lexeau » à l'INPI [1].

Une problématique de recherche-développement

Cette thèse a pour ambition d'inscrire la problématique du lexique de l'eau dans le développement d'une sémantique lexicale scientifiquement fondée, testée sur une analyse conceptuelle des « objets de connaissance » des professionnels du domaine et des journalistes qui en rendent compte au quotidien. On trouve ces professionnels dans le secteur de l'industrie et des services de l'eau quel que soit l'usage de la ressource, dans les services de l'état et des collectivités locales et dans les établissements publics, y compris en recherche-développement.

L'analyse conceptuelle débouche sur la mise en oeuvre d'une sémantique lexicale lorsque les concepts sont lexicalisés et définis en tant qu'unités lexicales, avec des exemples d'emploi, en s'appuyant sur des textes. L'analyse de corpus avec des logiciels toujours plus performants sur un éventail de langues de plus en plus étendu permet de réaliser ce travail. Cette thèse a pour objectif de baliser la « rencontre » sur le terrain du lexique entre le socle lexical tiré d'une ontologie conceptuelle, qui relève des professionnels du domaine, ingénieurs, techniciens et scientifiques, mais aussi des juristes et des *medias*, et l'ensemble plus mouvant, mais

1. Institut national de la protection industrielle

beaucoup plus riche, des « mots pour le dire » tirés des corpus de textes dans des pratiques universitaires et journalistiques.

Le premier chapitre de la thèse a pour objet de parcourir la trajectoire qui mène d'un réseau de concepts du domaine, édités dans une ontologie, aux unités lexicales inter-connectées d'un lexique stratifié, dotées d'une définition et d'exemples d'emploi validés par des textes. Les prélèvements d'eau, en tant que mouvements d'eau générateurs de flux d'origine anthropique sont utilisés pour illustrer cette trajectoire. Le modèle des mouvements d'eau et des flux est détaillé, avec un essai d'intégration dans l'ontologie des données disponibles sur les prélèvements d'eau effectués en 2016 pour la centrale nucléaire de Golfech. Ce parcours permet d'illustrer le passage de l'ontologie du domaine aux unités lexicales, la mobilisation des textes du lexique et l'origine des relations entre les unités lexicales.

Le deuxième chapitre est consacré aux fondements du projet LEXEAU®. Il s'ouvre sur un schéma d'ensemble du dispositif, conçu pour l'utilisateur comme un portail en accès libre avec une interface sur laquelle il peut formuler des requêtes en langage naturel. Le chapitre parcours le dispositif depuis l'autre côté, c'est à dire les corpus de textes qui contribuent à la consolidation et à l'enrichissement du lexique. Les « objets » et les outils de la linguistique de corpus mise en œuvre sont présentés sur des exemples, à travers la constitution des corpus, l'analyse par textométrie du contenu des documents sources du domaine et l'émergence d'un lexique de base formé de substantifs récurrents. Cette émergence débouche sur une première mise à jour des substantifs du lexique qui peut s'accompagner d'une mise à jour des concepts éditée dans l'ontologie du domaine.

Le troisième chapitre élargit la problématique de l'enrichissement du lexique aux syntagmes nominaux et à l'émergence de collocations et de noms propres lexicalisés, illustrés par des exemples de recherche des cooccurrences dans les textes des communications scientifiques en anglais de la conférence internationale *Floodrisk 2016* sur la question des inondations.

Le quatrième chapitre revient sur la question des « données » et du « discours » que l'on peut tirer d'une analyse par textométrie de corpus constitués dans deux strates discursives différentes sur le même thème. Le premier corpus a été constitué des articles du journal britannique *The Guardian* sélectionnés dans une base de presse sur le mot clef *flooding*, avec un contrôle de pertinence, et regroupés par tranches annuelles de février 2011 à février 2016. Le second est constitué de la directive européenne de 2007 sur la prévention des inondations dans la version anglaise. Une métrique comparative des graphes de locutions récurrentes construits sur les cooccurrences droite et gauche d'une base nominale est proposée à partir des résultats de l'analyse de ces cooccurrences dans les deux corpus.

Le cinquième chapitre porte sur la sémantique lexicale mise en œuvre dans l'élaboration du lexique. Il justifie l'introduction des quatre strates discursives dans le lexique. Dans la même section, il donne un exemple de la restitution lexicale dans deux strates distinctes des concepts d'aléas et de vigilance édités en classes d'évènements dans l'ontologie, et dans une troisième, la strate discursive courante, pour les catastrophes. La place de la strate discursive technico-scientifique est illustré par la restitution lexicale de modèles physiques comme la relation entre les pluies exceptionnelles sur un bassin versant et le débit de crue à l'exutoire. Les trois ontologies du domaine, du lexique et du discours sont présentées ensuite avec le réseau des relations conceptuelles entre les classes, aux frontières et en interne, qui gouvernent la restitution dans le lexique des unités lexicales du domaine avec ses propres textes. Les disciplines à l'oeuvre dans

la constitution du lexique de l'eau sont détaillées, avec un essai de nomenclature des thèmes du domaine. La dernière partie du chapitre revient sur le statut des unités lexicales, qui résultent de l'instanciation des lexèmes, et de leur fonction dans une strate discursive comme réceptacle du sens de ce qui se dit dans cet type de discours avec les mots et les expressions — les unités discursives — que l'on retrouve dans d'autres strates. Le statut et les fonctions des unités lexicales chez Anna Wierzbicka et chez Igor Mel'čuk sont revus par comparaison, à la lecture de quelques ouvrages de ces auteurs qui nous ont semblé traiter de points communs.

Les annexes regroupent des tableaux et des figures des chapitres 1 à 5, un ensemble d'annotations de l'ontologie du dispositif et un glossaire de mots et d'expressions employés dans la thèse dans un sens spécifique.

Avertissement au lecteur,

Le texte, les figures et les tableaux de ce document reproduisent à l'identique, après correction des fautes d'orthographe et des coquilles, le contenu de la thèse soutenue le 25 janvier 2019. Depuis cette date, le modèle du lexique et notamment la partition de la base de connaissance ont évolué, dans le cadre du projet Lexeau® parrainé par l'Académie de l'Eau. La strate discursive dite « Décisionnelle incitative » (DI) a été abandonnée, ramenant à trois le nombre de strates discursives, libellées « 1-TS Technico-scientifique », « 2-TA Technico-administrative » et « 3-C Courante ». La base de connaissance a été mise à jour sur de nouveaux thèmes et de nouveaux concepts, notamment en matière de développement durable. Le logiciel « Protégé » est resté l'outil d'édition de la maquette du projet et des exemples de contenu du lexique . La plupart des figures de la thèse on été éditées avec ce logiciel et sont datées du mois de leur édition. Elles doivent être considérées comme obsolètes. Elles peuvent être mises à jour à la demande avec une nouvelle légende.

Bordeaux, le 2 mars 2020

Jean-Louis Janin

E-mail : jeanlouisjanin@gmail.com

Chapitre 1

Un lexique du domaine de connaissance

1.1 Un lexique stratifié

Lorsqu'on parle d'un domaine spécifique, le lexique utilisé, lexique au sens large, composé des termes grammaticaux et des « lexèmes » (nous reviendrons sur cette notion, pour l'instant, en gros, les mots ayant un référent), devient composite ; il est formé de plusieurs strates qui s'imbriquent les unes aux autres :

- on ne peut élaborer un texte, tenir des propos oraux sans recourir à un lexique quotidien ; d'une part, les prépositions, les conjonctions et de façon générale les « mots-outils », les termes grammaticaux ne se répartissent pas en discours quotidien et en discours spécifique ; d'autre part, un texte ou une production orale constitué(e) uniquement de lexèmes spécifiques serait rapidement incompréhensible ;
- un texte ou une production orale spécifique (c'est-à-dire ne portant pas sur la quotidienneté) comporte bien évidemment une part de termes spécifiques en relation avec l'auditoire visé ; mais cet auditoire est rarement homogène, même lorsqu'il est composé de spécialistes du domaine ; il y a toujours des différences dans l'extension de l'emploi de tel terme, des métaphores autorisées ou non, etc. ;
- le langage humain permet des « extensions de sens », des modalités pour préciser le sens d'un terme, des façons de rester dans un certain flou qui autorise des approximations sans pour autant « sortir du cadre » ou des « expressions reçues » ; grâce à cette propriété, un texte ou une production orale peut correspondre à un auditoire plus ou moins étendu ;
- la notion de médiation scientifique — qui remplace aujourd'hui celle de vulgarisation scientifique — est plus explicite que celle qui l'a précédée : une médiation est un processus destiné à articuler deux plans différents de l'expression discursive, contrairement à une transduction qui suppose un rapport bijectif (par exemple, la « traduction » d'un mot écrit, c'est-à-dire une séquence de lettres, en morse, c'est-à-dire en une suite de traits et de points ou de sons longs et brefs) ; si la médiation n'est pas bijective, il faut trouver des procédés d'accommodements entre les deux plans discursifs que l'on essaie de corréler.

Ces strates (nous n'avons pas cherché à fournir une description rigoureuse, mais plutôt à illustrer la suite de nos propos) correspondent au sentiment qu'il existe des difficultés d'intercompréhension dès lors que l'on se situe dans un domaine doté d'un certain degré de technicité, comme ce sera le cas ici. Il y a un « point fixe », ou supposé tel, celui de la scientificité : un discours scientifique se caractérise par l'emploi de termes monosémiques partagés par une communauté donnée et circonscrite. Les autres auditoires correspondent à des emplois plus « malléables ». Dès lors, d'où part-on ? de la strate « la plus dure », celle du discours scientifique pour aller vers des zones plus « approximatives » ? ou ferons-nous le chemin inverse ? Nous verrons que cette question n'est pas aussi évidente qu'il y paraît. Il est aussi important de bien comprendre que « strate plus dure » ou strate plus malléable » ne forment pas des jugements de valeurs. Une « strate dure » permet de traiter des problèmes bien cernés au sein de situations bien délimitées ; en revanche, une « strate plus malléable » permet d'aborder des problèmes aux contours plus flous parce qu'inscrits dans des situations plus ou moins instables. Chaque « point de vue » a donc son intérêt qui n'est pas réductible à une sorte de « sous-domaine ». Nous recourrons à l'emploi d'ontologies (représentations structurées d'un domaine de connaissance), ce qui nous obligera à nous situer constamment par rapport à ces strates que nous répartirons en différents niveaux (ce point sera clarifié par la suite) : pour l'instant, technico-scientifique (TS), technico-administratif (TA), courant (C).

Mais nous ne pouvions envisager cette problématique discursive dans sa généralité — même en la restreignant à la problématique de la gestion de l'eau, notre domaine — au risque de n'énoncer que des vérités générales peu productives. C'est pourquoi nous allons d'abord nous centrer sur une notion donnée, celle de prélèvement d'eau, en l'introduisant dans l'ontologie du domaine avant de la propager dans le lexique avec des textes à l'appui. Nous présentons au préalable une partition de la base de connaissance qui permet d'isoler cette ontologie.

1.2 Une partition de la base de connaissance

Le lexique est construit sur une partition de la base de connaissance du domaine de l'eau, du lexique de ce domaine et du discours qui le vise (fig. 1.1), entre :

- une sous-base DOMAINE des entités typées, des entités génériques et des propriétés ;
- une sous-base LEXIQUE d'unités lexicales et de noms propres ;
- une sous-base DISCOURS d'unités discursives, de textes et de corpus.

La sous-base DOMAINE est composée d'entités « typées » et d'entités « génériques » qui permettent de prendre en compte les objets de connaissance du domaine de l'eau dans une ontologie de classes d'objets hiérarchisées. La différence entre entités typées et entités génériques est que les classes d'entités génériques ont un nombre limité d'instances, par exemple le mode d'irrigation (par aspersion, par gravité ou au goutte à goutte), par opposition aux classes d'entités typées, par exemple celle des réseaux d'irrigation (dans le monde ...). La sous-base de connaissance du DOMAINE comprend aussi les propriétés relationnelles et « à valeur » des individus de ces classes.

Les noms des classes d'entités typées évoquent des concepts du DOMAINE, ainsi que les noms et les instances des classes d'entités génériques et les noms des « propriétés nommées ». Les instances des classes d'entités typées font appel à des noms propres d'objets concrets. Les

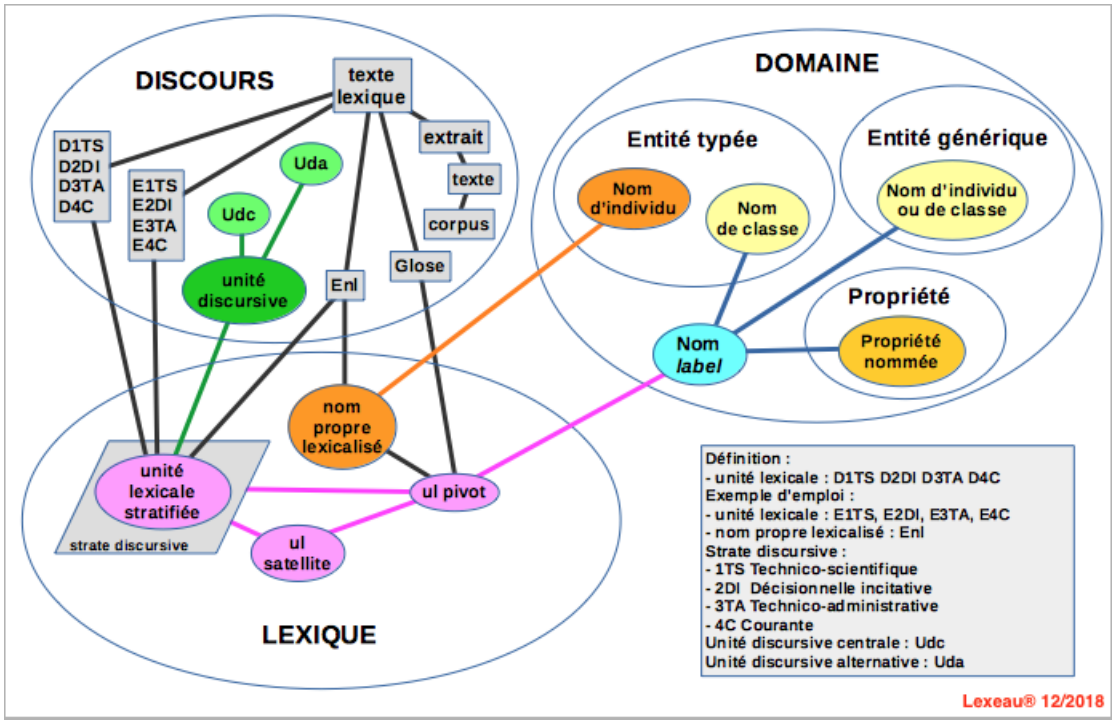


FIGURE 1.1 – Construction du lexique sur une partition de la base de connaissance

propriétés peuvent être des propriétés « relationnelles » (*object properties*) ou des propriétés « à valeur » (*data properties*). Certaines propriétés relationnelles font appel à des concepts éligibles dans le lexique sous une forme nominale (propriétaire, usager, opérateur, etc.). Il en est de même pour des propriétés « à valeur » (redevance, volume d'eau, débit maximum, etc.), sous réserve de leur attribuer une unité de valeur.

Les noms des concepts de l'ontologie sont transposés dans la sous-base LEXIQUE en tant qu'unités lexicales « pivot » rattachées à une strate discursive. Les individus des classes d'entités typées sont transposés dans le LEXIQUE sous forme de noms propres, « désignateurs rigides » de l'individu source. Pour éviter de confondre avec le même nom propre les « référents » de deux unités lexicales (la rivière Seine, le département de la Seine), leur libellé est le plus souvent construit en ajoutant au nom propre l'unité lexicale pivot qui reprend le nom de son type d'origine dans l'ontologie du domaine (type de lieu, de personne, de document, etc.). Dans tous les cas, les référents sont associés à une et une seule unité lexicale pivot. Ils forment une classe distincte de celle des noms propres sources. Ils sont associés à l'unité lexicale stratifiée à laquelle ils réfèrent. Nous verrons que cette unité lexicale est différente de l'unité lexicale pivot constitutive du nom propre lexicalisé s'il s'agit d'une unité satellite de l'unité pivot (cf. le référent du « Prélèvement d'eau soumis à redevance » dans la section 1.4).

La sous-base DISCOURS de la base comprend les unités discursives placées en entrée du lexique et les textes dont les extraits sont repris dans les définitions, les gloses et les exemples d'emploi du lexique. Parmi les unités discursives on distingue entre les unités discursives centrales, rattachées à une ou plusieurs unités lexicales (dont une unité pivot) et les unités

discursives alternatives, rattachées à une unité centrale et à une unité lexicale associée à l'unité discursive centrale dont elles reprennent le sens.

Les textes de la sous-base DISCOURS sont établis pour :

- définir une unité lexicale ;
- énoncer la glose d'une unité lexicale ;
- donner un exemple d'emploi d'une unité lexicale ;
- donner un exemple d'emploi d'un nom propre lexicalisé ;
- donner un exemple d'emploi d'une unité discursive alternative.

1.3 Modélisation des prélèvements d'eau

Cette modélisation a consisté à faire entrer les prélèvements d'eau dans des catégories d'objets plus générales et à établir des relations aussi générales que possible entre ces objets. Nous allons procéder par figures successives pour présenter le modèle, qui part des concepts les plus généraux. Parmi les entités hydrauliques du domaine de l'eau, nous distinguerons les mouvements d'eau et les flux. Un mouvement d'eau est une opération effectuée par l'homme que l'on peut associer à un flux d'eau, sur une période donnée. Ce flux sera qualifié d'anthropique pour le distinguer du flux d'un cours d'eau. Le modèle associe à chaque flux anthropique un mouvement d'eau « amont » qui génère le flux et un mouvement d'eau « aval » qui le récupère. Il n'y a pas de perte d'eau entre les deux mouvements d'eau.

La figure 1.2 présente au niveau des classes d'objets, dans l'ontologie du domaine, les relations entre les mouvement d'eau amont, les mouvement d'eau aval et les flux anthropiques. Le graphe a été établi avec la fonction « OntoGraf » de l'éditeur de l'ontologie. Les noms des relations curvilignes entre les classes ont été ajoutées sur fond vert au graphe original. Elles ont une forme standard où la classe « cible » de la relation est notée en abrégé et précède l'abrégé de la classe « source », placée entre parenthèses, sans autre notation si elle n'est pas indispensable.

Légende des relations

- flA(mam) : Flux anthropique généré par le mouvement d'eau amont ;
- mam(flA) : Mouvement d'eau amont qui génère le flux anthropique ;
- flA(mav) : Flux anthropique récupéré par le mouvement d'eau aval ;
- mav(flA) : Mouvement d'eau aval qui récupère le flux anthropique.

La figure 1.2 présente la hiérarchie des classes depuis la classe « Entité du domaine ». Cette classe est la plus générale dans l'ontologie du domaine. Elle est placée à la « racine » de l'arbre de cette partie de l'ontologie globale. Les relations hiérarchiques entre classes sont représentées par des traits rectilignes orientés de la classe vers la sous-classe, sans marquer par une flèche dans l'autre sens la relation inverse. La notation générique de classe à sous-classe est notée « s-c(c) ». Les relations hiérarchiques du graphe s'énoncent ainsi :

- Les entités typées forment une sous-classe de la classe des entités du domaine ;
- Les entités hydrauliques forment une sous-classe de la classe des entités typées ;
- Les mouvements d'eau forment une sous-classe de la classe des entités hydrauliques ;
- Les flux d'eau forment une sous-classe de la classe des entités hydrauliques ;

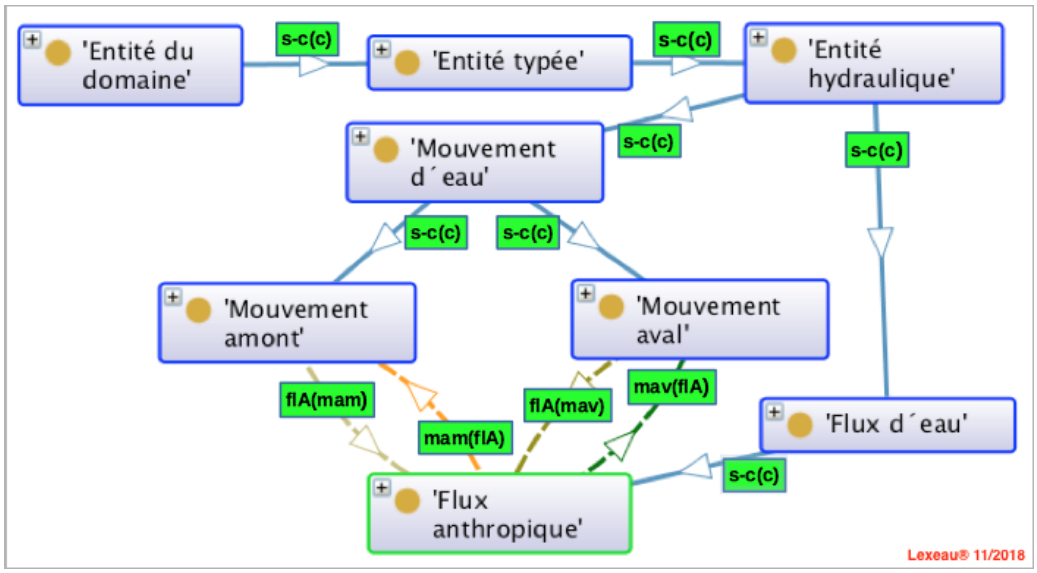


FIGURE 1.2 – Graphe relationnel des mouvements d'eau et des flux anthropiques

- Les flux anthropiques forment une sous-classe de la classe des flux d'eau ;
- Les mouvements d'eau amont forment une sous-classe de la classe des mouvements d'eau ;
- Les mouvements d'eau aval forment une sous-classe de la classe des mouvements d'eau.

Les relations inverses peuvent être formulées avec l'expression « est un/une » (*is a*) :

- Une entité typée est une entité du domaine ;
- Une entité hydraulique est une entité typée ;
- Un mouvement d'eau est une entité hydraulique ;
- Un flux d'eau est une entité hydraulique ;
- Un flux anthropique est un flux d'eau ;
- Un mouvement d'eau amont est un mouvement d'eau ;
- Un mouvement d'eau aval est un mouvement d'eau.

Le graphe relationnel des prélèvements d'eau en lien avec les flux d'eau générés est présenté dans la figure 1.3, suivie de la légende des relations spécifiques.

- prl(me) : prélèvement d'eau effectué dans la masse d'eau ;
- me(prl) : masse d'eau dans laquelle est effectué le prélèvement d'eau ;
- rst(me) : restitution d'eau effectuée dans la masse d'eau ;
- me(rst) : masse d'eau dans laquelle s'effectue la restitution d'eau ;
- ijn(rse) : injection d'eau effectuée dans le réseau d'eau ;
- rse(ijn) : réseau d'eau dans lequel s'effectue l'injection d'eau.

Les nouvelles entités introduites dans le modèle sont les masses d'eau et les réseaux d'eau, avec lesquels les prélèvements d'eau sont en relation par l'intermédiaire des flux qu'ils génèrent, en tant que mouvement amont et mouvements aval de ces flux. Un prélèvement d'eau sera suivi de la restitution de l'eau prélevée dans une masse d'eau ou de l'injection de l'eau prélevée dans un réseau d'eau. Il est modélisé avec le flux anthropique qu'il génère et le mouvement

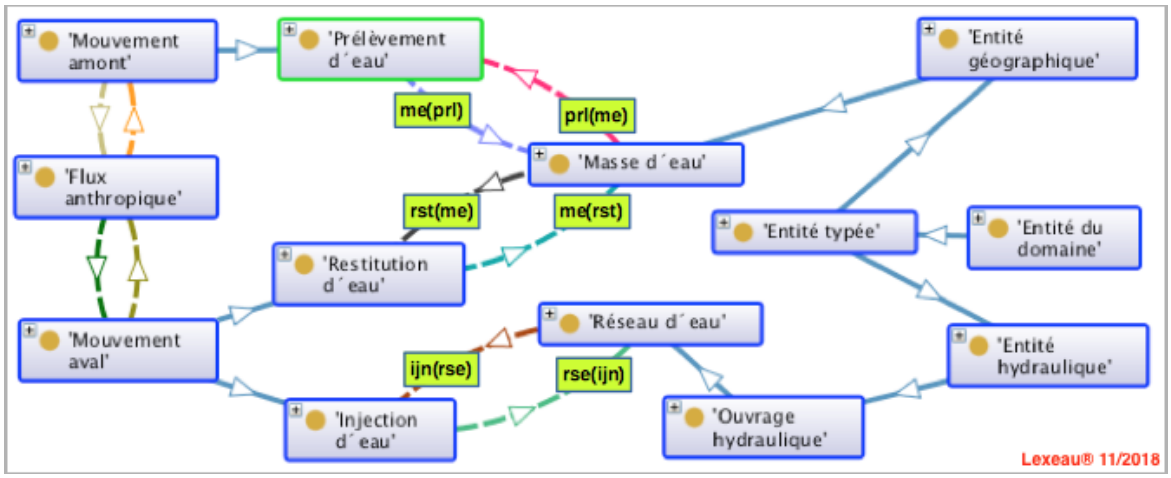


FIGURE 1.3 – Graphe relationnel des prélèvements d'eau en lien avec les flux anthropiques

d'eau qui récupère ce flux. Dans la hiérarchie des classes, une masse d'eau est une entité géographique, qui est elle-même entité typée. Un réseau d'eau est un ouvrage hydraulique, qui est lui-même une entité hydraulique. La période d'occurrence d'un flux anthropique est enregistrée au niveau du flux d'eau, comme le montre la figure 1.4. Dans la hiérarchie des classes, la période est une entité temporelle, qui est elle-même une entité typée.

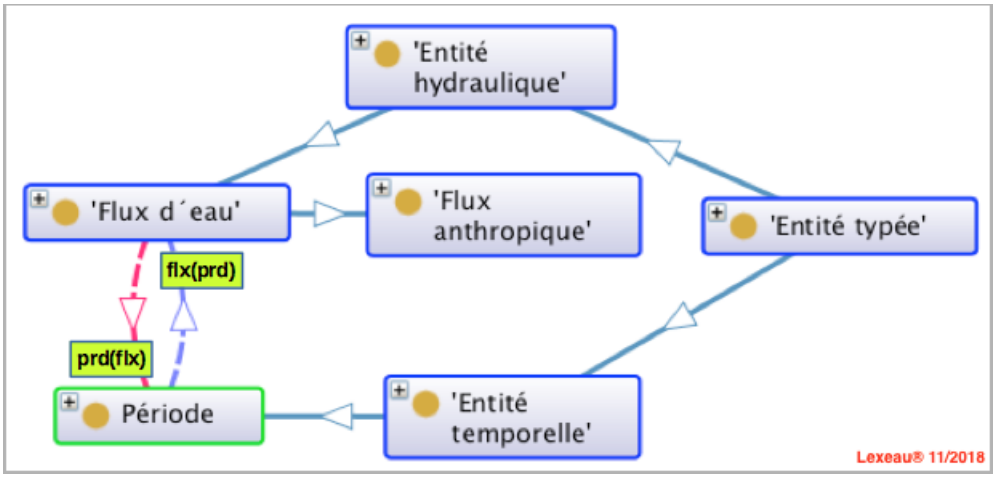


FIGURE 1.4 – Graphe relationnel de la période d'occurrence des flux anthropiques

La mesure du volume d'eau d'un flux anthropique généré par un prélèvement d'eau est assurée par un compteur d'eau spécifique installé sur l'ouvrage de prélèvement. Le graphe relationnel est présenté figure 1.5. Les nouvelles entités du modèle relationnel sont le compteur du prélèvement et l'ouvrage de prélèvement, considérés comme des ouvrages hydrauliques. Le compteur de prélèvement est associé, en tant que compteur d'eau volumétrique, au flux anthropique qu'il mesure, à travers la relation « fla(cev) ». Le prélèvement d'eau est effectué par un ouvrage de prélèvement, à travers la relation « prl(oPrl) ». Les points de prélèvement ont été introduits dans le modèle conceptuel des prélèvements comme des points matériels pour récupérer

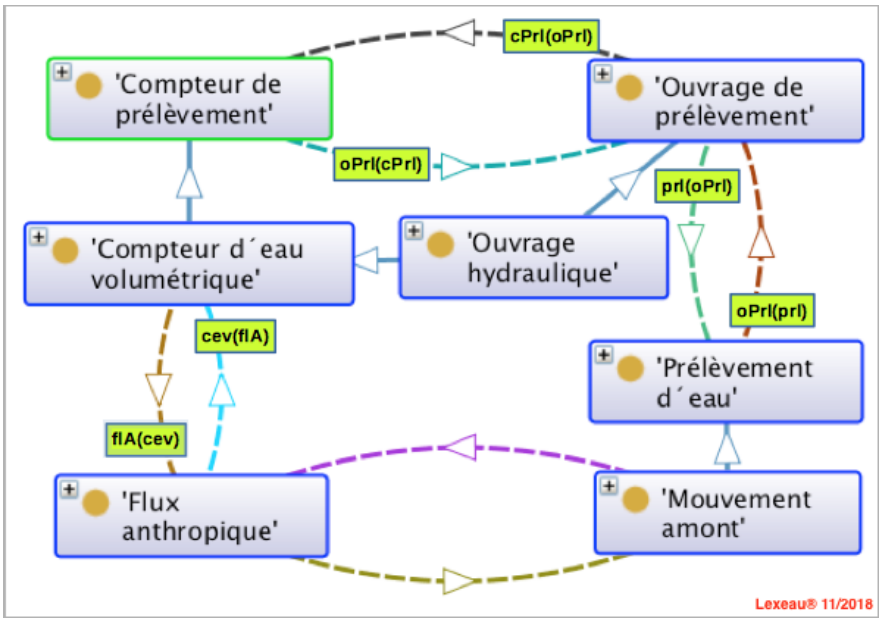


FIGURE 1.5 – Graphe relationnel du comptage des prélèvements d'eau

les données du SIE Adour-Garonne qui utilise cette notion pour localiser les prélèvements d'eau faisant l'objet d'une redevance de l'agence de l'eau du bassin. Le graphe relationnel est présenté dans la figure 1.6. Nous avons ajouté sur le graphe les identifiants de l'ouvrage de

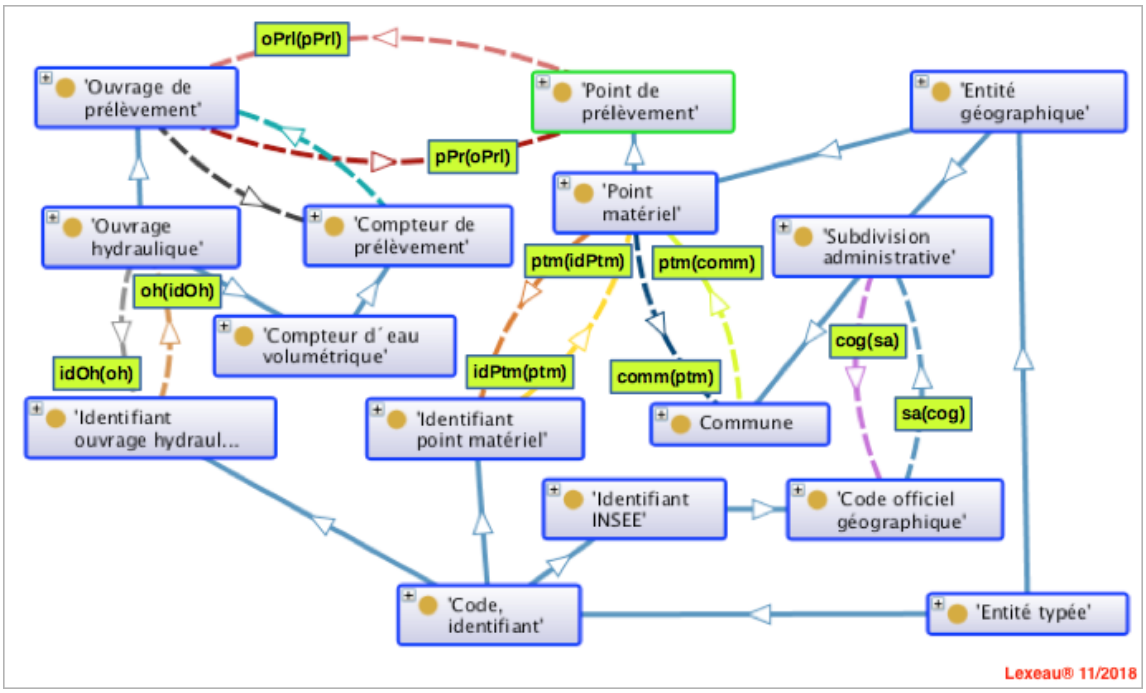


FIGURE 1.6 – Graphe relationnel des points de prélèvements

prélèvement et du compteur de prélèvement, en tant qu'ouvrage hydraulique, celui du point

de prélèvement, en tant que point matériel et celui de la commune, connu sous le nom de code officiel géographique (COG) de la subdivision administrative, attribué par l'INSEE.

L'implication humaine dans les prélèvements d'eau et l'usage de l'eau du flux anthropique ont été modélisés pour distinguer les différents acteurs associés aux prélèvements d'eau. Le graphe relationnel est présenté dans la figure 1.7.

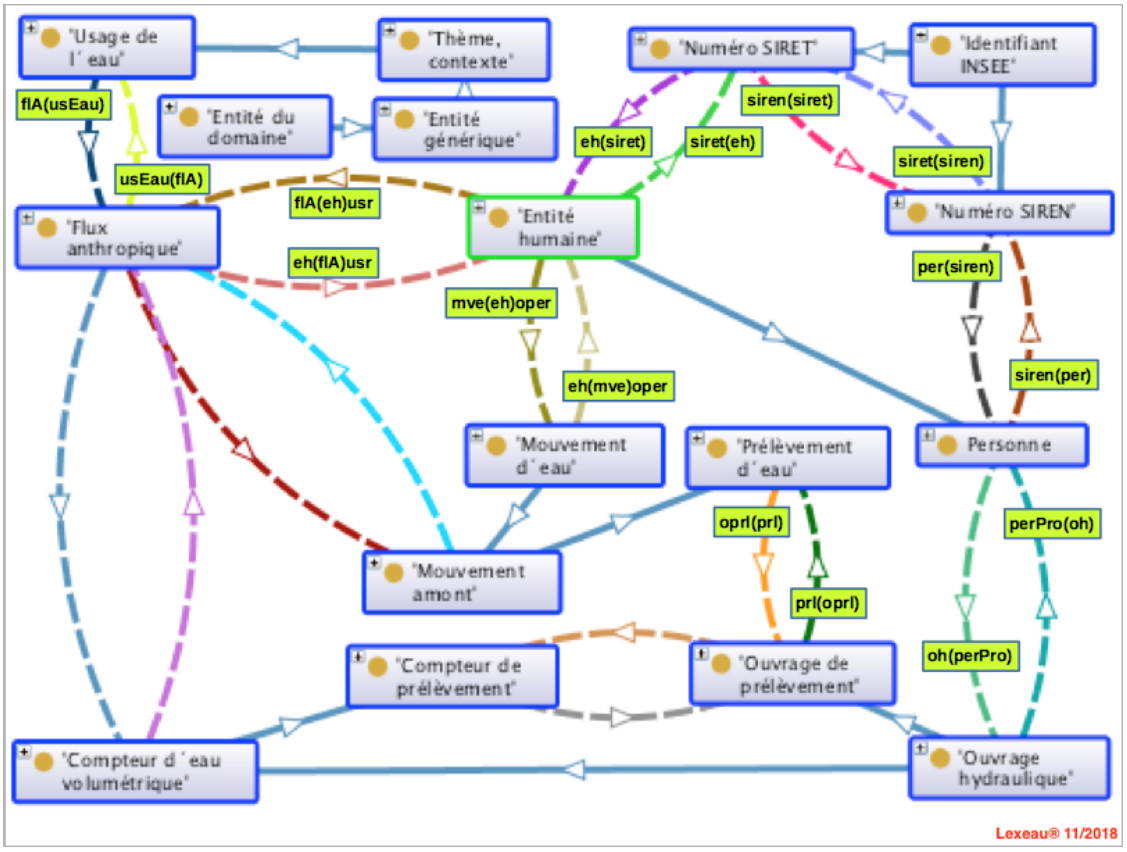


FIGURE 1.7 – Usage de l'eau prélevée et implication humaine dans les prélèvements

Le modèle permet de distinguer entre l'opérateur du prélèvement d'eau et l'usager de l'eau du flux anthropique généré par le prélèvement. L'opérateur et l'usager sont des « entités humaines », pour inclure les situations où ce ne sont sont des personnes avec la personnalité juridique, par exemple des services de l'état et des collectivités locales ou des établissements industriels d'entreprise. Les propriétaires (*owners*) des ouvrages hydrauliques sont par contre considérés comme des personnes physiques ou morales. Les numéros SIREN et SIRET sont respectivement les numéros des entreprises et des établissements d'entreprise gérés par l'INSEE dans le fichier SIRENE. Ces catégories ne figurent pas dans l'ontologie du domaine, ce qui explique leur enregistrement dans la classe (ou une sous-classe) des personnes et des entités humaines. L'usage de l'eau est une sous-classe de la classe « Thème, contexte » d'entités génériques. Les noms des instances sont utilisés de préférence aux noms des classes. La légende des relations nouvelles par rapport à celles des figures précédentes est la suivante :

- ehU(flA) : usager du flux anthropique ;
- flA(ehU) : flux anthropique utilisé par cet usager ;

- eh(mve)oper : entité humaine agissant comme opérateur du mouvement d'eau ;
- mve(eh)oper : mouvement d'eau mis en oeuvre par une entité humaine comme opérateur ;
- per(oh)pro : personne propriétaire de l'ouvrage hydraulique ;
- oh(per)pro : ouvrage hydraulique de la personne propriétaire ;
- siren(per) : numéro SIREN d'une entreprise attribué à la personne ;
- per(siren) : personne dotée du numéro SIREN d'une entreprise ;
- siret(eh) : numéro SIRET d'un établissement d'entreprise attribué à l'entité humaine ;
- eh(siret) : entité humaine dotée du numéro SIRET d'un établissement d'entreprise ;
- siren(siret) : numéro SIREN associé au numéro SIRET ;
- siret(siren) : numéro SIRET associé au numéro SIREN.

Parmi les propriétés à valeur (*data properties*) de l'ontologie du domaine (fig. 1.8), celles qui sont surlignées sont des propriétés d'un flux d'eau anthropique, avec les unités ou les valeurs suivantes :

- le volume d'eau écoulé sur la période, exprimé en m3 ;
- le stade final de l'usage de l'eau, exprimé en variable de Boole (*true/false*) ;
- le niveau collectif de l'usage de l'eau, exprimé en variable de Boole (*true/false*).

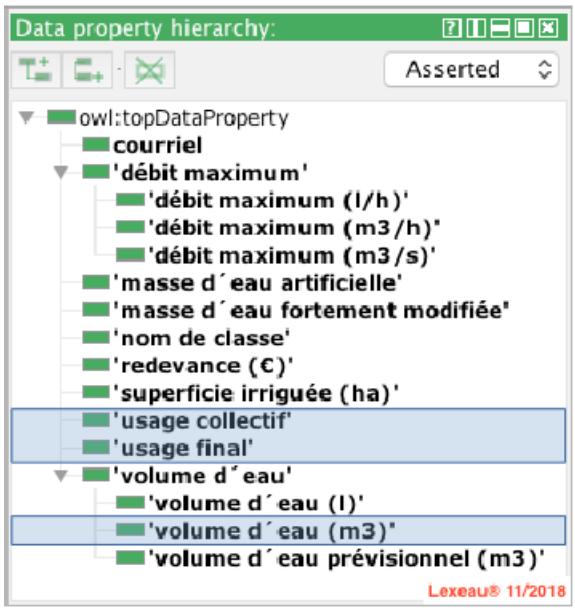


FIGURE 1.8 – Liste hiérarchisée des propriétés à valeur de l'ontologie du domaine

La liste des propriétés à valeur de la figure 1.8 ne préjuge pas de l'affectation de ces propriétés à telle ou telle classe de l'ontologie. Elle peut donc être complétée à tout moment, en évitant les redondances.

La figure 1.9 présente les classes créées comme super-classes de la classe « Flux anthropique » pour les propriétés directes des flux anthropiques, relationnelles et à valeur. On y retrouve les trois propriétés à valeur surlignées dans la figure 1.8 et les relations directes des figures 1.2, 1.5 et 1.7. Avec la hiérarchie des classes de l'ontologie du domaine, les flux anthropiques « héritent » des propriétés des super-classes « Flux d'eau » et « Entité typée ». La figure

FIGURE 1.9 – Propriétés directes des flux anthropiques

1.10 présente ces propriétés, où on retrouve la relation « prd(flx) » de la figure 1.4 qui permet d'attribuer une période d'occurrence à un flux d'eau anthropique. Les relations « ulPivot(nct) » et « np(ict) » correspondent respectivement à la transposition du nom de la classe « Flux anthropique » comme unité lexicale pivot du lexique et à celle d'une instance de cette classe comme nom propre.

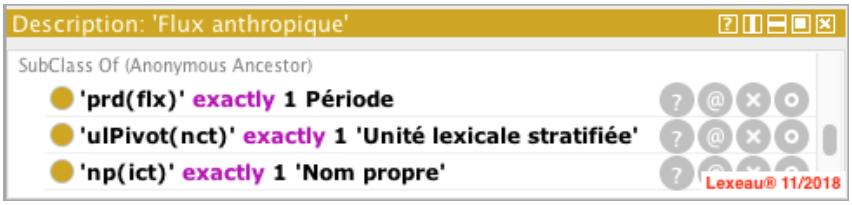


FIGURE 1.10 – Propriétés « héritées » des flux d'eau anthropiques

1.3.1 « Données » et « propriétés » dans l'ontologie du domaine

1.3.1.1 Éléments déictiques de l'ontologie du domaine et choix des propriétés

D'une façon générale l'ontologie du domaine de connaissance est une ontologie « située ». La création d'une classe d'entités typées revient à catégoriser un concept d'objets (abstrait) puis des objets concrets, les individus de la classe, que les linguistes appellent des « référents ». Une fois ces objets créés, on fait appel aux catégories spatiales et temporelles du domaine de connaissance pour les situer dans l'espace et dans le temps, ce que les linguistes appellent des « éléments déictiques » (<ici> et <maintenant>). Ceci explique la création dans l'ontologie des classes d'entités géographiques et d'entités temporelles déclinées en sous-classes.

Situer dans l'espace et dans le temps sans trop de précision les individus des classes d'entités typées peut se faire avec des propriétés relationnelles entre ces classes et des classes d'entités

géographiques et temporelles. Par contre, localiser un objet concret en coordonnées Lambert ou dater un événement en temps universel (jour, heure, minute et seconde) impose le recours à des propriétés « à valeur ».

1.3.1.2 Comment traduire *data properties* et *object properties*

L'introduction de l'anglais *data* dans le métalangage de l'éditeur d'ontologie introduit un risque de confusion en traduction mot à mot, compte tenu de la polysémie du mot « donnée » en français. Ce qui est visé par les *data properties* du logiciel « Protégé » correspond à ce que nous appellerons des propriétés « à valeur », un volume d'eau en m3, par exemple, propriété « à valeur » d'un flux anthropique, de nature différente d'une propriété « relationnelle », par exemple la propriété d'un flux anthropique d'être « généré par » un prélèvement.

Lorsque le débit d'un cours d'eau est évalué à 60 m3/s à un instant donné, le débit en m3/s est une « propriété à valeur » du cours d'eau à cet instant, la valeur de la propriété est 60. Le jour de la mesure du débit peut être une propriété relationnelle s'il est pris dans un calendrier associé aux relevés du débit. La date en temps universel de la mesure restera une propriété à valeur. La propriété du Tarn d'être un affluent de la Garonne est une propriété relationnelle (*object property*). Elle peut s'énoncer « est un affluent de ». Elle instaure dans la classe des rivières la possibilité d'une relation entre deux rivières, instanciée pour certains couples d'individus de cette classe. Une propriété relationnelle est parfois présentée comme une « relation » d'une classe avec elle-même ou avec une autre classe. On retiendra que cette relation s'établit toujours entre deux instances. Dans la suite de ce travail, nous traduirons systématiquement *data property* par propriété à valeur et *object property* par propriété relationnelle, en employant parfois la mot « relation » pour évoquer une propriété relationnelle.

La question des données s'est posée pour nous dans la confrontation entre deux systèmes d'information A et B, pour savoir si les données du système A se retrouvent ou pas dans les données du système B, et réciproquement, ce que nous avons testé dans l'essai d'intégration dans la base de connaissance du domaine de données sur les prélèvements d'eau tirées du système d'information sur l'eau du bassin Adour-Garonne.

1.4 Un essai d'intégration de données externes

Golfech est le nom de la commune du département du Tarn et Garonne sur laquelle est située la centrale nucléaire exploitée par EDF, avec un prélèvement d'eau important. Nous avons recherché les données sur les prélèvements annuels de la centrale dans le système d'information sur l'eau du bassin Adour-Garonne (SIE-AG). Ce système d'information est librement accessible au public. Il est mis en oeuvre par l'agence de l'eau Adour-Garonne à partir des données collectées pour la perception de la redevance sur les prélèvements d'eau dans la ressource et d'autres redevances, notamment sur les rejets d'eaux usées et de polluants. L'écran d'accueil du système d'information est présenté figure 1.11

Les données du SIE Adour-Garonne sur les prélèvements de la centrale nucléaire de Golfech en 2016 ont été transposées dans la base de connaissance. Le volume des prélèvements d'eau a été relevé dans un tableau du SIE reproduit par copie d'écran (fig. 1.12). Le point de prélèvement

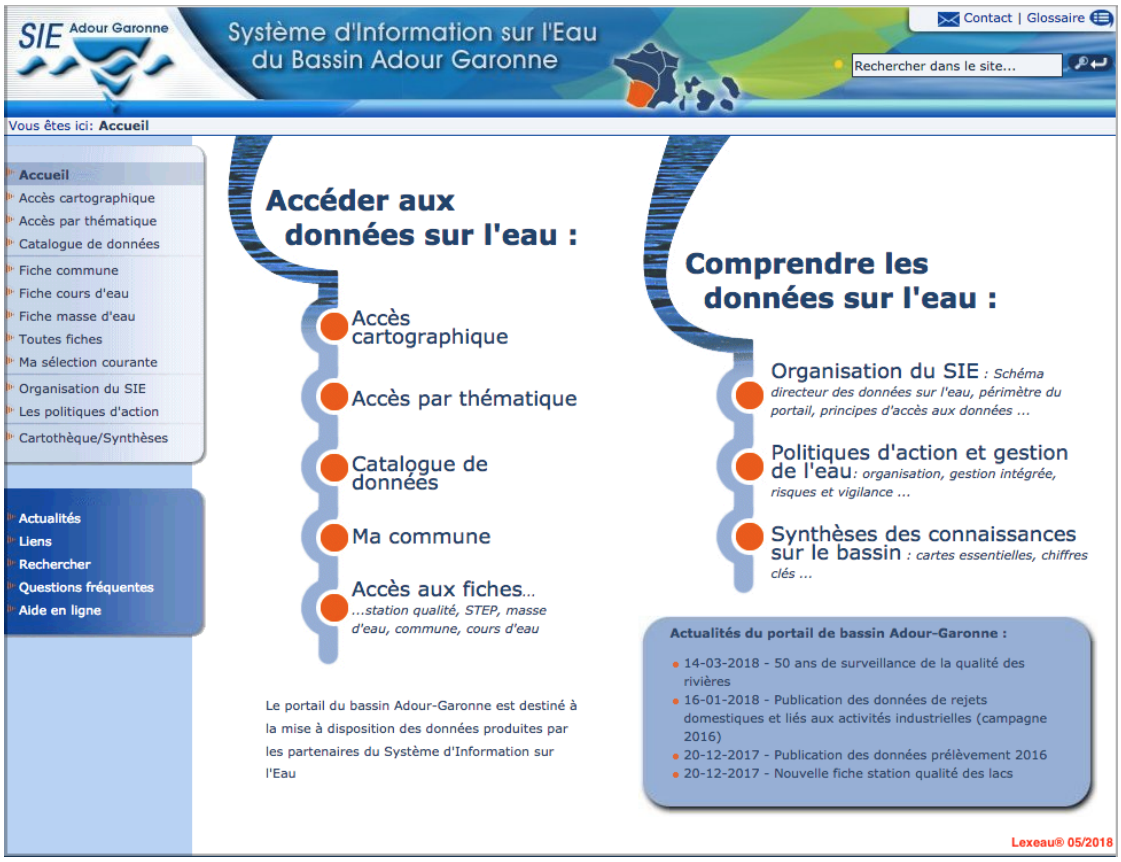
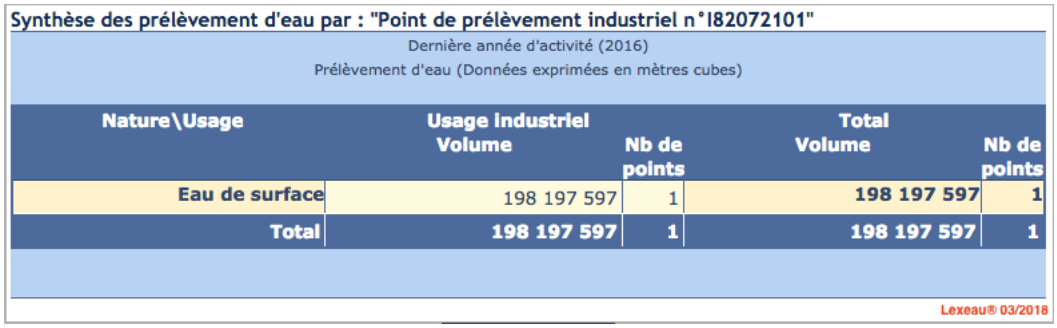


FIGURE 1.11 – Écran d'accueil du système d'information sur l'eau du bassin Adour-Garonne



Synthèse des prélèvement d'eau par : "Point de prélèvement industriel n°I82072101"

Dernière année d'activité (2016)

Prélèvement d'eau (Données exprimées en mètres cubes)

Nature\Usage	Usage industriel Volume	Nb de points	Total Volume	Nb de points
Eau de surface	198 197 597	1	198 197 597	1
Total	198 197 597	1	198 197 597	1

Lexeau® 03/2018

FIGURE 1.12 – Données du SIE Adour-Garonne sur le 'Point de prélèvement I82072101'

apparait sur la carte de la commune de Golfech paramétrée dans le SIE (fig. 1.13), à laquelle nous avons ajouté les encadrés.

La carte permet de voir la masse d'eau dans laquelle s'effectuent les prélèvements (le canal de Golfech) et le code européen de cette masse d'eau (FRFR920). Le numéro SIRET associé à « l'intervenant », au sens où l'entend le SIE Adour-Garonne, permet d'identifier la centrale de Golfech comme établissement de l'entreprise « Électricité de France », au sens du fichier SIRENE de l'INSEE. Avec la transposition des données du SIE dans la base de connaissance,

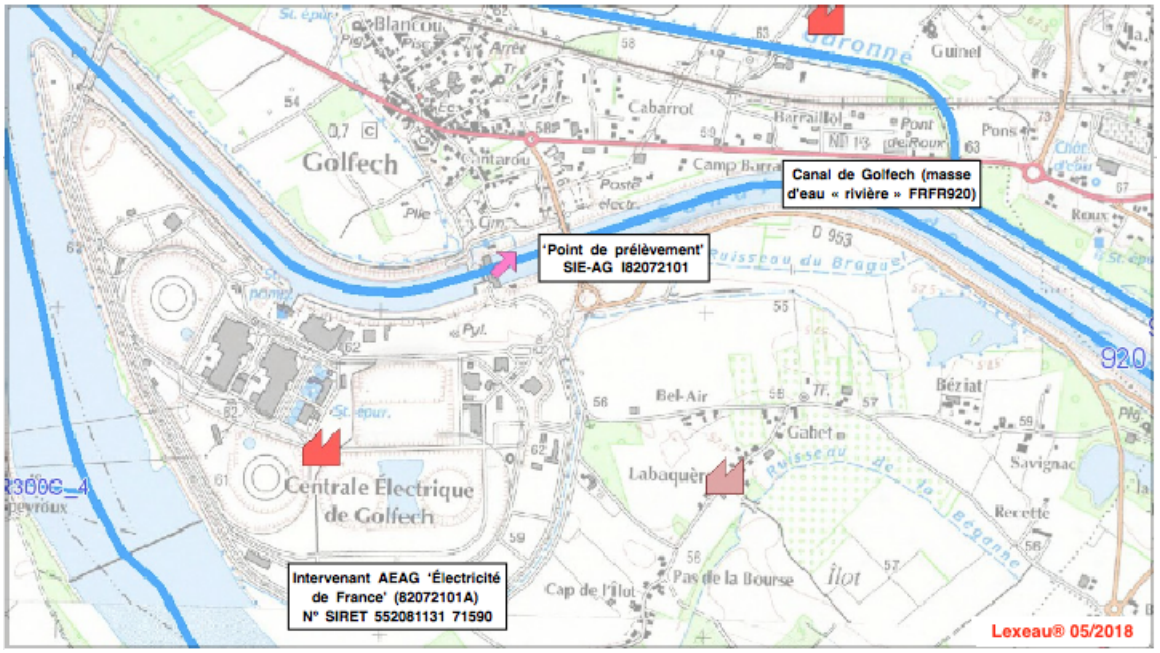


FIGURE 1.13 – Carte de la commune de Golfech éditée dans le SIE Adour-Garonne (extrait)

nous avons considéré l'entreprise « Électricité de France » comme l'« opérateur » des prélèvements tandis que la centrale à proprement parler a été considérée, en tant qu'installation industrielle, comme « usager » du flux anthropique.

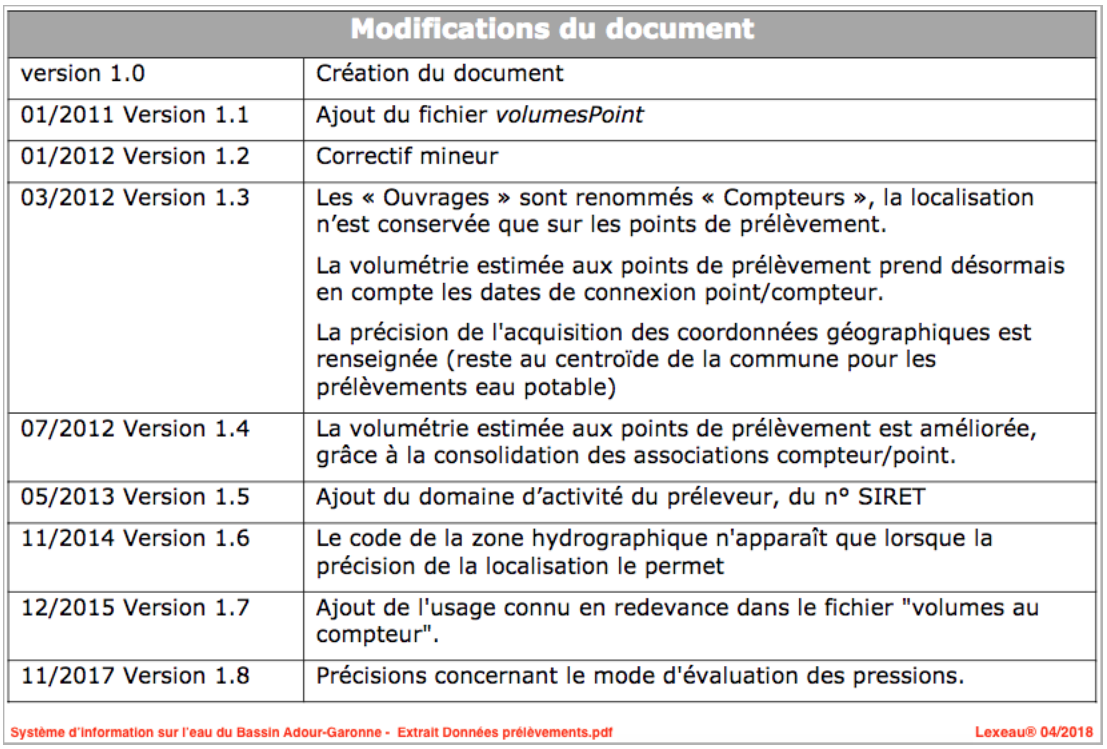
Parmi les propriété à valeur du flux anthropique, nous utilisons des variables de Boole — « vrai » (*'true'*) ou « faux » (*false*) — pour qualifier comme « final » le stade d'usage de l'eau du flux anthropique et comme « collectif » le niveau d'usage de ce flux.

Nous avons considéré que le stade d'usage de l'eau prélevée dans le canal de Golfech était final, ce qui demanderait à être confirmé. Le SIE Adour-Garonne ne donne aucune information sur l'ouvrage de prélèvement dans le canal de Golfech ni sur l'ouvrage d'injection de l'eau du flux dans le réseau d'eau industrielle de la centrale, ni a fortiori sur le point matériel d'injection et le compteur d'injection installé sur l'ouvrage d'injection. Ce déficit d'information n'exclut pas un piquage sur la conduite de refoulement des eaux prélevées pour un stockage temporaire d'une partie de l'eau prélevée dans une retenue d'eau. Si tel était le cas, il conviendrait de considérer deux prélèvements d'eau au lieu d'un seul, avec des volumes d'eau différents dans l'année dont nous ne connaissons actuellement que le total (198 197 597 m3) indiqué dans le SIE Adour-Garonne. Le premier prélèvement serait suivi d'une injection dans un réseau d'eau industrielle et le second suivi d'une restitution dans une masse d'eau. Nous retrouvons ici un élément déictique de l'interprétation de la propriété « stade d'usage de l'eau » : toute l'eau prélevée est-elle visée ici ? toute l'année ?

1.4.1 Modéliser les données source

L'introduction dans la base de connaissance des données du SIE Adour-Garonne sur les prélèvements d'eau pour la centrale de Golfech en 2016 a imposé une refonte de l'ontologie du

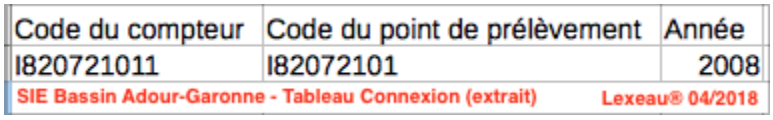
domaine pour modéliser des objets de connaissance du SIE à partir du vocabulaire du système d'information source et transposer les données correspondantes du SIE, notamment les identifiants mis en oeuvre. Cette refonte s'est appuyée sur une note explicative téléchargée dont la figure 1.14 présente un extrait de l'introduction. Il est dit que les volumes sont estimés aux



Modifications du document	
version 1.0	Création du document
01/2011 Version 1.1	Ajout du fichier *volumesPoint*
01/2012 Version 1.2	Correctif mineur
03/2012 Version 1.3	Les « Ouvrages » sont renommés « Compteurs », la localisation n'est conservée que sur les points de prélèvement. La volumétrie estimée aux points de prélèvement prend désormais en compte les dates de connexion point/compteur. La précision de l'acquisition des coordonnées géographiques est renseignée (reste au centroïde de la commune pour les prélèvements eau potable)
07/2012 Version 1.4	La volumétrie estimée aux points de prélèvement est améliorée, grâce à la consolidation des associations compteur/point.
05/2013 Version 1.5	Ajout du domaine d'activité du préleveur, du n° SIRET
11/2014 Version 1.6	Le code de la zone hydrographique n'apparaît que lorsque la précision de la localisation le permet
12/2015 Version 1.7	Ajout de l'usage connu en redevance dans le fichier "volumes au compteur".
11/2017 Version 1.8	Précisions concernant le mode d'évaluation des pressions.

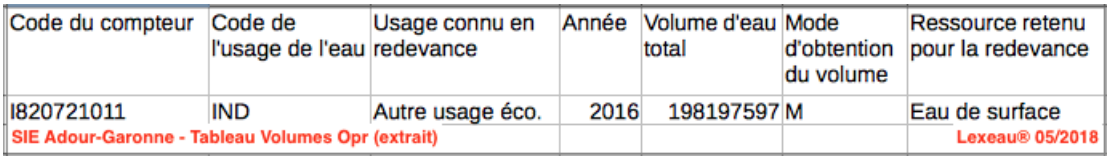
FIGURE 1.14 – Extrait de l'introduction de la note 'Données prélèvements' (extrait)

points de prélèvement, ce qui laisse entendre qu'un compteur peut être connecté à plusieurs points de prélèvement, conduisant à une estimation des volumes prélevés par une répartition entre ces points du volume mesuré au compteur. Le tableau téléchargé reproduit dans la figure 1.15 montre que le point de prélèvement qui nous intéresse est associé à un seul compteur.



Code du compteur	Code du point de prélèvement	Année
I820721011	I82072101	2008

FIGURE 1.15 – *Connexion* des points de prélèvement (extrait)



Code du compteur	Code de l'usage de l'eau	Usage connu en redevance	Année	Volume d'eau total	Mode d'obtention du volume	Ressource retenu pour la redevance
I820721011	IND	Autre usage éco.	2016	198197597	M	Eau de surface

FIGURE 1.16 – Volume d'eau mesuré au compteur EDF Golfech en 2016

Il est donc logique d'attribuer au flux anthropique de la base de connaissance le volume mesuré au compteur du tableau de la figure 1.16, extrait d'un tableau téléchargé depuis le SIE. L'extrait reproduit dans la figure 1.14 précise qu'un compteur identifié dans le SIE est assimilé à un ouvrage. Les points de prélèvement ont été intégrés dans la base de connaissance comme des points matériels, distincts des ouvrages de prélèvement, ce qui n'est pas le cas dans le SIE Adour-Garonne. Le modèle de notre base de connaissance « déborde » donc celui du SIE Adour-Garonne, dont il intègre les identifiants locaux que nous avons préfixés « SIE-AG » pour disposer d'un identifiant national. (cf. fig. 1.15).

1.4.2 Un déficit d'information

La figure 1.18 présente le résultat de notre essai d'intégration des données du SIE Adour-Garonne, avec les réserves faites sur la validité de notre l'hypothèse d'injection de l'eau faute de données complémentaires. Nous avons entouré en rouge les classes du modèle sans référent attesté (cf. 1.4.1), y compris l'ouvrage de prélèvement, confondu avec le « point de prélèvement ». Le rattachement du point de prélèvement à la masse d'eau « Canal de Golfech » s'est fait « à vue », grâce à son code repéré sur la carte du SIE (920) (cf. fig. 1.13). La photo de la figure 1.17 est tirée de l'encyclopédie Wikipedia, consultée le 08/11/2018 [1]

FIGURE 1.17 – La centrale nucléaire de Golfech vue d'Auvillar, une commune voisine.

1. Par Jack ma — Travail personnel, CC BY-SA 3.0, https ://commons.wikimedia.org/w/index.php ?curid=7709534

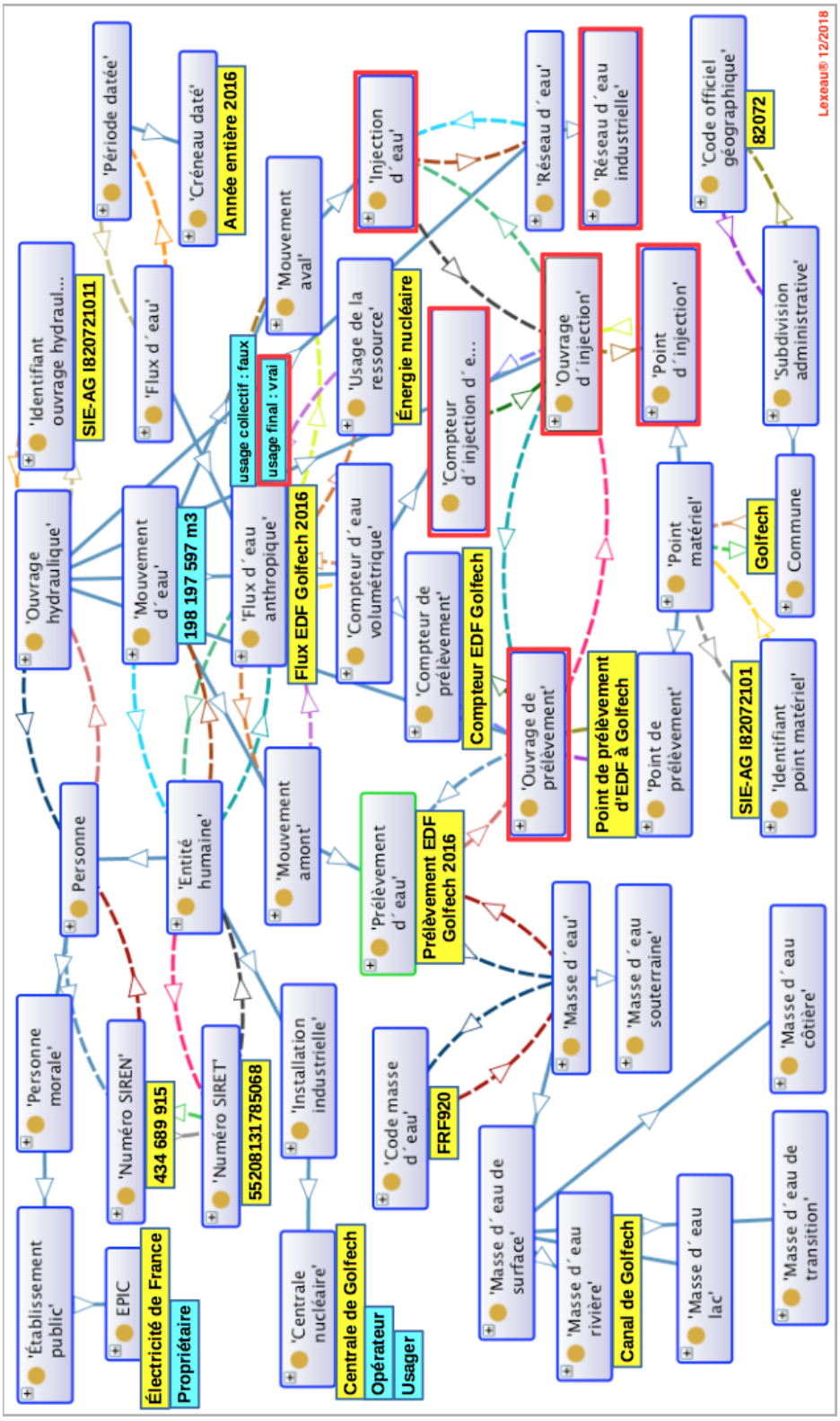


FIGURE 1.18 – Intégration dans l'ontologie du domaine des données du SIE Adour-Garonne sur les prélèvements effectués en 2016 par la centrale nucléaire EDF de Golfech

1.5 De l'ontologie du domaine aux unités lexicales

Une ontologie est une représentation structurée d'un domaine de connaissance (c'est d'ailleurs cette définition qui distingue radicalement la notion d'ontologie en informatique — notion récente — de la notion d'ontologie en philosophie qui traite de « la problématique de l'être » face à celle « du paraître » comme dans l'opposition kantienne entre noumènes et phénomènes). Cela explique qu'un domaine de connaissance puisse donner lieu à différentes ontologies (comme en médecine) en fonction du point de vue qui guide l'élaboration de cette représentation.

Une ontologie travaille de ce fait sur des concepts et non sur des termes d'un lexique. Avant d'aller plus avant, il faut s'entendre sur cette notion de « termes d'un lexique », d'une part parce que des données historiques croisent des données référentielles et conceptuelles et d'autre part parce que les terminologies conservent encore des différences notables.

Si nous prenons des termes concrets comme « siège », on constate rapidement que celui-ci correspond à la fois à des référents (d'ailleurs diversifiés : chaise, fauteuil, banc, etc.) et à un concept qui peut être caractérisé grosso modo par « objet pour s'asseoir ». On retrouve là la description componentielle (*component* signifiant composant constituant, en anglais) bien connue de la classe des sièges avec les co-hyponymes du terme « siège » (analyse déjà présente dans les arbres de Porphyre) et la distinction entre sèmes distinguant les différents types de sièges (« avec des bras », « avec un dossier », etc.) et le sème définitionnel de la classe ou archisème. Mais ce sème qui permet de définir la classe a un statut radicalement différent de celui des autres sèmes, ce que l'analyse componentielle ne met en lumière que sur le plan opératoire. Ce sème renvoie à un concept « objet permettant de s'asseoir ». Comme tous les concepts, celui-ci a un « noyau » (parfois nommé « prototype » ou « haut degré ») correspondant à ce qui a clairement statut de siège, et des zones plus ou moins floues de « pas tout à fait sièges » ou de « sièges pas pour tout le monde » (comme dans le dialogue imaginaire : « je refuse de m'asseoir sur ton pouf ; pour moi ce n'est pas un siège puisque je n'arrive pas à me relever ») avec une frontière séparant les « pas tout à fait sièges » des « non-sièges », frontière qui délimite à la fois le concept « être siège » et la classe des objets du monde nommés « sièges ». Ces concepts n'intéressaient pas a priori le linguiste parce qu'ils relevaient prioritairement de la psychologie. Mais l'émergence d'une linguistique englobant des aspects cognitifs — à partir des années 1970 — a transformé le panorama théorique. Auparavant, les linguistes nommaient « sens » cette appréhension conceptuelle.

Ces concepts relèvent de la pensée et sont présents dans certaines démonstrations en mathématique mais ils ne se formulent dans le langage humain que par le truchement des « mots ». La notion de mot — bien que d'usage courant — n'a rien d'évident, notamment d'une part du fait de facteurs historiques qui font qu'un « mot » peut se présenter sous la forme de « plusieurs mots » (d'où la notion de locution) ou être un composant d'une unité plus grande pourtant ressentie comme « véritable » unité, et d'autre part du fait de la complexité de la notion de sens due principalement aux « extensions de sens » dont parlent abondamment les dictionnaires, et bien que la notion de « mot » paraisse en même temps une notion utile.

De cette problématique sont nées différentes notions, celles de lexème, de lexie, de vocable qui permettent — selon les différentes théories — de nommer ce qui compose le lexique au niveau de différentes strates. Pour l'instant, nous ne pouvons entrer dans la complexité de ces

différentes dénominations. C'est pourquoi nous recourrons au terme « unité lexicale », tout en ayant conscience du flou résiduel que provoque cette décision. Nous n'aurons donc pour l'instant que deux modes d'appréhension des faits lexicaux, celui des concepts et celui des unités lexicales. Dans cette section, nous présentons sur l'exemple des prélèvements d'eau de la section 1.4 les trois sources du lexique dans l'ontologie du domaine et la transposition d'un concept de l'ontologie en unités lexicales *pivots* et *satellites*.

1.5.1 Les trois sources du lexique dans l'ontologie du domaine

Les noms des classes d'entités typées de l'ontologie du domaine constituent la première source du lexique. Ils renvoient à des concepts transposés dans le lexique s'ils sont pertinents dans une activité langagière du domaine, avec un sens bien défini dans la strate discursive qui caractérise cette activité. L'exemple de la figure 1.19 porte sur la classe des prélèvements d'eau, dont le nom est transposé avec le même libellé, au singulier, dans la strate discursive technico-scientifique, précédé de l'acronyme de sa strate. Le choix des noms des classes est

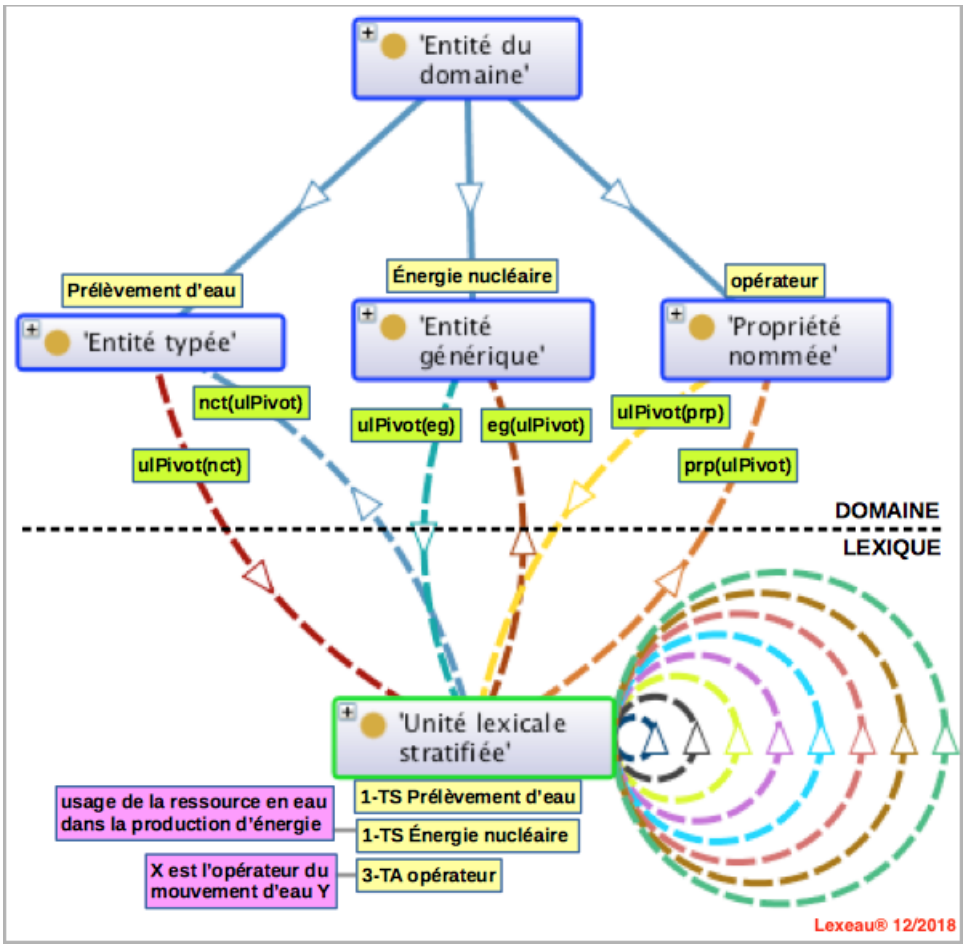


FIGURE 1.19 – Les trois sources du lexique dans l'ontologie du domaine

une étape essentielle dans la construction de l'ontologie du domaine. Cette construction s'effectue sur trois axes : l'axe hiérarchique, du général au particulier, l'axe relationnel, entre les

individus de deux classes ou les individus de la même classe, par exemple l'acronyme et sa forme développée, rangés par ordre alphabétique dans la même classe, et l'axe des propriétés « à valeur » des individus de la classe. On évitera de créer des classes d'entités typées qui ne font que regrouper des sous-classes sans entrer dans une ou plusieurs relations avec d'autres classes ou sans avoir en propre des propriétés transmises à leur sous-classes par « héritage ». Ce sont des classes « fantômes ». Cela ne sera pas toujours possible à un niveau de généralités élevé, comme par exemple la classe des entités du domaine et celle des entités du lexique, introduites dans l'ontologie globale du projet pour rendre compte de la partition du domaine de connaissance.

Les entités génériques constituent la deuxième source du lexique. Ces entités se définissent *a contrario* comme n'étant pas des entités typées. Nous avons créé deux classes d'entités génériques. La première, intitulée « Discipline, matière », porte sur les disciplines et les matières scientifiques ou non du domaine de l'eau, sans chercher à en faire une nomenclature, donc sans introduire une hiérarchie de sous-classes. Les instances de cette classe ne sont pas exclusives l'une de l'autre. L'objectif est de qualifier, si possible avec une seule instance, la discipline ou la matière principale d'un document. La deuxième classe, intitulée « Thème, contexte », poursuit un objectif plus ambitieux. Son développement a commencé par une réflexion sur les usages de l'eau, étendue ensuite au concept plus général du « rapport à l'eau », trop flou pour être choisi comme nom de classe mais qui sous-tend la construction d'une nomenclature sur plusieurs niveaux dont les postes ne se recoupent pas, et qui peut être révisée.

L'exemple de la figure 1.19 illustre cet objectif. Nous avons qualifié l'usage de la ressource dans la centrale nucléaire de Golfech en faisant référence à l'énergie nucléaire, instance de la classe « Production énergétique » parmi les classes de l'usage industriel de l'eau. La figure 1.20 présente les classes de l'usage industriel de la ressource.

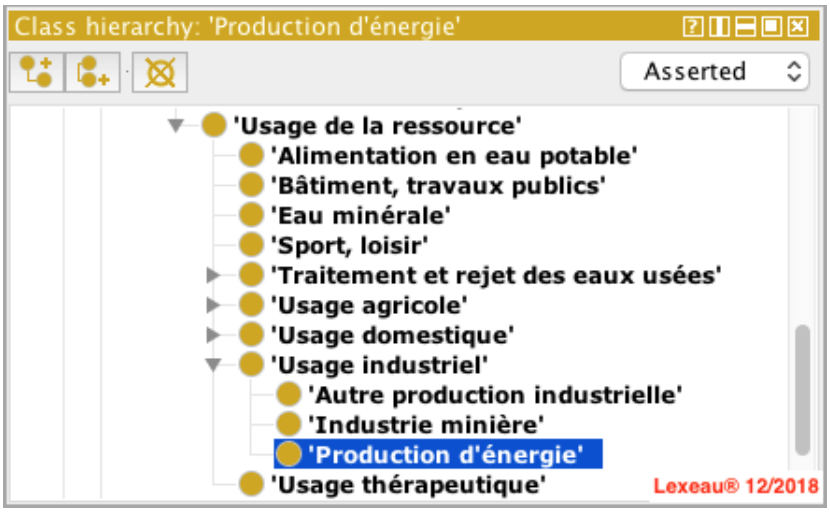


FIGURE 1.20 – Place de la production d'énergie dans le thème « Usage de la ressource »

La figure 1.21 présente les instances de la classe « Production d'énergie », parmi lesquelles on trouve l'instance « Énergie nucléaire ». Les noms de cette instance est transposé en l'état dans le lexique dans la strate technico-scientifique, ce qui donnée « 1-TS Énergie nucléaire » comme entrée, étant entendu qu'il ne s'agit pas de l'énergie nucléaire en général mais de

l'usage industriel de la ressource en eau dans la production d'énergie nucléaire, ce qui devra être précisé au début de l'article du lexique.

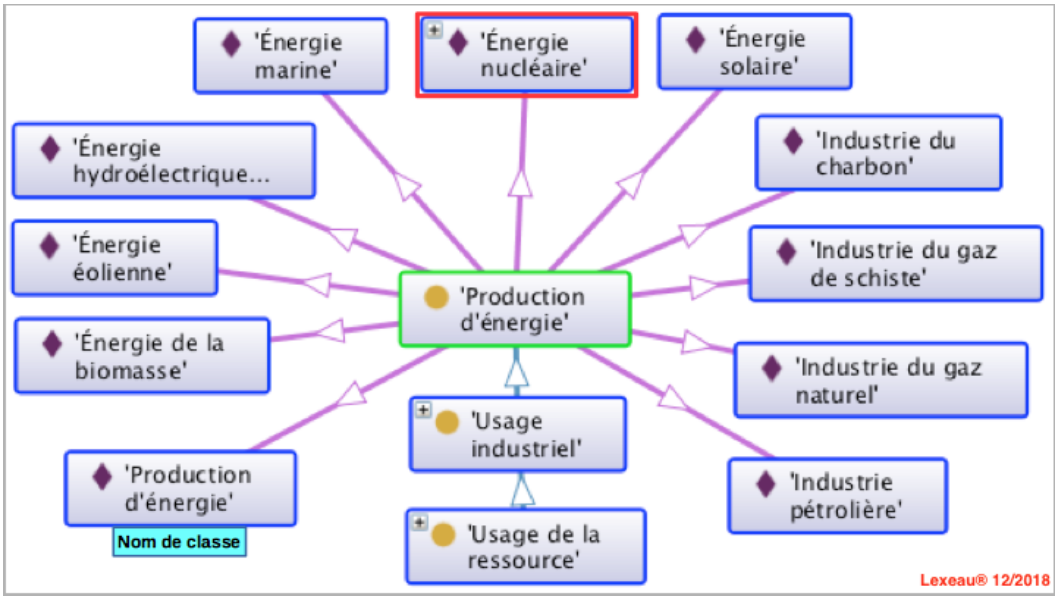


FIGURE 1.21 – Instances de la classe « Production d'énergie » (usage industriel de la ressource)

La troisième source de vocabulaire pour le lexique correspond à la mise en oeuvre de relations entre classes, sous la forme de propriété relationnelles, et dans l'attribution de propriétés à valeur pour certaines entités de l'ontologie. Le mot « opérateur » a-t-il vocation à figurer dans le lexique de l'eau ? Si oui, dans quelle strate discursive ? C'est un mot très employé dans le discours administratif sur l'eau. On parle d'opérateur ou d'opérateur délégué dans le cas d'une station d'épuration, d'un systèmes d'assainissement urbain, etc.

Dans l'état actuel de l'ontologie du domaine, le mot est employé pour désigner X (une entité humaine) comme « opérateur » d'un mouvement d'eau. Il entre de ce fait dans la classe des propriétés nommées de l'ontologie avant d'être transposé en entrée dans la strate discursive technico-administrative du lexique sous la forme « 3-TA opérateur », en précisant son emploi dans l'ontologie du domaine, où « X est l'opérateur du mouvement d'eau Y », au début de l'article du lexique. Si l'ontologie s'enrichit d'autres emplois, ils seront ajoutés en début d'article. Par convention, ce type d'entrée dans le lexique commence par une minuscule pour marquer son origine dans l'ontologie du domaine. Le problème est le même pour la transposition des propriétés « à valeur » dans le lexique, qui jouent un rôle très important dans le domaine de l'eau. Elles seront introduites dans l'ontologie en tant que propriété nommée avec une ou plusieurs unités de valeur possibles, en rappelant dans l'article qui leur sera consacré les classes d'objets de l'ontologie où elles sont employées et la ou les unités de valeur de leur emploi. On notera sur cet exemple la différence de traitement des articles du lexique en fonction de l'origine des entrées dans l'ontologie du domaine. L'idée générale est que l'article du lexique consacré à ce type d'unité lexicale, qualifié de« pivot » doit refléter la place du concept lexicalisé dans l'ontologie au moment de sa transposition, ce qui revient à prévoir la mise à jour de l'article dans le cas d'une mise à jour de l'ontologie du domaine dans l'emploi du concept.

1.5.2 Des unité lexicales « pivots » et « satellites »

En France, la locution « prélèvement d'eau » est interprétée de façon différente par la police de l'eau qui délivre une « autorisation de prélèvement » et par l'agence de l'eau qui émet un titre de perception de la redevance pour « prélèvement d'eau sur la ressource ». Pour la police de l'eau, on ne sait pas quel est stade d'usage de l'eau du prélèvement qui fait l'objet d'une autorisation ou d'une déclaration préalable. Il peut être « final » ou « non final, lorsque l'eau prélevée n'est pas utilisée immédiatement mais stockée dans une retenue d'eau, par exemple. Pour les agences de l'eau, la redevance n'est due que pour un usage final, par exemple lorsque l'eau prélevée est injectée dans un réseau de distribution d'eau potable, de façon à ne pas taxer deux fois la même eau. Pour prendre en compte cette distinction dans le lexique, nous avons introduit deux locutions, pour qu'il n'y ait pas d'ambiguïté sur la signification d'une locution trop vague, pour un auditoire comprenant des « pétitionnaires » (d'une demande d'autorisation) et des « redevables ». Les entrées du lexique sont respectivement, « Prélèvement d'eau au sens de la police de l'eau » et « Prélèvement d'eau soumis à redevance ». Ces deux unités lexicales relèvent d'un discours technico-administratif et sont rattachées à cette strate discursive.

Nous avons traduit « prélèvement d'eau » par « *water withdrawal* » dans l'ontologie du domaine et conservé cette traduction dans la strate technico-scientifique du lexique. Comme nous l'avons vu, la locution est trop vague pour être transposée en français dans la strate discursive technico-administrative, ce qui n'est pas le cas pour l'anglais.

Les textes officiels britanniques ne font pas la distinction, comme en France, entre police de l'eau et agences de l'eau en matière de prélèvements d'eau, ce qui interdit de traduire en anglais les entrées adoptées dans la strate discursive technico-administrative. La locution *water abstraction* est employée dans le textes officiels de façon explicite. Elle est retenue en anglais dans la strate discursive technico-administrative comme *3-TA Water abstraction*, avec « 3-TA Prélèvement d'eau » dans la version française, ce qui permet de la relier dans cette version, sans la définir, à ses hyponymes « 3-TA Prélèvement d'eau au sens de la police de l'eau » et « 3-TA Prélèvement d'eau soumis à redevance », lesquels seront suivis dans la version anglaise du lexique de la mention « @fr » pour marquer que ces unités lexicales n'existent qu'en français.

La recherche d'un emploi courant de « prélèvement d'eau » sur le web a débouché sur un site bilingue de l'OCDE sur les prélèvements d'eau en volume d'eau, avec l'emploi de « *water abstraction* » en variante de « *water withdrawal* ». L'unité lexicale « prélèvement d'eau/*water withdrawal* » a été ajoutée dans la strate discursive courante comme unité lexicale satellite de l'unité lexicale pivot de même libellés dans la strate discursive technico- scientifique, avec l'ajout d'une unité discursive alternative « prélèvement d'eau/*water abstraction*, associée à l'unité lexicale courante, pour prendre en compte cette variante en anglais.

La transposition du concept « Prélèvement d'eau » de l'ontologie du domaine dans le lexique est présentée dans la figure 1.22. Les unités lexicales stratifiées issues de la transposition figurent sur fond jaune sous la classe de leur strate discursive. Les relations de transposition d'un concept dans une unité lexicale pivot puis, à partir de celle-ci, dans des unités lexicales satellites sont notées sur fond vert. La légende des relations est la suivante, avec l'abrégé « ul » pour « unité lexicale » :

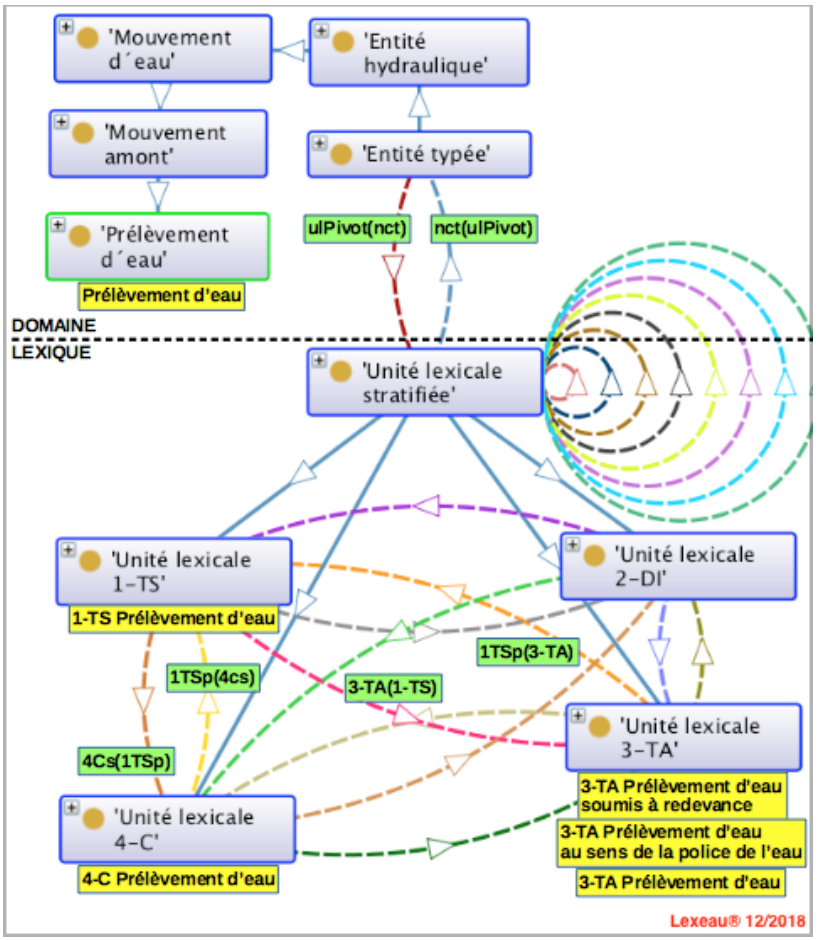


FIGURE 1.22 – Transposition du concept « Prélèvement d'eau » dans le lexique

- ulPivot(nct) : unité lexicale pivot issue du nom d'une classe d'entités typées ;
- nct(ulPivot) : nom de la classe d'entités typées à l'origine de l'unité lexicale pivot ;
- ulPivot(eg) : ul pivot issue d'une entité générique (nom ou instance d'une classe) ;
- 3TAs(1TSp) : ul technico-administrative satellite de l'ul technico-scientifique pivot ;
- 1TSp(3TAs) : ul technico-scientifique pivot de l'ul technico-administrative satellite ;
- 4Cs(1TSp) : ul courante satellite de l'ul technico-scientifique pivot ;
- 1TSp(4Cs) : ul technico-scientifique pivot de l'ul courante satellite.

La règle de la transposition des concepts dans le lexique depuis l'ontologie du domaine est qu'un concept de cette ontologie — nom d'une classe d'entités typées, nom ou instance d'une classe d'entités génériques, nom issu d'une propriété — est transposé dans une unité lexicale qualifiée de « pivot ». La relation fonctionne dans l'autre sens, entre une unité lexicale pivot et le concept source de l'ontologie du domaine de connaissance.

L'élaboration du lexique se poursuit donc au delà des unités lexicales « pivot » avec l'introduction, dans d'autres strates discursives, d'unités lexicales « satellite », en ce sens qu'elle gravitent autour de l'unité pivot sans être transposées du domaine de connaissance. Leur libellé est identique ou proche de celui de l'unité pivot, dans une acception différente. Ainsi les

« prélèvements d'eau » des tableaux de données édités par l'OCDE [2] sont en fait des volumes d'eau, et non des opérations humaines [3]. Le site de l'OCDE propose la variante *water abstraction* de l'unité lexicale courante *water withdrawal* [4]. Nous verrons comment cette variante est introduite en entrée du lexique comme unité discursive alternative.

En conclusion, les principes généraux appliqués pour introduire dans le lexique les unités lexicales issues du concept de prélèvement d'eau sont les suivants :

- Il est possible de construire une ontologie conceptuelle du domaine de l'eau dans laquelle des locutions utilisées dans ce domaine trouveront leur place, sous la forme de concepts qui couvrent sans redondance les « objets du monde » du domaine, y compris les entités humaines, et ce en tant que noms des classes d'entités typées ou génériques, instances des classes d'entités génériques et propriétés relationnelles ou à valeur.
- L'analyse des discours du domaine permet de ranger des unités lexicales de libellé identique ou proche dans différentes strates discursives, dans la mesure où ces discours se distinguent dans le choix des locutions, leur agencement et le sens qu'elles prennent pour l'auditoire, pour éviter des ambiguïtés et des incompréhensions. Le discours courant, celui des *media* du quotidien (presse, radios et télévision) correspond à une strate discursive. Il se distingue du discours technico-administratif et du discours technico-scientifique.
- Les sources primaires du lexique sont les concepts de l'ontologie du domaine, transposés en tant qu'unités lexicales « pivot » dans la strate discursive qui correspond au contexte de leur introduction dans l'ontologie.
- Les unités lexicales pivot sont la source d'un enrichissement du lexique en leur associant des unités lexicales « satellite », d'un libellé identique ou proche, dans d'autres strates discursives.
- Les unités lexicales sont préfixées par l'acronyme de leur strate discursive. Dans des textes, elles prennent la forme d'unités discursives qui vont figurer en entrée du lexique sous une forme grammaticale conventionnelle avec le même libellé, sans préfixe.

1.6 Des unités discursives en entrée du lexique

Les unités discursives sont, pour l'utilisateur, les entrées alphabétiques du lexique stratifié. Les unités lexicales ont comme préfixe le sigle de leur strate discursive, suivi d'un libellé qui doit varier lorsque plusieurs unités lexicales satellites de la même unité pivot sont dans la même strate discursive, comme nous l'avons vu avec les prélèvements d'eau dans la strate discursive technico-administrative. Les unités discursives centrales reprennent à l'identique les libellés en français et en anglais de l'unité lexicale pivot sans son préfixe. Elles donnent ainsi accès aux unités lexicales dans la forme grammaticale conventionnelle des entrées d'un dictionnaire [5]. Une unité lexicale peut être ainsi rattachée à plusieurs unités discursives. Une unité discursive centrale est rattachée à au moins une unité lexicale stratifiée et une ou plusieurs unités lexicales

2. https ://data.oecd.org/fr/water/prelevements-d-eau.htm

3. "Les prélèvements d'eau désignent les volumes d'eau douce extraits définitivement ou temporairement d'une source souterraine ou de surface et transportés sur leur lieu d'usage"

4. "*Water withdrawals, or water abstractions, are defined as freshwater taken from ground or surface water sources, either permanently or temporarily, and conveyed to a place of use*"

5. À l'infinitif pour les verbes et au masculin singulier, sauf exception, pour les noms et les adjectifs.

satellites. Il reste la question des variantes des unités discursives, comme nous l'avons vu sur le site de l'OCDE avec la variante *water abstraction* de l'unité *water withdrawal* lexicalisée dans la strate discursive courante. Nous avons créé, à côté de la classe des unités discursives centrales, une classe d'unités discursives alternatives, qui ont vocation à figurer comme telles en entrée du lexique. Tout comme les unités discursives centrales, chaque unité discursive alternative est rattachée à une unité lexicale. Le graphe relationnel de la figure 1.23 présente un exemple d'unité discursive alternative en anglais.

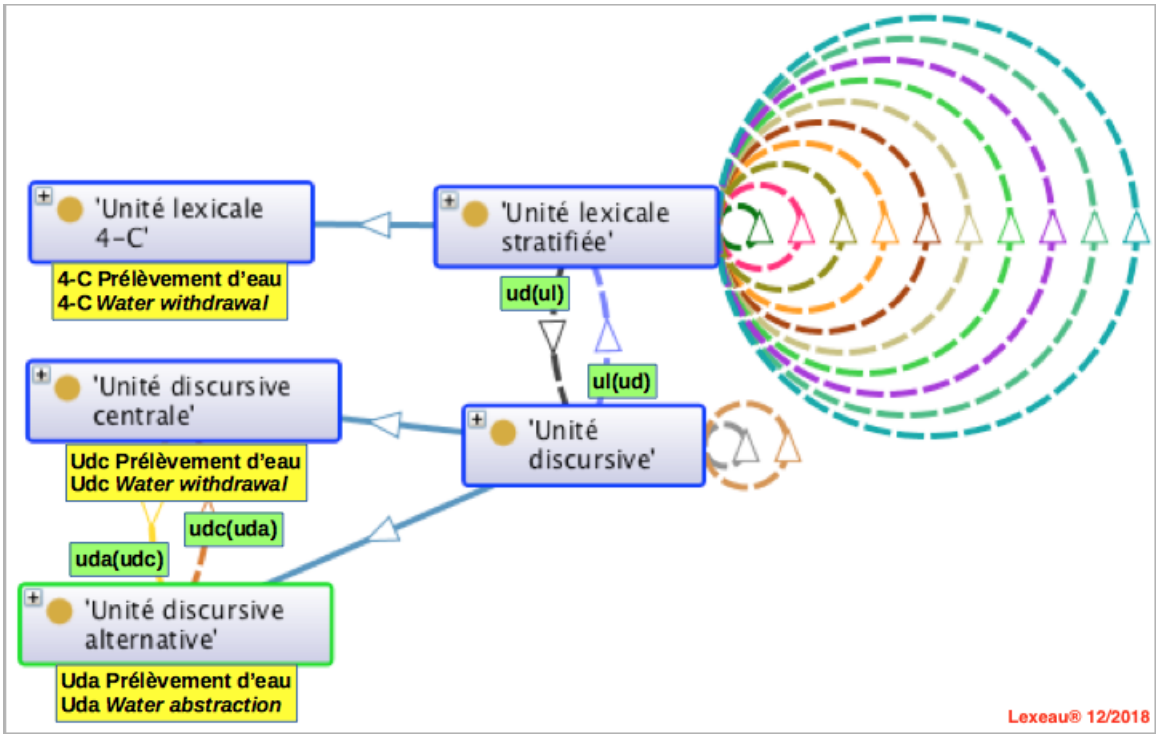


FIGURE 1.23 – Exemple d'une unité discursive alternative en anglais

Les relations du graphe se lisent ainsi :

- ud(ul) : unité discursive de rattachement de l'unité lexicale pivot ;
- ul(ud) : unité lexicale rattachée à l'unité discursive ;
- uda(udc) : unité discursive alternative de l'unité discursive centrale ;
- udc(uda) : unité discursive centrale de l'unité discursive alternative ;

1.7 Acronymes et noms propres

La question des acronymes se pose pour tous les libellés de la base de connaissance. Nous l'avons traitée pour les unités lexicales, les noms propres et les unités discursives.

1.7.1 Des acronymes associés à leur forme développée

Chaque acronyme est associé à une ou plusieurs formes développées dans la classe où il figure par une relation directe et inverse. Le nombre de classes où peuvent figurer des acronymes avec leur forme développée est limité à trois (cf. fig. 1.27). La forme développée des acronymes employés seuls dans d'autres classes sera signalée dans une annotation (*comment*). Les propriétés relationnelles entre les classes où les deux formes sont autorisées portent toujours sur la forme développée. Si une forme développée est transférée sous la même forme d'une classe A à une classe B, la forme développée dans B reprendra l'acronyme de la forme développées dans A. Les acronymes sont donc traités comme des entités à part entière du domaine, du lexique ou du discours. Les règles ci-dessus devront être testées avec un suivi des acronymes de la base de connaissance pour en contrôler le nombre et la pertinence.

La figure 1.24 donne, parmi les entités typées de l'ontologie du domaine, l'exemple des classes d'établissements publics reprenant les acronymes utilisés en France. Parmi les instances de la classe des EPIC (Établissements publics à caractère industriel et commercial) de la figure 1.25, on trouve BRGM, Électricité de France et EPIC. La présence d'EPIC dans la liste s'explique par la nécessité de le faire figurer avec les instances de la classe qui porte son nom pour établir le lien avec l'unité lexicale pivot du même nom, en tant que nom d'une classe d'entités typées de l'ontologie du domaine. La forme développée de BRGM est indiquée en commentaire, dans les annotations de l'entité (fig. 1.26). L'acronyme EPIC est transposé dans le lexique, dans la strate discursive technico-administratives. L'acronyme BRGM est transposé dans le lexique comme nom propre lexicalisé.

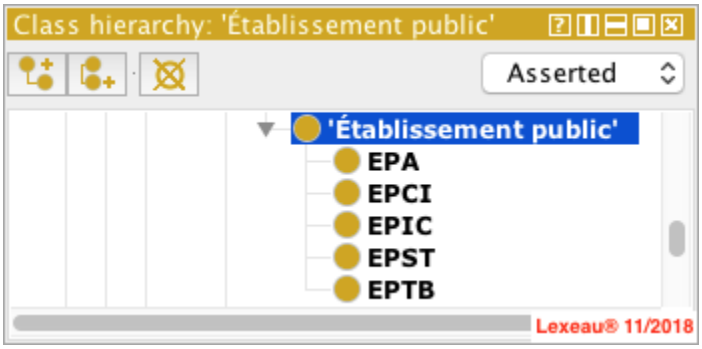


FIGURE 1.24 – Classe et sous-classes des établissements publics dans l'ontologie du domaine

FIGURE 1.25 – Instances de la classe EPIC dans l'ontologie du domaine

Les relations étiquetées sur fond vert des graphes relationnels de la figure 1.27 présentent la

FIGURE 1.26 – Forme développée de l'acronyme BRGM en annotation de l'instance typée

prise en compte des acronymes dans les unité lexicales, les noms propres, les noms propres lexicalisés et les unités discursives. Nous reviendrons sur les autres relations de la figure.

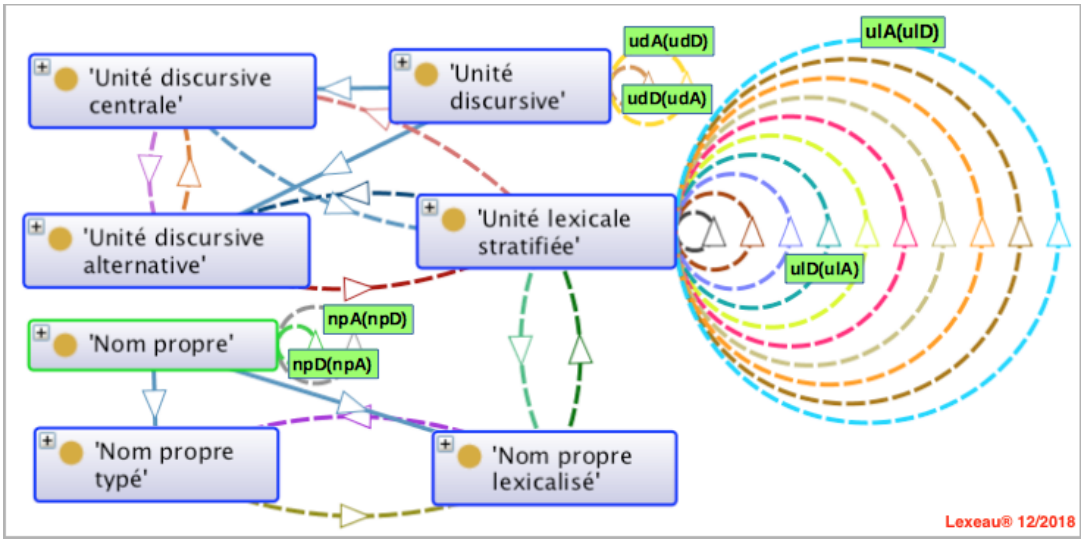


FIGURE 1.27 – Trois classes d'acronymes et de formes développées

1.7.2 Lexicalisation des noms propres typés

On connait l'importance des noms propres dans le domaine de l'eau, pour nommer des entités géographiques, des personnes, des événements, etc. Les noms des individus des classes d'entités typées peuvent être transposés dans la classe des noms propres lexicalisés. La difficulté pratique avec les noms propres et que le libellé d'un nom propre typé peut renvoyer à plusieurs « objets du monde », la « Marne » comme rivière ou comme département français, « Verdun » comme ville ou comme la bataille de la guerre de 14-18, etc. La classe de noms propres lexicalisés du lexique permet de dissiper toute ambiguïté sur l'objet désigné. Les instances de cette classe sont construites sur l'unité lexicale pivot qui renvoie au nom de la classe source d'entités typées

où figure l'instance typée et sur un ou plusieurs noms propres typés. La lexicalisation d'un nom propre consiste à le rattacher à une unité lexicale pivot et à un ou plusieurs noms propres typés.

La figure 1.28 présente l'exemple du prélèvement d'eau effectué à Golfech en 2016 pour la centrale nucléaire d'EDF, que nous avons nommé « Prélèvement EDF Golfech 2016 ». Le nom propre lexicalisé est construit sur les noms propres typés « EDF », « Golfech » et « 2016 » et sur l'unité lexicale pivot «1-TS Prélèvement d'eau ». L'objet visé par cette désignation peut être un « prélèvement soumis à redevance », doté des attributs utiles à l'agence de l'eau Adour-Garonne pour établir un titre de redevance, mais seulement dans l'extrait du texte qui en explique l'emploi dans le système d'information de l'agence. Le nom propre reste dans sa strate discursive d'origine. Ce caractère « rigide » du nom propre renvoie à la notion

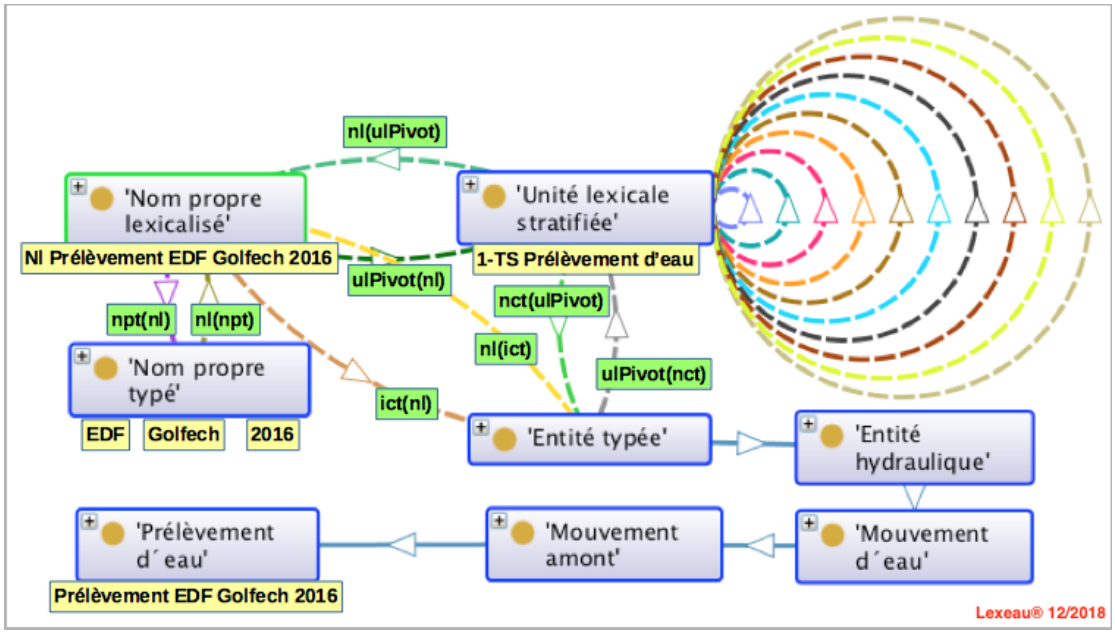


FIGURE 1.28 – Graphe relationnel des noms propres lexicalisés avec un exemple

de « désignateur rigide » développée par Samuel Kripke ([Kripke, 1982, Kripke, 1980], ici « Nl Prélèvement EDF Golfech 2016 » et à la distinction que faisait Frege avant lui entre la « dénotation » de l'objet (Nl Prélèvement EDF Golfech 2016) et le « sens » dans un contexte donné. En philosophie du langage, Bertrand Russell s'est opposé à cette distinction dans sa théorie des « descriptions finies » développée après un article célèbre [Russel, 1905]. Il faudrait revenir sur cette théorie, qui ne nous parait pas incompatible avec la façon dont les noms propres sont traités dans le lexique, en prenant leur origine dans l'ontologie du domaine avec les attributs de leur classe d'origine. La figure 1.28 illustre la formation d'un nom propre composite sur plusieurs noms propres typés, eux mêmes associés à d'autres noms propres lexicalisés construits sur d'autres unités lexicales pivot (« 3-TA EPIC », « 1-TS Centrale nucléaire » et « 1-TS Année »).

1.8 Des textes pour les unités lexicales

En dehors des exemples d'emploi des noms propres lexicalisés, les définitions, les exemples d'emploi et les gloses forment les trois classes de textes du lexique qui concernent les unités lexicales. La figure 1.29 présente le graphe relationnel des textes du lexique rattachés aux unités lexicales.

Nous avons ajouté au graphe édité dans l'ontologie les unités lexicales dont les définitions et les gloses sont présentées dans cette section (sur fond jaune) et les unités lexicales pivot mentionnées dans ces gloses et auxquelles les gloses sont reliées (sur fond orange clair).

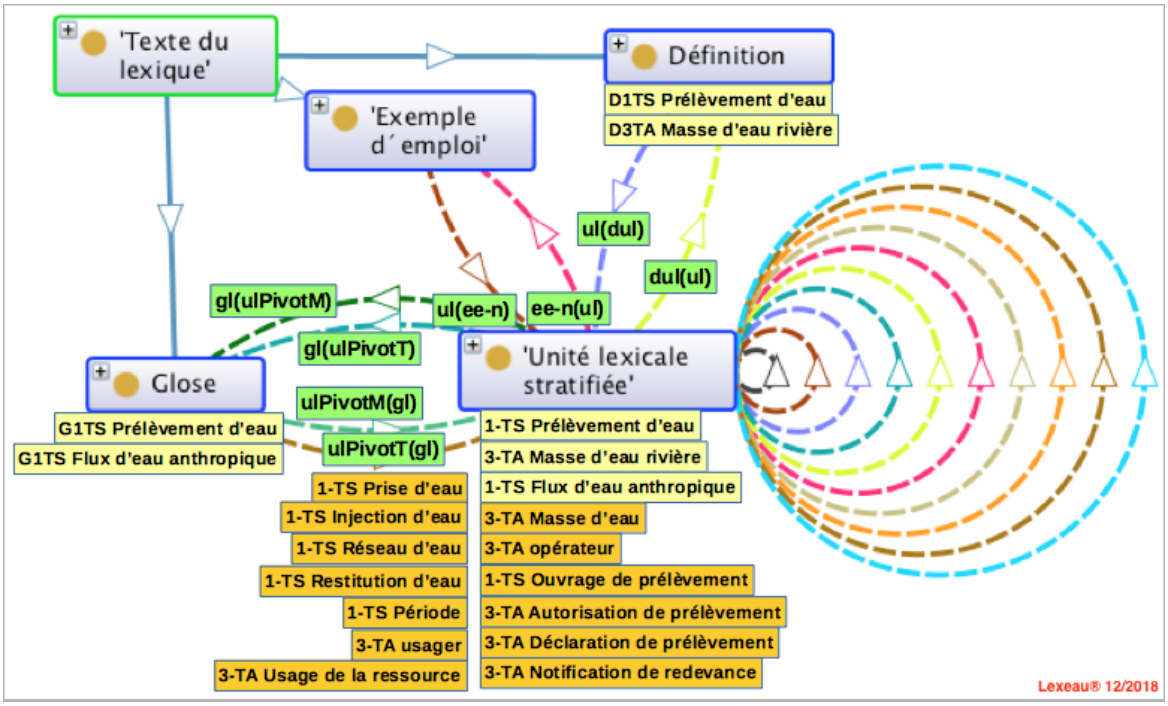


FIGURE 1.29 – Graphe relationnel des textes des unités lexicales

Légende :

- dul(ul) : définition de l'unité lexicale stratifiée ;
- ul(dul) : unité lexicale définie ;
- ee-n(ul) : exemple d'emploi numéroté de l'unité lexicale ;
- ul(ee-n) : unité lexicale de l'exemple d'emploi numéroté ;
- gl(ulPivotT) : glose de l'unité lexicale pivot en tant que « tête » de la glose ;
- ulPivot(gl) : unité lexicale pivot « tête » de la glose ;
- gl(ulPivotM) : glose faisant mention de l'unité lexicale pivot ;
- ulPivotM(gl) : unité lexicale pivot mentionnée dans la glose.

1.8.1 Des définitions et des gloses pour les unités lexicales pivot

Les définitions des prélèvements d'eau sont reportées en annexe (section A.1.1).

La glose d'une unité lexicale pivot est rédigée au vu du graphe relationnel « de proximité » de la classe dont elle est issue, sur la base des classes en relation immédiate avec la classe source dans l'ontologie du domaine. On utilisera les unités lexicales pivot sans leur préfixe. Les propriétés nommées issues des relations de proximité sont intégrées dans l'énoncé. Le graphe de proximité de la classe source de l'unité lexicale pivot « 1-TS Prélèvement d'eau » est présenté dans la figure 1.30. Les liens hiérarchiques ne sont pas pris en compte et ne

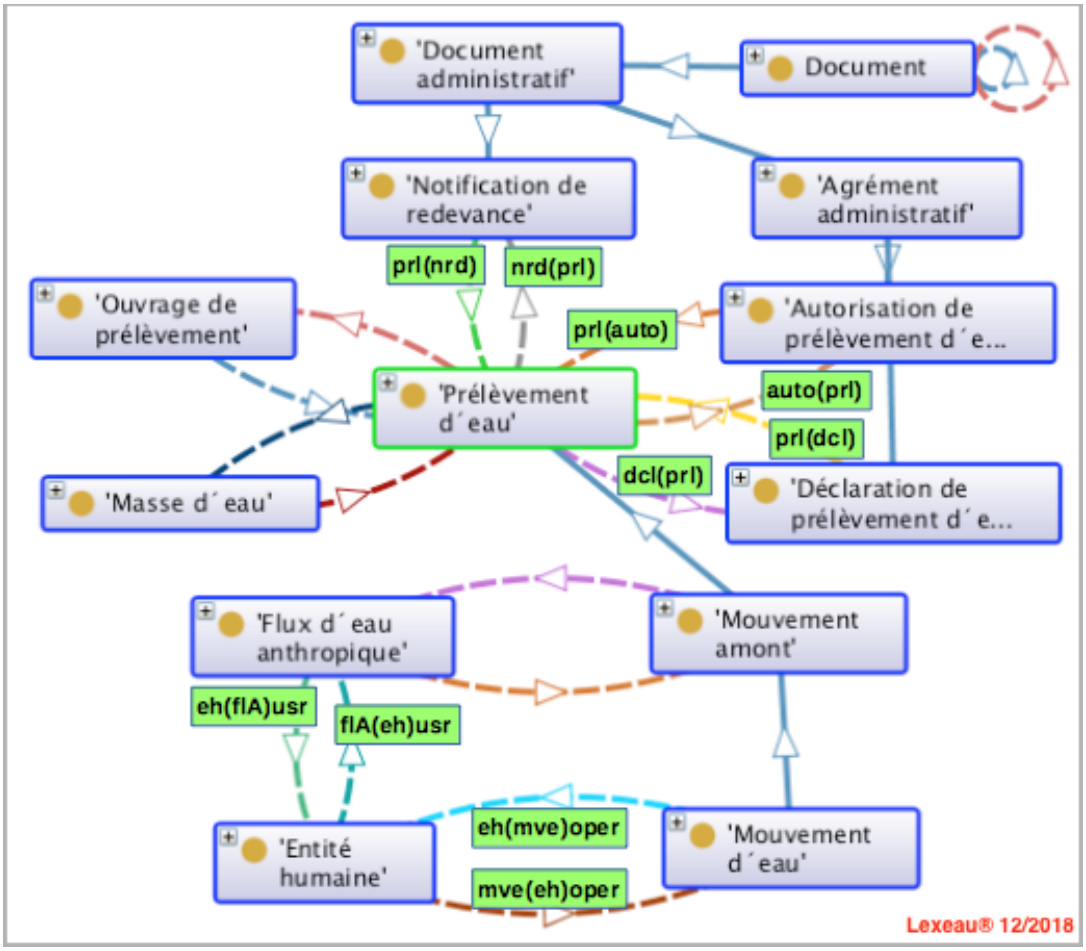


FIGURE 1.30 – Graphe de proximité de la classe « Prélèvement d'eau »

comptent pas dans la recherche des concepts proches. L'unité lexicale pivot est la « tête » de la glose principale « G1TS Prélèvement d'eau », présentée dans la figure 1.31.

Une glose est « secondaire » pour une unité lexicale si elle est mentionnée dans la glose principale d'une autre unité lexicale. Elle apporte des associations complémentaires. La période d'occurrence d'un prélèvement d'eau est ainsi celle qui figure dans la glose principale du flux d'eau anthropique généré (fig. 1.32). Associer la période aux mouvements d'eau aurait été une erreur de conception de l'ontologie du domaine, avec un flux généré par un prélèvement sur

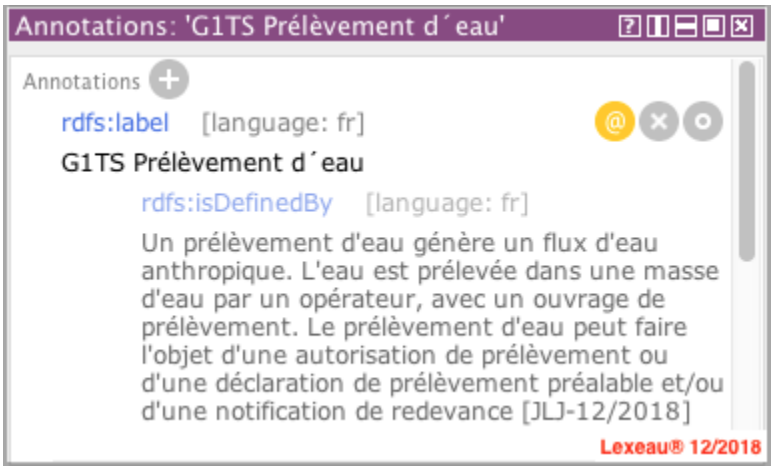


FIGURE 1.31 – Glose principale de l'unité lexicale pivot « 1-TS Prélèvement d'eau »

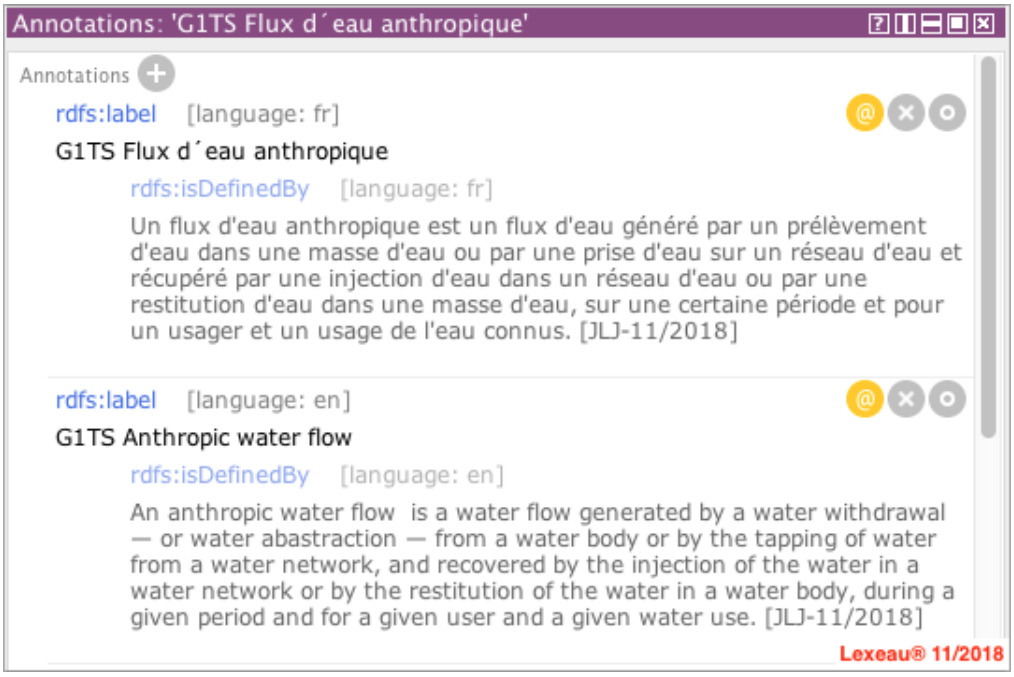


FIGURE 1.32 – Glose de l'unité lexicale pivot « 1-TS Flux d'eau anthropique »

une période donnée puis restitué dans une masse d'eau ou injecté dans un réseau sur une autre période.

La figure 1.33 donne la définition des masses d'eau rivière dans le lexique, tirée d'un extrait de l'article 2 de la directive cadre sur l'eau consacré aux définitions de la terminologie employée.

La remontée à la source de cette définition est présenté dans la figure 1.34.

FIGURE 1.33 – Définition de « 3-TA Masse d'eau rivière » dans la directive cadre sur l'eau

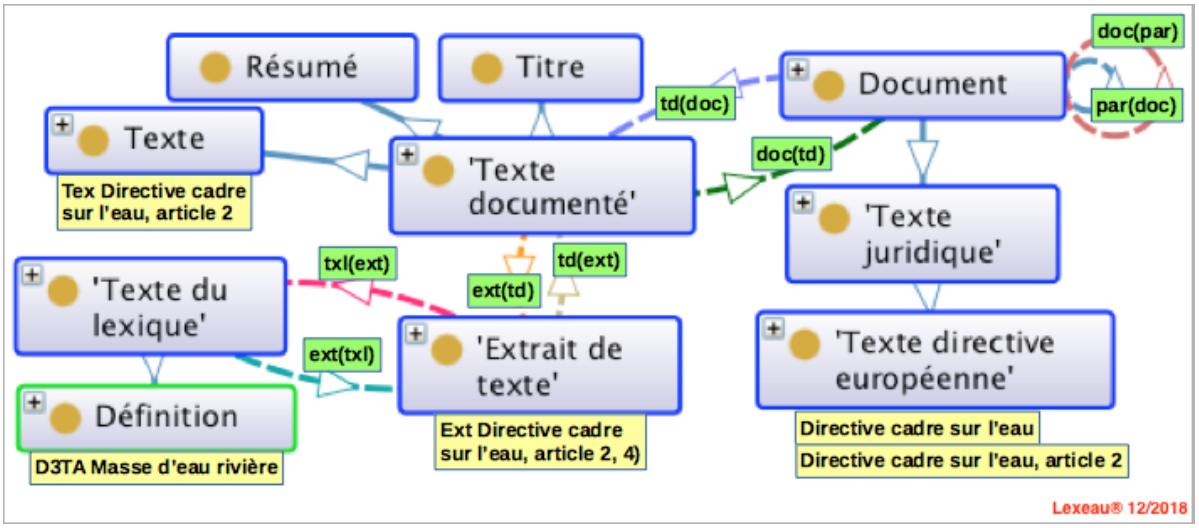


FIGURE 1.34 – Graphe relationnel de la définition des masses d'eau rivière

1.8.2 Des exemples d'emploi du lexique

La figure 1.35 présente l'exemple d'emploi du nom propre lexicalisé « Enl-1 Électricité de France ». Le texte est extrait d'un article de Wikipedia référencé par l'URL du site source et par les initiales de l'auteur de la sélection, suivies de la date de la sélection.

La figure 1.36 présente un emploi de l'unité lexicale alternative « Uda Prélèvement d'eau » tiré du site de l'OCDE où elle apparait, texte utilisé pour définir son unité lexicale de rattachement (4-C Prélèvement d'eau/*4-C Water withdrawal*). [*exemple obsolète JLJ-17/12/2018*]

FIGURE 1.35 – Exemple d'emploi du nom propre lexicalisé « Électricité de France »

FIGURE 1.36 – Exemple d'emploi de l'unité discursive alternative « Uda Prélèvement d'eau »

1.9 Des relations lexicales d'origine conceptuelle

Les relations hiérarchiques entre classes et les propriétés à valeur sont transposées de l'ontologie du domaine dans le lexique sous la forme de relations lexicales. Les relations hiérarchiques entre classes prennent la forme, dans le lexique, de relations d'hyperonymie et d'hyponymie — comme les appellent les linguistes — entre les unités lexicales pivot issues des noms des classes. Les propriétés « à valeur » des entités typées prennent la forme de variables numériques et de variables de Boole (de type « vrai/faux ») qui s'appliquent aux noms propres lexicalisés rattachés aux unités lexicales transposées. La figure 1.37 présente les relations introduites dans le lexique.

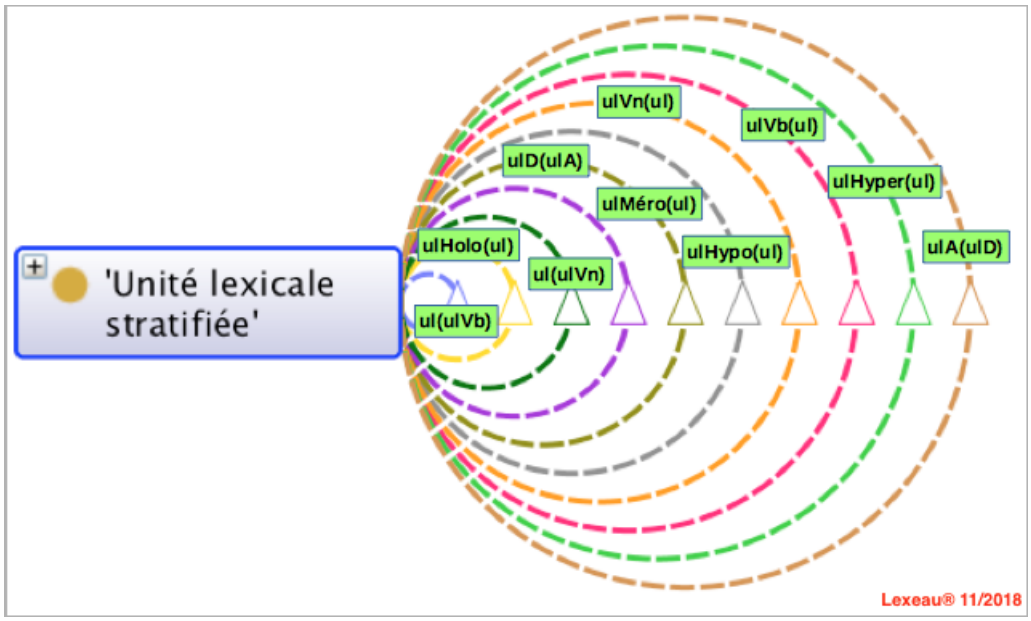


FIGURE 1.37 – Graphe des relations entre unités lexicales de la même strate discursive

Légende :

- ulA(ulD) : acronyme de l'unité lexicale ;
- ulD(ulA) : forme développée de l'unité lexicale acronyme ;
- ulHyper(ul) : hyperonyme de l'unité lexicale ;
- ulHypo(ul) : hyponyme de l'unité lexicale ;
- ulMéro(ul) : méronyme de l'unité lexicale ;
- ulHolo(ul) : holonyme de l'unité lexicale ;
- ulVn(ul) : variable numérique de l'unité lexicale ;
- ul(ulVn) : unité lexicale dotée de la variable numérique ;
- ulVb(ul) : variable booléenne de l'unité lexicale ;
- ul(ulVb) : unité lexicale dotée de la variable booléenne.

La figure 1.38 présente la hiérarchie des masses d'eau en tant qu'entités géographiques. Les hyponymes des unités lexicales « 3-TA Masse d'eau » et « 3-TA Masse d'eau de surface » sont présentés dans les fenêtres d'édition du lexique dans les figures 1.39 et 1.40.

Dans ce chapitre, nous avons illustré, sur l'exemple des prélèvements d'eau et d'une façon plus générale des mouvements d'eau et des flux anthropiques — ce qui est souvent présenté comme « le petit cycle de l'eau » — la construction d'une ontologie des objets de connaissance du domaine de l'eau. Nous avons montré comment s'opère la transposition des concepts de l'ontologie dans un lexique stratifié et comment cette stratification imposée fonde et simplifie le jeu des acceptions d'une entrée du lexique par rapport à la pratique dictionnairique traditionnelle. Nous avons montré comment les définitions et des exemples d'emploi habituels sont complétés par des gloses qui parcourent les liens conceptuels entre unités lexicales, par des exemples d'emploi des variantes lexicales et par un traitement approprié des noms propres et des acronymes, particulièrement développés dans le domaine. Les relations lexicales héritées des relations conceptuelles ont pu être illustrées.

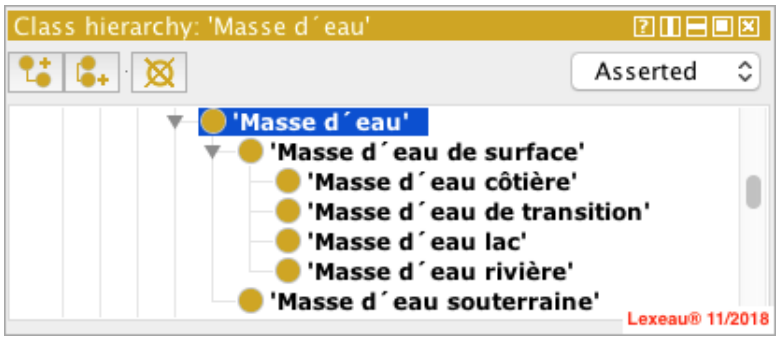


FIGURE 1.38 – Hiérarchie des masses d'eau dans l'ontologie du domaine

FIGURE 1.39 – Hyponymes de l'unité lexicale « 3-TA Masse d'eau »

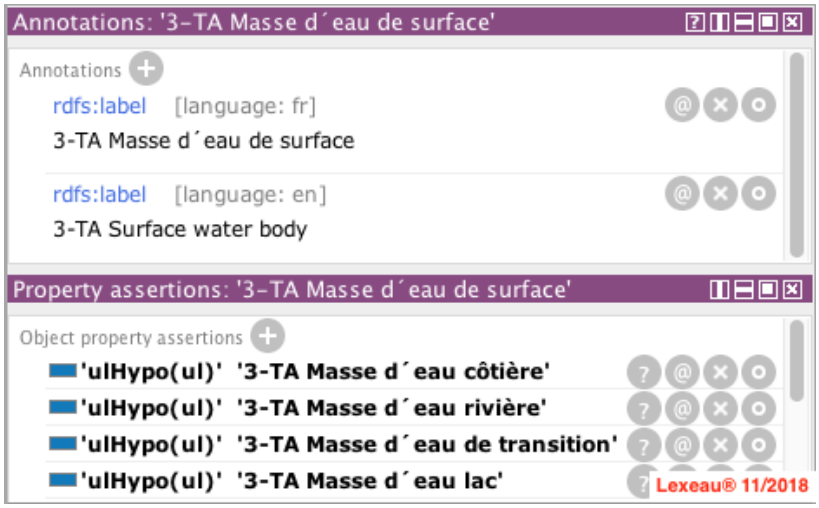


FIGURE 1.40 – Hyponymes de l'unité « 3-TA Masse d'eau de surface »

Cette émergence du lexique depuis l'ontologie du domaine reste à tester dans l'autre sens, à partir des textes, ainsi que la « rencontre » des entrées du lexique qui émergent de l'analyse des textes et des entrées transposées depuis l'ontologie. Ce sera l'objet du chapitre 2.

Chapitre 2

Les fondements du projet LEXEAU®

Le projet LEXEAU® est un projet de lexique bilingue stratifié des textes et des données sur l'eau. Il a pour objet de mettre à la disposition du grand public, des media et des acteurs publics et privés du domaine de l'eau un lexique en français et en anglais développé sur une base de connaissance du domaine et validé par des textes extraits de documents référencés. L'objectif est de faciliter l'inter-compréhension entre spécialistes et non spécialistes de telle ou telle question en prenant en compte la différentiation des usages du lexique sur la base des strates discursives (cf. section 5.1). Le développement d'une ontologie des objets de connaissance du domaine a été illustré dans le chapitre 1, avec la modélisation des mouvements d'eau et des flux d'origine anthropique et la transposition des concepts édités dans l'ontologie en unités lexicales pivot, associées le cas échéant à des unités lexicales satellites. L'état actuel de la base est présenté page 150 dans la section 5.2.

Le présent chapitre traite des fondements du développement du lexique à partir des textes qui relèvent du domaine de l'eau, par extraction de syntagmes nominaux récurrents d'un corpus de textes et ajout de nouvelles entrées ou consolidation d'entrées existantes avec des textes, en lien avec l'ontologie du domaine. Nous parlerons d'« intégration lexicale » pour ces ajouts. La question des cooccurrences et l'intégration de syntagmes nominaux comme entrées à part entière ou construits sur un lexique de base est discutée au chapitre 3. Le lexique est ainsi le point de rencontre des textes et des concepts, soit en lexicalisant des concepts issus d'un processus de modélisation du domaine de connaissance, soit en partant des textes pour faire émerger des unités lexicales associées directement ou indirectement à des concepts. Cette rencontre s'effectue dans une ontologie globale qui réunit une ontologie du domaine de connaissance de l'eau, une ontologie du lexique de l'eau et une ontologie du discours sur l'eau.

Le schéma d'ensemble du dispositif est présenté dans la section 2.1 avec l'interface utilisa-teur. La section 2.2 précise comment la linguistique de corpus est mise en œuvre dans le dispositif, en associant aux documents des corpus textuels. Cette mise en œuvre passe par la structuration des corpus en sous-corpus et par leur partition en sous-corpus jointifs qui les recouvrent entièrement. La textométrie, en tant qu'outil de la linguistique de corpus, est pré-sentée dans la section 2.3 avec ses fonctionnalités (dénombrement des occurrences, recherche

des « concordances », recherche des co-occurrents), et ce après l'attribution d'une étiquette morphosyntaxique et d'un lemme [1] à chaque « mot » du texte. L'organisation du répertoire des sources, la confection du fichier texte à importer sans l'habillage du document source et le choix du module d'import sont discutés dans cette section. Trois exemples de corpus sont présentés dans la section 2.4. Ils permettent d'illustrer des situations très différentes de développement du lexique à partir des textes. L'élaboration d'un lexique de base est présentée dans la section 2.5.

2.1 Le schéma d'ensemble du dispositif

Pour l'utilisateur, le lexique se présente comme un portail du web mettant à sa disposition une interface interactive accessible en langage naturel. Le schéma d'ensemble du dispositif est présenté figure 2.1.

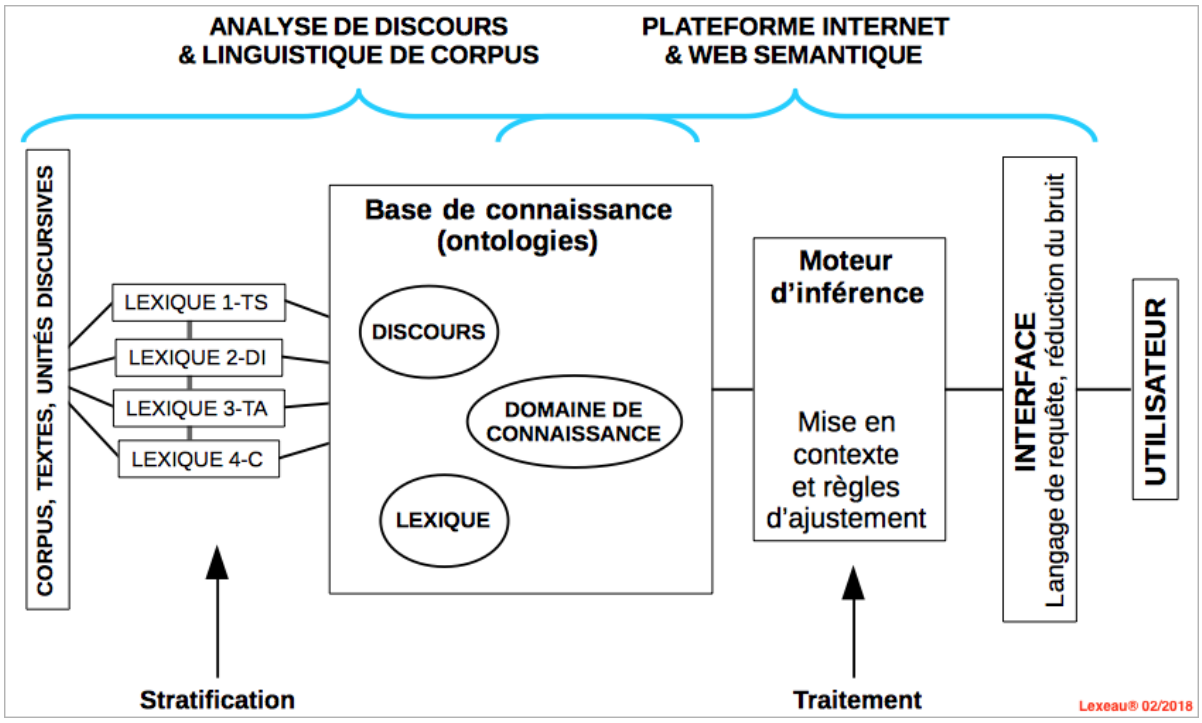


FIGURE 2.1 – Schéma d'ensemble du dispositif

La partie gauche rend compte de l'émergence d'unités discursives récurrentes dans les textes du domaine regroupés en corpus, sur lesquelles va s'effectuer la stratification. Leur sens est restitué dans un lexique stratifié, relié aux concepts de l'ontologie du domaine de connaissance par les unités lexicales « pivots ». Ces unités jouent un rôle privilégié, du fait de la stabilité de leur définition et de leur emploi, par comparaison avec les unités « satellites » qui leur sont associées dans d'autres strates discursives. Les unités lexicales satellites n'ont pas de concept homologue dans l'ontologie du domaine. Un tel concept ferait double emploi avec le

1. Forme grammaticale conventionnelle d'un mot lorsqu'il est placé en entrée dans un dictionnaire

concept source de l'unité pivot à laquelle elles sont rattachées. Les « lexiques » de la figure 2.1 correspondent aux emplois potentiels des unités discursives en tant qu'unités lexicales dans quatre strates discursives (qui seront justifiées et développées au chapitre 5.1) :

1-TS Technico-scientifique (articles, communications et ouvrages scientifiques)

2-DI Décisionnelle *incitative* (propositions et décisions en amont du judiciaire)

3-TA Technico-administrative *normative* (lois, directives, activité judiciaire)

4-C Courante (presse quotidienne, radio, télévision)

La strate discursive d'une unité lexicale pivot est le plus souvent déterminée par le document source de sa définition ou de ses emplois récurrents.

La partie centrale du dispositif est occupée par la base de connaissance. Celle-ci est développée sous la forme d'une ontologie globale composée de trois sous-ontologies (DOMAINE, LEXIQUE et DISCOURS), ce qui correspond à la partition de la base de connaissance de la figure 1.1.

La figure 2.2 [*à revoir JLJ-17/12/2018*] présente les trois classes d'entités de l'ontologie globale. La classe *« Thing »*, à la « racine » de l'ontologie globale considérée comme un arbre, regroupe tous les « objets de connaissance » du projet. Ces trois classes forment les trois

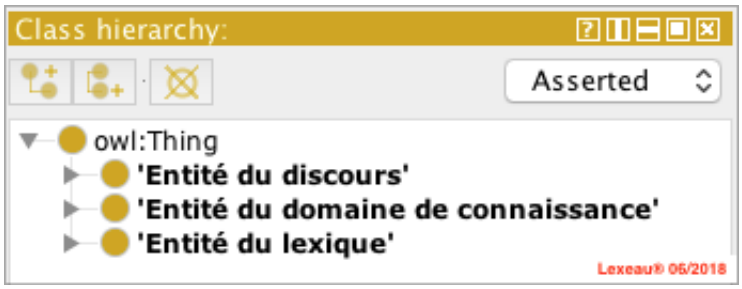


FIGURE 2.2 – Trois classes d'entités à la racine de l'ontologie de base de connaissance

branches de l'ontologie globale ; chacune des classes est à la racine d'une ontologie spécifique. L'existence de relations entre des classes des trois branches nous a conduit à élaborer simultanément les trois ontologies spécifiques.

La partie droite du schéma de la figure 2.1 porte sur la consultation du lexique par l'utilisateur dans une interface de requêtes. Ces requêtes sont traitées dans le moteur d'inférence, qui peut accéder au contenu des ontologies et aux textes des corpus placés en entrée du dispositif.

2.2 Une linguistique de corpus

2.2.1 Documents, textes documentés et corpus textuels

La distinction opérée dans le projet entre les objets de connaissance du domaine de l'eau et les objets de connaissance du discours sur l'eau nous a amené à distinguer entre les documents, objets de connaissance du domaine, répertoriés dans l'ontologie du domaine en classes et sous-classes, et les textes tirés de ces documents — titres, résumés et textes proprement dits —

reproduits dans les classes de l'ontologie du discours. Les textes sont traités en linguistique de corpus par l'intermédiaire de corpus textuels qui restituent partiellement l'activité langagière. Le graphe relationnel des corpus textuels est présenté figure 2.3.

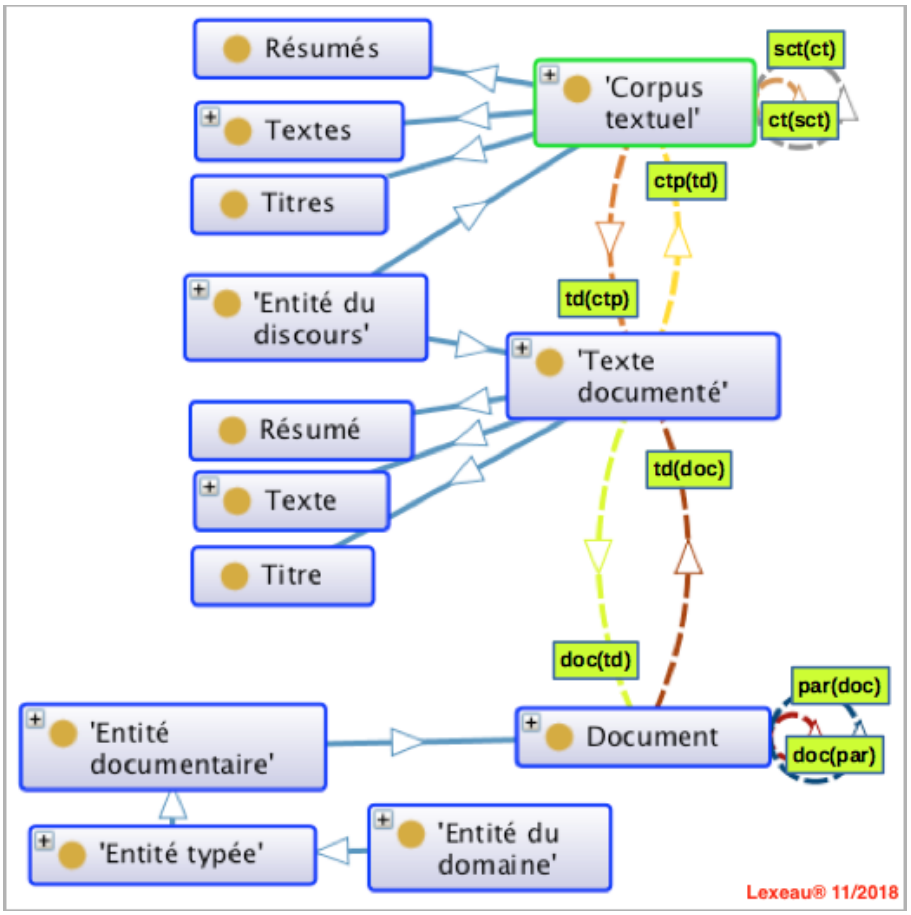


FIGURE 2.3 – Graphe relationnel des corpus textuels

Légende

- ct(sct) : corpus textuel du sous-corpus ;
- sct(ct) : sous-corpus textuel du corpus textuel ;
- td(ctp) : texte documenté du corpus textuel le plus proche ;
- ctp(td) : corpus textuel le plus proche du texte documenté ;
- td(doc) : texte documenté du document ;
- par(doc : partie du document ;
- doc(par) : document partitionné.

Dans ce graphe, les relations de corpus à sous-corpus sont les relations de proximité immédiate que l'on rencontre dans un arbre de classes et sous-classes. Il en est de même pour les parties d'un document, considérées comme des documents à part entière. Les textes documentés sont rattachés au corpus — ensemble de textes — qui leur est le plus proche.

2.2.1.1 Sous-corpus et partition d'un corpus textuel

Un sous-corpus textuel est défini comme un sous-ensemble des textes documentés du corpus qui ne le recouvre pas entièrement. La création des sous-corpus est libre, y compris en créant des sous-corpus qui ont des éléments en commun, pour étudier différents regroupements de textes. En pratique, on travaillera sur des partitions de corpus, organisées dans un arbre hiérarchique renversé, avec le corpus à la « racine » de l'arbre et les textes aux extrémités, comme des « feuilles » de l'arbre, sur des branches qui peuvent être plus ou moins ramifiées.

La partition d'un corpus est définie comme un ensemble de sous-corpus disjoints qui recouvrent entièrement le corpus partitionné. Les requêtes de dénombrement des occurrences d'une expression dans un corpus, par exemple, peuvent être lancées sur la partition du corpus. Les résultats de la requête se présentent sous la forme d'un tableau où les sous-corpus de la partition figurent en colonne.

Pour distinguer entre la partition d'un corpus en chacun de ses textes et les partitions avec au moins un sous-corpus constitué de plusieurs textes, nous distinguerons une classe des partitions « fine » de corpus. Le graphe relationnel est présenté dans la figure 2.4.

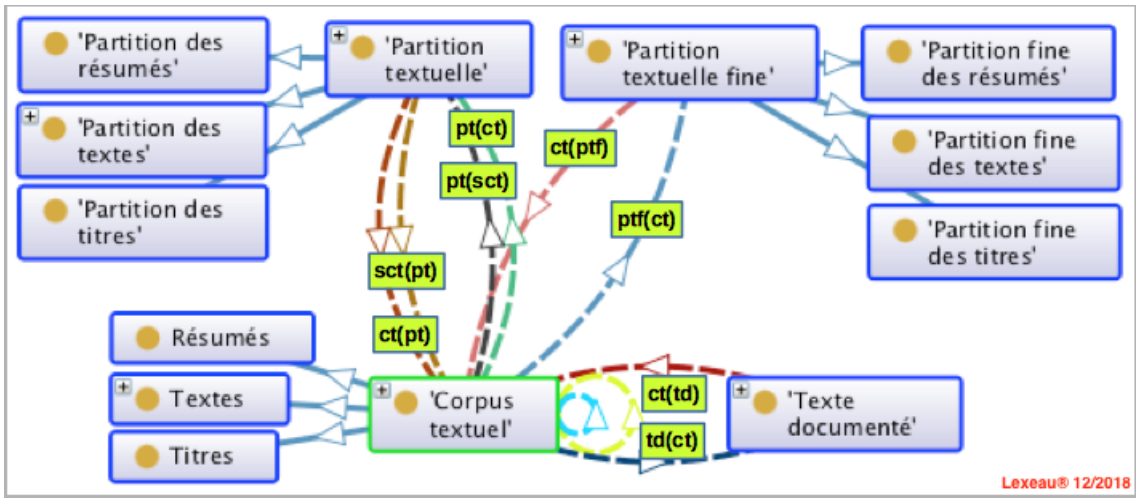


FIGURE 2.4 – Graphe relationnel des partitions textuelles

- ct(pt) : corpus partitionné ;
- pt(ct) : partition du corpus ;
- sct(pt) : sous-corpus de la partition ;
- pt(scp) : partition de rattachement du sous-corpus ;
- ct(ptf) : corpus partitionné en textes documentés ;
- ptf(ct) : partition du corpus en textes documentés ;
- ct(td) : corpus incluant le texte documenté ;
- td(ct) : texte documenté du corpus.

Les trois types de textes d'un document (titre, résumé et texte proprement dit) se retrouvent dans les corpus et les partitions. Les textes consacrés aux auteurs et aux références bibliographiques sont traités de façon détaillée dans l'ontologie du domaine.

2.2.2 Des textes et du hors-texte pour le lexique

Le travail sur les textes débouche sur des extraits repris dans le lexique comme définitions ou comme exemples d'emploi. Les éléments hors-textes dans les documents source qui accompagnent ces extraits doivent être sauvegardés car ils sont susceptibles d'illustrer les articles du lexique. Il s'agit de figures (schémas, courbes), de photos, de tableaux de chiffres et de formules mathématiques. La figure 2.5 présente le graphe relationnel des textes et des hors-textes du lexique. Nous avons fait figurer sur la figure 2.5 la classe des URL associés aux documents.

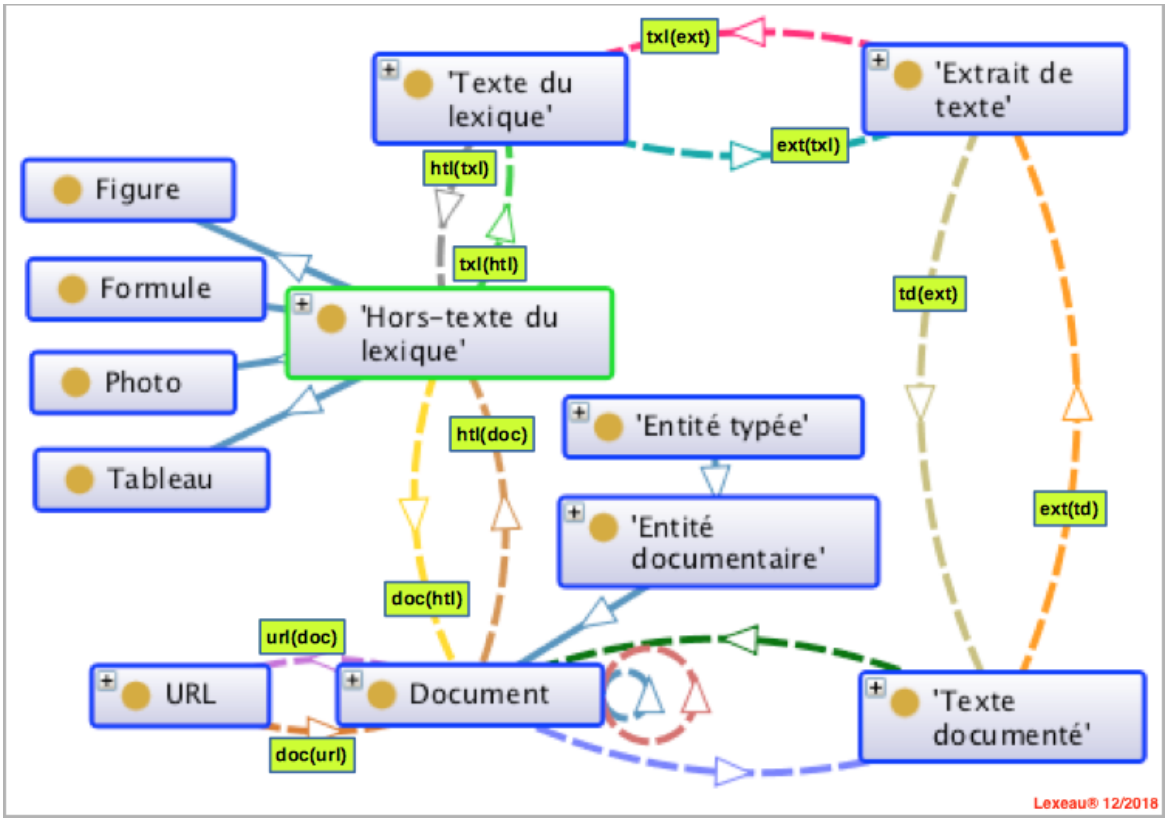


FIGURE 2.5 – Graphe relationnel des textes et des hors-textes du lexique

- ext(td) : texte extrait du texte documenté ;
- td(ext) : texte documenté source de l'extrait de texte ;
- txl(ext) : texte du lexique issu de l'extrait de texte ;
- ext(txl) : extrait de texte correspondant au texte du lexique ;
- htl(txl) : hors-texte associé au texte du lexique ;
- txl(htl) : texte du lexique associé au hors-texte ;
- doc(htl) : document source du hors-texte du lexique ;
- htl(doc) : hors-texte du lexique issu du document ;
- url(doc) : URL du document source ;
- doc(url) : document source accessible à cet URL.

2.2.2.1 Exploitation des contenu hors corpus dans l'ontologie du domaine

Le contenu des documents exploité dans l'ontologie du domaine n'est pas considéré comme du texte documenté pour éviter toute redondance avec le traitement des textes en corpus. Cette exploitation porte notamment sur les auteurs et les références bibliographiques. Le graphe relationnel complet est présenté dans la figure 2.6.

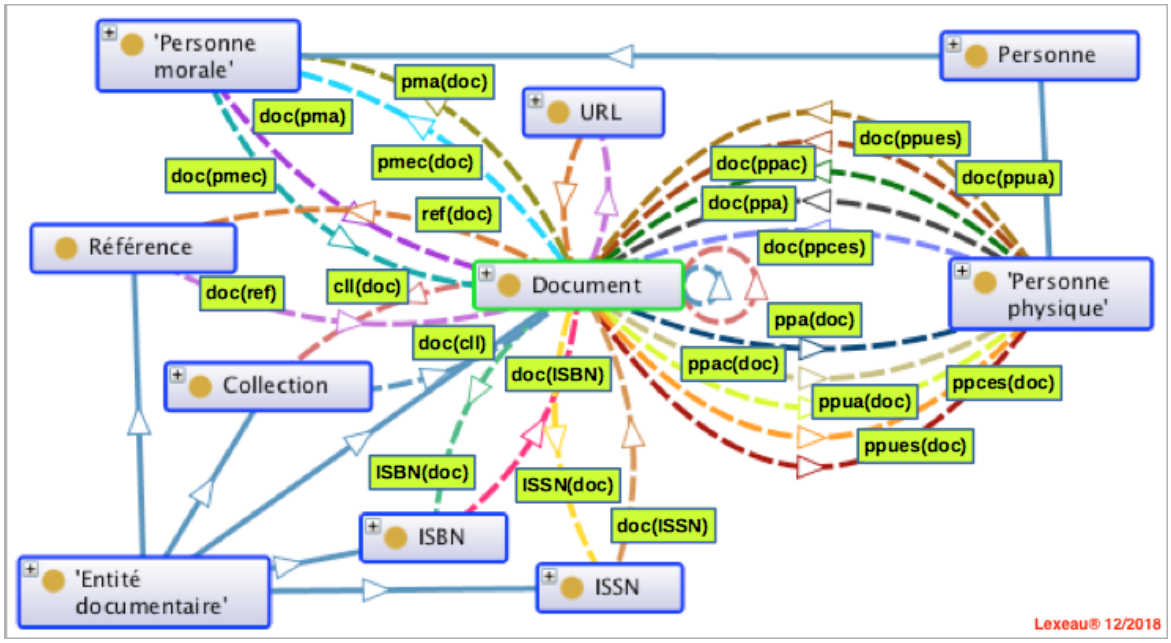


FIGURE 2.6 – Graphe relationnel de l'exploitation des documents dans l'ontologie du domaine

- ppa(doc) : personne physique, auteur du document, parmi d'autres auteurs ;
- ppua(doc) : personne physique, unique auteur du document ;
- ppac(doc) : personne physique, auteur « correspondant » du document ;
- ppues(doc) : personne physique, unique éditeur scientifique du document ;
- ppces(doc) : personne physique, co-éditeur scientifique du document ;
- pma(doc) : personne morale, auteur du document ;
- pmec(doc) : personne morale, éditeur commercial du document ;
- ref(doc) : référence bibliographique du document ;
- cll(doc) : collection dont fait partie le document ;
- ISBN(doc) : numéro ISBN du document (International Standard Book Number) ;
- ISSN(doc) : numéro ISSN du document (International Standard Serial Number).

Chaque auteur est introduit comme personne physique dans l'ontologie. Le cadre de travail est introduit comme personne morale, avec un lien sur la personne physique. Ce cadre est assimilé à un « employeur », par exemple le département de l'université où le doctorant prépare sa thèse ou l'association où l'auteur est salarié ou bénévole. L'auteur « correspondant » est introduit en l'état. La bibliographie qui regroupe les références en fin de texte relève de l'ontologie si les publications référencées y sont enregistrées. Les références bibliographiques sont alors appariées si leur formalisme est différent et celle de l'ontologie est mise en relation avec le document où apparait son homologue.

2.3 La textométrie, pour analyser des contenus contrôlés

La textométrie dans le projet LEXEAU consiste à faire émerger d'un texte issu d'un document des syntagmes nominaux et des extraits de texte susceptibles de trouver leur place dans le lexique. La lemmatisation des mots du texte précède cette recherche. Elle consiste à mettre chaque « mot » du texte dans une classe qui correspond à une « partie du discours » ou *pos* (acronyme de l'anglais *part of speech*). Chaque *pos* — nous adopterons cet acronyme anglophone au masculin pour « partie du texte » — est identifié par une étiquette morphosyntaxique.

À chaque mot est attribué un « lemme » (*lemma*), c'est à dire une forme grammaticale conventionnelle du mot dans la classe du *pos*, soit, en français, le masculin singulier pour les noms et les adjectifs et l'infinitif pour les verbes. Le logiciel TreeTagger, développé par l'Université de Leipzig, est mis en oeuvre à cet effet dans le logiciel TXM développé par l'ENS de Lyon que nous utilisons. La recherche du lemme au lieu du « mot » (*word*) permet de limiter le nombre de formes qui émergent du texte, en masquant des différences de forme grammaticale.

Le jeu des étiquettes morphosyntaxiques auxquelles sont rattachés les lemmes déborde le champ des mots du langage. Il intègre les nombres cardinaux, les signes de ponctuation et d'une façon générale tous les caractères non verbalisés que l'on peut trouver dans un texte.

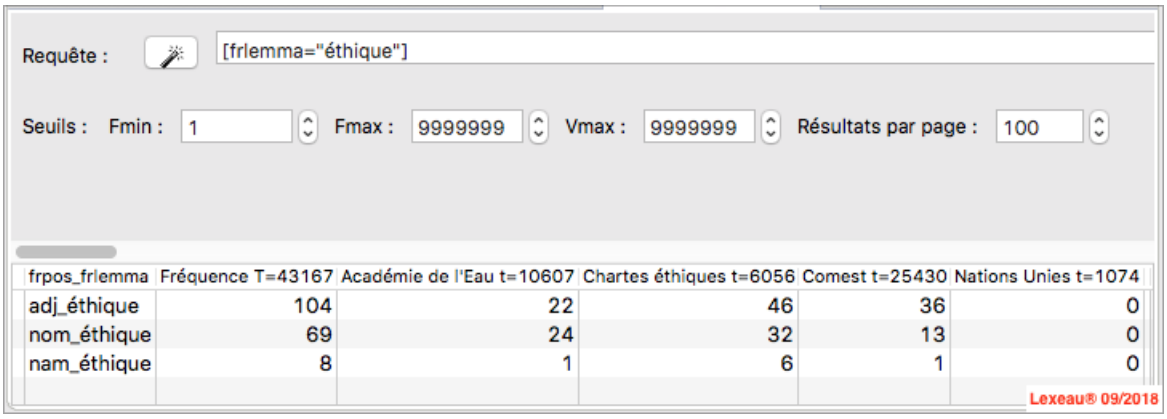
Le travail de textométrie doit s'effectuer sur des textes « propres », c'est à dire débarrassés de leur habillage dans les documents source (renvois, notes, en-têtes, etc.) pour ne pas fausser l'interprétation des résultats. Les textes sont regroupés en corpus textuels dans un « répertoire des sources » depuis lequel les fichiers sont importés, et ce dans un format de fichier accepté par le module d'importation. Cela suppose une structuration pertinente des répertoires source, où vont cohabiter les fichiers à importer et les fichiers des documents source. Cela passe aussi par le choix d'un format de fichier pour les documents source et pour les fichiers importés. Après de nombreux essais, nous avons retenu le format *pdf*, ou éventuellement *rtf*, pour les documents source et le format *odt* pour les fichiers importés, ce qui nous a permis de mettre au point une chaine de traitement pour extraire au format *odt* des textes « propres », avec un premier traitement des fichiers *pdf* des documents sources pour en enlever tous les éléments hors texte, avec leur titre (figures, formules, tableaux, etc.).

2.3.1 Dénombrer les « réalisations » d'une unité linguistique

Dans TXM, la fonction « index » du menu génère une requête qui portera sur une entité du texte pour en dénombrer les réalisations (la « fréquence »), par exemple un « mot » ou un lemme. Il est possible d'éditer le mot (*word*) ou le lemme, *frlemma* ou *enlemma*, selon la langue du texte, avant ou après son étiquette morphosyntaxique (*frlpos* ou *enpos*), ce qui permet de caractériser l'emploi du mot ou du lemme dans le texte. Il est possible d'éditer le mot après son lemme pour avoir le détail de l'emploi du lemme dans le texte. L'édition du lemme, au lieu du mot, et de son *pos* en édition réduit le nombre de lignes du tableau, sans changer le nombre de réalisations. Un lemme figure autant de fois dans un texte que le mot auquel il est associé. Les entités du texte peuvent être beaucoup plus complexes (suite de mots ou de lemmes dans un texte, mot ou lemme dans une liste de choix possibles, etc.). Les entités

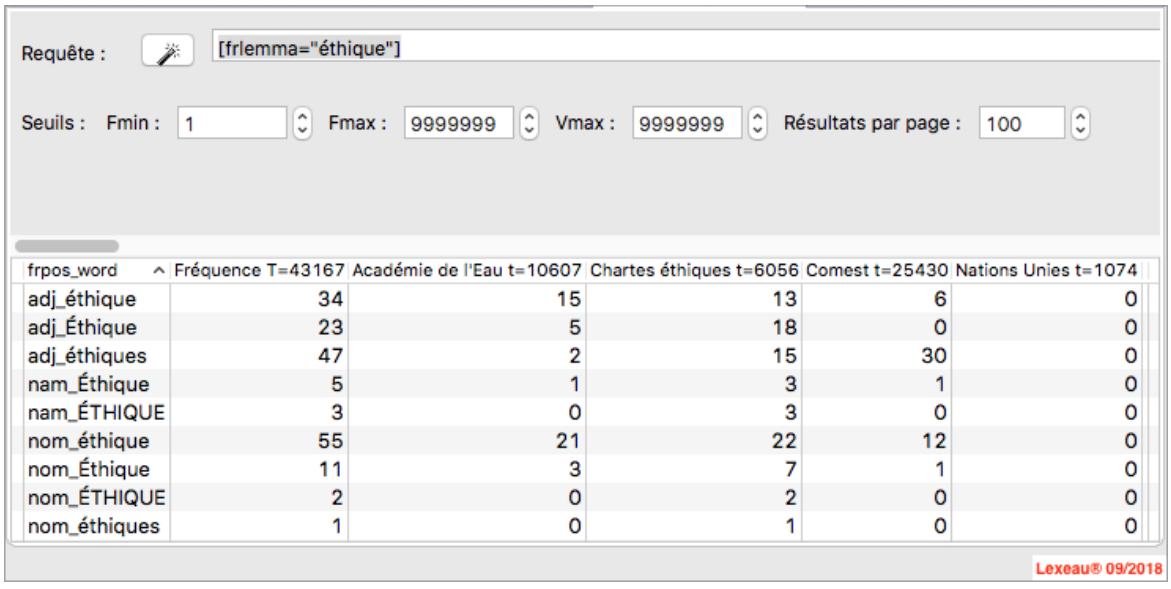
présentes dans le texte sont rangées dans l'ordre décroissant de leur « fréquence » (nombre de réalisations dans le texte), laquelle peut être comparée au nombre de réalisations des « mots » du texte lemmatisé, notée « Fréquence T » dans l'entête du tableau de résultats.

Les figures 2.7 et 2.8 présentent les résultats de la recherche dans les textes du corpus « Ethique », sur lequel nous allons revenir, des réalisations du lemme « éthique », édité en tant que lemme et en tant que mot précédé de son étiquette morphosyntaxique. La requête a été lancée sur une partition du corpus selon l'origine des textes.



frpos_frlemma	Fréquence T=43167	Académie de l'Eau t=10607	Chartes éthiques t=6056	Comest t=25430	Nations Unies t=1074
adj_éthique	104	22	46	36	0
nom_éthique	69	24	32	13	0
nam_éthique	8	1	6	1	0

FIGURE 2.7 – Réalisations étiquetées du lemme « éthique » dans une partition du corpus



frpos_word	Fréquence T=43167	Académie de l'Eau t=10607	Chartes éthiques t=6056	Comest t=25430	Nations Unies t=1074
adj_éthique	34	15	13	6	0
adj_Éthique	23	5	18	0	0
adj_éthiques	47	2	15	30	0
nam_Éthique	5	1	3	1	0
nam_ÉTHIQUE	3	0	3	0	0
nom_éthique	55	21	22	12	0
nom_Éthique	11	3	7	1	0
nom_ÉTHIQUE	2	0	2	0	0
nom_éthiques	1	0	1	0	0

FIGURE 2.8 – Réalisations étiquetées des mots du lemme « éthique » sur la même partition

Le dénombrement des mots étiquetés réalisés recoupe bien celui des lemmes étiquetés, avec un total de 181 réalisations. Le tableau des réalisations du lemme compte 3 lignes alors que celui de la réalisation des mots en compte 9. L'édition de l'étiquette morphosyntaxique (*frpos*) avant le lemme en distingue l'emploi comme adjectif, comme nom et comme nom propre. Ce dernier emploi (ligne nam_éthique du tableau 2.7) ne correspond pas à un emploi avéré

dans les textes. Il a été détecté automatiquement par le logiciel de lemmatisation. Le nom propre existe en tant que titre de l'ouvrage de Spinoza (« L'Éthique »), par exemple, ce qui explique une telle annotation dans l'entraînement du logiciel 'TreeTagger', reprise ici par erreur. L'emploi de l'adjectif « éthique » domine largement sur les autres emplois, sauf pour le corpus des textes de l'Académie de l'Eau. Le nombre total des réalisations du lemme et du mot s'établit à 181 dans les deux tableaux.

La réalisation du lemme « ethique » sans accent a été testée. L'extrait correspondant de la charte éthique de l'entreprise SAUR édité dans TXM est présenté dans la figure 2.9, au dessus du résultat de la requête. Le lemme sans accent est étiqueté comme nom propre, avec un « E » majuscule pour le mot.

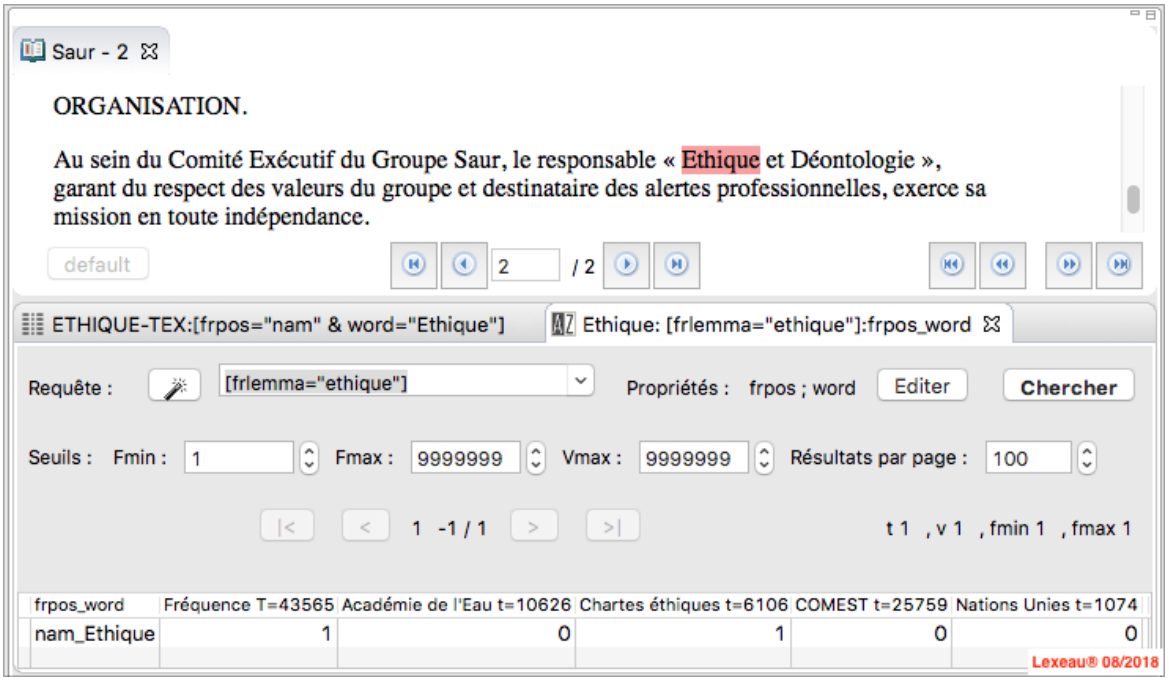


FIGURE 2.9 – Une réalisation du lemme « ethique » sans accent

L'édition du texte qui précède et qui suit une réalisation, ici celle du lemme « ethique » vient après une recherche des concordances de cette réalisation.

2.3.2 Rechercher et éditer des « concordances »

Les « concordances » permettent de présenter sur une ligne les contextes gauche et droit des réalisations d'une entité textuelle dans le texte. Suivant l'édition et la langue du texte, cette entité peut être un mot ou une suite de mots (*word*), un lemme ou une suite de lemmes (*frlemma* ou *enlemma*), une *pos* ou une suite de *pos*. L'entité textuelle est placée au milieu de la ligne, en position de pivot. Cette recherche s'effectue à partir des résultats d'une requête sur index, sur chaque ligne du tableau pour laquelle on souhaite obtenir la ou les concordances.

Les figures 2.10 et 2.11 présentent les résultats de cette recherche dans le texte de MJG importé dans TXM. Les 5 réalisations de la forme « Éthique » de l'adjectif dans le tableau 2.10 apparaissent dans les 5 lignes du tableau de la figure 2.11.

Il est possible d'éditer dans TXM le paragraphe où figure le « pivot » de la ligne de concordance, lequel est surligné comme le montre la figure 2.12.

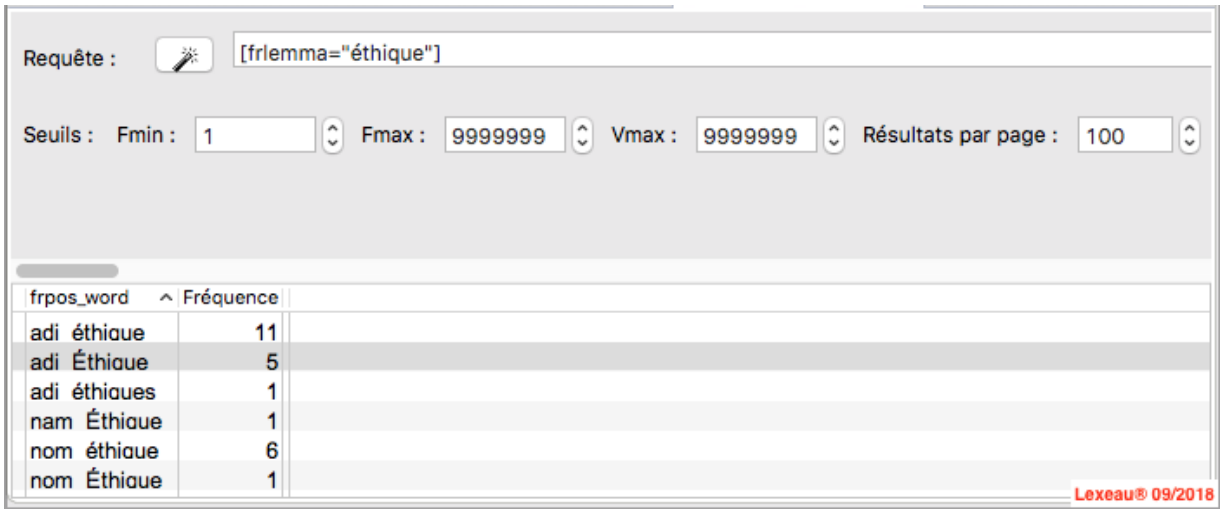


FIGURE 2.10 – Réalisations étiquetées des mots du lemme « éthique » dans le texte MJG

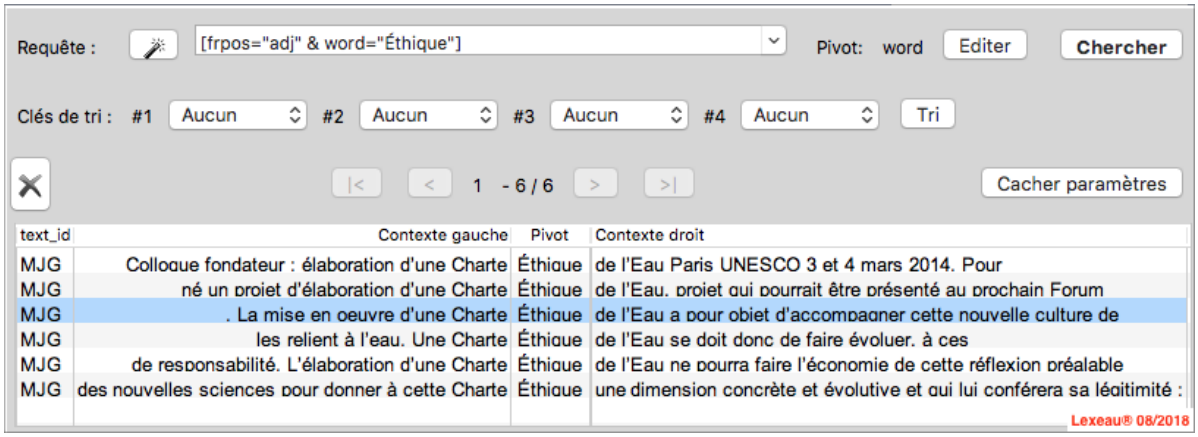


FIGURE 2.11 – Concordances de l'adjectif « Éthique » dans le texte MJG

2.3.3 Trouver des cooccurrences significatives

Dans Wikipedia, une cooccurrence (ou co-occurrence) est définie comme « la présence simultanée de deux ou de plusieurs mots (ou autres unités linguistiques) dans le même énoncé (la phrase, le paragraphe, l'extrait) » [2]. Cette recherche peut être précédée d'une recherche « à

2. https://fr.wikipedia.org/wiki/

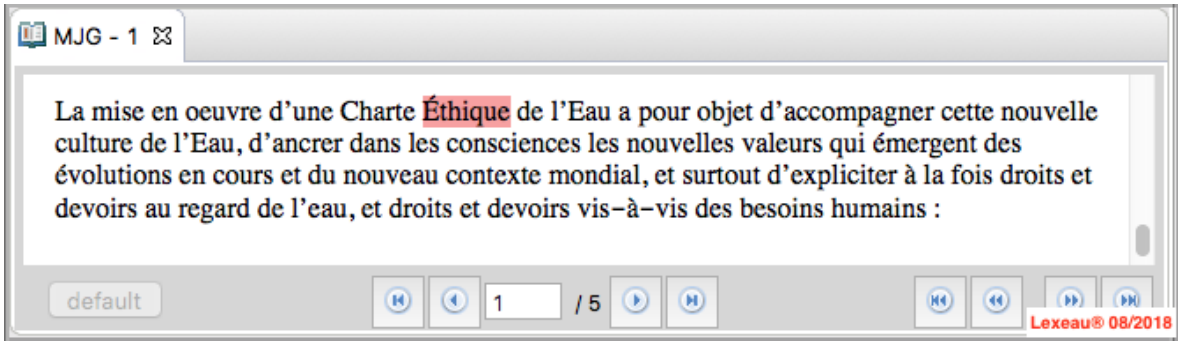


FIGURE 2.12 – Édition dans TXM du paragraphe associé à une concordance

vue » sur les lemmes à gauche et à droite du pivot dans un tableau des concordances des lemmes les plus fréquents. Les figures 2.13 et 2.14 présentent ces lemmes pour la résolution 64/292 de l'Assemblée Générale des Nations Unies et les concordances du lemme « eau ».

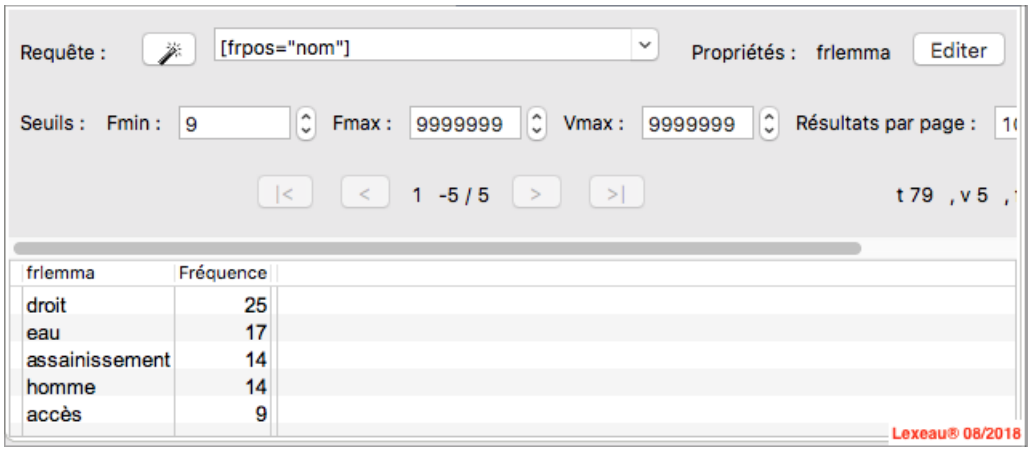


FIGURE 2.13 – Lemmes présents 9 fois ou plus dans la résolution des Nations Unies

La recherche des cooccurrents significatifs à droite ou à gauche d'une unité linguistique s'effectue avec TXM dans une limite donnée de la « distance » entre les cooccurrents, exprimée en nombre d'unités de part et d'autre de l'unité pivot qui fait l'objet de la requête. Ces distances limites définissent une « fenêtre » de recherche des cooccurrents significatifs qui est déplacée dans le texte. Les distances sont laissées au choix de l'utilisateur, entre 0 et 10 unités à droite et/ou à gauche. La cooccurrence est jugée « significative » si elle ne peut être considérée comme le fruit du « hazard ». Un « indice » de spécificité est calculé automatiquement et doit dépasser un certain seuil, égal à 2 par défaut. Deux autres seuils de cooccurrence significative, de valeur 2 par défaut, sont fixés pour la « fréquence » du cooccurrent (nombre de réalisations) et sa « cofréquence » avec l'unité linguistique pivot. Sans entrer dans le calcul de l'indice de spécificité, il est d'autant plus élevé que la probabilité que la cooccurrence ne soit pas le fruit du hasard est élevée. La figure 2.15 présente les résultats de la recherche des cooccurrents significatifs du lemme « eau » à une distance de 4 lemmes en contextes gauche et droit dans la résolution des Nations Unies.

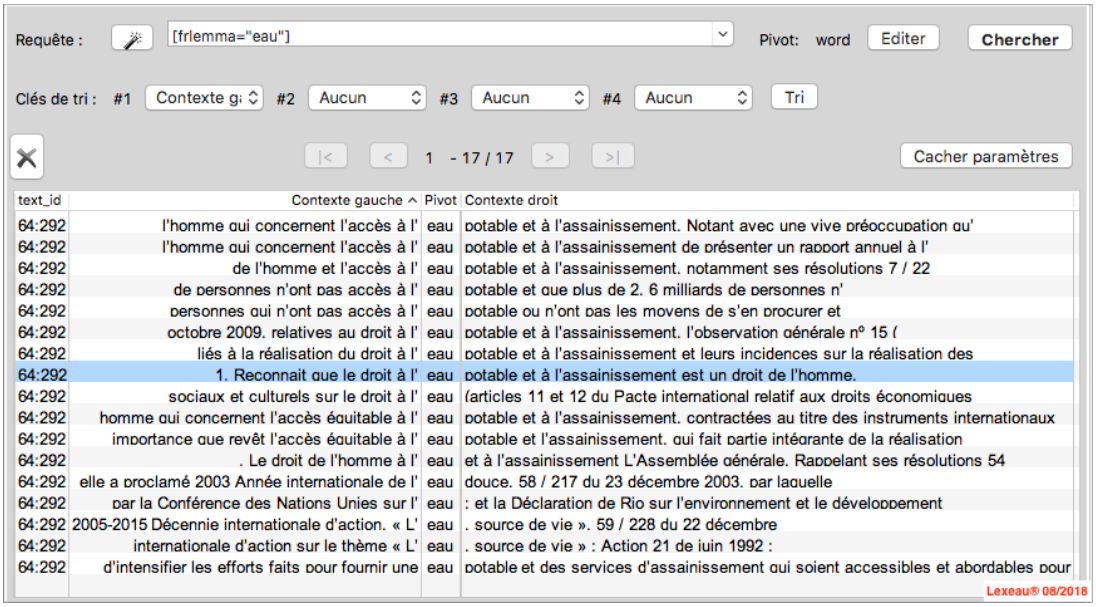


FIGURE 2.14 – Concordances du lemme « eau » dans la résolution des Nations Unies

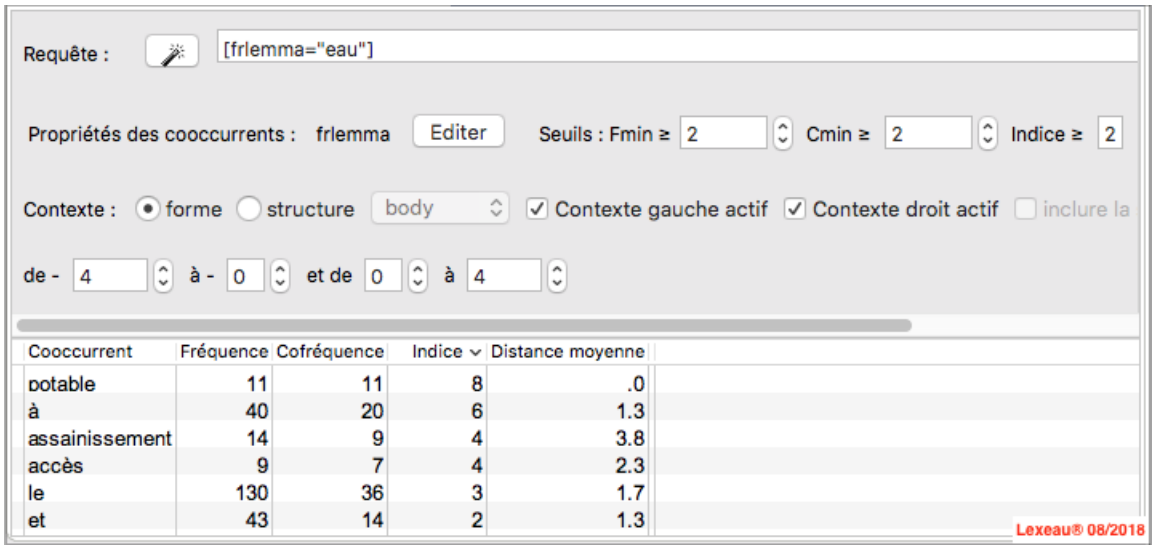
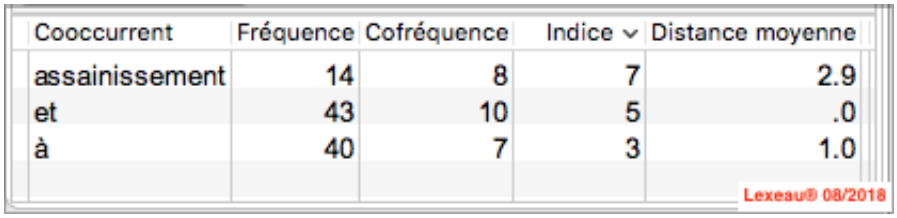


FIGURE 2.15 – Cooccurrents du lemme « eau » sur 4 lemmes à G et à D (rés. 64/292)

On constate une cooccurrence immédiate du lemme « potable » sur la droite, ce qui incite à poursuivre la recherche de cooccurrents sur la droite du lemme « potable », à 3 lemmes de distance (fig. 2.16). On met ainsi en place une certaine heuristique de recherche.

La recherche des cooccurrents à gauche débouche sur les résultats des figures 2.18 et 2.19. Les résultats des requêtes des tableaux des figures 2.18 et 2.19 incitent à tester la présence des expressions « accès à l'eau potable et à l'assainissement » et « droit à l'eau potable et à l'assainissement » dans la résolution. La requête sur l'index correspondant a donné 3 réalisations de chaque expression. La recherche des concordances de la 2ème expression est présentée figure 2.20, ce qui conduit à l'édition de la 2ème ligne dans TXM, où l'expression est



Cooccurrent	Fréquence	Cofréquence	Indice	Distance moyenne
assainissement	14	8	7	2.9
et	43	10	5	.0
à	40	7	3	1.0

Lexeau® 08/2018

FIGURE 2.16 – Cooccurrents du lemme « potable » sur 4 lemmes à droite (rés. 64/292)

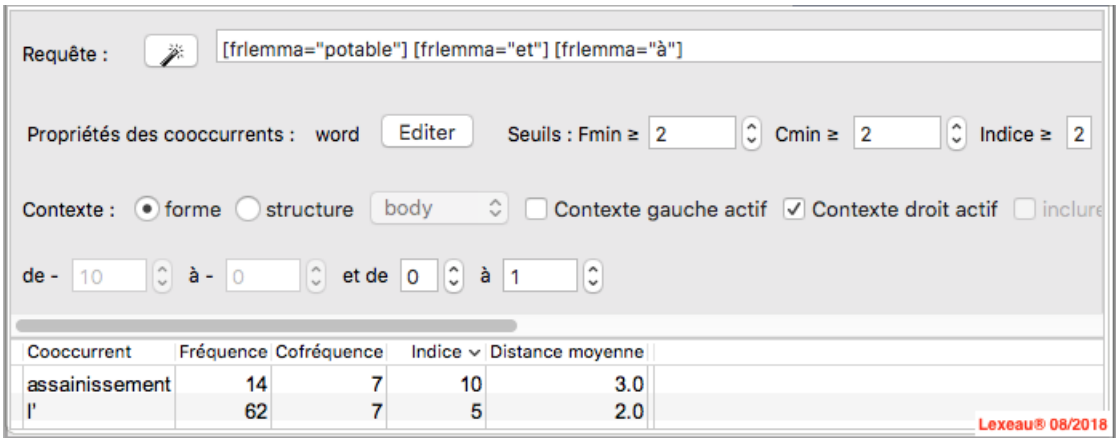


FIGURE 2.17 – Cooccurrents de la suite « potable-et-à » sur 2 mots à droite (rés. 64/292)

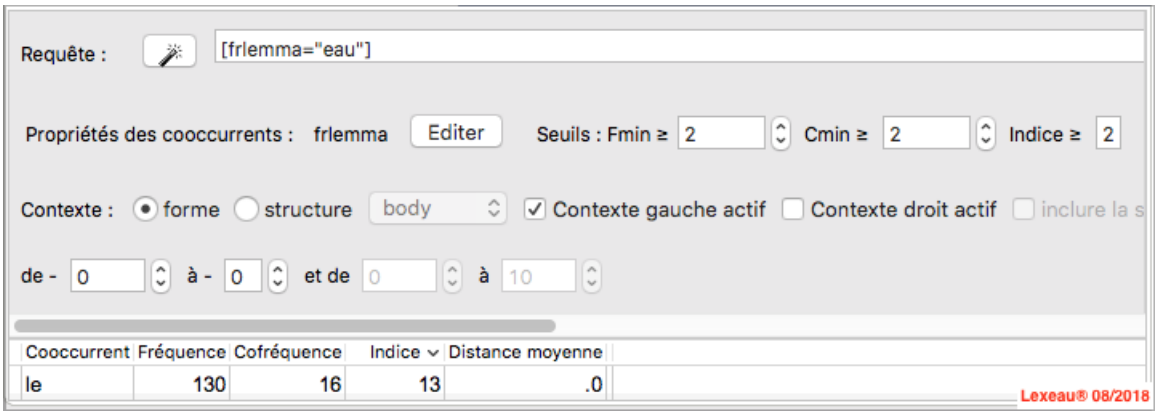


FIGURE 2.18 – Cooccurrents du lemme « eau » sur 1 mot à gauche (rés. 64/292)

surlignée (fig. 2.21). Ces résultats nous montrent qu'un syntagme tel que « droit à l'eau et à l'assainissement », détecté par une recherche itérative des cooccurrents du lemme récurrent « eau », peut faire l'objet d'une requête sur index, suivie d'une recherche des concordances du syntagme en position de pivot qui débouche sur une édition d'un extrait du texte où il est surligné.

Dans cet exemple, le syntagme est apparu sur une ligne des concordances du lemme « eau » présentées figure 2.14. L'intérêt de la recherche sur le syntagme construit à partir de l'étude itérative des cooccurrents est de dénombrer les réalisations du syntagme et d'en trouver toutes

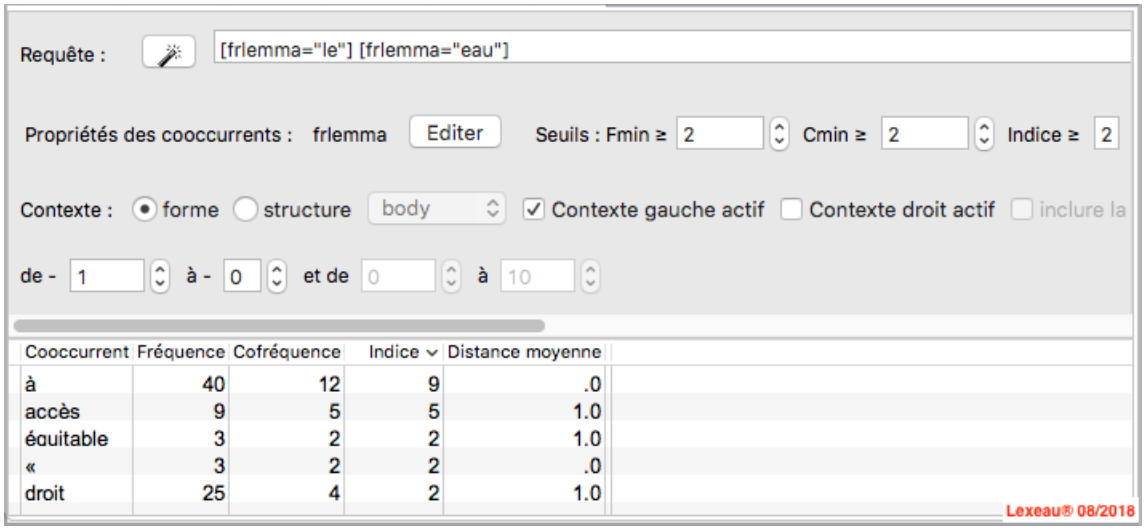


FIGURE 2.19 – Cooccurrents de la suite « le-eau » sur 2 mots à gauche (rés. 64/292)

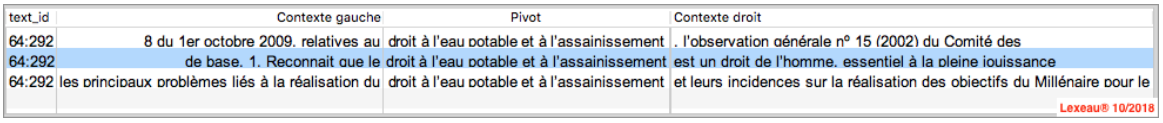


FIGURE 2.20 – Concordances de la 2ème expression dans TXM (Résolution des Nations Unies)

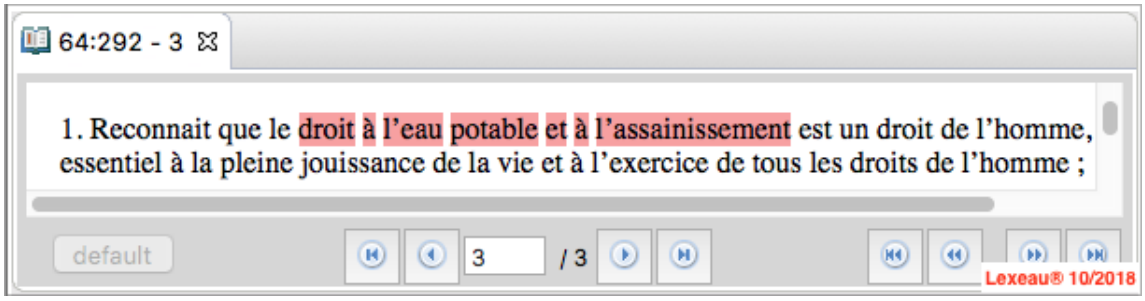


FIGURE 2.21 – Édition de la 2ème expression dans TXM (Résolution des Nations Unies)

les concordances.

2.3.4 « Lemmatiser » les « mots »

Un « mot » est entendu comme une chaîne de caractères précédée et/où suivie d'un blanc dans le texte, y compris les signes de ponctuation, les nombres et les symboles. Chaque « mot » d'un texte est rattaché à une « partie du discours », *part of speach* ou *pos* dans la langue du texte et doté d'une étiquette morphosyntaxique (*tag*). La lemmatisation consiste à attribuer à chaque « mot » un « lemme » (*lemma*), c'est à dire soit une base soit une forme conventionnelle du « mot » lorsqu'il est rattaché à cette partie du discours. Cette opération est effectuée avec un logiciel spécifique entraîné à reconnaître les *pos* et les lemmes sur des textes annotés. Les jeux

d'étiquettes doivent être étudiés pour chaque langue car ils ne se recoupent pas d'une langue à l'autre, comme nous allons le voir pour les nombres.

2.3.4.1 Les étiquettes morphosyntaxiques du français

Les tableaux 2.1 et 2.2 présentent les étiquettes morphosyntaxiques du français intégrées dans TXM avec 'TreeTagger'. Ces tableaux correspondent aux étiquettes des mots lorsque le texte est importé dans un fichier au format *txt* Unicode UTF-8 avec le module d'importation « TXT + CSV » de TXM. Les mêmes étiquettes sont utilisées en minuscules pour les *pos* des mots lorsque le texte est importé dans un fichier au format *odt* avec le module d'importation « ODT/DOC/RTF + CSV ». Le tableau correspondant est reporté en annexe (B.1.1).

Certains *pos* et certains lemmes peuvent réserver des surprises. Le lemme « @card@ » vise *a priori* les nombres entiers naturels exprimés en chiffres et dotés de l'étiquette « NUM » ou « num ». La figure 2.22 présente une exception, avec le début du dénombrement des « mots » du lemme « @card@ », par ordre alphabétique de leur *pos* dans le corpus « ETHIQUE » et l'emploi de 2013 comme abréviation, dont le contexte est présenté dans la figure 2.23.

frpos_frlemma_word	Fréquence T=43167	Académie de l'Eau t=10607	Chartes éthiques t=6056	Comest t=25430	Nations Unies t=1074
abr_@card@_2013	1	1	0	0	0
num_@card@_000	2	0	0	2	0
num_@card@_017	1	0	0	1	0
num_@card@_1	20	3	1	14	2

Lexeau® 10/2018

FIGURE 2.22 – Réalisations des « mots » du lemme « @card@ » par *pos* (corpus ETHIQUE)

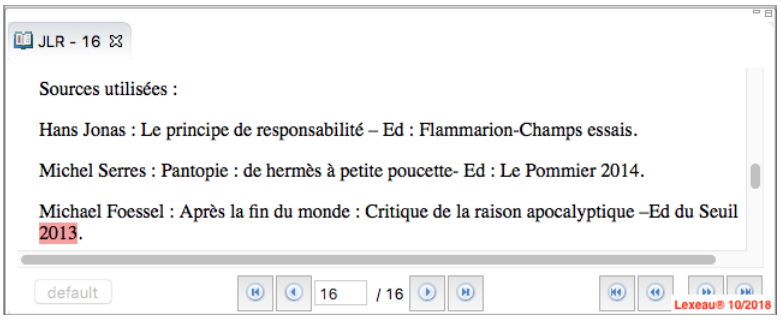


FIGURE 2.23 – Contexte du « mot » 2013 étiqueté comme abréviation (corpus ETHIQUE)

En dehors du lemme « @card@ », les mots qui portent l'étiquette « NUM » sont les adjectifs numéraux cardinaux et ordinaux (deux, premier, etc.), comme le montre le résultat de la recherche des lemmes de ce *pos* dans la partition du corpus « Ethique » (fig. 2.24). Dans le jeu d'étiquettes morphosyntaxiques, la fonction « numérale » des mots qui correspondent à des nombres l'emporte sur la fonction adjectivale, qu'ils soient cardinaux ou ordinaux.

Requête : [frlemma=".*" & frlemma!="@card@" & frpos="num"]

Seuils : Fmin : 1 Fmax : 9999999 Vmax : 9999999 Résultats par page : 100

|< < 1 -28 / 28 > >|

frlemma	Fréquence T=43167	Académie de l'Eau t=10607	Chartes éthiques t=6056	Comest t=25430	Nations Unies t=1074
deux	21	6	1	14	0
premier	20	6	1	13	0
un	17	3	2	12	0
trois	11	0	3	8	0
3ème	7	0	0	7	0
cinq	7	0	0	7	0
quatre	6	0	3	3	0
dix	5	0	0	5	0
second	5	0	1	4	0
c	3	1	0	2	0
deuxième	3	0	0	3	0
six	3	0	0	3	0
vingt	3	1	0	2	0
sixième	2	1	0	0	1
10ème	1	0	0	1	0
19ème	1	1	0	0	0
1er	1	0	0	0	1
1er|1ère	1	0	0	1	0
20e	1	1	0	0	0
20ème	1	1	0	0	0
5ème	1	1	0	0	0
6ème	1	0	0	1	0
seize	1	0	0	1	0
soixante	1	0	0	0	1
trente-sept	1	0	0	1	0
troisième	1	0	1	0	0
xixe	1	0	0	1	0
xxie	1	1	0	0	0

Lexeau® 10/2018

FIGURE 2.24 – Lemmes étiquetés « num » hors lemme « @card@ » (corpus ETHIQUE)

TABLE 2.1 – Les étiquettes morphosyntaxiques du français dans TXM (début)

ABR	abréviation
AD	adjectif
ADV	adverbe
DET :ART	article
DET :POS	pronom possessif (ma,ta...)
INT	interjection
KON	conjonction
NAM	nom propre
NOM	nom

2.3.4.2 Les étiquettes morphosyntaxiques de l'anglais

Le jeu des étiquettes morphosyntaxiques pour l'anglais est plus fourni que le jeu correspondant pour le français. Il exploite les particularités de cette langue (verbes modaux, verbes *be* et *have*). Les tableaux 2.3, 2.4 et 2.5 sont extraits du tableau *Tree Tagger Tag Set (58 tags)* que

TABLE 2.2 – Les étiquettes morphosyntaxiques du français dans TXM (fin)

NUM	numeral
PRO	pronom
PRO :DEM	pronom démonstratif
PRO :IND	pronom indéfini
PRO :PER	pronom personnel
PRO :POS	pronom possessif (mien, tien...)
PRO :REL	pronom relatif
PRP	préposition
PRP :det	préposition plus article (au, du, aux, des)
PUN	ponctuation
PUN :cit	ponctuation de citation
SENT	balise de phrase
SYM	symbole
VER :cond	verbe au conditionnel
VER :futu	verbe au futur
VER :impe	verbe à l'impératif
VER :impf	verbe à l'imparfait
VER :infi	verbe à l'infinitf
VER :pper	verbe au participe passé
VER :ppre	verbe au participe présent
VER :pres	verbe au présent
VER :simp	verbe au passé simple
VER :subi	verbe au subjonctif imparfait
VER :subp	verbe au subjonctif présent

propose l'Université de l'État du Washington aux USA [3] reporté en annexe par copie d'écran sur le site source (B.1.2). Nous avons regroupé dans le tableau 2.4 les étiquettes des nombres cardinaux (en chiffres et en lettres), des marques de fin phrase (*sentence end*), des symboles, des connecteurs non verbalisés (*general joiner*) et des symboles monétaires. Le corpus FLR-1 de la conférence FLOODRISK 2016 a été utilisé pour analyser les *pos* réalisés. Nous avons identifié 6 *pos* non répertoriés dans le jeu numéro 2, édités dans la figure 2.25.

Le *pos* « CD » du tableau 2.4 est dédié aux seuls nombres cardinaux, contrairement à l'étiquette « num » en français (cf. fig 2.24). La recherche des lemmes associés au *pos* « cd » sur la partition du corpus FLR-1 montrent que le lemme « @card@ » n'inclut pas les chiffres de 1 à 9 (fig. 2.26). Les résultats de la requête sur les lemmes « first » et « second » montrent que ces nombres ordinaux ne sont pas toujours des adjectifs (fig. 2.27). Le *pos do* dans le tableau source de l'annexe B.1.2 n'est pas opérationnel dans TXM et n'a pas été repris dans le tableau 2.5 (jeu numéro 3). En conclusion, ce test des *pos* réalisés en français et en anglais a montré qu'ils sont propres à une langue et que la lemmatisation des mots peut réserver des surprises en termes de *pos*.

3. `https://courses.washington.edu/hypertxt/csar-v02/penntable.html`

TABLE 2.3 – Étiquettes morphosyntaxiques pour l'anglais (jeu 1)

POS Tag	Description	Example
CC	coordinating conjunction	and, but, or, &
EX/ex	existential there	*there* is
FW	foreign word	d'oeuvre
IN	preposition/subord. conj.	in, of, like, after, whether
IN/*that*	complementizer	that
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list maker	(1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NP	proper noun, singular	John
NPS	proper noun, plural	Vikings

TABLE 2.4 – Étiquettes morphosyntaxiques pour l'anglais (jeu 2)

POS Tag	Description	Example
CD	cardinal number	1, three
SENT	end punctuation	?, !, .
SYM	symbol	@, +, *, ^, |, =
:	general joiner	;, -, –
$	currency symbol	$, £

Édition : enpos_enlemma_word	F	Thème 01	Thème 02	Thème 03	Thème 04
,_,_,	10812	1993	858	3372	4589
``_"_"	51	6	1	22	22
"_'_'	3	1	0	0	2
"_`_`	21	3	0	9	9
"_'_'	27	2	0	13	12
"_"_"	49	9	1	19	20
(_»_»	4	0	0	2	2
(_(_(	3216	655	224	983	1354
(_{_{	2	0	1	0	1
)_«_«	4	0	0	2	2
)_)_)	3286	666	227	995	1398
)_}_}	2	0	1	0	1
#_#_#	4	0	0	2	2
					Lexeau® 10/2018

FIGURE 2.25 – Autres *pos* réalisés dans le corpus FLR-1 édités avec lemmes et « mots »

Requête : [enpos="cd"]

Seuils : Fmin : 1 Fmax : 9999999 Vmax : 9999999 Résultats par page : 100

1 -32 / 32

enlemma	Fréquence T=304301	Thème 01 t=61931	Thème 02 t=22317	Thème 03 t=101567	Thème 04 t=118486
@card@	3967	708	375	1164	1720
1	934	230	75	312	317
2	778	180	47	231	320
3	646	153	38	199	256
5	524	116	56	167	185
4	452	96	45	146	165
0	358	79	43	103	133
two	350	63	43	101	143
6	347	67	28	125	127
one	347	80	36	76	155
7	230	50	19	67	94
8	195	39	25	53	78
three	172	31	13	57	71
9	164	36	16	51	61
four	88	16	7	34	31
five	21	5	1	10	5
billion	15	3	0	1	11
million	15	0	0	2	13
seven	14	3	1	4	6
six	13	3	1	1	8
zero	10	3	2	2	3
ten	8	2	2	1	3
hundred	6	2	1	0	3
eight	4	0	0	0	4
thousand	4	1	0	0	3
fifty	2	0	0	1	1
ii	2	0	0	1	1
twenty	2	1	0	1	0
eighty	1	0	1	0	0
nine	1	0	0	0	1
seventeen	1	0	0	1	0
thirty	1	0	0	0	1

Lexeau® 10/2018

FIGURE 2.26 – Lemmes étiquetés « cd » dans la partition du corpus FLR-1 par thème

enlemma_enpos	Fréquence T=304301	Thème 01 t=61931	Thème 02 t=22317	Thème 03 t=101567	Thème 04 t=118486
first_jj	199	20	18	69	92
second_jj	106	8	14	34	50
first_rb	57	10	5	18	24
second_rb	7	1	0	4	2
first_np	5	2	0	2	1
second_nns	4	0	0	0	4
second_nn	3	1	0	1	1

Lexeau® 10/2018

FIGURE 2.27 – *Pos* des lemmes « first » et « second » dans la partition FLR-1

TABLE 2.5 – Étiquettes morphosyntaxiques pour l'anglais (jeu 3)

POS Tag	Description	Example
PDT	predeterminer	*both the boys*
POS	possessive ending	friend's
PP	personal pronoun	I, he, it
PP$	possessive pronoun	my, his
RB	adverb	however, usually, here, not
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give up
TO	*to*	*to go, to him*
UH	interjection	uhhuhhuhh
VB	verb be, base form	be
VBD	verb be, past	was|were
VBG	verb be, gerund/participle	being
VBN	verb be, past participle	been
VBZ	verb be, pres 3rd p. sing.	is
VBP	verb be, pres non-3rd p.	am|are
VH	verb have, base form	do
VHD	verb have, past	had
VHG	verb have, grund/participle	having
VHN	verb have, past participle	had
VHZ	verb have, pres 3rd p. sing.	has
VHP	verb have, pres non-3rd per.	have
VV	verb, base form	take
VVD	verb, past tense	took
VVG	verb, gerund/participle	taking
VVN	verb, past participle	taken
VVP	verb, present, non-3rd p.	take
VVZ	verb, present 3rd p. sing.	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-adverb	where, when

2.3.5 Structurer le répertoire des sources

Les corpus traités par textométrie doivent être importés depuis un « répertoire des sources ». Ce répertoire se présente comme un « dossier », au sens du terme en bureautique informatisée, avec éventuellement des sous-dossiers, où sont stockés les documents source, le plus souvent des fichiers au format .pdf, avec les fichiers au format *txt*, avec le codage Unicode UTF-8 des caractères, ou *odt*, suivant le choix entre les deux modules d'import de TXM que nous avons testés. Le module sélectionne les fichiers à importer en fonction de son extension « .* ». L'expérience montre que le contenu du répertoire peut évoluer dans le temps avec la suppression, la mise à jour ou l'ajout de documents qui se répercutent sur les corpus importés, dont le nommage dans TXM prend toute son importance au fil des travaux de textométrie, qui peuvent s'étaler sur plusieurs mois, voire sur plusieurs années, par exemple pour des travaux sur une base de presse.

Le travail de textométrie conduit à structurer le corpus des textes importés en sous-corpus et en partitions. Chaque texte est doté d'un identifiant (« id ») qui reprend le nom du fichier dans le répertoire des sources. Il est essentiel de pouvoir confronter les résultats de la textométrie affichés en édition dans le logiciel, en l'occurrence TXM, pour un extrait du texte importé, et le contenu original du document source ou figure le même extrait. Pour effectuer au mieux cette confrontation, nous avons choisi de structurer le répertoire des sources à l'image de la structuration anticipée du corpus traité par textométrie. La mise à jour, la suppression ou l'ajout de nouveaux documents et/ou textes dans le répertoire des sources conduit à importer un nouveau corpus dans le logiciel qui vient écraser le précédent, s'il porte le même nom, sans nuire à la logique des traitements à venir. Il est essentiel de sauvegarder les requêtes au format texte, voire de les archiver, en lien avec les tableaux de résultats exportés au format *csv* ou *tsv*, pour y revenir d'une session à l'autre. Le logiciel ne propose aucune procédure de sauvegarde en fin de session.

Le choix du module d'import s'effectue dans le logiciel de textométrie. Nous discuterons ici du choix entre le format *txt* avec le codage Unicode UTF-8 et le format *odt* » de Libre Office, dans le cas de l'extraction du texte d'un document source au format *pdf*, ce qui exclut les fichiers « image » sauvegardés en pdf et le contenu protégé de certains *e-book*.

Dans le cas d'une publication scientifique, en plus du texte de la publication proprement dit, le document peut contenir des textes plus courts, titre et résumé, par exemple, qu'il peut être intéressant et plus rapide d'analyser dans des corpus distincts de corpus des textes proprement dits.

Nous avons testé différents modes d'extraction des textes des documents source et différents choix du module d'importation avant d'optimiser une chaine de traitement et d'importation depuis un répertoire des sources adapté à la chaîne.

2.3.6 Extraire et contrôler les textes avant importation

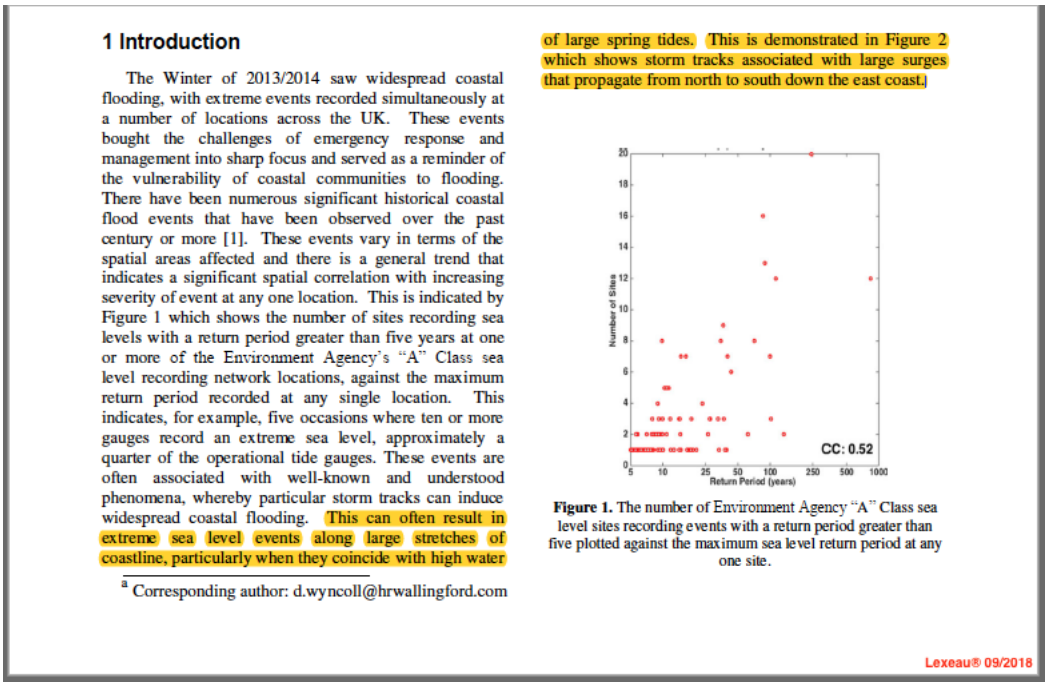
Le choix de l'importation des textes au format *odt* de préférence à l'importation au format *txt*, utilisé pour les premiers essais de textométrie en 2016, s'est imposé au bout du compte, en même temps que celui d'une chaîne de traitement des fichiers source au format *pdf*, le plus souvent utilisé pour les publications scientifiques. La première importation au format *txt*

des communications à la conférence FLOODRISK 2016 s'est faite la même année de façon expéditive, par un « copié-collé » depuis les fichiers *pdf* dans des fichiers *odt* sauvegardés au format *txt* avec un codage Unicode UTF-8 des caractères.

La fabrication d'un premier texte brut au format *pdf* à partir du fichier *pdf* source consiste à supprimer :

- le résumé, s'il existe ; il peut être intégré dans un autre corpus ;
- les figures, tableaux et formules mathématiques (titres et numéros compris) ;
- les références en bas de page et en fin de publication ;
- les hauts et bas de page répétitifs, les numéros de page ;
- les notes en bas de page ou en marge ;
- les caractères de renvoi sur des éléments supprimés s'ils sont collés aux mots du texte.

Le logiciel « Adobe Acrobat » (sur abonnement) a été utilisé pour effectuer ces opérations. Le suffixe « -tex » a été utilisé pour sauvegarder le fichier *pdf* modifié et obtenir par un « copié/collé » le fichier *odt*, importé après relecture. Le choix du module d'import ODT dans TXM s'est imposé en comparant l'édition à l'écran des textes bruts importés avec les modules ODT et TXT. La comparaison est présentée sur les figures 2.28 (document source), 2.29 (import TXT) et 2.30 (import ODT). Le même fichier « 01001-tex.pdf » est à l'origine des textes importés. Le texte brut importé dans TXM avec le module TXT a été fabriqué en exportant avec « Adobe Acrobat » le fichier « 01001-tex.pdf » dans l'application « Word » (fichier *docx*) et en sauvegardant le texte brut au format *txt* (codage unicode UTF-8).



1 Introduction

The Winter of 2013/2014 saw widespread coastal flooding, with extreme events recorded simultaneously at a number of locations across the UK. These events bought the challenges of emergency response and management into sharp focus and served as a reminder of the vulnerability of coastal communities to flooding. There have been numerous significant historical coastal flood events that have been observed over the past century or more [1]. These events vary in terms of the spatial areas affected and there is a general trend that indicates a significant spatial correlation with increasing severity of event at any one location. This is indicated by Figure 1 which shows the number of sites recording sea levels with a return period greater than five years at one or more of the Environment Agency's "A" Class sea level recording network locations, against the maximum return period recorded at any single location. This indicates, for example, five occasions where ten or more gauges record an extreme sea level, approximately a quarter of the operational tide gauges. These events are often associated with well-known and understood phenomena, whereby particular storm tracks can induce widespread coastal flooding. This can often result in extreme sea level events along large stretches of coastline, particularly when they coincide with high water

[a] Corresponding author: d.wyncoll@hrwallingford.com

of large spring tides. This is demonstrated in Figure 2 which shows storm tracks associated with large surges that propagate from north to south down the east coast.

Figure 1. The number of Environment Agency "A" Class sea level sites recording events with a return period greater than five plotted against the maximum sea level return period at any one site.

FIGURE 2.28 – Extrait de la page 1 du document source

L'édition de l'extrait de la page 1 avec le module d'importation ODT se rapproche le plus de l'édition originale. L'extrait importé avec le module TXT introduit un retour à la ligne lors du

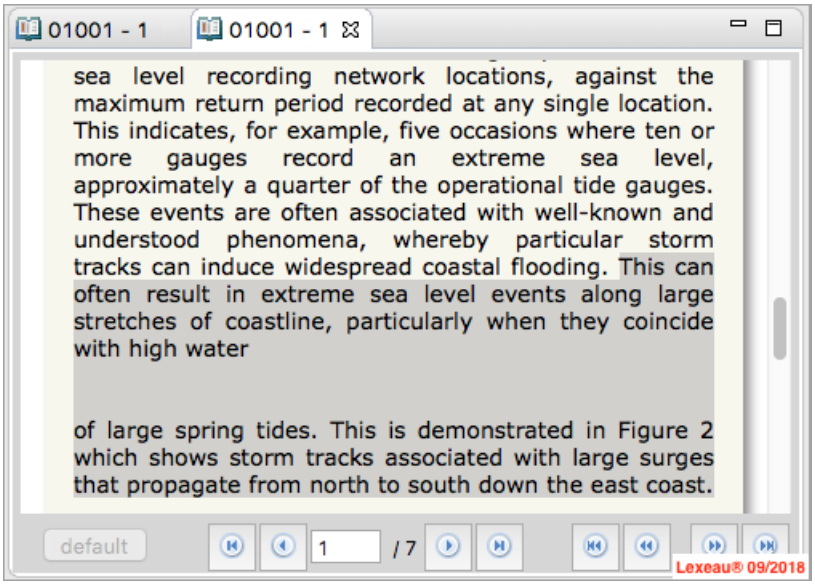


FIGURE 2.29 – Vue dans TXM du même extrait importé avec le module TXT

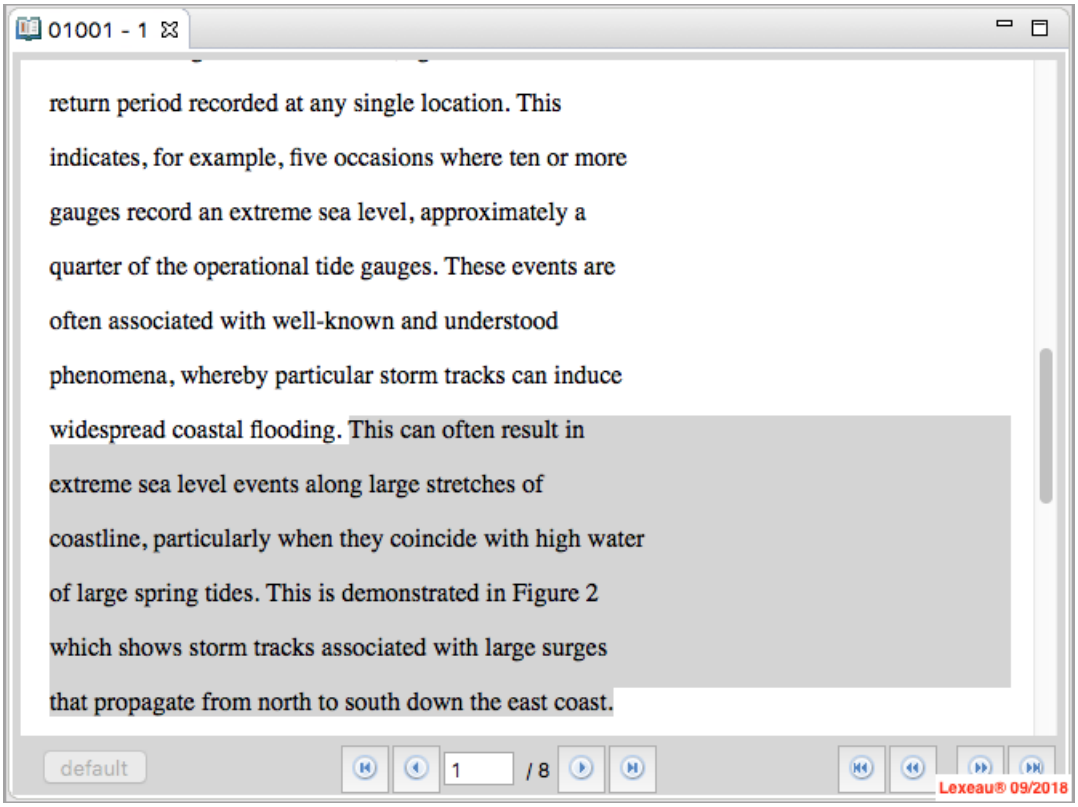


FIGURE 2.30 – Vue dans TXM du même extrait du texte importé avec le module ODT

changement de colonne du texte, ce qui n'est pas le cas avec le module d'importation ODT. On peut repérer aussi un retour ligne prématuré dans l'édition en import « TXT », avec la

phrase « This indicates, for example, ... » qui débute sur la ligne précédente dans l'édition originale et dans l'édition « ODT ». Cette différence s'est avérée essentielle pour pouvoir passer facilement de l'édition d'un extrait de texte dans TXM à l'extrait correspondant publié sur deux colonnes, avec des figures et des tableaux, dans le document original.

2.4 Trois exemples de corpus

2.4.1 Un petit corpus sur le thème de l'eau et l'éthique

Le premier exemple de corpus est celui d'un petit corpus exploratoire créé sur le thème « Eau et éthique » à partir de 8 documents, corpus que nous avons restructuré en 13 textes pour tenir compte du caractère hétérogène des documents source, avec 3 textes de petite taille de l'Académie de l'Eau, un rapport de 56 pages d'une sous-commission de la COMEST (Commission Mondiale d'Éthique des Connaissances Scientifiques et Techniques de l'UNESCO), 3 chartes éthiques d'entreprises du domaine et la résolution 64/292 adoptée le 28/07/2010 par l'assemblée générale des Nations Unies, intitulée « Le droit de l'homme à l'eau et à l'assainissement ». Le rapport de la COMEST de 56 pages sur une gouvernance « éthique » de l'eau a été scindé en 6 textes, pour séparer les textes de synthèse (préface et introduction du rapport) des cinq exemples de gouvernance. Dans la liste ci-dessous, les documents d'origine sont regroupés par origine. Leur titre est suivi entre parenthèses d'un titre abrégé repris comme identifiant du texte importé dans TXM. Les noms des sous-corpus créés dans TXM sont écrits en caractères gras.

Textes de l'Académie de l'Eau

- L'Eau et l'Éthique : ce qu'enseigne l'Histoire (JLO)
- Dénouer les liens idéologiques entre éthique, environnement, sciences et croissance (JLR)
- Pour une Charte éthique de l'Eau (MG)

Rapport complet de la Comest

- Meilleures pratiques éthiques - L'eau et la gouvernance (Comest-rapport)
- Histoire du lac Biwa, ou comment les Japonais maîtrisent leur interaction avec la nature (Comest-ex-1)
- En Afrique du Sud, un nouveau cadre légal pour de bonnes pratiques (Comest-ex-2)
- L'eau des forêts sacrées, des atouts traditionnels pour le développement durable aux Philippines (Comest-ex-3)
- Paroles de femmes pour plus d'équité dans les Andes (Comest-ex-4)
- Biorégions et gouvernance transnationale, des solutions pour résoudre les conflits (Comest-ex-4)

Chartes éthiques d'entreprise

- Charte éthique Nantaise des eaux services (Nantaise)
- Charte éthique du groupe (Saur)
- Charte éthique (Suez)

Résolution 64/292 de l'Assemblée Générale des Nations Unies

- Le droit de l'homme à l'eau et à l'assainissement (Nations Unies)

La figure 2.31 présente la structuration en sous-corpus et en partitions du corpus ETHIQUE-TEX dans TXM, avec les titres créés après l'importation des 13 textes. Ce corpus figure à côté du corpus ETHIQUE-TTR des titres des documents.

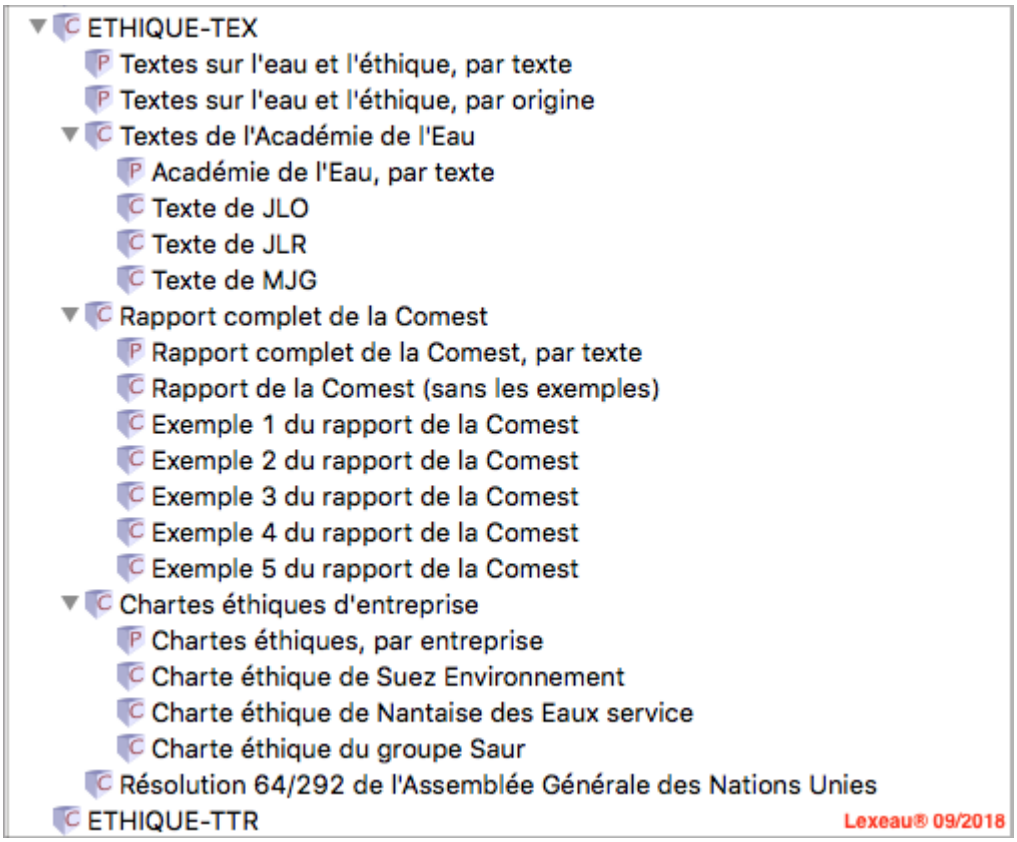


FIGURE 2.31 – Corpus des textes sur l'eau et l'éthique importés dans TXM

2.4.1.1 Des adjectifs qualificatifs dans les textes

Au delà du constat du rôle de l'adjectif « éthique » dans les textes présenté ci-dessus (cf. figure 2.7), nous avons recherché les adjectifs les plus fréquents et leurs associations avec des substantifs qui les précèdent dans les textes. Ces substantifs qualifiés caractérisent la thématique du corpus.

Les 19 adjectifs avec un fréquence minimum de 30 sont présentés dans l'ordre alphabétique dans le tableau obtenu sur la partition du corpus par catégorie de document dans la figure 2.32. La recherche des combinaisons du type adjectif + nom a fait apparaitre 24 combinaisons, avec une fréquence minimum de 4 (fig. 2.33). Ces résultats montrent l'importance des adjectifs dans le discours sur l'eau et l'éthique. Au cours de nos recherches, il nous est apparu une production non négligeable de textes en anglais sur ce thème, notamment [Groenfeld, 2013], et sur des thèmes proches comme celui du « formatage » socio-politique de notre rapport à

l'eau ([Linton, 2010]). L'analyse textométrique est à engager dans la langue de Shakespeare car le mot « ethics », dans la tradition protestante, y a une résonance collective, voire politique, qu'il n'a pas ou peu en français. La résolution des Nations Unies de 2010, sans citer le mot, nous semble avoir élargi la sémantique du thème au delà de la sphère privée.

frlemma	Fréquence T=43167	Académie de l'Eau t=10607	Chartes éthiques t=6056	Comest t=25430	Nations Unies t=1074
autre	81	26	2	53	0
bon	36	6	4	26	0
civil	35	4	1	28	2
culturel	49	6	0	40	3
durable	40	18	4	17	1
économique	34	16	0	15	3
éthique	104	22	46	36	0
grand	42	17	1	24	0
humain	58	25	3	30	0
important	33	10	0	23	0
international	46	17	1	16	12
local	50	6	6	38	0
mondial	30	10	2	17	1
national	58	3	1	54	0
naturel	43	11	1	31	0
nouveau	72	43	3	26	0
politique	47	8	3	35	1
social	52	13	2	34	3
traditionnel	30	0	0	30	0

Lexeau® 09/2018

FIGURE 2.32 – Adjectifs récurrents du corpus Ethique-tex (Fmin=30)

frlemma	Fréquence T=43167	Académie de l'Eau t=10607	Chartes éthiques t=6056	Comest t=25430	Nations Unies t=1074
activité humain	8	7	0	1	0
besoin humain	4	3	0	1	0
charte éthique	24	9	15	0	0
communauté local	5	0	0	5	0
convention international	5	3	0	1	1
crise économique	4	4	0	0	0
culture traditionnel	6	0	0	6	0
démarche éthique	4	0	4	0	0
développement durable	23	8	4	10	1
développement économique	5	1	0	4	0
dignité humain	8	2	1	5	0
diversité culturel	16	0	0	16	0
droit humain	7	0	0	7	0
niveau local	5	1	0	4	0
niveau national	4	0	0	4	0
organisation éthique	6	0	6	0	0
pratique éthique	11	0	2	9	0
principe éthique	8	0	4	4	0
réflexion éthique	4	4	0	0	0
ressource naturel	19	4	0	15	0
science social	4	2	0	2	0
société civil	31	4	1	26	0
valeur éthique	11	1	0	10	0
volonté politique	7	0	0	7	0

Lexeau® 09/2018

FIGURE 2.33 – Substantifs « qualifiés » du corpus Ethique-tex (Fmin=4)

2.4.2 Un corpus structuré de textes scientifiques

Le deuxième exemple de corpus est celui des communications scientifiques présentées et débattues en anglais à la conférence internationale FLOODRISK 2016 (*3rd European Conference on Flood Risk Management*) qui s'est tenue à Lyon du 17 au 21 octobre 2016 sur la question des inondations, traitée en 24 thèmes regroupés dans 6 thèmes directeurs (tableaux 2.6 et 2.7). La mise à disposition des textes par l'Irstea peu avant la conférence, suivie sur place, a

TABLE 2.6 – Les 24 thèmes de la conférence FLOOD*risk* 2016

Thème d.	Thème	Titre du thème
Thème 1	Thème 01	Probability of floods and storms
	Thème 02	Climate change
	Thème 03	Performance and behaviour of flood defences
	Thème 04	Hazard analysis and modelling
Thème 2	Thème 05	Physical, economic and environmental consequences
	Thème 06	Loss-of-life estimation and modelling
	Thème 07	Critical infrastructure and cascading impacts
Thème 3	Thème 08	Vulnerability and societal resilience
	Thème 09	Risk perception and communication
	Thème 10	Hazard and risk mapping
	Thème 11	Hazard and risk mapping
Thème 4	Thème 12	Long-term protection and prevention measures
	Thème 13	Non-structural measures and instruments
	Thème 14	Management and maintenance of infrastructure
	Thème 15	Public education and engagement
	Thème 16	Learning from past events
Thème 5	Thème 17	Disaster management and recovery
	Thème 18	Forecasting and warning
	Thème 19	Evacuation and emergency management planning
Thème 6	Thème 20	Policy appraisal, investment planning and decision making tools
	Thème 21	Adaptation to long term change
	Thème 22	Residual risk and insurance
	Thème 23	EU Floods Directive and international basins
	Thème 24	Ethics, equity and social responsibility

TABLE 2.7 – Les six thèmes directeurs de la conférence FLOOD*risk* 2016

Thème directeur	Titre du thème directeur
Thème 1	Characterising the flood hazard
Thème 2	Characterising the consequences
Thème 3	Characterising the flood risk
Thème 4	Risk reduction and management
Thème 5	Flood event management
Thème 6	Decision making, policy and governance

débouché sur un premier travail de textométrie. Le contenu textuel de chaque fichier source au format *pdf* a été sélectionné en l'état, puis collé dans un fichier *odt* et sauvegardé au format *txt* (codage Unicode UTF-8) pour constituer l'ensemble FLOODRISK-2016 dans le répertoire des sources, importé en corpus sous le même nom avec le module TXT .

Un deuxième travail de textométrie plus approfondi a été engagé en 2018 sur les communications relevant du 1er thème directeur (thèmes 01 à 04), avec importation des fichiers textes au format *odt*. Trois corpus ont été créés dans TXM, composés respectivement des résumés (FLR-1- RES), des textes proprement dits (FLR-1-TEX) et des titres des communications (FLR-1-TTR).

Les 4 corpus importés sur 3 ans dans TXM sont présentés dans la figure 2.34, avec le détail des sous-corpus et les partitions de 1er niveau pour les corpus de résumés, de textes et de titres

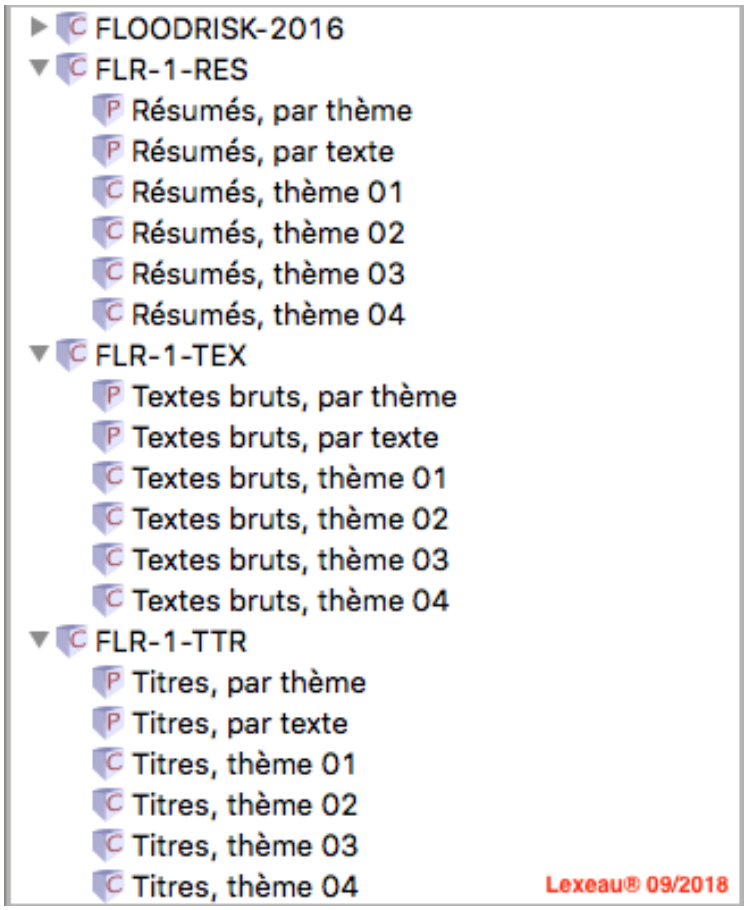


FIGURE 2.34 – Quatre corpus créés dans TXM sur la conférence FLOODRISK 2016

Nous avons voulu comparer l'importation en mode « express » de 2016 et l'importation contrôlée de 2018. La figure 2.35 présente le nombre de « mots » par *pos* dans le corpus FLOODRISK-2016-1, constitué pour la circonstance, et dans le corpus FLR-1 importé en 2018. Ce nombre passe de 412 891 à 304 301. La réduction de la taille des textes est due principalement à la suppression des résumés et des références bibliographique dans les textes importés en 2018.

Parmi les *pos* les plus courants, surlignés en orange clair et en vert, on constate une diminution importante des adjectifs (JJ, JJR, JJS, jj, jjr, jjs), des noms communs (NN, NNS, nn, nns) et des noms propres (NP, NPS, np, nps). L'importation contrôlée évite la redondance

FLOODRISK 2016-1 (textes « pdf », import TXT)						FLR-1 (textes bruts, import ODT)					
	T 412891						T 304301				
enpos	F	N tex.	enpos	F	N tex.	enpos	F	N tex.	enpos	F	N tex
CC	10685	68	VBD	2072	67	cc	8559	68	vbd	1965	67
CD	21924	68	VBG	90	38	cd	9672	68	vbg	86	37
DT	36721	68	VBN	569	62	dt	33253	68	vbn	542	62
EX	228	59	VBP	2194	68	ex	225	58	vbp	2122	68
FW	441	45	VBZ	4187	68	fw	169	35	vbz	4078	68
IN	40045	68	VH	158	49	in	33919	68	vh	159	49
IN/that	902	68	VHD	71	32	in/that	901	68	vhd	69	31
JJ	29233	68	VHG	28	22	jj	23570	68	vhg	29	23
JJR	1076	68	VHN	4	4	jjr	1082	68	vhn	4	4
JJS	416	64	VHP	468	65	jjs	385	64	vhp	454	64
LS	2349	68	VHZ	635	68	ls	1381	68	vhz	605	66
MD	2327	68	VV	5211	68	md	2298	68	vv	4819	68
NN	74334	68	VVD	1496	68	nn	58569	68	vvd	1339	68
NNS	24351	68	VVG	5693	68	nns	20752	68	vvg	4831	68
NP	36058	68	VVN	11614	68	np	13620	68	vvn	10717	68
NPS	340	58	VVP	1387	68	nps	111	37	vvp	1276	68
PDT	154	44	VVZ	3107	68	pdt	154	45	vvz	2858	68
POS	211	47	WDT	1577	68	pos	148	40	wdt	1519	68
PP	2011	68	WP	87	34	pp	1969	68	wp	79	33
PP$	712	66	WP$	21	12	pp$	677	65	wp$	22	13
RB	8533	68	WRB	844	67	rb	8296	68	wrb	784	67
RBR	453	66	,	18603	68	rbr	456	65	,	10812	68
RBS	230	53	:	3038	68	rbs	206	50	:	1595	67
RP	329	61	''	139	30	rp	306	59	''	100	26
SENT	27334	68	(	6025	68	sent	13838	68	(	3222	68
SYM	7161	68	)	6124	68	sym	4182	68	)	3292	68
TO	6910	68	#	26	7	to	6317	68	#	4	3
UH	21	14	``	60	19	uh	3	3	``	51	15
VB	1844	68	$	30	10	vb	1833	68	$	17	7
										Lexeau® 09/2018	

FIGURE 2.35 – Dénombrement des *pos* réalisés dans FLOODRISK 2016-1 et FLR-1

d'une analyse textométrique des résumés et une inflation de noms propres due aux références bibliographiques.

2.4.3 Des articles sélectionnés dans une base de presse

La problématique du troisième exemple de corpus est la constitution d'un corpus d'articles sélectionnés dans une base de presse. Le travail a porté sur des articles du journal britannique *The Guardian* sélectionnés dans la base de presse FACTIVA. Dans un premier temps (novembre 2016), le corpus a été constitué sur 5 ans par sélection dans FACTIVA des articles filtrés sur le mot « flooding » et téléchargéss par lots au format *rtf*, après un contrôle succinct

de leur pertinence. Le corpus a été constitué par sélection du texte des lots, sauvegardé au format *txt* (codage unicode UTF-8) avant importation dans TXM avec le module TXT. Ce corpus a permis d'analyser les différences entre les expressions employées dans un media anglais sur les inondations et dans le texte la directive européenne [4], importée dans les mêmes conditions dans un corpus distinct, et d'étudier la restitution des résultats dans un module graphique.

Un 2ème corpus d'articles du Guardian a été constitué dans FACTIVA sur le mois de janvier 2018, avec le même filtre, après un contrôle rigoureux de la pertinence de chaque article affiché à l'écran. La figure 2.36 illustre le processus de sélection dans la base de presse.

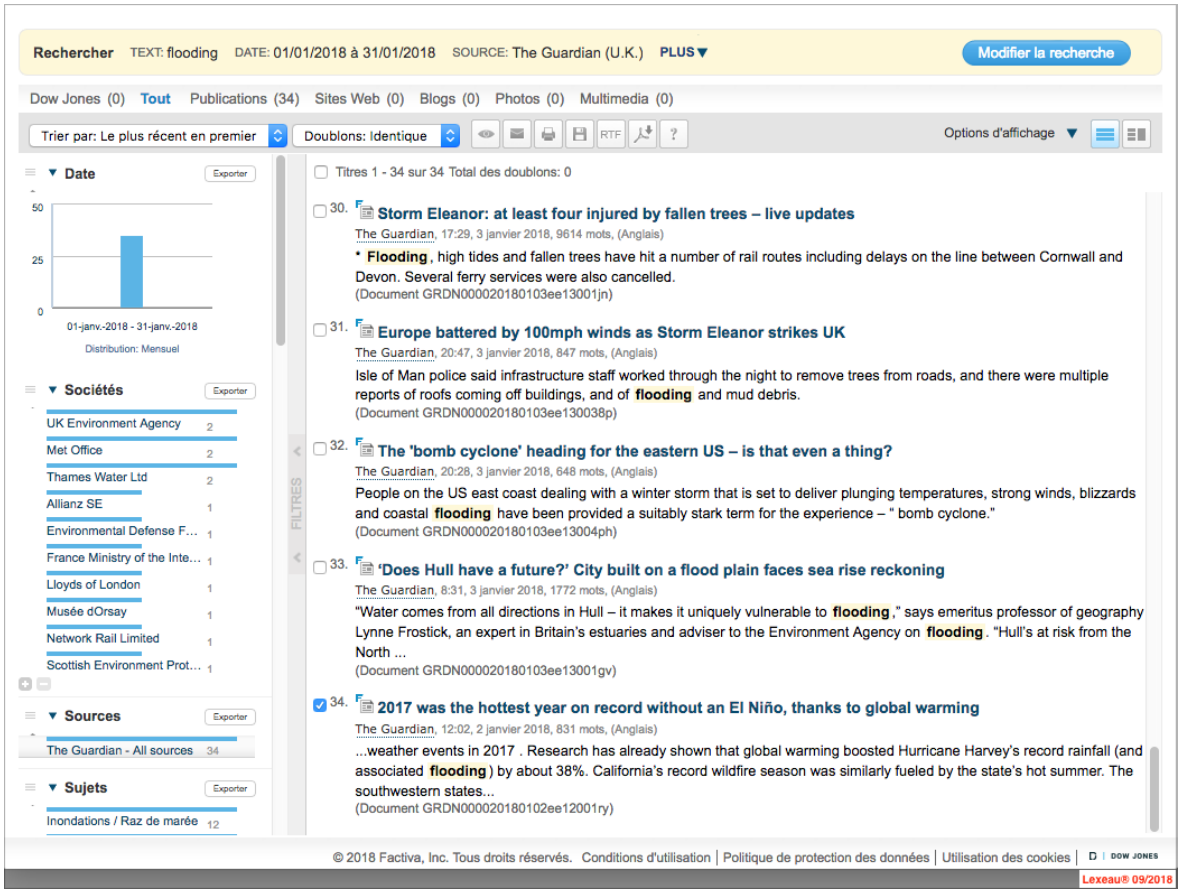


FIGURE 2.36 – Sélection des articles du Guardian de janvier 2018 dans FACTIVA

Le contenu des fichiers téléchargés individuellement au format *pdf* a été traité et sauvegardé au format *odt*. L'opération avait pour but de tester une méthode de constitution dans le temps d'un corpus d'articles tirés d'une base de presse. L'analyse de l'indexation de la rubrique éditoriale des articles a permis de tester l'évolution des rubriques qui traitent des inondations entre 2012 et 2018. Pour conserver le lien entre les textes du corpus dans TXM et les fichiers source dans FACTIVA, l'identifiant temporel du téléchargement est associé à celui de l'article dans la base, du type « GRDN0000xxx », pour identifier les fichiers **.pdf* et **-tex.pdf* de

4. DIRECTIVE 2007 / 60 / EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 23 October 2007 on the assessment and management of flood risks

la chaîne des traitements. L'identifiant de l'article dans FACTIVA est ensuite utilisé pour identifier les articles importés dans TXM.

Le texte à importer dans TXM a été sélectionné avec Adobe Acrobat en intégrant une partie de l'indexation des éléments de l'article, soient « SE » pour la rubrique de l'article dans le journal, « HD » pour son titre, « BY » pour son ou ses auteurs, « PD » pour la date de publication, « ET » pour l'heure de publication et l'index « AN », avant la mention « Document » suivie de l'identifiant de l'article. La fonte Arial 10 a été appliquée aux textes importés pour retrouver en édition les retours à la ligne des articles originaux.

La figure 2.37 présente la structure du corpus des 23 articles importés depuis FACTIVA parmi les 34 articles extraits dans la base sur le mot *flooding* pour le mois de janvier 2018. Chaque sous-corpus correspond à un article, classé dans l'ordre alphabétique de l'identifiant par la date que nous leur avons attribué (« id » dans TXM). Cet exemple montre qu'il est possible

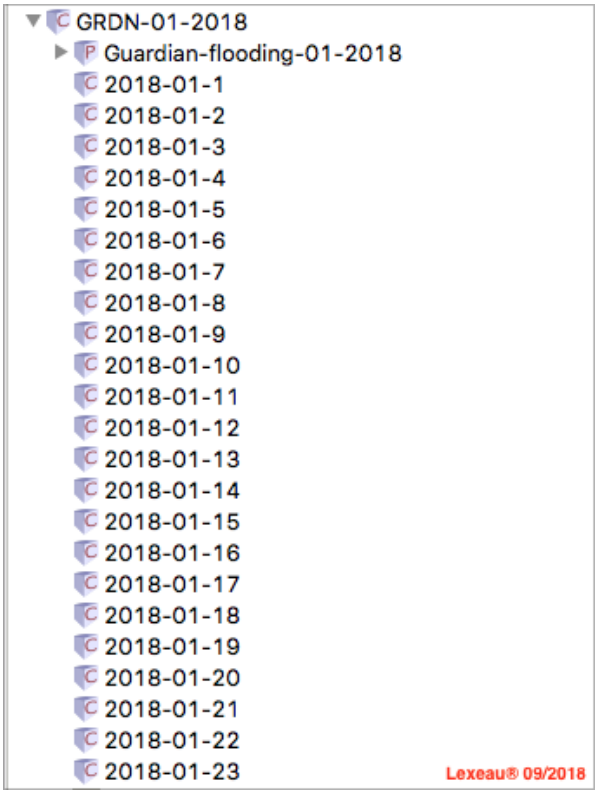


FIGURE 2.37 – Corpus des articles de janvier 2018 du Guardian dans TXM

de créer un corpus d'articles de presse dans TXM par extraction sélective contrôlée d'articles de la base de presse sur une base temporelle, téléchargement des fichiers et extraction avec Adobe Acrobat des textes importés dans TXM.

Cette extraction sélective d'articles du *Guardian* sur un filtre et une période donnés et la validation de la chaine de traitement des fichiers téléchargés avant importation dans TXM suggèrent d'automatiser des étapes de cette chaine pour constituer des corpus d'articles sur des durées plus longues, avec des mises à jour mensuelles et annuelles, sur des thèmes bien identifiés, comme nous l'avons fait sur le thème des inondations avec le mot *flooding*.

2.5 Élaborer un lexique de base

2.5.1 Une base nominale paradigmatique ?

L'élaboration d'un lexique de base du domaine de l'eau s'est faite tout naturellement sur une base nominale, parfois dérivée d'une base verbale. Prélèvement, restitution, injection ou prise d'eau sont des « déverbaux [5] », dérivés des verbes prélever, restituer, injecter ou prendre ... de l'eau. Il serait redondant de faire figurer ces verbes en entrée du lexique. Les verbes pleuvoir, neiger, inonder sont évoqués dans le lexique par des classes d'évènements extrêmes comme une pluie (un épisode pluvieux), un chute de neige, une inondation. Si une base nominale peut évoquer un verbe, en dehors d'un déverbal, avec les couples « barrage-barrer » et « flux-s'écouler », cela n'est pas toujours le cas. Quel verbe associer à « mouvement d'eau » (« *mouvementer de l'eau » ?), à catastrophe, à substance chimique ? Il semble bien qu'une base nominale se prête mieux à l'élaboration du lexique de l'eau qu'une base verbale.

Pour la version française du lexique de l'eau, il nous semble qu'un verbe serait soit redondant soit hors sujet en entrée du lexique, tout en restant fort utile pour les textes du lexique. Ceci n'exclut pas des exceptions, y compris dans la version anglaise du lexique.

Une autre question se pose sur la nature « taxonomique » ou « paradigmatique » de la base nominale du lexique, au sens de la commutabilité d'un nom de la base avec un autre nom de la base dans un énoncé. Ceci ne fait pas de doute lorsque le nom est issu du nom d'une classe d'entités typées, qui peut commuter sur l'axe syntagmatique (« la pluie/l'inondation a commencé à 10h30 ı»). La réponse est la même pour les unités lexicales pivot issues des entités génériques (« Le professeur enseigne l'hydrologie/l'histoire », « J'ai un système d'irrigation par aspersion/au goutte à goutte »).

La réponse est différente pour les propriétés relationnelles transposées dans le lexique, en plus petit nombre. Dans le modèle des mouvements d'eau et des flux anthropique, une personne peut agir comme « opérateur d'un mouvement d'eau » ou comme « usager d'un flux d'eau ». Dire qu'elle agit comme « *opérateur d'un flux » est grammaticalement correct mais n'a pas de sens. Même résultat pour la commutation du sujet d'un verbe « *l'usager effectue le mouvement d'eau » n'a pas de sens. Le constat est le même pour les propriétés « à valeur ». Une masse d'eau artificielle peut éventuellement être fortement modifiée, mais ces deux entrées du lexique ne sont pas interchangeables. On ne peut pas dire « *une masse d'eau artificielle est une masse fortement modifiée », pas plus que « *Je paie une redevance de 30 m3/h » ou « *La rivière en crue avait un débit de 500 € » ou encore « *Le débit de la rivière en crue s'exprime en € ». Les entrées du lexique tirées d'une propriété de l'ontologie du domaine ne peuvent commuter.

2.5.2 Des noms propres et des substantifs pour le lexique

Nous allons maintenant nous intéresser aux difficultés qu'éprouve le profane à la lecture des communications de la conférence FLOODRISK 2016, à commencer par les acronymes, pas

5. Rodney Huddleston [Huddleston, 1988] consacre la section 6-12 de son cours d'introduction à la grammaire anglaise à la nominalisation des verbes et des adjectifs avec des exemples de déverbaux (*deverbals*).

toujours accompagnés de leur forme développée (cf. 1.7.1), en passant par les graphiques techniques et les variables énigmatiques des formules mathématiques, mais surtout les noms propres et les substantifs dont on aimerait qu'ils figurent dans lexique en rapport avec d'autres entrées qui les accompagnent.

La recherche des noms propres d'auteurs, de méthodes ou de modèles de référence est abordée en 2.5.2.1 sur un exemple d'auteurs et un exemple de méthode. Elle sera reprise en 3.4.1 avec la recherche des noms propres associés à « modèle » dans une collocation.

La recherche de substantifs à intégrer dans le lexique précède celle des syntagmes et des cooccurrences, qui fera l'objet du chapitre 3. Il s'agit ici d'identifier des substantifs récurrents avec un contenu scientifique propre à les faire entrer dans le lexique. Nous en donnons des exemples avec la recherche et l'intégration des substantifs récurrents, en procédant successivement à des recherches sur seuils de fréquence, dans une partition ordinaire de corpus (cf. 2.5.2.2), puis à des recherches par texte sur la partition fine d'un corpus (2.5.2.3, 2.5.2.4, 2.5.2.5). Un exemple de recherche du sens d'un substantif destiné à entrer dans le lexique est donné *in fine*.

2.5.2.1 Des auteurs, des méthodes et des modèles

Un des obstacles majeurs à la compréhension des textes est le caractère énigmatique de certains noms propres, qui visent des auteurs de référence et des méthodes et modèles connus dans les matières scientifiques sous-jacentes du texte, par exemple le calcul des probabilités dans les communications rattachées au 1er thème directeur de la conférence. Ce thème est axé sur la probabilité d'occurrence des événements extrêmes à l'origine des inondations (pluies, crues, vents forts, tempêtes en mer, rupture de barrages, de digues, etc.).

La recherche directe des auteurs de référence parmi les noms propres du texte est inopérante. Ils sont noyés dans le texte avec d'autres noms propres, de lieux, d'entreprises, etc., et parfois avec des « mots » étiquetés comme des noms propres lorsqu'ils commencent par une majuscule.

Nous avons testé la recherche avec les noms cités dans l'article de l'encyclopédie Wikipedia consacré à la loi d'extremum généralisée, qui nous dit : « En probabilité et statistique, la loi d'extrémum généralisée est une famille de lois de probabilité continues qui servent à représenter des phénomènes de valeurs extrêmes (minimum ou maximum). Elle comprend la loi de Gumbel, la loi de Fréchet et la loi de Weibull, respectivement lois d'extrémum de type I, II et III. ».

La recherche sur ces trois noms a montré que les noms de Gumbel et de Weibull figurent dans certaines communications du corpus, sans pouvoir les afficher à l'écran compte tenu la dimension du tableau qui reprend en colonne toutes les communications, y compris celles où les noms ne figurent pas, ni connaître immédiatement le nombre de communications affichant une fréquence non nulle.

Les huit communications concernées ont été obtenues avec une macro de dénombrement des « en ayant », intitulée « PartitionIndex2NParts », obligeamment écrite à notre demande par Serge Heiden [6]. Le tableau des résultats de cette macro est exporté depuis TXM dans un fichier

6. Responsable du projet TXM conduit avec une équipe au sein du laboratoire IHRIM UMR 5317 de l'ENS de Lyon.

csv où il peut être exploité sur tableur. Le résultat de cette exploitation est présenté dans la figure 2.38.

A	B	C	D	E	F	G	H	I	J	K
	T	1003	1004	1005	1013	1014	2001	2002	3004	NPARTS
	304301	2403	4303	4689	2583	3569	6612	7550	5518	
	F									
Gumbel	38	0	0	10	7	6	5	10	0	5
Weibull	6	1	0	2	0	0	0	0	3	3
COLMARGINS		1	0	12	7	6	5	10	3	
										Lexeau® 09/2108

FIGURE 2.38 – Huit textes avec une/des références à Gumbel et/ou Weibull (corpus FLR-1)

Ce résultat montre que le travail de textométrie permet de trouver dans les textes le nom d'un d'auteur de référence. Il sera classé comme un « nom propre potentiel » avant d'être transféré dans le lexique comme « nom propre typé » associé à un « nom propre lexicalisé » après mise à jour de l'ontologie du domaine et du lexique.

Le souvenir scolaire d'une « méthode de Monte Carlo » en calcul des probabilités a été confirmé par l'encyclopédie Wikipedia, consultée le 08/09/2018, qui nous dit « Le terme méthode de Monte-Carlo, ou méthode Monte-Carlo, désigne une famille de méthodes algorithmiques visant à calculer une valeur numérique approchée en utilisant des procédés aléatoires, c'est-à-dire des techniques probabilistes. » Ceci nous a conduit à vérifier d'abord que cette suite de deux noms propres figurait bien pour 28 occurrences dans le corpus FLR-1 des communications rattachées aux thèmes 01, 02, 03 et 04 du 1er thème directeur. Pour trouver les communications où cette méthode est appliquée, nous avons utilisé la macro présentée ci-dessus, lancée sur la partition fine du corpus par texte. La figure 2.39 présente le tableau des résultats.

A	B	C	D	E	F	G	H	I	J	K	L	M
	T	1001	1002	1006	1007	1010	1011	2004	3009	3019	4028	NPARTS
Édition : word	304301	3371	7626	6666	2710	3223	7486	3989	6927	4435	1973	
	F											
Monte Carlo	28	3	4	5	4	1	1	1	1	1	7	10
COLMARGINS		3	4	5	4	1	1	1	1	1	7	
												Lexeau® 09/2018

FIGURE 2.39 – Les 10 communications où figure le nom propre « Monte Carlo »

2.5.2.2 Extraire les substantifs récurrents

La recherche des substantifs récurrents dans un corpus partitionné en sous-corpus s'opère en deux étapes. La première étape consiste à faire des recherches distinctes dans le corpus et dans ses sous-corpus en choisissant dans chaque requête un seuil de fréquence pour obtenir à peu près le même nombre de substantifs. La deuxième étape consiste à constituer une liste agrégée des substantifs obtenus dans les requêtes de la première étape et à faire une recherche

globale des occurrences des substantifs de la liste agrégée sur la même partition du corpus. Pour écrire cette requête, les résultats des requêtes de la première étape sont exportés dans un tableur dont les fonctionnalités sont utilisées pour créer la liste agrégée en ligne, par collage spécial avec transposition de la colonne des substantifs dans une ligne de cellules qui seront fusionnées. La requête à introduire dans le logiciel de textométrie est préparée dans un éditeur de texte pour remplacer les espaces entre les substantifs de la liste agrégée par le caractère du « ou » logique du logiciel (« | »). La réalisation de la deuxième étape sur la partition donne des fréquences plus élevées que celles de la première étape. Ces deux étapes peuvent être répétées, par itération, sur des niveaux plus profonds de la hiérarchie des partitions du corpus.

La figure 2.40 présente côte à côte, dans l'ordre alphabétique, les substantifs les plus fréquents du corpus FLR-1 et de ses sous-corpus FLR-01 à FLR-04 de la partition du corpus FLR-1. Dans cette 1ère étape, les seuils de fréquence ont été choisi pour obtenir environ 20 substantifs dans chacune des 5 requêtes. Les 5 listes de substantifs récurrents de la figure 2.40 ont été

FLR-1 (fmin=528)		FLR-01 (fmin=130)		FLR-02 (fmin=57)		FLR-03 (fmin=197)		FLR-04 (fmin=171)	
enlemma	F	enlemma	F	enlemma	F	enlemma	F	enlemma	F
analysis	738	analysis	202	analysis	57	analysis	307	analysis	172
area	708	approach	149	area	90	assessment	200	approach	188
datum	924	area	130	change	66	breach	418	area	345
event	945	catchment	200	climate	98	datum	200	catchment	202
failure	570	datum	274	copula	58	embankment	262	damage	201
flood	1 751	distribution	202	datum	81	erosion	356	datum	369
flow	762	event	369	event	87	failure	466	event	382
levee	646	flood	388	height	129	flood	544	flood	771
level	945	flow	170	level	190	flow	225	flow	360
method	748	level	316	model	125	levee	573	level	173
model	1 561	method	290	parameter	75	level	266	method	235
parameter	551	model	242	period	90	model	535	model	659
probability	584	parameter	167	result	61	probability	230	rainfall	231
result	590	period	153	sea	108	risk	251	result	248
risk	631	probability	269	study	85	slope	229	risk	284
study	559	rainfall	177	surge	93	soil	205	river	171
time	574	value	267	time	61	system	266	study	217
value	713	variable	142	water	75	time	199	time	223
water	954	water	233	wave	132	water	397	value	243
wave	528	wave	138	year	85	wave	197	water	249

Lexeau® 09/2018

FIGURE 2.40 – Substantifs les plus fréquents dans FLR-1 et FLR-01 à FLR-04

fusionnées dans une liste « agrégée » de 42 substantifs, reprise dans la requête de réalisation de ces substantifs présentée dans la figure 2.41. Cette requête a été lancée sur les partitions

```
AZ [enpos="nn|nns" & enlemma="analysis|approach|area|assessment|breach|
catchment|change|climate|copula|damage|datum|distribution|embankment|erosion|
event|failure|flood|flow|height|levee|level|method|model|parameter|period|
probability|rainfall|result|risk|river|sea|slope|soil|study|surge|system|time|value|
variable|water|wave|year"]:enlemma
```

Lexeau® 09/2018

FIGURE 2.41 – Requête par index de la fréquence des 42 substantifs de la liste agrégée

du corpus FLR-1 par thème et par texte en regroupant les résultats dans un seul tableau. . Ils

FLR-1, partition par thème	T	Thème 01	Thème 02	Thème 03	Thème 04	Nb de	Nb de
42 substantifs (début)	304 301	61 931	22 317	101 567	118 486	thèmes	textes
	F						
analysis	738	202	57	307	172	4	63
approach	479	149	35	107	188	4	50
area	708	130	90	143	345	4	62
assessment	302	39	4	200	59	4	41
breach	491	0	1	418	72	3	17
catchment	413	200	11	0	202	3	22
change	230	40	66	78	46	4	44
climate	150	11	98	11	30	4	26
copula	103	45	58	0	0	2	4
damage	290	4	11	74	201	4	32
datum	924	274	81	200	369	4	65
distribution	378	202	55	60	61	4	48
embankment	293	3	1	262	27	4	19
erosion	440	10	0	356	74	3	23
event	945	369	87	107	382	4	56
failure	570	68	0	466	36	3	32
flood	1 751	388	48	544	771	4	67
flow	762	170	7	225	360	4	54
height	261	39	129	69	24	4	29
levee	646	1	0	573	72	3	21
level	945	316	190	266	173	4	61
method	748	290	32	191	235	4	60
						Lexeau® 10/2018	

FIGURE 2.42 – Fréquence des 42 substantifs et thèmes/textes « en ayant » (début)

FLR-1, partition par thème	T	Thème 01	Thème 02	Thème 03	Thème 04	Nb de	Nb de
42 substantifs (fin)	304 301	61 931	22 317	101 567	118 486	thèmes	textes
	F						
model	1 561	242	125	535	659	4	63
parameter	551	167	75	165	144	4	55
period	427	153	90	55	129	4	52
probability	584	269	47	230	38	4	41
rainfall	434	177	22	4	231	4	33
result	590	87	61	194	248	4	65
risk	631	88	8	251	284	4	51
river	425	106	17	131	171	4	53
sea	238	87	108	13	30	4	23
slope	313	41	7	229	36	4	35
soil	298	11	11	205	71	4	29
study	559	96	85	161	217	4	63
surge	175	53	93	21	8	4	15
system	441	47	5	266	123	4	53
time	574	91	61	199	223	4	63
value	713	267	53	150	243	4	62
variable	279	142	33	81	23	4	35
water	954	233	75	397	249	4	62
wave	528	138	132	197	61	4	30
year	418	117	85	70	146	4	51
						Lexeau® 10/2018	

FIGURE 2.43 – Fréquence des 42 substantifs et thèmes/textes « en ayant » (fin)

sont présentés dans les figures 2.42 et 2.43. La plupart des 42 substantifs sont présents dans les 4 thèmes. La présence dans les textes est variable, avec 4 textes pour « copula ».

2.5.2.3 Une première émergence du lexique des textes

Pour mieux apprécier le partage des substantifs entre les textes, la recherche s'est poursuivie sur la partition d'un corpus par textes, en deux étapes, avec un choix analogue de seuils de fréquence que pour la partition d'un corpus partitionné en sous-corpus.

Pour ne pas alourdir la présentation des résultats, la recherche a été effectuée sur le corpus FLR-02, compte tenu du petit nombre de communications. Elle a été décomposée en deux étapes pour travailler sur des tableaux de résultats ne dépassant pas 20 lignes. Les résultats sont présentés dans la figure 2.44, ce qui a permis de constituer une première liste agrégée des 51 substantifs (fig. 2.45).

FLR-02 (fmin=57)		FLR-02001 (fmin=16)		FLR-02002 (fmin=27)		FLR-02003 (fmin=12)		FLR-02004 (fmin=13)	
enlemma	F	enlemma	F	enlemma	F	enlemma	F	enlemma	F
analysis	57	analysis	18	area	48	accumulation	12	analysis	16
area	90	change	16	copula	54	area	28	approach	20
change	66	climate	25	dependence	36	change	18	change	16
climate	98	coast	21	distribution	41	climate	34	climate	13
copula	58	datum	34	estimate	29	drought	20	condition	14
datum	81	flood	16	event	33	event	33	datum	13
event	87	level	60	function	33	flood	25	defence	25
height	129	model	50	grid	38	hazard	16	event	19
level	190	period	25	height	112	impact	18	level	51
model	125	resolution	20	level	76	loss	15	line	13
parameter	75	result	21	parameter	60	model	36	model	18
period	90	sea	27	period	31	period	12	period	22
result	61	station	30	point	41	property	12	rate	17
sea	108	storm	21	probability	34	rainfall	22	result	23
study	85	study	46	sea	53	scenario	18	return	17
surge	93	surge	64	time	34	study	15	scenario	20
time	61	tide	32	trend	35	watershed	22	sea	27
water	75	water	32	value	27	year	32	water	14
wave	132			variable	27			wave	33
year	85			wave	96			year	22
								Lexeau® 09/2018	

FIGURE 2.44 – Substantifs les plus fréquents dans FLR-02 et chacun de ses textes (étape 1)

accumulation analysis approach area change climate coast condition copula datum defence dependence distribution drought estimate event flood function grid hazard height impact level line loss model parameter period point probability property rainfall rate resolution result return scenario sea station storm study surge tide time trend value variable water watershed wave year

Lexeau® 10/2018

FIGURE 2.45 – Liste alphabétique des 51 substantifs les plus fréquents (FLR-02 et textes)

Les substantifs de cette liste qui ne figurent pas dans la liste agrégée des 42 substantifs les plus fréquents du corpus FLR-1 partitionné par thème sont surlignés en jaune foncé.

La liste agrégée a été introduite dans une requête sur la partition fine du corpus, suivie d'un

traitement à l'export pour dénombrer les textes « en ayant ». Les résultats sont présentés dans les tableaux successifs des figures 2.46 et 2.47. Les substantifs sont listés dans l'orde alphabétique.

Partition FLR-02 par texte	T	02001	02002	02003	02004	Nb de
Liste alphabétique des 51	22 317	6 612	7 550	4 166	3 989	textes
substantifs (début)	F					
accumulation	12	0	0	12	0	1
analysis	57	18	20	3	16	4
approach	35	3	8	4	20	4
area	90	6	48	28	8	4
change	66	16	16	18	16	4
climate	98	25	26	34	13	4
coast	35	21	13	0	1	3
condition	28	2	2	10	14	4
copula	58	4	54	0	0	2
datum	81	34	23	11	13	4
defence	26	1	0	0	25	2
dependence	39	3	36	0	0	2
distribution	55	12	41	1	1	4
drought	20	0	0	20	0	1
estimate	35	4	29	0	2	3
event	87	2	33	33	19	4
flood	48	16	3	25	4	4
function	34	0	33	0	1	2
grid	46	5	38	1	2	4
hazard	33	15	1	16	1	4
height	129	12	112	0	5	3
impact	33	5	4	18	6	4
level	190	60	76	3	51	4
line	27	3	10	1	13	4
loss	15	0	0	15	0	1
model	125	50	21	36	18	4
					Lexeau® 09/2018	

FIGURE 2.46 – Fréquence des 51 substantifs et textes « en ayant » du corpus FLR-02 (début)

La fréquence la plus élevée a été surlignée en vert sur chaque ligne des deux tableaux, ce qui permet de repérer en ligne le texte à consulter en priorité pour tel ou tel substantif et d'avoir la répartition des substantifs entre les textes en consultant la dernière colonne. Cette coloration des cases livre en colonne un « profil » plus ou moins riche du texte pour l'emploi des substantifs en ligne.

Les résultats montrent la place importante du texte 02002 dans l'emploi des substantifs les plus fréquents et le caractère spécifique du texte 02003 où se trouvent les 6 substantifs présents dans un seul texte.

Dès ce stade de l'émergence des substantifs les plus fréquents dans les textes, certains d'entre

Partition FLR-02 par texte	T	02001	02002	02003	02004	Nb de
Liste alphabétique des 51	22 317	6 612	7 550	4 166	3 989	textes
substantifs (fin)	F					
parameter	75	7	60	1	7	4
period	90	25	31	12	22	4
point	47	1	41	1	4	4
probability	47	5	34	2	6	4
property	12	0	0	12	0	1
rainfall	22	0	0	22	0	1
rate	20	1	0	2	17	3
resolution	27	20	2	5	0	3
result	61	21	10	7	23	4
return	48	15	10	6	17	4
scenario	46	7	1	18	20	4
sea	108	27	53	1	27	4
station	31	30	1	0	0	2
storm	54	21	26	2	5	4
study	85	46	19	15	5	4
surge	93	64	22	2	5	4
tide	36	32	1	0	3	3
time	61	14	34	5	8	4
trend	44	7	35	2	0	3
value	53	13	27	7	6	4
variable	33	2	27	1	3	4
water	75	32	21	8	14	4
watershed	22	0	0	22	0	1
wave	132	1	96	2	33	4
year	85	15	16	32	22	4
Total	2 909	723	1 214	476	496	
					Lexeau® 09/2018	

FIGURE 2.47 – Fréquence des 51 substantifs et textes « en ayant » du corpus FLR-02 (fin)

eux n'appartiennent pas au langage courant et/ou sont peu partagés par les auteurs.

La question est de savoir s'ils peuvent trouver leur place dans le lexique. Elle se pose différemment selon qu'ils appartiennent ou non au langage courant. Le substantif *copula* en est un exemple. Il n'appartient pas au langage courant, avec un emploi limité à 2 textes dans ce corpus, dont 54 emplois pour le texte 02002 (cf. fig. 2.46).

Nous reviendrons sur cet exemple (cf. 2.6.1), qui soulève la question de l'extension du champ du lexique à des mots ou des expressions difficiles à définir et peu partagés par les auteurs.

2.5.2.4 Finaliser la recherche des substantifs

La poursuite de la recherche de substantifs pertinents au delà des 51 substantifs s'effectue par itération, avec comme précédemment le tableau de la figure 2.48 des substantifs les plus fréquents dans le corpus du thème 02 et ses 4 textes, sur des seuils de fréquence nettement plus faibles que ceux de la figure 2.44.

FLR-02 (fmin2 =18)		FLR-02001 (fmin2 =7)		FLR-02002 (fmin2 =9)		FLR-02003 (fmin2 =6)		FLR-02004 (fmin2 =6)	
enlemma	F	enlemma	F	enlemma	F	enlemma	F	enlemma	F
century	20	accuracy	7	century	14	catchment	11	adaptation	7
design	19	calculation	7	design	14	cost	6	crest	8
effect	24	calibration	9	effect	10	damage	6	estuary	6
exceedance	18	comparison	9	exceedance	18	disaster	8	example	6
factor	20	factor	15	half	12	dryness	7	formula	7
increase	35	gauge	15	interval	10	increase	10	frontage	11
location	23	increase	14	location	13	insurance	6	infrastructure	8
maxima	27	information	11	marginals	9	order	7	measure	6
method	32	map	10	maxima	12	overflow	6	number	8
number	18	maxima	15	method	12	rain	6	option	7
order	28	method	14	order	16	region	6	railway	9
part	23	observation	10	profile	11	river	10	range	6
range	19	part	10	range	11	scheme	7	rise	7
rise	19	performance	9	scale	11	simulation	11	seawall	7
scale	18	rise	10	shape	11	soil	11	structure	11
section	20	series	11	structure	19	surface	11	table	7
simulation	27	set	9	variation	16	temperature	6	train	6
structure	31	simulation	10	window	14	variation	10	wall	7
variation	29			work	14			wind	6
work	18								

Lexeau® 09/2018

FIGURE 2.48 – Substantifs les plus fréquents dans FLR-02 et dans ses textes (étape 2)

Ce tableau permet d'établir une deuxième la liste agrégée des 66 substantifs supplémentaires les plus fréquents (fig. 2.49).

accuracy adaptation calculation calibration catchment century comparison cost crest damage design disaster dryness effect estuary example exceedance factor formula frontage gauge half increase information infrastructure insurance interval location map marginals maxima measure method number observation option order overflow part performance profile railway rain range region rise river scale scheme seawall section series set shape simulation soil structure surface table temperature train variation wall wind window work

Lexeau® 10/2018

FIGURE 2.49 – Deuxième liste agrégée des 66 substantifs dans FLR-02 et dans ses textes

Cette liste permet d'écrire la requête de recherche dans la partition du corpus FLR-02 par texte des fréquences des 66 substantifs supplémentaires et de traiter à l'export, comme précédemment, les résultats de cette requête pour obtenir le nombre de textes « en ayant ». Les résultats de la requête sont présentés en quatre tableaux successifs qui regroupent les substantifs en fonction de leur présence dans les textes (fig. 2.50, 2.51, 2.52 et 2.53).

FLR-02, partition par texte	FLR-02	02001	02002	02003	02004
2ème sélection	T				
14 substantifs dans les 4 textes	22 317	6 612	7 550	4 166	3 989
	F				
effect	24	6	10	4	4
increase	35	14	8	10	3
information	15	11	1	2	1
location	23	6	13	1	3
measure	17	5	5	1	6
method	32	14	12	3	3
number	18	3	2	5	8
order	28	4	16	7	1
part	23	10	6	5	2
range	19	1	11	1	6
rise	19	10	1	1	7
simulation	27	10	1	11	5
variation	29	1	16	10	2
wind	16	6	2	2	6
				Lexeau® 10/2018	

FIGURE 2.50 – Substantifs de la 2ème liste agrégée présents dans les 4 textes (FLR-02)

FLR-02, partition par texte	FLR-02	02001	02002	02003	02004
2ème sélection	T				
14 substantifs dans 3 textes sur 4	22 317	6 612	7 550	4 166	3 989
	F				
calculation	12	7	0	1	4
calibration	11	9	1	0	1
century	20	2	14	4	0
damage	11	0	1	6	4
formula	10	1	2	0	7
half	15	2	12	1	0
region	13	5	0	6	2
river	17	4	0	10	3
scale	18	2	11	5	0
section	20	9	4	0	7
set	12	9	0	2	1
structure	31	0	19	1	11
table	9	1	0	1	7
work	18	2	14	2	0
				Lexeau® 10/2018	

FIGURE 2.51 – Substantifs de la 2ème liste agrégée présents dans 3 textes sur 4 (FLR-02)

Les résultats de cette 2ème étape montrent un renforcement important des substantifs présents dans moins de 4 textes par rapport aux résultats obtenus avec la 1ère liste agrégée pour des fréquences plus importantes. Ces résultats permettent d'approcher le vocabulaire spécifique de chaque texte. On constate le rôle prépondérant des communications 02002 et 02001 pour les substantifs plus ou moins partagés. En dehors de *exceedanc*, *map* et *marginals*, les substantifs de la figure 2.53 sont propres aux communications 02003 et 02004. Les résultats des deux étapes peuvent être regroupés dans une liste de 117 substantifs (51 + 66).

FLR-02, partition par texte	FLR-02	02001	02002	02003	02004
2ème sélection	T				
16 substantifs dans 2 textes sur 4	22 317	6 612	7 550	4 166	3 989
	F				
accuracy	8	7	0	0	1
comparison	10	9	0	1	0
cost	7	0	0	6	1
design	19	0	14	0	5
example	9	3	0	0	6
factor	20	15	0	5	0
gauge	17	15	0	0	2
infrastructure	10	2	0	0	8
interval	11	1	10	0	0
maxima	27	15	12	0	0
observation	14	10	4	0	0
overflow	7	0	0	6	1
performance	11	9	0	0	2
profile	13	0	11	0	2
shape	14	3	11	0	0
window	15	0	14	0	1
				Lexeau® 10/2018	

FIGURE 2.52 – Substantifs de la 2ème liste agrégée présents dans 2 textes sur 4 (FLR-02)

FLR-02 partition par texte	FLR-02	02001	02002	02003	02004
2ème sélection	T				
22 substantifs dans un seul texte	22 317	6 612	7 550	4 166	3 989
	F				
adaptation	7	0	0	0	7
catchment	11	0	0	11	0
crest	8	0	0	0	8
disaster	8	0	0	8	0
dryness	7	0	0	7	0
estuary	6	0	0	0	6
exceedance	18	0	18	0	0
frontage	11	0	0	0	11
insurance	6	0	0	6	0
map	10	10	0	0	0
marginals	9	0	9	0	0
option	7	0	0	0	7
railway	9	0	0	0	9
rain	6	0	0	6	0
scheme	7	0	0	7	0
seawall	7	0	0	0	7
series	11	11	0	0	0
soil	11	0	0	11	0
surface	11	0	0	11	0
temperature	6	0	0	6	0
train	6	0	0	0	6
wall	7	0	0	0	7
				Lexeau® 10/2018	

FIGURE 2.53 – Substantifs de la 2ème liste agrégée présents dans 1 texte sur 4 (FLR-02)

2.5.2.5 Une première sélection du lexique de base

L'entrée d'un substantif dans le lexique de l'eau stratifié est acquise si son lemme correspond à celui d'une unité lexicale existante, partagé par celui de l'unité discursive centrale associée, et s'il peut être rattaché à la strate discursive de l'unité lexicale dans la même acception, ce que les textes qui ont conduit à sa sélection permettent de vérifier. Si cette dernière condition n'est pas réunie et si le contexte témoigne d'un emploi spécifique et stable dans le domaine de l'eau dans une autre strate discursive, l'intégration du substantif dans le lexique peut être envisagée comme nouvelle unité lexicale satellite de l'unité lexicale pivot existante, à la condition d'y figurer avec au minimum une définition. Si le lemme du substantif candidat ne correspond pas à celui d'une unité discursive centrale mais qu'il partage la définition d'une unité lexicale existante dans un emploi constant dans le domaine de l'eau, il peut être intégré dans le lexique comme nouvelle unité discursive alternative. Il reste une quatrième possibilité d'intégration du substantif dans une strate discursive avec une acception stable en tant que nouvelle unité lexicale pivot, ce qui suppose de nouveaux concepts dans l'ontologie ou de nouveaux liens de concepts existants avec le lexique. Les cinq possibilités d'intégration se présentent ainsi, avec un code de couleurs en cas d'intégration effective ou potentielle :

1. pas d'intégration dans le lexique de l'eau ;
2. vert : unité discursive centrale et unité lexicale existantes ;
3. cyan : nouvelle unité discursive alternative et unité lexicale existante ;
4. jaune : unité discursive centrale et nouvelle unité lexicale satellite ;
5. magenta : nouvelle unité lexicale pivot ou satellite (mise à jour de l'ontologie).

La figure 2.54 présente avec un code des couleurs, dans l'ordre alphabétique, les possibilités d'intégration des 117 substantifs de la sélection finale.

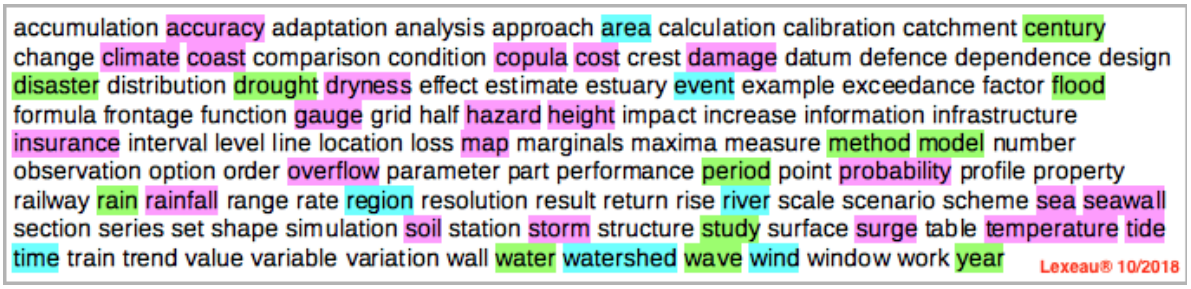


FIGURE 2.54 – Intégration dans le lexique des substantifs sélectionnés du corpus FLR-02

La requête sur la partition du corpus FLR-02 par texte issue de l'addition des listes agrégées 1 et 2 donne les fréquences des 117 substantifs dans les tableaux des figures 2.55 à 2.59 avec un report des couleurs de la figure 2.54.

Les figures 2.60 et 2.61 présentent en deux tableaux successifs le résultat de l'intégration des substantifs dans le lexique jusqu'à l'objet source de l'unité lexicale dans l'ontologie du domaine. Le tableau de la figure 2.62 présente les possibilités d'intégration des 117 substantifs les plus fréquents dans le lexique en l'état au 28/09/2018.

FLR-02, partition par texte	T	02001	02002	02003	02004
Sélections 1 et 2	22 317	6 612	7 550	4 166	3 989
47 noms dans les 4 textes (début)	F				
analysis	57	18	20	3	16
approach	35	3	8	4	20
area	90	6	48	28	8
change	66	16	16	18	16
climate	98	25	26	34	13
condition	28	2	2	10	14
datum	81	34	23	11	13
distribution	55	12	41	1	1
effect	24	6	10	4	4
event	87	2	33	33	19
flood	48	16	3	25	4
grid	46	5	38	1	2
hazard	33	15	1	16	1
impact	33	5	4	18	6
increase	35	14	8	10	3
information	15	11	1	2	1
level	190	60	76	3	51
line	27	3	10	1	13
location	23	6	13	1	3
measure	17	5	5	1	6
method	32	14	12	3	3
model	125	50	21	36	18
number	18	3	2	5	8
order	28	4	16	7	1
				Lexeau® 10/2018	

FIGURE 2.55 – Fréquence et intégration des substantifs présents dans les 4 textes (début)

Partition FLR-02 par texte	T	02001	02002	02003	02004
Listes agrégées 1 et 2	22 317	6 612	7 550	4 166	3 989
47 noms dans les 4 textes (fin)	F				
parameter	75	7	60	1	7
part	23	10	6	5	2
period	90	25	31	12	22
point	47	1	41	1	4
probability	47	5	34	2	6
range	19	1	11	1	6
result	61	21	10	7	23
return	48	15	10	6	17
rise	19	10	1	1	7
scenario	46	7	1	18	20
sea	108	27	53	1	27
simulation	27	10	1	11	5
storm	54	21	26	2	5
study	85	46	19	15	5
surge	93	64	22	2	5
time	61	14	34	5	8
value	53	13	27	7	6
variable	33	2	27	1	3
variation	29	1	16	10	2
water	75	32	21	8	14
wave	132	1	96	2	33
wind	16	6	2	2	6
year	85	15	16	32	22
				Lexeau® 10/2018	

FIGURE 2.56 – Fréquence et intégration des substantifs présents dans les 4 textes (fin)

FLR-02, partition par texte	T	02001	02002	02003	02004	
Sélections 1 et 2	22 317	6 612	7 550	4 166	3 989	
21 noms dans 3 textes sur 4	F					
calculation	12	7	0	1	4	3
calibration	11	9	1	0	1	3
century	20	2	14	4	0	3
coast	35	21	13	0	1	3
damage	11	0	1	6	4	3
estimate	35	4	29	0	2	3
formula	10	1	2	0	7	3
half	15	2	12	1	0	3
height	129	12	112	0	5	3
rate	20	1	0	2	17	3
region	13	5	0	6	2	3
resolution	27	20	2	5	0	3
river	17	4	0	10	3	3
scale	18	2	11	5	0	3
section	20	9	4	0	7	3
set	12	9	0	2	1	3
structure	31	0	19	1	11	3
table	9	1	0	1	7	3
tide	36	32	1	0	3	3
trend	44	7	35	2	0	3
work	18	2	14	2	0	3
					Lexeau® 10/2018	

FIGURE 2.57 – Fréquence et intégration des substantifs présents dans 3 textes sur 4

FLR-02, partition par texte	T	02001	02002	02003	02004	
Sélections 1 et 2	22 317	6 612	7 550	4 166	3 989	
21 noms dans 2 textes sur 4	F					
accuracy	8	7	0	0	1	2
comparison	10	9	0	1	0	2
copula	58	4	54	0	0	2
cost	7	0	0	6	1	2
defence	26	1	0	0	25	2
dependence	39	3	36	0	0	2
design	19	0	14	0	5	2
example	9	3	0	0	6	2
factor	20	15	0	5	0	2
function	34	0	33	0	1	2
gauge	17	15	0	0	2	2
infrastructure	10	2	0	0	8	2
interval	11	1	10	0	0	2
maxima	27	15	12	0	0	2
observation	14	10	4	0	0	2
overflow	7	0	0	6	1	2
performance	11	9	0	0	2	2
profile	13	0	11	0	2	2
shape	14	3	11	0	0	2
station	31	30	1	0	0	2
window	15	0	14	0	1	2
					Lexeau® 10/2018	

FIGURE 2.58 – Fréquence des substantifs présents dans 2 textes sur 4

FLR-02, partition par texte	T	02001	02002	02003	02004	
Sélections 1 et 2	22 317	6 612	7 550	4 166	3 989	
28 noms dans 1 texte sur 4	F					
accumulation	12	0	0	12	0	1
adaptation	7	0	0	0	7	1
catchment	11	0	0	11	0	1
crest	8	0	0	0	8	1
disaster	8	0	0	8	0	1
drought	20	0	0	20	0	1
dryness	7	0	0	7	0	1
estuary	6	0	0	0	6	1
exceedance	18	0	18	0	0	1
frontage	11	0	0	0	11	1
insurance	6	0	0	6	0	1
loss	15	0	0	15	0	1
map	10	10	0	0	0	1
marginals	9	0	9	0	0	1
option	7	0	0	0	7	1
property	12	0	0	12	0	1
railway	9	0	0	0	9	1
rain	6	0	0	6	0	1
rainfall	22	0	0	22	0	1
scheme	7	0	0	7	0	1
seawall	7	0	0	0	7	1
series	11	11	0	0	0	1
soil	11	0	0	11	0	1
surface	11	0	0	11	0	1
temperature	6	0	0	6	0	1
train	6	0	0	0	6	1
wall	7	0	0	0	7	1
watershed	22	0	0	22	0	1
					Lexeau® 10/2018	

FIGURE 2.59 – Fréquence et intégration des substantifs présents dans 1 seul texte

L'intégration effective des substantifs est précisée dans les tableaux des figures 2.60 et 2.61, selon le type d'intégration. L'intégration des substantifs *Area* et *Time* n'a pas été jugée

Substantifs présents dans les 4 textes				
Nlle unité discursive alternative (anglais)	Unité discursive centrale (anglais, préfixe Udc)	Unité lexicale de rattachement (anglais)	Classe (C) ou instance de classe (I) dans l'ontologie du domaine (anglais/français)	
Area	Aucune	Aucune		Aucune
Event	Extreme event	1-TS Extreme event	C	Extreme event/Événement extrême
Time	Aucune	Aucune		Aucune
Wind	Strong wind	1-TS Strong wind	C	Strong wind/Vent fort
	Flood	1-TS Flood	C	Flood/Inondation
	Method	1-TS Method	C	Method/Méthode
	Model	1-TS Model	C	Model/Modèle
	Period	1-TS Period	C	Period/Période
	Study	1-TS Study	C	Study/Étude
	Water	1-TS Water	I	Water/Eau [classe : Substance chimique]
	Wave	1-TS Wave	C	Wave/Vague
	Year	1-TS Année	C	Year/Année
				Lexeau® 11/2018

FIGURE 2.60 – Intégration dans le lexique des substantifs présents dans les 4 textes

possible, compte tenu de la polysémie du premier et du problème de traduction du second en français. Le premier ne correspond à aucune classe d'entités géographiques. Le second est introduit dans l'ontologie du domaine au sens de « temps universel » comme classe d'enti-

Substantifs présents dans moins de 4 textes				
Nlle unité discursive alternative (anglais)	Unité discursive centrale (anglais, préfixe Udc)	Unité lexicale de rattachement (anglais)	Classe (C) ou instance de classe (I-C) dans l'ontologie du domaine (anglais/français)	
Région @fr	Région INSEE @fr	3-TA Région INSEE @fr	C	Région INSEE @fr/Région INSEE
River	Watercourse	1-TS Watercourse	C	Watercourse/Cours d'eau
Watershed	Drainage basin	1-TS Drainage basin	C	Drainage basin/Bassin versant
	Century	1-TS Century	C	Century/Siècle
	Disaster	3-TA Disaster	C	Disaster/Catastrophe
	Drought	1-TS Drought	C	Drought/Sécheresse
	Rain	1-TS Rain	C	Rain/Pluie
				Lexeau® 11/2018

FIGURE 2.61 – Intégration dans le lexique des substantifs présents dans moins de 4 textes

tés temporelles libellée UTC dans les deux langues, sans créer de nouvelle unité discursive alternative.

Le mot « région » utilisé en France comme subdivision administrative n'a pas d'équivalent en anglais avec *region* où l'on trouve d'autres mots (*province*, *county*, *district*, etc.). Il a té introduit comme unité discursive alternative monolingue, rattachée à l'unité lexicale monolingue « 3-TA Région INSEE » issue d'une classe monolingue dans l'ontologie du domaine.

Résultats de l'intégration dans le lexique des substantifs sélectionnés (FLR-02)	Présence dans les textes				
	4 sur 4	3 sur 4	2 sur 4	1 sur 4	Ensemble
Pas d'intégration	31	14	16	17	78
Unités discursives et lexicales exisantes	8	1	0	3	12
Nouvelles unités discursives alternatives	2	2	0	1	5
Intégration envisagée après mise à jour	6	4	5	7	22
Ensemble des substantifs	47	21	21	28	117
					Lexeau® 11/2018

FIGURE 2.62 – Résultats de l'intégration des 117 substantifs dans le lexique (16/11/2018)

2.6 Une mise à jour lexicale et conceptuelle

La mise à jour lexicale et conceptuelle est la phase finale de la rencontre des textes et du lexique. Elle porte aussi bien sur un ajout de noms propres que d'unités du lexique. Elle s'accompagne de la mise à jour de l'ontologie du domaine. Elle peut se faire aussi par l'introduction de syntagmes dans le lexique en lieu et place de leur base nominale, classée comme unité récurrente. Cette question est abordée dans le chapitre 3. Le modèle des unités récurrentes regroupe les noms et les syntagmes susceptibles d'entrer dans le lexique. Il comprend les bases nominales des syntagmes récurrents ou lexicalisés, lesquelles permettent d'accéder à ces entités.

Le graphe relationnel des unités récurrentes et des noms propres potentiels qui émergent de l'analyse textométrique est présenté dans la figure 2.63 [*à revoir JLJ-17/12/2018*]. Dans ce graphe, la mise à jour s'effectue par transfert de noms propres et d'unités récurrentes d'une classe dans une autre, comme nouvelle entité, en lien avec des entités homologues de l'ontologie du domaine (cf. fig. 2.64). Un nom propre récurrent est transféré comme nom propre

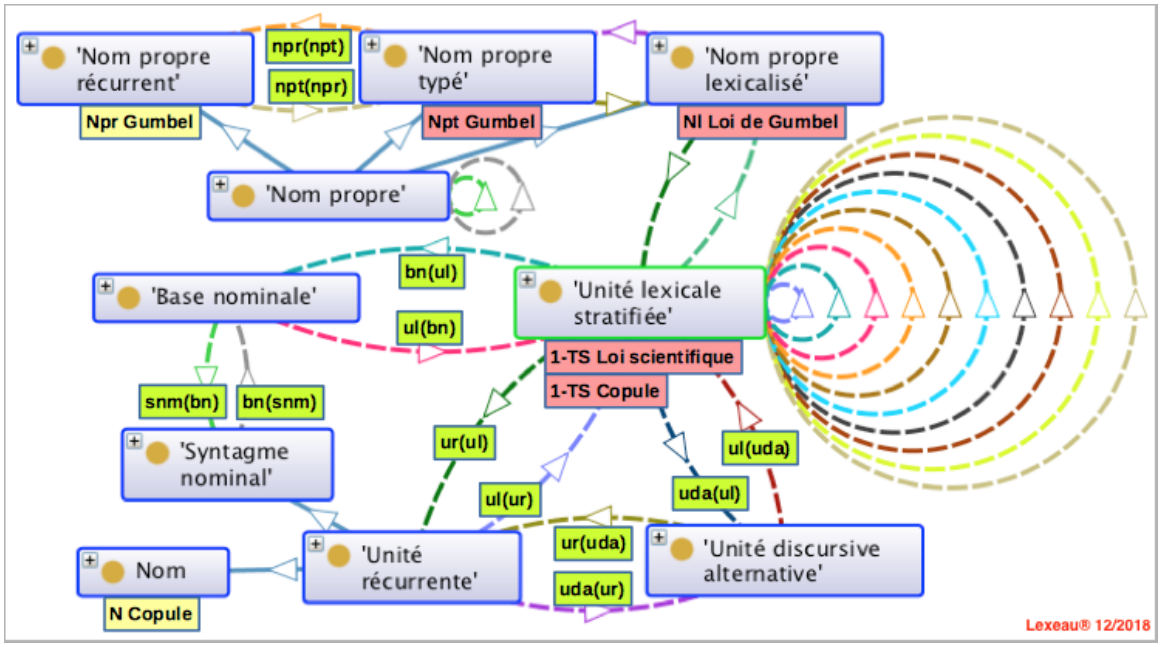


FIGURE 2.63 – Lexicalisation des noms, des syntagmes nominaux et des noms propres

lexicalisé s'il recoupe un nom propre typé sans lien avec un nom propre déjà lexicalisé et trouve son homologue dans l'ontologie du domaine. Des noms et des syntagmes récurrents sont transférés comme unité lexicales, en tant qu'unité lexicale pivot ou satellite dotée d'une entité source directe ou indirecte dans l'ontologie du domaine. L'unité récurrente peut aussi être transférée comme unité discursive alternative rattachée à une unité lexicale. Le changement du libellé préfixé au cours du transfert est illustré dans les exemples du graphe, avant le transfert (sur fond jaune) et après le transfert (sur fond magenta).

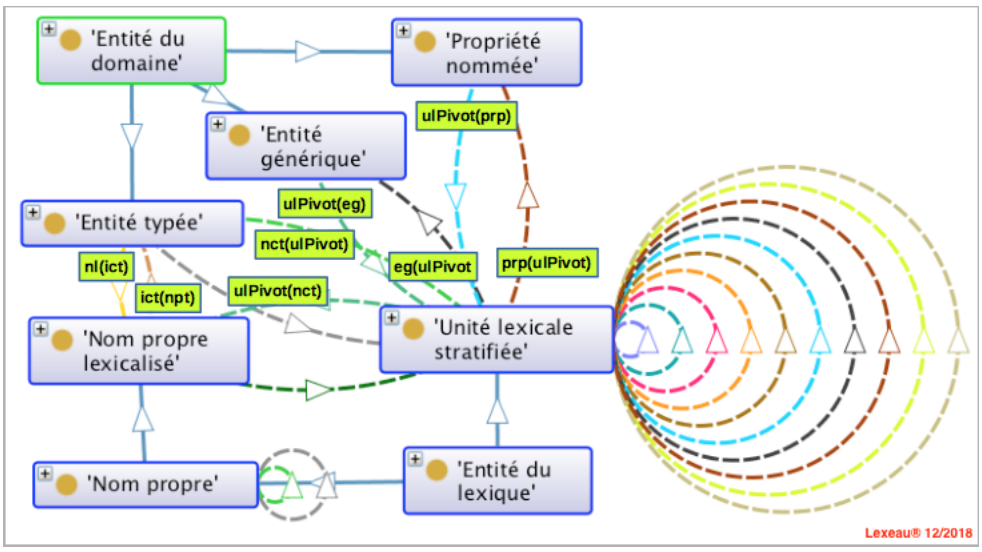


FIGURE 2.64 – Graphe de la « remontée » des entités du lexique dans l'ontologie du domaine

2.6.1 Des définitions adaptées au domaine de connaissance

Le choix des unités lexicales pivot du lexique en lien avec un concept du domaine pose la question de l'extension du lexique dans le périmètre d'un domaine de connaissance raisonnable. Le point de vue adopté est celui d'un observateur qui voudrait comprendre comment le langage savant et le langage courant « parlent » de l'eau, avec parfois les mêmes mots qui ne veulent pas dire la même chose, et parfois d'autres mots qui ne sont employés que par un petit nombre de spécialistes, sur des problématiques finales que tout le monde peut comprendre, avec des outils de connaissance très spécialisés. S'il n'est pas utile de connaître la mécanique pour conduire une voiture, c'est parce que les garagistes sont là pour l'entretenir et remplacer les pièces défectueuses. La « mécanique » du langage expose les locuteurs à des « pannes » de compréhension de leur auditoire auxquelles ils remédient le plus souvent en évitant les termes techniques et scientifiques qui fâchent. Tout le monde reconnait pourtant que les bons professeurs sont ceux qui font comprendre à leur auditoire les choses réputées difficiles à comprendre avec les mots les plus simples.

Sur le plan de son extension, le caractère encyclopédique du lexique de l'eau est incontestable, en termes de disciplines concernées et de thèmes traités. Il ne s'agit pas d'une encyclopédie de la connaissance, à l'image de l'encyclopédie Wikipedia, dont le lexique est d'abord tributaire, dans la mesure où cette encyclopédie a remplacé, pour ceux de ma génération, le grand Larousse du 20ème siècle, voire l'Encyclopedia Universalis. Le concept d'un « wikimedia de l'eau » serait à creuser, avec l'introduction des strates discursives pour élargir le public, le recours à une ontologie des objets de connaissance pour maîtriser la thématique et le recours au traitement automatique des textes pour valider les entrées. La problématique due la définition de « Copule » dans la strate technico-scientifique permet d'illustrer ce qui précède.

Le substantif *copula* apparait 54 fois dans le texte 02002 et 4 fois dans le texte 02001. Une recherche complémentaire sur la partition du corpus FLR-1 par texte montre qu'il apparait dans 4 textes du corpus avec une fréquence non négligeable dans le texte 01011 (fig. 2.65).

Partition de FLR-1 par texte	FLR-1	01005	01011	02001	02002	Nb de
Lemme réalisé	T					textes
	304 301	4 689	7 486	6 612	7 550	
	F					
copula	103	13	32	4	54	4
					Lexeau® 09/2018	

FIGURE 2.65 – Fréquence du lemme « *copula* » dans le corpus FLR-1

L'encyclopédie Wikipedia a été consultée le 21/09/2018 dans les deux langues. Elle donne une définition en anglais [7] et en français [8] reproduite dans les figures 2.66 et 2.67.

L'introduction de « Copule/*Copula* » dans la strate discursive technico-scientifique du lexique devra être précédée de l'introduction du concept bilingue dans l'ontologie du domaine comme source d'une nouvelle unité lexicale pivot. Ceci implique un travail plus large sur les entrées les plus courantes du calcul des probabilités telles que « Période de retour/*Return period* » et « Densité de probabilité/*Probability density* », et sur des noms propres lexicalisés tels que

7. `https://en.wikipedia.org/wiki/Copula_(probability_theory)`
8. `https://fr.wikipedia.org/wiki/Copule` pour accéder à l'article sur la copule en mathématiques.

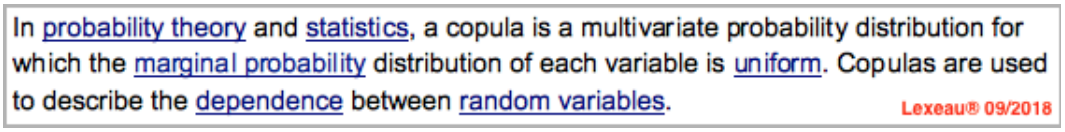


FIGURE 2.66 – Début de l'article de Wikipedia sur le substantif « Copula »

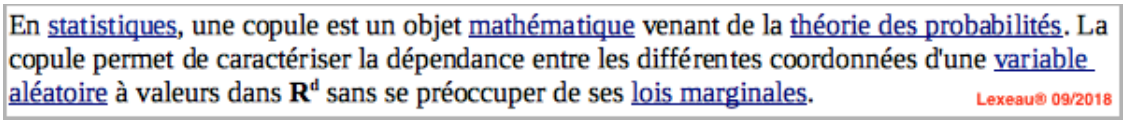


FIGURE 2.67 – Début de l'article de Wikipedia sur la copule en mathématiques

« Copule de Gumbel/*Gumbel's copula* » ou « Loi de Gumbel/*Gumbel law* », rencontré dans le résumé du texte 01014. Comment délimiter le périmètre de ce travail ?

Pour revenir aux articles de Wikipedia, avec une définition en forme de renvois sur d'autres termes, suivie de développements mathématiques, il nous semble que ce type de définition ne peut être retenu pour le lexique. L'article de Wikipedia ne peut que compléter, par un renvoi, une définition propre au lexique, tirée ou inspirée d'un ouvrage de référence. La définition du lexique tenterait d'expliquer en quelques phrases pourquoi et comment le concept est utilisé dans la pratique des ingénieurs et dans la recherche, et ce pour calculer la probabilité d'un évènement extrême du domaine. L'ajout de « Calcul des probabilités » comme entité générique dans la classe des disciplines et des matières du domaine de l'eau est un préalable à la délimitation du périmètre conceptuel de l'ontologie (fig. 2.68).

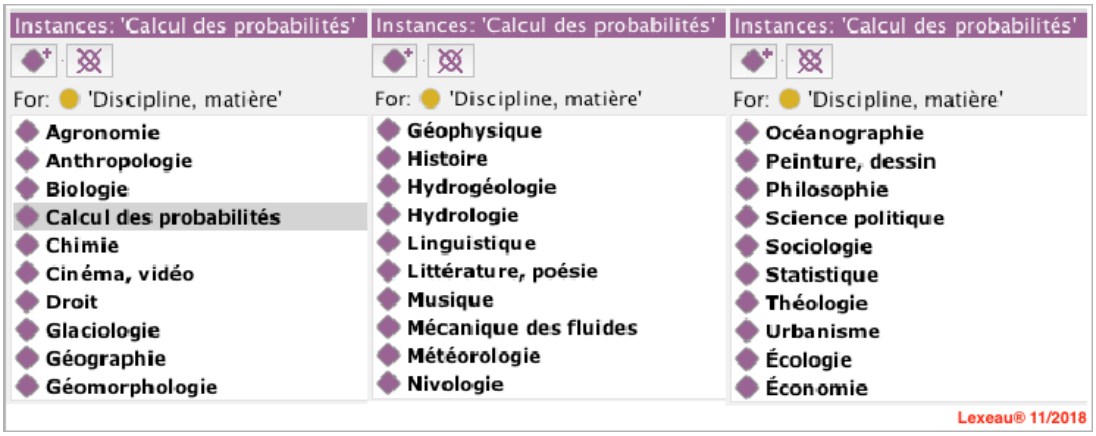


FIGURE 2.68 – Le calcul des probabilités parmi les disciplines et les matières du lexique

Il reste à introduire le concept de copule dans l'ontologie, ce qui est possible avec une sous-classe « Copule » de la classe « Objet mathématique » dans l'arbre des entités typées du domaine, avec comme instance la copule de Gumbel. Nous avons profité de cette mise à jour de l'ontologie pour introduire une classe des lois scientifiques parmi les entités typées, avec la loi de Gumbel comme instance.

La définition de la loi de Gumbel pose les mêmes problèmes que celle des copules, mais le texte de l'article de Wikipedia soulève une autre difficulté, à la lecture de l'article, considéré

comme une ébauche dans la discussion. L'utilisation de cette loi dans la prévision des crues est évoquée dans le texte surligné en jaune de la figure 2.69, qui présente des extraits du début de l'article [9], consulté le 19/11/2018.

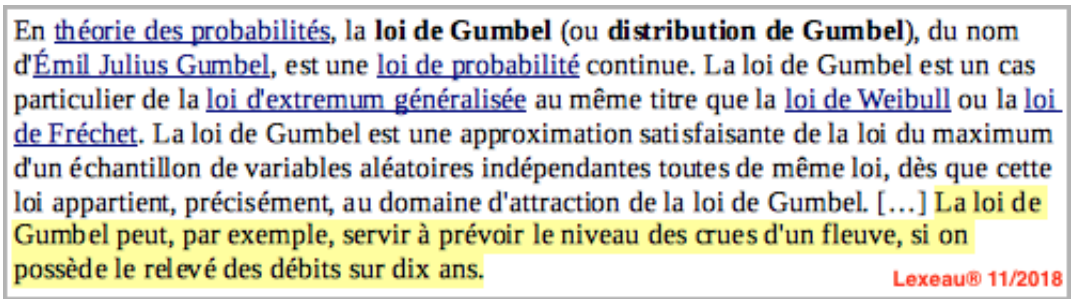


FIGURE 2.69 – Extrait d'un article de Wikipedia consacré à la loi de Gumbel

La question qui doit être posée est celle de la durée de retour de cette prévision des crues. 10 ans, 100 ans, 1000 ans ? Une autre question à poser aux climatologues est de savoir de quelle manière et avec quelle amplitude l'application de cette loi à des relevés de débit sur dix ans est biaisée par le processus du réchauffement climatique. Peut-on soutenir que la durée de retour d'une crue d'une ampleur déterminée évaluée de la sorte n'est pas sous évaluée ? De combien ? On voit que la mise à jour lexicale et conceptuelle ne peut passer à côté des questions que peuvent se poser les non spécialises de la prévision des crues sur la pratique des ingénieurs et de scientifiques d'aujourd'hui en la matière.

2.7 Conclusion du chapitre 2

Ce chapitre a montré comment le projet LEXEAU s'inscrit dans une linguistique de corpus et utilise les méthodes de la textométrie pour faire émerger des substantifs récurrents d'un corpus de textes, après un contrôle préalable de la pertinence des documents sources et du contenu des fichiers, purgé des éléments hors textes ou récurrents et de l'habillage du texte dans les documents source. Il a montré comment les noms propres sont intégrés dans le lexique des noms communs à partir des entités typées du domaine et comment effectuer sur un corpus de textes scientifiques une première sélection de substantifs et de noms propres, à la rencontre du contenu du lexique pour faire ressortir les unités potentielles du lexique, dont la mise à jour passe par celle de l'ontologie du domaine. La question du périmètre cognitif des définitions a été illustrée sur un exemple, et celle de l'application des « lois de probabilité » dans le contexte déterministe des climatologues. Ces questions soulignent l'enjeu de la pertinence du lexique dans ce siècle des *fake news*. Le chapitre 3 va élargir aux syntagmes et aux cooccurrences la rencontre des textes et du lexique.

9. https ://fr.wikipedia.org/wiki/Loi_de_Gumbel

Chapitre 3

Des syntagmes et des cooccurrences

Ce chapitre traite de la sélection des « agencements de mots » en vue de leur intégration dans le lexique. Les sections 3.1 et 3.2 sont consacrées à l'extraction et à la sélection de suites de noms et de substantifs qualifiés, sans présumer de la base nominale des syntagmes. Cette recherche est faite en anglais, où les syntagmes nominaux (*nominal phrases*) sont des suites de noms, ce qui n'est pas le cas en français, ou des noms précédés d'un adjectif. Les sélections sont effectuées sur un corpus et ses sous-corpus avec des seuils de fréquence appropriés. La méthode vise à optimiser la pertinence des résultats en faisant jouer la spécificité des sous-corpus. Elle peut être appliquée sur de très gros corpus. La section suivante **3.3** poursuit sur un exemple la recherche au niveau de chaque texte, à partir de syntagmes « ciblés » dont la base nominale est connue. La section 3.4 qui clôt le chapitre revient sur la question des collocations et traite sur un exemple de la question des noms propres en collocation.

3.1 Extraction et sélection primaire des suites de noms

La recherche et l'intégration dans le lexique de suites de noms est effectuée pour l'anglais sur des suites de 2 à 4 noms, compte tenu de la possibilité de construire dans cette langue des syntagmes nominaux où les noms qui précèdent le dernier nom dans la suite, ou « base nominale » du syntagme, ont une fonction adjectivale. Les scientifiques utilisent cette faculté de la langue anglaise, sans équivalent en français, pour nommer les objets de leurs recherches. La recherche textométrique débouche sur des suites classées par fréquence décroissante, en utilisant un seuil de fréquence approprié pour que le nombre de suites ne dépasse pas un certain plafond. Nous avons fixé ce plafond à 20 dans la recherche des suites de noms dans le corpus FLR-02 et les 4 textes qui le composent. Les suites obtenues sur les 5 corpus sont regroupées pour obtenir une liste agrégée de suites. Les noms de ces suites sont extraits dans une liste des noms pour construire la requête qui débouche sur une sélection finale de syntagmes. Ces syntagmes sont partagés par 2, 3 ou 4 textes ou figurent seulement dans un texte (syntagmes « spécifiques »). À l'issue de la recherche pour chaque taille de syntagme, les syntagmes qui s'avèrent être des artefacts de la textométrie sont éliminés « à vue », en éditant les lignes de concordances et si nécessaire des extraits surlignés du texte source.

3.1.1 Suites de deux noms

Le recherche des suites « nom+nom» sur la partition FLR-02 et sur chacun des 4 textes avec un seuil de fréquence approprié a permis d'obtenir le tableau de synthèse du résultat des 5 requêtes. Les suites de noms sélectionnées sont classées dans l'orde alphabétique dans la figure 3.1.

FLR-02 (fmin=10)	
enlemma	F
climate change	46
copula function	11
dependence parameter	12
dependence structure	17
distribution function	13
exceedance probability	17
grid point	36
level height	43
level rise	13
return period	45
sea level	83
shape parameter	10
storm surge	44
study area	12
surge height	10
tide gauge	10
time window	10
water level	57
wave height	69

FLR-02003 (fmin=3)	
enlemma	F
catchment area	3
climate change	16
climate condition	4
flood event	5
hazard model	4
impact model	8
rain accumulation	5
rainfall accumulation	4
rainfall datum	4
return period	6
river overflow	3
surface run-off	4

FLR-02001 (fmin=3)	
enlemma	F
climate change	7
climate datum	3
flood hazard	7
hazard map	6
level rise	7
return period	15
sea level	15
storm surge	18
surge height	10
surge level	7
tide gauge	8
time period	7
water level	29
wind speed	4

FLR-02004 (fmin=3)	
enlemma	F
adaptation measure	5
change scenario	5
climate change	12
crest level	4
defence structure	3
level condition	4
level rise	6
look-up table "	3
railway line	5
return period	17
sea level	21
storm surge	3
water level	9
wave condition	3
wave height	3
wave transformation	3

FLR-02002 (fmin=6)	
enlemma	F
climate change	11
copula function	11
dependence parameter	12
dependence structure	17
design event	9
distribution function	13
exceedance probability	17
grid point	35
height datum	9
level height	43
probability isoline	6
return period	7
sea level	47
shape parameter	9
storm surge	21
study area	9
time window	10
water level	19
wave height	66

Lexeau® 09/2018

FIGURE 3.1 – Sélection des suites « nom+nom » dans FLR-02 et dans chacun des textes

Les 53 suites agrégées de la figure 3.1 sont présentées dans les cellules du tableau de la figure 3.2. Les 71 noms tirés des 53 suites de la figure 3.2 sont présentés par ordre alphabétique dans la liste de la figure 3.3. débouchent sur les résultats de la recherche des suites « nom+nom » par combinaison deux à deux des noms de la liste, soient 110 suites (fig. 3.4 et 3.5).

adaptation measure	air pressure	catchment area	change scenario	climate change
climate condition	climate datum	climate reanalysis	compensation scheme	copula function
crest level	defence structure	dependence parameter	design event	distribution function
emission scenario	exceedance probability	flood event	flood hazard	grid point
hazard map	hazard model	height datum	impact model	interpolation technique
level condition	level height	level rise	look-up table	probability isoline
railway line	rain accumulation	rainfall datum	return period	revetment slope
river overflow	sea level	shape parameter	storm surge	study area
surface run-off	surge height	surge level	ten-year threshold	tide gauge
time period	time window	water level	wave characteristic	wave condition
wave height	wave transformation	wind speed		

Lexeau® 09/2018

FIGURE 3.2 – Les 53 suites « nom+nom » sélectionnées dans les textes et le corpus FLR-02

accumulation adaptation air area catchment change characteristic climate compensation condition copula crest datum defence dependence design distribution emission event exceedance flood function gauge grid hazard height impact interpolation isoline level line look-up map measure model overflow parameter period point pressure probability railway rain rainfall reanalysis return revetment rise river run-off scenario scheme sea shape slope speed storm structure study surface surge table technique ten-year threshold tide time transformation water wave wind window

Lexeau® 09/2018

FIGURE 3.3 – Les 71 noms extraits des 53 suites « nom+nom » sélectionnées

Partition FLR-02 par texte	T	02001	02002	02003	02004	Nb de
25 « nom+nom » partagés	22 317	6 612	7 550	4 166	3 989	textes
	F					
change impact	2	1	0	0	1	2
change scenario	6	1	0	0	5	2
climate change	46	7	11	16	12	4
climate condition	5	0	1	4	0	2
climate model	4	2	1	1	0	3
emission scenario	6	0	1	0	5	2
flood event	6	0	1	5	0	2
flood hazard	8	7	0	1	0	2
grid point	36	1	35	0	0	2
level datum	3	1	0	0	2	2
level rise	13	7	0	0	6	2
probability distribution	5	2	3	0	0	2
return period	45	15	7	6	17	4
river flood	3	2	0	1	0	2
sea level	83	15	47	0	21	3
shape parameter	10	1	9	0	0	2
storm surge	44	18	21	2	3	4
study area	12	1	9	0	2	3
surge level	9	7	1	0	1	3
tide gauge	11	9	0	0	2	2
time period	9	7	1	0	1	3
water level	57	29	19	0	9	3
wave height	70	0	67	0	3	2
wave model	3	0	1	0	2	2
wave period	5	0	4	0	1	2
Total des fréquences	501	133	239	36	93	

Lexeau® 10/2018

FIGURE 3.4 – Les 25 suites de deux noms partagées entre des textes de FLR-02

Dans les tableaux bruts de résultats, les fréquences sont présentées sur la même ligne pour le corpus et les textes, avec le nombre de textes concernés par chaque suite, ce qui a permis d'isoler les 85 syntagmes « spécifiques » sur tableur et de faire la synthèse des fréquences obtenues dans la figure 3.5. Après un contrôle des concordances, les syntagmes « artificiels » issus de la textométrie ont été surlignés en rouge dans les tableaux des figures 3.4 et 3.5.

Texte 02001		Texte 02002		Texte 02003		Texte 02004	
air pressure	5	copula function	12	catchment area	3	adaptation measure	5
climate datum	3	copula parameter	2	climate scenario	1	crest height	1
climate reanalysis	3	dependence function	1	compensation scheme	4	crest level	4
flood map	1	dependence parameter	12	hazard model	4	defence crest	1
gauge datum	1	dependence structure	17	impact model	8	defence level	1
hazard level	1	design event	9	overflow model	2	defence line	1
hazard map	6	design water	1	rain accumulation	5	defence structure	3
hazard study	2	distribution function	13	rainfall accumulation	4	design point	1
level return	1	exceedance probability	17	rainfall datum	4	flood water	1
pressure datum	1	height datum	9	rainfall event	1	interpolation technique	3
surge height	10	height event	1	river overflow	3	level change	1
surge return	1	level height	43	run-off model	2	level condition	4
tide height	1	model grid	1	sea surface	1	look-up table	5
tide level	1	parameter copula	1	surface area	1	period event	1
tide model	2	parameter distribution	1	surface run-off	4	period scenario	2
wind speed	4	probability isoline	6	ten-year threshold	3	railway line	5
		return level	2			revetment slope	3
		sea storm	1			sea defence	1
		surge climate	2			sea water	1
		surge event	1			storm event	1
		surge model	1			surge rise	1
		surge parameter	1			transformation model	1
		time window	10			wave characteristic	3
		wave climate	5			wave condition	3
		wave datum	1			wave impact	1
		wave event	4			wave transformation	3
						wind datum	2
						Lexeau® 10/2018	

FIGURE 3.5 – Les 85 suites spécifiques de deux noms dans les textes de FLR-02

3.1.2 Suites de trois noms

La recherche des suites de type « nom+nom+nom » sur le même corpus a débouché sur le tableau de synthèse de la figure **3.6**.

Les 37 suites de 3 noms sélectionnées sont présentées dans des cellules distinctes recopiées en ligne dans la figure 3.7. Comme dans le cas des suites de deux noms, un travail combiné sur tableur et en traitement de texte permet d'établir la liste des noms qui composent ces 37 suites avec des séparateurs entre les noms (fig. 3.8). Cette liste est utilisée pour rechercher les suites obtenues par composition 3 par 3 des noms qu'elle contient. Les résultats de la requête lancée sur la partition du corpus FLR-02 débouche sur 58 suites présentes dans les textes. Ils sont présentés dans les figures 3.9 et 3.10, avec 4 suites partagées et les 54 suites spécifiques des 4 textes. Les suites « artificielles » issues du traitement par textométrie sont surlignées en rouge. La figure 3.10 montre la variété des syntagmes propres aux 4 textes

FLR-02 (fmin=3)		FLR-02001 (fmin=2)		FLR-02002 (fmin=2)	
enlemma	F	enlemma	F	enlemma	F
climate change scenario	6	flood hazard map	4	climate change effect	2
exceedance probability estimate	3	sea level rise	7	exceedance probability estimate	3
exceedance probability isoline	4	storm surge height	3	exceedance probability isoline	4
flood hazard map	4	tide gauge measurement	2	level height datum	6
level height datum	6			order polynomial trend	2
sea level condition	4			scale parameter variation	2
sea level height	42			sea level height	42
sea level rise	12			storm surge climate	2
storm surge height	3			storm surge extreme	2
water level estimate	4			water level estimate	3
wave height datum	3			wave height datum	3
wave height estimate	4			wave height estimate	4
wave height maxima	3			wave height extreme	2
FLR-02003 (fmin=1)		FLR-02004 (fmin=2)		wave height marginals	2
enlemma	F	enlemma	F	wave height maxima	3
clay soil content	1	climate change scenario	5	wave height quantiles	2
drought impact model	2	return period scenario	2		
flood hazard model	1	rock revetment slope	2		
greenhouse gas concentration	1	sea level condition	4		
greenhouse gas emission	1	sea level rise	5		
insurance impact model	1				
moisture variation range	1				
return period loss	1				
river flood zone	1				
river overflow model	2				
sea surface temperature	1				
soil moisture variation	2				
surface run-off model	1				

Lexeau® 09/2018

FIGURE 3.6 – Sélection des suites de trois noms dans FLR-02 et dans les 4 textes

clay soil content	climate change effect	climate change scenario	drought impact model
exceedance probability estimate	exceedance probability isoline	flood hazard map	flood hazard model
greenhouse gas concentration	greenhouse gas emission	insurance impact model	level height datum
moisture variation range	order polynomial trend	return period loss	return period scenario
river flood zone	river overflow model	rock revetment slope	scale parameter variation
sea level condition	sea level height	sea level rise	sea surface temperature
soil moisture variation	storm surge climate	storm surge extreme	storm surge height
surface run-off model	tide gauge measurement	water level estimate	wave height datum
wave height estimate	wave height extreme	wave height marginals	wave height maxima
wave height quantiles			

Lexeau® 09/2018

FIGURE 3.7 – Les 37 suites de 3 noms sélectionnées dans FLR-02 et dans les 4 textes

.

change|clay|climate|concentration|condition|content|datum|drought|effect|emission|estimate|
exceedance|extreme|flood|gas|gauge|greenhouse|hazard|height|impact|insurance|isoline|level|
loss|map|marginals|maxima|measurement|model|moisture|order|overflow|parameter|period|
polynomial|probability|quantiles|range|return|revetment|rise|river|rock|run-off|scale|scenario|sea|
slope|soil|storm|surface|surge|temperature|tide|trend|variation|water|wave|zone

Lexeau® 09/2018

FIGURE 3.8 – Liste des 59 noms, avec séparateurs, des 37 suites de 3 noms sélectionnées

Partition FLR-02 par texte 4 suites de 3 noms partagées	FLR-02 F	02001	02002	02004	Nb de textes
climate change impact	2	1	0	1	2
climate change scenario	6	1	0	5	2
sea level rise	12	7	0	5	2
water level estimate	4	1	3	0	2
					Lexeau® 10/2018

FIGURE 3.9 – Les 4 suites « nom+nom+nom » partagées entre les textes de FLR-02

Texte 02001		Texte 02002		Texte 02003		Texte 02004	
flood hazard map	4	climate change effect	2	clay soil content	1	return period scenario	2
level return period	1	exceedance probability estimate	3	drought impact model	2	rock revetment slope	2
river flood hazard	1	exceedance probability isoline	4	flood hazard model	1	sea level condition	4
river flood map	1	level height datum	6	greenhouse gas concentration	1	sea level datum	1
storm surge height	3	level height extreme	1	greenhouse gas emission	1	sea water level	1
storm surge return	1	level height range	1	insurance impact model	1	storm surge rise	1
surge return period	1	maxima wave height	1	moisture variation range	1	surge level rise	1
tide gauge measurement	2	order polynomial trend	2	return period loss	1		
water level datum	1	order polynomial variation	1	river flood zone	1		
water level return	1	scale parameter variation	2	river overflow model	2		
		sea level height	42	sea surface temperature	1		
		storm surge climate	2	soil moisture variation	2		
		storm surge extreme	2	surface run-off model	1		
		storm surge level	1				
		storm surge model	1				
		storm surge parameter	1				
		water level range	1				
		wave height datum	3				
		wave height estimate	4				
		wave height extreme	2				
		wave height marginals	2				
		wave height maxima	3				
		wave height quantiles	2				
		wave period quantiles	1				
						Lexeau® 10/2018	

FIGURE 3.10 – Les 54 suites « nom+nom+nom » spécifiques des textes de FLR-02

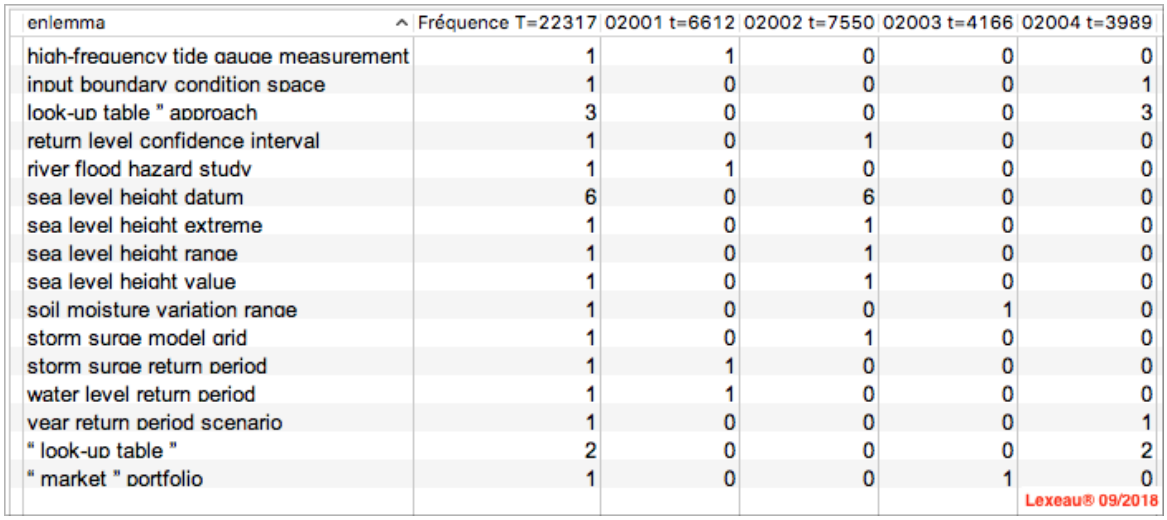
3.1.3 Suites de quatre noms

Les suites de 4 noms ont été recherchées avec la même méthode.

Les premier résultats obtenus sur la partition du corpus FLR-02, présentés dans la figure 3.11, font apparaitre une difficulté, avec la sélection de guillemets incomplets vus comme des lemmes par le lemmatiseur de TXM (TreeTagger). Ils ont été éliminés des requêtes lancées sur chacun des 4 textes pour construire le tableau de synthèse de la figure ce qui ramène les résultats de la recherche à 15 suites de 4 noms, avec deux syntagmes entre de « vrais » guillemets.

Les suites de la figure 3.11 sont toutes spécifiques, avec un seuil de fréquence de 1, ce qui permet de considérer le tableau de la figure 3.12 comme un tableau final, sans avoir à recomposer des suites de 4 noms comme dans les suites de taille plus réduite. Nous n'avons pas trouvé de syntagme « artificiel » dans ces suites de 4 noms.

Cette recherche « en aveugle » des syntagmes nominaux reste conditionnée par le choix des seuils de fréquence, sauf pour les suites de 4 noms.



enlemma	Fréquence T=22317	02001 t=6612	02002 t=7550	02003 t=4166	02004 t=3989
high-frequency tide gauge measurement	1	1	0	0	0
input boundary condition space	1	0	0	0	1
look-up table " approach	3	0	0	0	3
return level confidence interval	1	0	1	0	0
river flood hazard study	1	1	0	0	0
sea level height datum	6	0	6	0	0
sea level height extreme	1	0	1	0	0
sea level height range	1	0	1	0	0
sea level height value	1	0	1	0	0
soil moisture variation range	1	0	0	1	0
storm surge model grid	1	0	1	0	0
storm surge return period	1	1	0	0	0
water level return period	1	1	0	0	0
year return period scenario	1	0	0	0	1
" look-up table "	2	0	0	0	2
" market " portfolio	1	0	0	1	0

FIGURE 3.11 – Recherche des suites de 4 noms dans la partition FLR-02

Texte 02001		Texte 02002	
enlemma	F	enlemma	F
high-frequency tide gauge measurement	1	return level confidence interval	1
river flood hazard study	1	sea level height datum	6
storm surge return period	1	sea level height extreme	1
water level return period	1	sea level height range	1
		sea level height value	1
		storm surge model grid	1
Texte 02003		Texte 02004	
enlemma	F	enlemma	F
soil moisture variation range	1	input boundary condition space	1
" market " portfolio	1	year return period scenario	1
		" look-up table "	2

FIGURE 3.12 – Suites spécifiques de 4 noms des textes du corpus FLR-02

L'interprétation des résultats de cette section passe par un travail de regroupement des syntagmes de toute taille sur leur base nominale, dont il est possible d'établir la liste « à la main ». Nous avons abordé ce travail dans une recherche « a priori » des syntagmes nominaux construits sur une base nominale connue, présentée dans la section **3.3**.

Avant d'y venir, nous présentons dans la section 3.2 les résultats de la recherche « en aveugle » des substantifs qualifiés où l'adjectif précède immédiatement la base nominale ou en est séparée par un ou plusieurs noms.

Ceci permettra d'établir une liste complète des bases nominales obtenues « en aveugle » sur l'ensemble des syntagmes nominaux recherchés, avec ou sans adjectif, et de rapprocher cette liste des entrées du lexique et de la sélection de substantifs comme unité lexicales potentielles du lexique élaborée dens la section 2.5 (cf. fig. 2.54).

3.2 Extraction et sélection primaires des substantifs qualifiés

Les substantifs qualifiés que nous avons choisi d'extraire des textes en anglais sont les syntagmes nominaux du type « adjectif+nom », « adjectif+nom+nom » et « adjectif+nom+nom+nom ». Les résultats obtenus sont présentés pour chaque type de syntagme. Les adjectifs comparatifs et superlatifs sont inclus dans la requête.

3.2.1 Syntagmes nominaux du type « adjectif+nom »

La même requête a été lancée sur chacun des quatre textes du corpus avec des seuils de fréquence adaptés. Les résultats de ces 5 requêtes sont regroupés dans la figure 3.13.

Corpus FLR-02 (fmin=7)	
17 syntagmes	F
21st century	10
annual maxima	11
coastal area	12
coastal frontage	7
extreme value	17
extreme water	8
extreme wave	11
future climate	11
joint exceedance	16
joint probability	10
marginal distribution	13
monthly maxima	7
polynomial trend	7
present work	13
second half	8
significant wave	7
simulated year	10

Texte 02001 (fmin=3)	
21 syntagmes	F
100-year level	3
100-year surge	5
annual maxima	4
coastal flood	3
coastal segment	3
daily maxima	4
delft3d model	3
different part	3
extreme surge	4
extreme water	6
future climate	5
good accuracy	3
harmonic analysis	3
high tide	3
hydrodynamic model	5
individual station	3
isostatic adjustment	5
mean sea	3
monthly maxima	6
skew surge	6
tidal amplitude	5

Texte 02002 (fmin=5)	
17 syntagmes	F
21st century	8
annual maxima	7
archimedean copula	6
coastal area	12
extreme value	13
extreme wave	11
first half	5
fourth order	5
joint exceedance	16
joint probability	8
linear trend	5
marginal distribution	13
polynomial trend	7
present work	13
second half	7
selected area	5
selected profile	6

Texte 02003 (fmin=3)	
13 syntagmes	F
annual loss	3
climatic condition	4
current climate	3
current study	4
different scale	3
extreme event	4
financial impact	4
future climate	3
hourly rainfall	3
hydrographic region	3
long return	3
natural disaster	6
simulated year	10

Texte 02004 (fmin=3)	
9 syntagmes	F
coastal frontage	7
offshore wave	4
different return	3
extreme value	3
large number	3
nearshore point	3
significant wave	3
sloping structure	3
vertical wall	3

Lexeau® 10/2018

FIGURE 3.13 – Syntagmes « adjectif+nom » les plus fréquents dans FLR-02 et ses textes

Le sélection des adjectifs issue de cette première recherche de substantifs qualifiés permet

d'établir une requête présentée dans la figure 3.14. Le lancement de cette requête sur la parti-

```
[enpos="jj|jjr|jjs" & enlemma="100-year|21st|annual|archimedean|climatic|
coastal|current|daily|delft3d|different|extreme|financial|first|fourth|future|good|
harmonic|high|tide|hourly|hydrodynamic|hydrographic|individual|isostatic|
joint|large|linear|long|marginal|mean|monthly|natural|nearshore|offshore|
polynomial|present|second|selected|significant|simulated|skew|sloping|tidal|
vertical"] [enpos="nn|nns" & enlemma=".*"]                    Lexeau® 10/2018
```

FIGURE 3.14 – Requête sur les syntagmes « adjectif+nom » avec les adjectifs de la figure 3.13

tion par texte du corpus FLR-02 avec un seuil de fréquence de 3 permet de récupérer tous le résultats du tableau de la figure 3.13 et d'obtenir d'autres syntagmes de fréquence inférieure aux seuils de fréquence du tableau, soient 110 syntagmes au total. Les résultats sont présentés dans les tableaux des syntagmes partagés et spécifiques des figures 3.15 et 3.16.

Adjectif+nom	Textes du corpus FLR-02				Nb de
24 syntagmes partagés	02001	02002	02003	02004	textes
21st century	0	8	2	0	2
annual maxima	4	7	0	0	2
coastal defence	1	0	0	2	2
coastal flood	3	0	0	1	2
current study	0	0	4	0	1
daily maxima	4	0	0	0	1
different return	2	0	0	3	2
extreme event	0	1	4	0	2
extreme sea	1	0	0	2	2
extreme storm	1	3	0	0	2
extreme value	0	13	1	3	3
extreme water	6	1	0	1	3
extreme weather	2	0	0	1	2
financial impact	0	0	4	0	1
first half	1	5	0	0	2
future climate	5	2	3	1	4
high water	1	2	0	2	3
joint probability	0	8	0	2	2
large flood	1	0	2	0	2
large number	1	0	0	3	2
mean sea	3	2	0	1	3
monthly maxima	6	1	0	0	2
second half	1	7	0	0	2
significant wave	0	4	0	3	2
				Lexeau® 11/2018	

FIGURE 3.15 – Syntagmes « adjectif+nom » partagés par les textes (adjectifs fig. 3.13)

Texte 02001	
14 syntagmes spécifiques	F
100-year level	3
100-year surge	5
coastal segment	3
delft3d model	3
different part	3
extreme surge	4
good accuracy	3
harmonic analysis	3
high tide	3
hydrodynamic model	5
individual station	3
isostatic adjustment	5
skew surge	6
tidal amplitude	5

Texte 02002	
21 syntagmes spécifiques	F
archimedean copula	6
coastal area	12
extreme wave	11
fourth order	5
high estimate	4
joint exceedance	16
joint return	4
linear trend	5
marginal distribution	13
marginal estimate	3
marginal parameter	4
marginal wave	4
polynomial trend	7
present work	13
second order	3
selected area	5
selected grid	3
selected location	3
selected profile	6
significant increase	3
significant trend	3

Texte 02003	
9 syntagmes spécifiques	F
annual loss	3
climatic condition	4
current climate	3
different scale	3
hourly rainfall	3
hydrographic region	3
long return	3
natural disaster	6
simulated year	10

Texte 02004	
5 syntagmes spécifiques	F
coastal frontage	7
offshore wave	4
nearshore point	3
sloping structure	3
vertical wall	3

Lexeau® 10/2018

FIGURE 3.16 – Syntagmes « adjectif+nom » spécifiques (textes FLR-02, adjectifs fig. 3.13)

Les syntagmes artificiels ont été détectés en lisant à l'écran les concordances des syntagmes et si nécessaire l'extrait du texte source, sans quitter l'application TXM. Ils apparaissent sur fond rouge dans les figures 3.15 et 3.16.

Ces syntagmes sont des formes tronquées, non anaphoriques, de syntagmes plus développés du type « adjectif+nom+nom » ou « adjectif+nom+nom=nom ». Leur repérage se fait « à vue » par reconnaissance du syntagme englobant, ce qui requiert une pratique suffisante de ces constructions courantes en anglais technique et scientifique et une compréhension *a minima* du thème traité dans le texte.

Les fréquences les plus élevées des syntagmes partagés apparaissent sur fond vert dans la figure 3.15, ce qui permet de visualiser le « profil » des textes qui les emploient le plus.

3.2.2 Substantifs qualifiés du type « adjectif + nom + nom »

Les syntagmes du type « adjectif + nom + nom » correspondent à des substantifs qualifiés par un adjectif et par un nom qui suit l'adjectif. La recherche et la sélection de ces syntagmes s'effectuent de façon analogue à celles des syntagmes du type « adjectif + nom » en recherchant les syntagmes « les plus fréquents » dans le corpus FLR-02 et dans chacun de ses textes sur la base d'un seuil de fréquence. Les résultats sont présentés dans la figure 3.17. Les adjectifs des syntagmes sélectionnés, qualifiés de « récurrents »(fig. 3.18), vont permettre de construire la requête de recherche finale des syntagmes « adjectif+nom+nom » qui les contiennent, présentée figure 3.19.

Corpus FLR-02 (fmin=4)	
11 syntagmes	F
different return period	5
extreme storm surge	4
extreme value analysis	4
extreme water level	8
extreme wave height	7
joint exceedance probability	16
joint return period	4
likely design event	4
marginal wave height	4
mean sea level	6
significant wave height	5

Texte 02001 (fmin=2)	
5 syntagmes	F
different return period	2
euro-cordex climate simulation	2
extreme water level	6
mean sea level	3
regional frequency analysis	2

Texte 02002 (fmin=3)	
15 syntagmes	F
appropriate copula function	3
extreme storm surge	3
extreme value analysis	3
extreme value model	3
extreme wave height	7
fourth order polynomial	3
joint exceedance probability	16
joint probability analysis	3
joint return period	4
likely design event	4
likely wave event	3
marginal wave height	4
maximum wave height	3
selected grid point	3
significant wave height	4

Texte 02003 (fmin=2)	
4 syntagmes	F
annual soil moisture	2
future climate condition	2
hourly rainfall datum	2
long return period	3

Texte 02004 (fmin=2)	
6 syntagmes	F
coastal infrastructure resilience	2
different return period	3
extreme sea level	2
low return period	2
medium emission scenario	2
traditional look-up table	2

Lexeau® 10/2018

FIGURE 3.17 – Syntagmes « adjectif+nom+nom » les plus fréquents, corpus et textes FLR-02

appropriate coastal different euro-cordex extreme fourth future hourly joint likely long marginal
maximum mean medium regional selected significant traditional
Lexeau® 10/2019

FIGURE 3.18 – Adjectifs récurrents du corpus FLR-02 et de ses textes (syntagmes figure 3.17)

```
[enpos="jj|jjr|jjs" & enlemma="appropriate|coastal|different|euro-cordex|extreme|fourth|future|hourly|joint|
likely|long|marginal|maximum|mean|medium|regional|selected|significant|traditional"] [enpos="nn|nns &
enlemma="analysis|climate|condition|copula|datum|design|emission|event|exceedance|frequency|function|
grid|height|infrastructure|level|look-up|model|order|period|point|polynomial|probability|rainfall|resilience|
return|scenario|sea|simulation|storm|surge|table|value|water|wave"] [enpos="nn|nns &
enlemma="analysis|climate|condition|copula|datum|design|emission|event|exceedance|frequency|function|
grid|height|infrastructure|level|look-up|model|order|period|point|polynomial|probability|rainfall|resilience|
return|scenario|sea|simulation|storm|surge|table|value|water|wave"]
```
Lexeau® 10/2018

FIGURE 3.19 – Recherche des syntagmes « adjectifs récurrent + nom + nom » (requête)

Les résultats de cette requête sont présentés dans la figure 3.20 pour les syntagmes partagés, avec sur fond vert les occurrences les plus élevées, et dans la figure 3.21 pour les syntagmes spé-

cifiques. Après contrôle des concordances, 3 syntagmes spécifiques du texte 02002 apparaissent comme artificiels et surlignés en rouge.

Syntagmes « adjectif+nom+nom »		Textes du corpus FLR-02				Nb de
9 syntagmes partagés	F	02001	02002	02003	02004	textes
different return period	5	2	0	0	3	2
extreme sea level	3	1	0	0	2	2
extreme storm surge	4	1	3	0	0	2
extreme value analysis	4	0	3	0	1	2
extreme water level	8	6	1	0	1	3
future climate condition	3	0	1	2	0	2
mean sea level	6	3	2	0	1	3
regional frequency analysis	3	2	1	0	0	2
significant wave height	5	0	4	0	1	2
					Lexeau® 10/2018	

FIGURE 3.20 – Syntagmes partagés « adjectif récurrent + nom + nom », textes FLR-02

Texte 02001		Texte 02002	
4 syntagmes spécifiques	F	15 syntagmes spécifiques	F
euro-cordex climate simulation	2	appropriate copula function	3
future surge level	1	different copula function	1
long simulation period	1	extreme storm surge	3
regional sea level	1	extreme wave event	1
Texte 02003		extreme wave height	7
3 syntagmes spécifiques	F	fourth order polynomial	3
extreme rainfall event	3	joint exceedance probability	16
hourly rainfall datum	2	joint probability analysis	3
long return period	1	joint return period	4
Texte 02004		likely design event	4
3 syntagmes spécifiques	F	likely wave event	3
coastal infrastructure resilience	2	marginal wave height	4
medium emission scenario	2	maximum wave height	3
traditional look-up table	2	regional climate model	1
		selected grid point	3
		Lexeau® 10/2018	

FIGURE 3.21 – Syntagmes spécifiques « adjectif récurrent + nom + nom », textes FLR-02

3.2.3 Syntagmes du type « adjectif + nom + nom + nom »

La recherche des syntagmes du type « adjectif + nom + nom + nom » sur le corpus FLR-02 et sur ses 4 textes a produit 52 syntagmes spécifiques dont nous avons supprimé 2 syntagmes comportant des caractères isolés habituellement en couple. Le tableau de la figure 3.22 présente une synthèse des résultats obtenus avec les 4 requêtes. L'étude des concordances conduit à disqualifier deux syntagmes sur fond rouge du texte 02002 car *likely* est toujours précédé par *most* dans ce texte.

Texte 02001		Texte 02002	
5 syntagmes	F	28 syntagmes	F
extreme water level estimate	1	" design event "	1
future climate change impact	1	annual maxima wave height	1
high storm surge frequency	1	associated sea level height	1
regional sea level rise	1	concomitant sea level height	2
skew surge level surgep	1	corresponding sea level height	1
		extreme storm surge climate	1
Texte 02003		extreme storm surge event	1
7 syntagmes	F	extreme value model result	1
" target " year	1	extreme wave height quantiles	1
30-year return period loss	1	fourth order polynomial trend	1
annual soil moisture variation	2	fourth order polynomial variation	1
in-house flood hazard model	1	high water level estimate	1
in-house surface run-off model	1	high wave height quantiles	1
so-called " market "	1	joint exceedance probability estimate	2
uniform clay soil content	1	joint exceedance probability estimation	1
		joint exceedance probability isoline	4
Texte 02004		joint probability distribution function	1
10 syntagmes	F	likely design event method	1
" look-up table "	1	likely wave event range	1
2d wave transformation model	1	marginal wave height estimate	3
coastal flood forecasting system	1	maximum likelihood estimation procedure	1
different climate change scenario	1	maximum wave height datum	1
future climate change impact	1	mean sea level variability	1
high water spring value	1	reliable exceedance probability estimate	1
long term resilience strategy	1	second order polynomial trend	1
mean sea level rise	1	significant wave height estimate	1
total sea level rise	1	simultaneous sea level height	1
traditional look-up table approach	1	univariate t student distribution	1
		Lexeau® 10/2018	

FIGURE 3.22 – Syntagmes spécifiques « adjectif + nom + nom + nom », textes FLR-02

3.3 Sélection finale des unités récurrentes

3.3.1 Des substantifs ou des bases nominales ?

Après la sélection d'unités récurrentes présentée à la fin des sections 2.5, 3.1 et 3.2, le rapprochement des bases des syntagmes nominaux et des substantifs permet d'analyser la construction des syntagmes sur leur base. Cette construction peut amener des syntagmes à prendre la place de leur base dans le lexique. Ce procédé typique des langages spécialisés permet aux spécialistes de détourner avec un syntagme le sens courant de sa base nominale en utilisant un modificateur simple ou composite. Il peut arriver que la base nominale figure aussi dans le lexique en tant que substantif et qu'un syntagme construit sur cette base y figure aussi. La construction de syntagmes en poupées russes peut donner lieu au même procédé de détournement du sens d'un syntagme classé comme base nominale avec un modificateur qui le fait entrer dans le lexique, alors que sa base nominale n'y figure pas.

La figure 3.23 présente le résultat du rapprochement dans l'ordre alphabétique des bases nominales des syntagmes récurrents et des substantifs de la figure 2.54, avec le même code des couleurs, complété par du gris pour les bases nominales absentes de la liste des substan-

tifs.

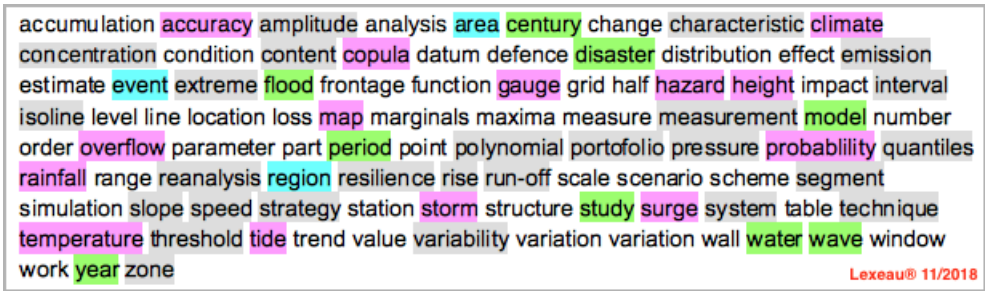


FIGURE 3.23 – Rapprochement des bases nominales et des substantifs sélectionnés

3.3.2 Sélection des syntagmes nominaux par leur base

Jusqu'à présent, la sélection des syntagmes nominaux a été faite *ex nihilo* sur des seuils de fréquence. La sélection ciblée va se faire à partir des bases nominales appariées avec les substantifs issus de la première sélection. La recherche portera sur les syntagmes dont la base est connue, avec un classement ultérieur des syntagmes qui dépend du type de substantif apparié à la base.

Lorsque le substantif apparié à la base figure dans le lexique (couleur verte), il s'agit d'identifier les syntagmes nominaux susceptibles d'être ajoutés au lexique ou d'être transférés comme collocation basée sur une unité lexicale.

Lorsque le substantif apparié figure dans le lexique comme unité discursive alternative (couleur cyan), nous avons les mêmes possibilités de transfert que dans le cas précédent, à ceci près que l'unité lexicale « base » de la collocation est celle à laquelle l'unité discursive alternative est rattachée. La situation est la même pour les syntagmes dont la base est appariée avec une unité lexicale satellite (couleur jaune).

L'appariement de la base nominale avec un substantif classé comme unité récurrente (couleur magenta) permet de lister les syntagmes concernés et de choisir les syntagmes à classer comme récurrents. Le listage permet aussi de voir si le substantif isolé a bien un sens pertinent pour le lexique ou s'il y a lieu de le remplacer par un syntagme plus pertinent, ce qui n'a pas été fait lors de la première sélection. Quel que soit le résultat — invalidation du substantif comme nom récurrent ou confirmation de son statut — l'entrée dans le lexique des syntagmes classés comme récurrents peut être envisagée comme précédemment (collocation ou unité lexicale), au titre d'une mise à jour du lexique et de l'ontologie du domaine.

Dans toutes les situations d'appariement, la recherche des syntagmes nominaux dont la base nominale est connue se fera avec une série de requêtes qui reconstituent les conditions de construction des syntagmes les plus courantes dans la langue considérée. Nous sommes partis des requêtes de recherche « en aveugle » effectuées dans les sections précédentes. L'ajout des noms propres dans la recherche des cooccurrents de la base nominale ouvre d'autres perspectives dont nous reparlerons dans la section et en ajoutant une requête avec un lemme nominal après celui de la base, sans dépasser un total de 4 lemmes successifs par requête, dont un seul adjectif.

La figure 3.24 présente le tableau des 12 requêtes correspondantes. Chaque requête vise un type de syntagme où figure la base nominale. Le type de syntagme recherché est surligné en orange clair. Il est reporté dans le tableau après le numéro de la requête et avant le texte proprement dit, dans le langage CQL des requêtes par indexation de TXM. La base nominale à spécifier avant de lancer la requête figure entre crochets dans ce texte.

FIGURE 3.24 – Requêtes de recherche des syntagmes nominaux à partir de leur base nominale

La requête 1 porte sur la base, c'est à dire sur le substantif apparié. Réalisée sans seuil de fréquence, cette requête permet de rechercher de façon exhaustive, par concordance, des emplois autonomes du substantif dans les textes avec un sens qui lui est propre, et de confirmer ainsi de sa qualité d'unité lexicale potentielle si ces emplois sont attestés. Cette recherche du sens est facilitée par la lecture à l'écran du « voisinage textuel » du substantif dans une des pages du texte créées par le logiciel après l'importation du corpus. Nous définissons le voisinage textuel du substantif, et d'une façon plus générale du « pivot » de la concordance, lequel est surligné dans la page qui s'affiche, comme un passage bien délimité de cette page contenant au moins une phrase et ne dépassant pas un paragraphe. Le « voisinage textuel » correspond au concept d'« extrait de texte » de l'ontologie du discours, comme source d'un texte du lexique, par exemple une définition d'unité lexicale.

Le test de la sélection des syntagmes nominaux sur leur base nominale a été réalisé sur la base *flood*. Le nom figure dans le lexique comme unité lexicale bilingue de la strate discursive technico-scientifique (1-TS Inondation/*Flood*). Les syntagmes nominaux extraits du corpus

sans seuil de fréquence sont présentés figures 3.25 et 3.26, avec les *pos* des lemmes.

FLR-02, partition par texte	T	02001	02002	02003	02004	Nb de	Requête
Syntagmes de la base « flood »	22 317	6 612	7 550	4 166	3 989	textes	TXM
43 syntagmes étiquetés (début)	F						
coastal_jj flood_nn	3	2	0	0	1	2	5
coastal_jj flood_nn forecasting_nn	1	0	0	0	1	1	6
coastal_jj flood_nn hazard_nn	1	1	0	0	0	1	6
coastal_jj flood_nns	1	1	0	0	0	1	5
extreme_jj flood_nns	1	0	0	1	0	1	5
flash_jj flood_nns	1	0	0	1	0	1	5
flash_nn flood_nns	3	1	0	2	0	2	2
flood_nn	37	12	3	18	4	4	1
flood_nn discharge_nn	1	0	1	0	0	1	3
flood_nn event_nn	1	0	0	1	0	1	3
flood_nn event_nns	5	0	1	4	0	2	3
flood_nn extent_nns	1	0	0	0	1	1	3
flood_nn forecasting_nn	1	0	0	0	1	1	3
flood_nn hazard_nn	8	7	0	1	0	2	3
flood_nn increase_nns	1	0	0	1	0	1	3
flood_nn map_nns	1	1	0	0	0	1	3
flood_nn part_nn	1	0	0	1	0	1	3
flood_nn peak_nn	1	0	1	0	0	1	3
flood_nn risk_nn	1	0	0	0	1	1	3
flood_nn water_nn	1	0	0	0	1	1	3
flood_nn zone_nn	1	0	0	1	0	1	3
							Lexeau® 10/2018

FIGURE 3.25 – Syntagmes nominaux étiquetés de la base « flood », corpus FLR-02 (début)

FLR-02, partition par texte	T	02001	02002	02003	02004	Nb de	Requête
Syntagmes de la base « flood »	22 317	6 612	7 550	4 166	3 989	textes	TXM
43 syntagmes étiquetés (fin)	F						
flood_nn zone_nns	1	1	0	0	0	1	3
flood_nns	11	4	0	7	0	2	1
flood_np	4	0	0	3	1	2	1
flood_np hazard_nn	1	0	0	1	0	1	3
flood_np risk_np	1	0	0	0	1	1	3
hamburg_np flood_nn	1	1	0	0	0	1	2
in-house_jj flood_nn	1	0	0	1	0	1	5
in-house_jj flood_nn hazard_nn	1	0	0	1	0	1	6
large_jj flood_nn	1	0	0	1	0	1	5
large_jj flood_nn event_nns	1	0	0	1	0	1	6
large_jj flood_nns	2	1	0	1	0	2	5
major_jj flood_nn	1	0	0	1	0	1	5
major_jj flood_nn event_nns	1	0	0	1	0	1	6
national_np flood_np	1	0	0	0	1	1	2
national_np flood_np risk_np	1	0	0	0	1	1	4
north_np sea_np flood_nn	1	1	0	0	0	1	7
numerous_jj flash_nn flood_nns	1	0	0	1	0	1	10
river_nn flood_nn	3	2	0	1	0	2	2
river_nn flood_nn hazard_nn	1	1	0	0	0	1	4
river_nn flood_nn map_nns	1	1	0	0	0	1	4
river_nn flood_nn zone_nn	1	0	0	1	0	1	4
sea_np flood_nn	1	1	0	0	0	1	2
							Lexeau® 10/2018

FIGURE 3.26 – Syntagmes nominaux étiquetés de la base « flood », corpus FLR-02 (fin)

Le logiciel de textométrie a détecté 8 emplois, qui se recoupent, de la base flood comme nom propre dans des syntagmes nominaux. Ils sont surlignés dans la figure 3.26. Ces emploie s'expliquent pour 3 d'entre eux par la référence dans l'introduction du texte 02004 au titre d'une étude réalisée pour l'Agence de l'Environnement britannique. La figure 3.27 présente l'extrait du texte édité dans TXM. Les 3 autres emplois correspondent à des intertitres du

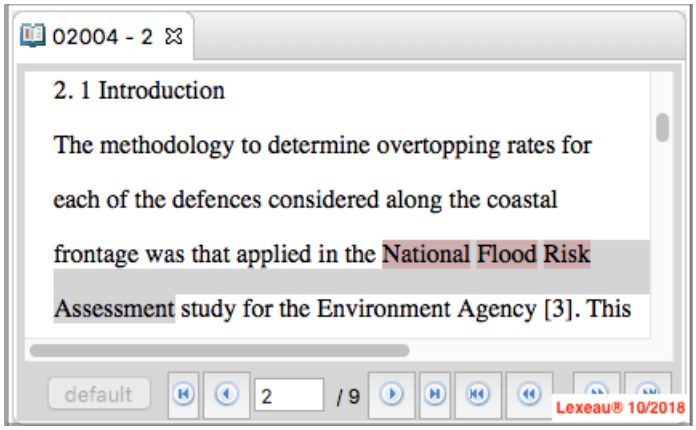


FIGURE 3.27 – Édition de *National Flood Risk Assessment dans le texte source*

texte 02003 (*3. 2 Flood hazard model*, *4. 1. 1 Flood* et *4. 2. 1 Flood*).

L'analyse des résultats des figures 3.25 et 3.26 montre que 9 syntagmes étiquetés sur 43 sont partagés, ce qui souligne, à travers le choix de la base nominale, l'intérêt de la recherche pour caractériser l'approche spécifique des communications présentées à FLOODRISK 2016 sur la question des inondations.

Les syntagmes ont été regroupés sans les étiquettes morphosyntaxiques des lemmes. Ils sont présentés par ordre alphabétique dans le tableau de la figure 3.28, sans les emplois de la base nominale comme nom propre. Les noms sont notés avec une majuscule et soulignés en gras. Les adjectifs sont notés en caractères gras. L'adjectif « flash » figure bien dans une phrase du texte 02003 (*In this study, floods include flash floods and surface runoff as well as riverine overflow.*).

Nous avons surligné en rouge les syntagmes tronqués que sont *Hambourg flood* pour *1962 Hamburg flood* dans le texte 02001, et, dans le même texte, *Sea flood* pour *North Sea flood* ainsi que *North Sea flood* pour *1953 North Sea flood*. Il en est de même pour les syntagmes *in house flood* et *in-house flood hazard*, qui renvoient au syntagme complet *in-house flood hazard model*, dans le texte 02003, et pour le syntagme *river flood*, qui renvoie aux syntagmes complets *river flood hazard* et *river flood map* dans le texte 02001 et *river flood zone*, dans le texte 02003.

Ce premier tri des syntagmes montre que les syntagmes tronqués ne renvoient pas toujours à des syntagmes du tableau faute d'avoir pris en compte d'autres combinaisons de nom et d'adjectifs, voire autres étiquettes morphosyntaxique. Dans notre exemple, il aurait fallu prendre en compte les syntagmes « adjectif + base + nom + nom » et « nombre + nom + base + nom », ce qui ne présente pas de difficulté.

Un deuxième tri des syntagmes construits sur la base est illustré dans la figure 3.28 par les

FLR-02, partition par texte	T	02001	02002	02003	02004	Nb de	Nb de	Requête
Base « flood » (hors nom propre)	22 317	6 612	7 550	4 166	3 989	textes	lemmes	TXM
32 syntagmes « bruts »	F							
flood	48	16	3	25	4	4	1	1
coastal flood	4	3	0	0	1	2	2	5
extreme flood	1	0	0	1	0	1	2	5
flash flood	3	1	0	2	0	2	2	2
flash flood	1	0	0	1	0	1	2	5
flood discharge	1	0	1	0	0	1	2	3
flood event	6	0	1	5	0	2	2	3
flood extent	1	0	0	0	1	1	2	3
flood forecasting	1	0	0	0	1	1	2	3
flood hazard	8	7	0	1	0	2	2	3
flood increase	1	0	0	1	0	1	2	3
flood map	1	1	0	0	0	1	2	3
flood part	1	0	0	1	0	1	2	3
flood peak	1	0	1	0	0	1	2	3
flood water	1	0	0	0	1	1	2	3
flood zone	2	1	0	1	0	2	2	3
Hamburg flood	1	1	0	0	0	1	2	2
in-house flood	1	0	0	1	0	1	2	5
large flood	3	1	0	2	0	2	2	5
major flood	1	0	0	1	0	1	2	5
river flood	3	2	0	1	0	2	2	2
Sea flood	1	1	0	0	0	1	2	2
coastal flood forecasting	1	0	0	0	1	1	3	6
coastal flood hazard	1	1	0	0	0	1	3	6
in-house flood hazard	1	0	0	1	0	1	3	6
large flood event	1	0	0	1	0	1	3	6
major flood event	1	0	0	1	0	1	3	6
North Sea flood	1	1	0	0	0	1	3	7
numerous flash flood	1	0	0	1	0	1	3	10
river flood hazard	1	1	0	0	0	1	3	4
river flood map	1	1	0	0	0	1	3	4
river flood zone	1	0	0	1	0	1	3	4
								Lexeau® 10/2018

FIGURE 3.28 – Syntagmes nominaux sans étiquettes (base *flood* hors noms propres, FLR-02)

couleurs jaune et magenta. La couleur verte est réservée à des syntagmes déjà lexicalisés. Le choix de la couleur correspond au choix de faire entrer certains syntagmes dans le lexique comme unité lexicale (couleur magenta), ou comme collocation insérée dans les textes du lexique (couleur jaune). Les expressions qui ne présentent pas d'intérêt ne sont pas surlignées, en remontant si nécessaire au texte source (fig. 3.29).

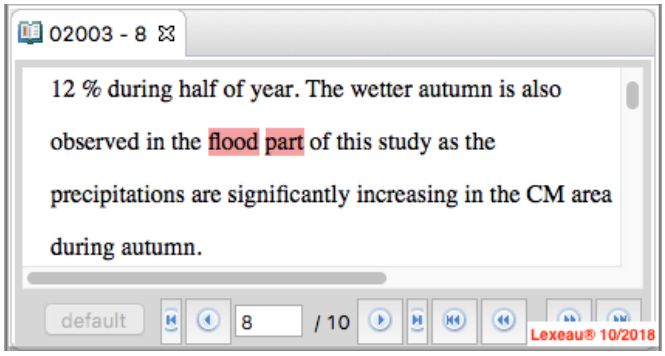


FIGURE 3.29 – Emploi du syntagme *flood part* dans le texte 02003

Le choix des unités lexicales à transférer dans le lexique passe par un examen des unités récurrentes et un choix des éléments de l'ontologie et du lexique à ajouter pour finaliser le transfert. Ce travail doit être effectué dans les deux langues. Pour éviter les erreurs et les ambiguïtés de traduction, il conviendrait de prévoir un contrôle croisé en langue native, ce qui s'applique ici, l'anglais étant une langue seconde pour une bonne partie des auteurs des textes présentés à la conférence FLOODRISK 2016 et pour l'auteur de ces propositions de création de nouvelles entrées.

Le choix d'une strate discursive de rattachement du syntagme intégré comme unité lexicale est nécessaire. Un tel choix n'est pas souhaitable s'il s'agit des collocations car ce serait nier toute possibilité de passage d'une strate discursive à une autre en s'appuyant sur des collocations. Si un rattachement semble nécessaire pour un syntagme considéré comme une collocation, on devra s'interroger sur la possibilité de le transférer dans le lexique comme unité lexicale stratifiée à part entière en changeant son statut.

3.4 Cooccurrences et collocations

L'analyse des syntagmes nominaux récurrents nous a conduit à qualifier de collocation les cooccurrences privilégiées qui ne peuvent être considérés comme des unités lexicales mais qui peuvent être intégrées dans les articles du lexique. Dans notre cas, il s'agit de collocations « nominales ». Ells se présentent comme la combinaison d'une base nominale et d'un collocatif. Les noms propres n'entrent pas dans le champ des collocations, bien qu'ils apparaissent en cooccurrence avec des subtantifs. Les liens avec le lexique sont établis depuis la classe des noms propres lexicalisés (cf. 3.4.1. Il est possible, par contre, d'étendre la notion de collocation nominale à des cooccurrences de plus de deux mots. La figure 3.30 [*à revoir JLJ/17/12/2018*] présente le graphe relationnel des collocations.

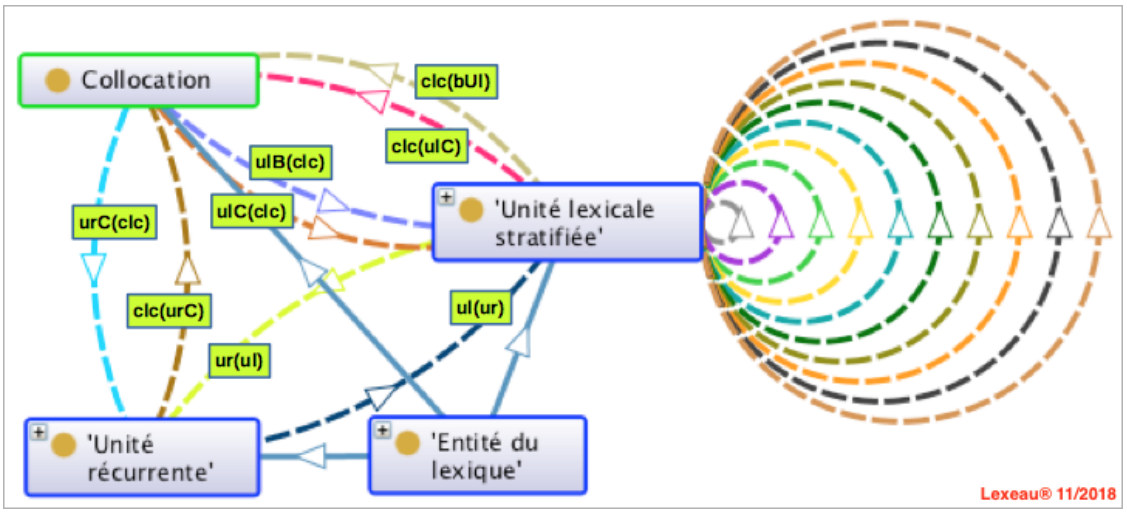


FIGURE 3.30 – Graphe relationnel des collocations

- ulB(clc) : base lexicale de la collocation ;
- clc(bUl) : collocation dont la base est l'unité lexicale ;

- ulC(clc) : unité lexicale employée comme collocatif de la collocation ;
- clc(ulC) : collocation dont l'unité lexicale est le collocatif ;
- urC(clc) : unité récurrente employée comme collocatif de la collocation ;
- clc(urC) : collocation dont l'unité récurrente est le collocatif ;
- ur(ul) : unité récurrente transférée comme unité lexicale ;
- ul(ur) : unité lexicale issue du transfert de l'unité récurrente.

Dans une collocation, le collocatif peut ne pas être une unité lexicale. Il peut alors être une unité récurrente non transférée dans le lexique. Les syntagmes nominaux récurrents « coastal flood forecasting/prévision de submersion côtière » et « *river flood map*/carte d'inondation de la rivière » sont des exemple de collocations. Ces syntagmes ont été transférés dans la classe des collocations depuis la classe des syntagmes récurrents (cf. fig. 2.63 Graphe relationnel des unités récurrentes), avec respectivement comme base nominale les unités lexicales« *1-TS Forecasting*/1-TS Prévision » et « *1-TS Map*/1-TS Carte » et comme collocatifs les syntagmes (« *Nmp Coastal flood*/Snm Submersion côtière » et *Nmp River flood*/Snm Inondation de rivière ». Ces syntagmes récurrents sont transférés dans le lexique dans la strate technico-scientifique car ils correspondent à des classes d'inondations dans l'ontologie du domaine et ne sont pas considérés comme des collocations. L'adectif *coastal*/côtier » n'est pas introduit dans le lexique. Le substantif « *river*/rivière » est transféré de la classe des noms récurrent dans celle des unités discursives alternative, rattaché à l'unité lexicale « *1-TS Watercourse*/Cours d'eau ».

Dans les langages spécialisés, les collocations ne sont pas les formes les plus fréquentes de syntagmes récurrents. À titre d'exemple, il nous a semblé plus approprié de transférer les syntagmes récurrents *Nmp Sea storm* et *Nmp Storm surge* dans le lexique, ce qui permet de les définir et de donner des exemples d'emploi. Le graphe relationnel de la figure 3.31 [*à revoir JLJ/17/12/2018*] illustre ces transferts.

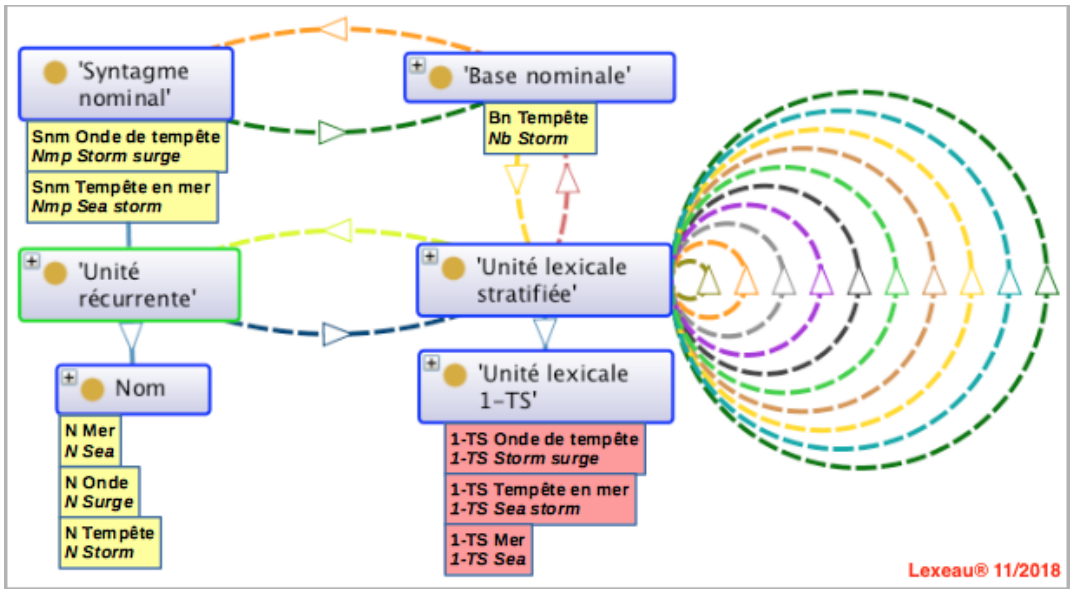


FIGURE 3.31 – Graphe relationnel du transfert de *storm surge* dans le lexique

L'étude de la cooccurrence d'un nom propre et d'une unité lexicale est traitée en dehors des

collocations. La classe des noms propres lexicalisés joue un rôle de classe d'accueil des noms propres et de lien avec le lexique de façon analogue au rôle des collocations nominales pour les noms communs dont ils sont la base. Nous allons voir dans la section 3.4.1 que l'enjeu cognitif peut être très important.

3.4.1 Recherche et identification des modèles numériques

La base nominale lexicalisée « model » permet de retrouver des modèles numériques nommés dans le texte. Ce sont autant de « boîtes noires » dont il faut comprendre le rôle *a minima*, en renvoyant sur des textes de référence, au delà du voisinage textuel du syntagme extrait du texte. Il est possible de retrouver les noms propres qui figurent dans des syntagmes nominaux de la base *model* avec les requêtes de la figure 3.24. Un extrait des syntagmes utiles avec les étiquettes morphosyntaxiques des lemmes est présenté dans la figure 3.32. Les résultats complets de la recherche sont donnés en annexe.

Les éditions du texte source avec TXM dans le voisinage des syntagmes, ou à défaut depuis le fichier pdf de la communication, ont permis de préciser le rôle des modèles mentionnés dans les textes. Il a été presque toujours été possible de trouver sur internet un site de présentation du modèles et parfois un libellé complet de l'acronyme qui le désigne. Les résultats de ces investigations sont présentés dans l'ordre alphabétique des noms des modèles.

Dans le cas d'une mention « adjectivée » du modèle, ce qui est le cas de *delft3d*, nous avons conservé l'adjectif comme désignateur rigide du modèle, c'est à dire comme nom propre. Le voisinage textuel de la première occurrence du syntagme est présenté dans la figure 3.33. Le syntagme récurrent est surligné.

Un extrait de la présentation du modèle Delft3D sur le site de Deltares [1], consulté le 13/10/2018, est présenté figure 3.34. Cet organisme néerlandais est bien connu pour ses recherches appliquées dans le domaine de l'eau et du sous-sol, aux Pays Bas et dans de nombreux pays. Ces indications justifient l'introduction de Delft3D dans le domaine comme nom propre d'une suite de modèles, y compris un modèle de calcul des vagues.

L'édition du voisinage textuel du syntagme étiqueté *agency_ np model_ nn* surligné dans la figure 3.35) donne l'exemple d'un syntagme tronqué. Le syntagme complet n'a pas sa place parmi les noms propres associés au lexique. Il n'a pas été possible de trouver d'autres informations sur le modèle cité entre parenthèses, ni d'accéder sur internet à la publication référencée avec le numéro 18 : *Mott MacDonald (2012) Exe Estuary Mapping & Modelling study. For the Environment Agency.*.

Le voisinage textuel du syntagme ARPEGE-Climat model présenté figure 3.36 nous éclaire sur le modèle développé par le Centre National de Recherches Météorologiques. [2].

Faute d'un voisinage explicatif du syntagme EURO-CORDEX model dans le texte source (FLR-02001), un extrait du guide d'emploi du modèle européen de données de projections du climat, consulté sur internet [3] le 13/10/2018, est présenté figure 3.37.

1. Url : https ://oss.deltares.nl/web/delft3d/about
2. Url : http ://www.umr-cnrm.fr/spip.php ?article124
3. Url : https ://www.euro-cordex.net/imperia/md/content/csc/cordex/euro-cordex-guidelines-version1.0-2017.08.pdf

FLR-02, partition par texte	T	02001	02002	02003	02004	Nb de	Nb de	Numéro
Base « model » : syntagmes nominaux étiquetés	22 317	6 612	7 550	4 166	3 989	textes	lemmes	requête
Extrait : recherche du nom des modèles	F							
model_np	2	0	2	0	0	2	1	1
agency_np model_nn	1	0	0	0	1	2	2	2
arpege–climat_np model_nn	3	0	0	3	0	1	2	2
delft3d_jj model_nn	3	3	0	0	0	1	2	5
euro-cordex_np model_nn	1	1	0	0	0	1	2	2
hydraulic_np model_nns	1	0	0	0	1	1	2	2
isba_np model_nn	1	0	0	1	0	1	2	2
mesoscale_np model_np	1	0	0	1	0	1	2	2
model_nn cfflood_np	1	1	0	0	0	1	2	3
model_nn eu-dem_np	1	1	0	0	0	1	2	3
model_nn swan_np	1	0	1	0	0	1	2	3
pot_np model_nns	1	0	1	0	0	1	2	2
swan_np model_nn	2	0	0	0	2	1	2	2
ccr_np impact_nn model_nns	2	0	0	2	0	1	3	7
ccr's_np impact_nn model_nns	1	0	0	1	0	1	3	7
exe_np estuary_np model_nn	1	0	0	0	1	1	3	7
mesoscale_np model_np version_nn	1	0	0	0	1	1	3	4
swan_np wave_nn model_nn	1	0	0	0	1	1	3	7
tpxo8_np tide_nn model_nn	1	1	0	0	0	1	3	7
in-house_jj flood_nn hazard_nn model_nn	1	0	0	1	0	1	4	12
in-house_jj surface_nn run-off_nn model_nn	1	0	0	1	0	1	4	12
teign_np estuaries_np hydraulic_np model_nns	1	0	0	0	1	1	4	10

Lexeau® 10/2018

FIGURE 3.32 – Noms propres des modèles dans les textes du corpus FLR-02

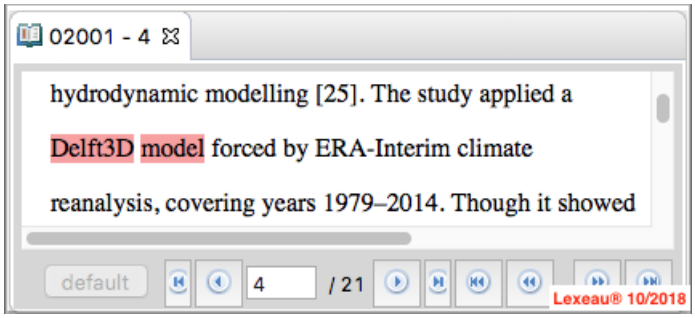


FIGURE 3.33 – Voisinage textuel de la collocation *Delft3D model* (texte FLR-02001)

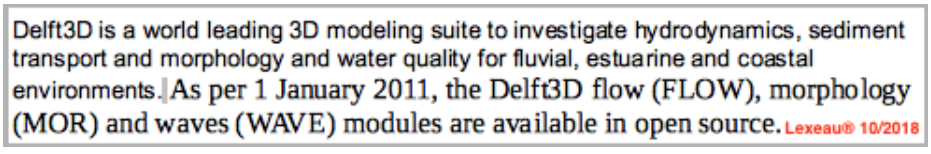


FIGURE 3.34 – Présentation du modèle Delft3D sur le site de Deltarès

Le syntagme à retenir parmi les trois syntagmes de la figure 3.32 qui contiennent Hydraulic model est a priori le plus long (*Teign Estuaries Hydraulic model*), avec une suite de 4 mots. Son voisinage textuel est présenté figure 3.38. Le pluriel de *model* semble indiquer que l'estuaire de la rivière *Teign*, *Teign Estuary*, a fait l'objet d'un modèle hydraulique utilisé aussi pour

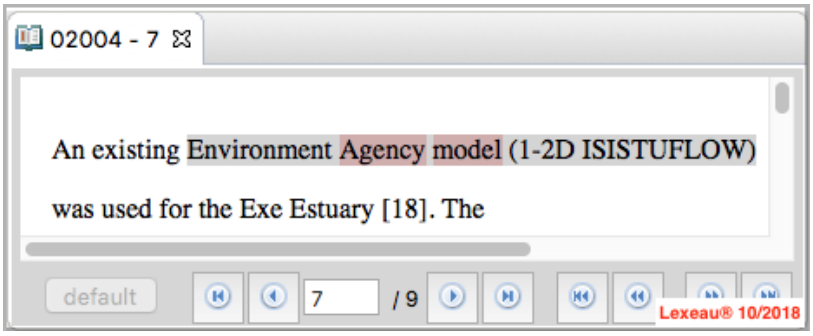


FIGURE 3.35 – Voisinage textuel du syntagme *Agency model* (texte 02004)

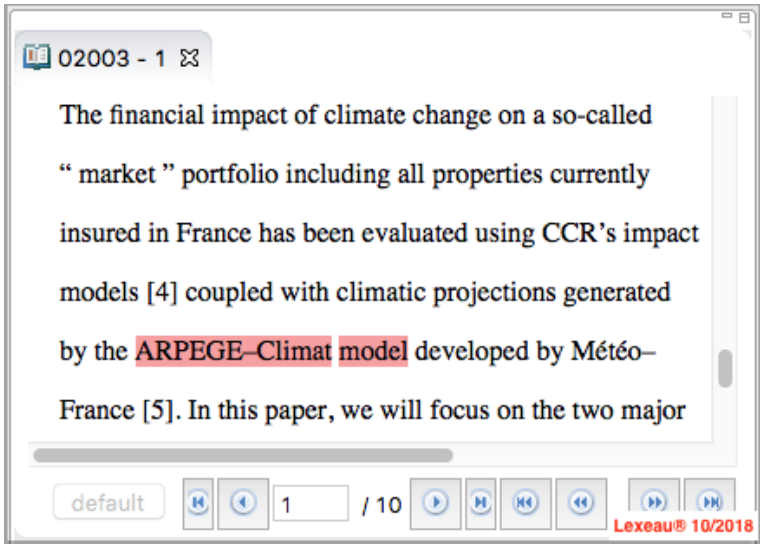


FIGURE 3.36 – Voisinage textuel du syntagme *ARPEGE-Climat model* (texte 02003)

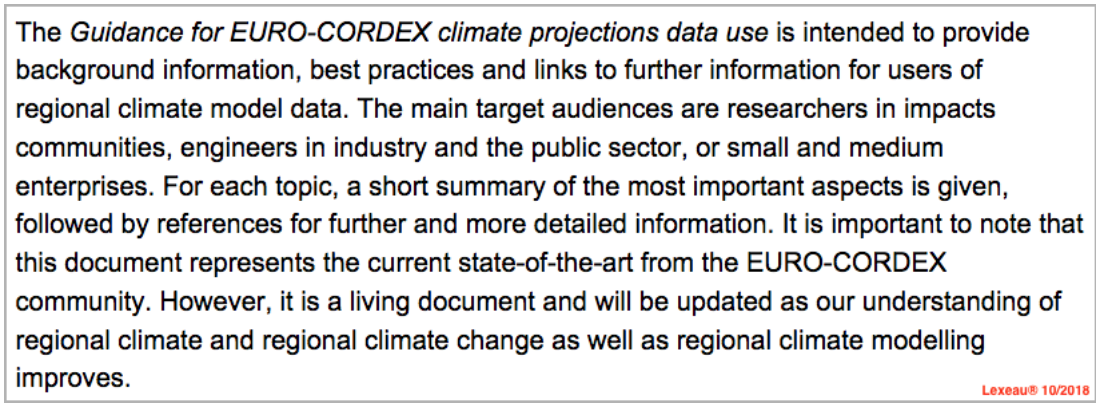


FIGURE 3.37 – Guide d’utilisation des données d’EURO-CORDEX (extrait)

l’estuaire de la rivière *Exe*. Le nom du modèle n’est pas précisé.

Le voisinage textuel du syntagme *Isba model* est présenté figure 3.39. Le texte de la figure

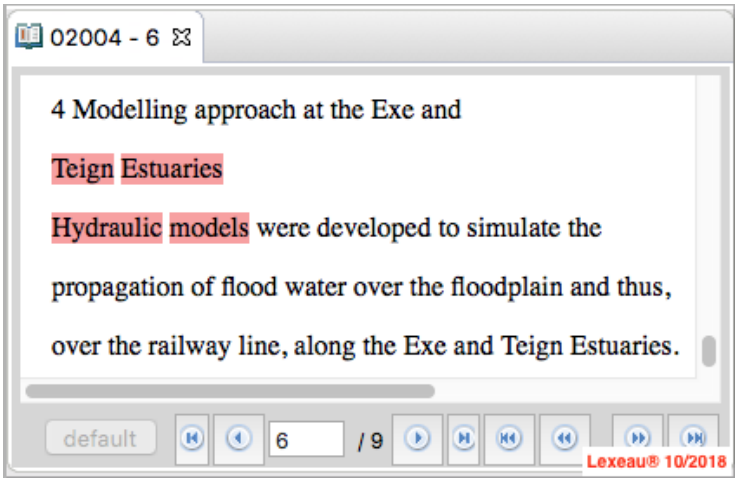


FIGURE 3.38 – Voisinage textuel du syntagme *Teign Estuaries Hydraulic models*

justifie l'introduction du nom du modèle parmi les noms propres associés au lexique du domaine. La forme développée de l'acronyme est *Interaction Sol-Biosphère-Atmosphère*, comme le précise l'article du site internet du Centre National de Recherches Météorologiques consacré au modèle [4], consulté le 13/10/2018. L'acronyme SWI signifie *Soil Wetnesss Index*, comme l'indique le guide méthodologique de l'agence de l'eau Rhône-Méditerranée sur la sécheresse, consulté sur internet le même jour [5].

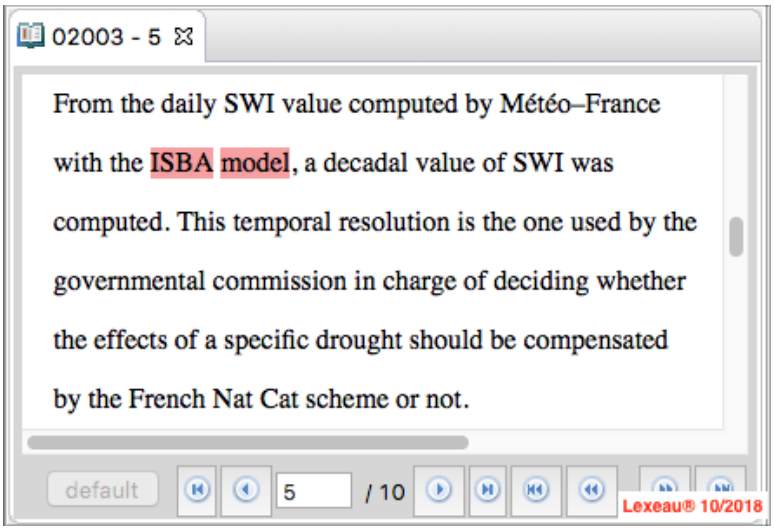


FIGURE 3.39 – Voisinage textuel du syntagme *Isba model* (texte 02003)

Les syntagmes *Mesoscale model* et *Mesoscale model version* sont des syntagmes tronqués, ce que nous montre le voisinage textuel de la figure 3.40, avec le syntagme complet *Mesoscale model version 4* dont l'acronyme est MM4. Nous retiendrons Mesoscale comme nom propre, plutôt que l'acronyme de la version 4 du modèle.

4. https ://www.umr-cnrm.fr/spip.php ?article146
5. http ://www.rhone-mediterranee.eaufrance.fr/docs/infos-secheresse/guide-methodo/indicateur_swi.pdf

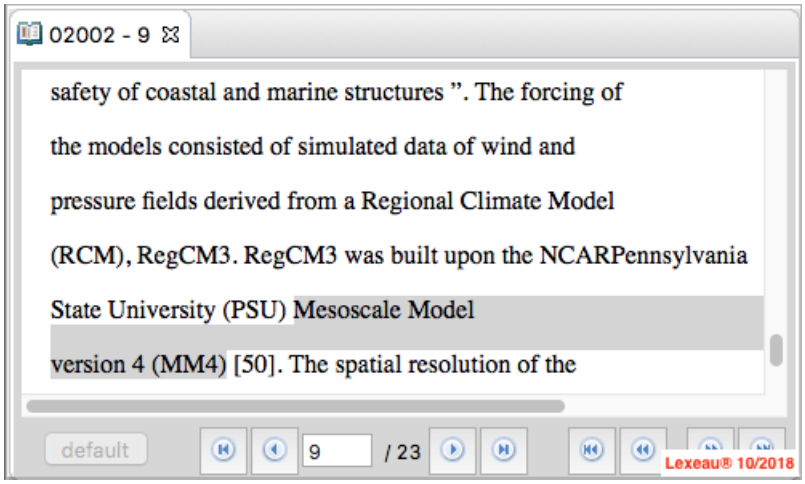


FIGURE 3.40 – Voisinage textuel du syntagme *Mesoscale model* (texte 02002)

La recherche du voisinage textuel du syntagme *model Cfflood* s'est faite dans le fichier *pdf* du document source car l'édition du texte dans TXM aurait du se faire sur deux pages. Ce voisinage est présenté figure 3.41.

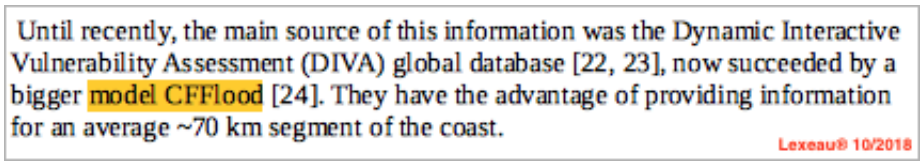


FIGURE 3.41 – Voisinage textuel du syntagme *model Cfflood* (texte 02001)

Le libellé complet de l'acronyme — *Coastal Fluvial Flood* — figure dans le résumé de l'article cité dans le texte 02001. L'article a été publié dans la revue *Climatic Change* aux éditions *Springer Netherlands*. Le résumé est accessible sur le site de l'éditeur [6], consulté le 14/10/2018, avec la référence *Mokrech, M., Kebede, A.S., Nicholls, R.J. et al. Climatic Change (2015) 128 : 245. https ://doi.org/10.1007/s10584-014-1298-6*. L'accès à l'article complet est payant (plus de 40 euros), avec pour titre *An integrated approach for assessing flood impacts due to future climate and socio-economic conditions and the scope of adaptation in Europe.*

Le voisinage textuel du modèle EU-DEM présenté dans la figure 3.42 nous renvoie sur le site de l'AEE (Agence Européenne de l'Environnement) [7], qui diffuse gratuitement sous cet acronyme un modèle digital de surface de l'occupation du sol, élaboré par le *Copernicus Land Monitoring Service - EU-DEM.*

La présentation du modèle sur le site de l'AEE est reprise dans la figure 3.43. Les deux exemples précédents montrent la sophistication des recherches sur le thème des inondations côtières dans le contexte du réchauffement climatique. Il permet de souligner la différence entre l'accès payant à un article publié en 2015 aux éditions Springer et la diffusion gratuite depuis 2017 d'un modèle numérique par l'Agence Européenne de l'Environnement.

6. https ://link.springer.com/article/10.1007/s10584-014-1298-6#citeas
7. https ://www.eea.europa.eu/data-and-maps/data/copernicus-land-monitoring-service-eu-dem

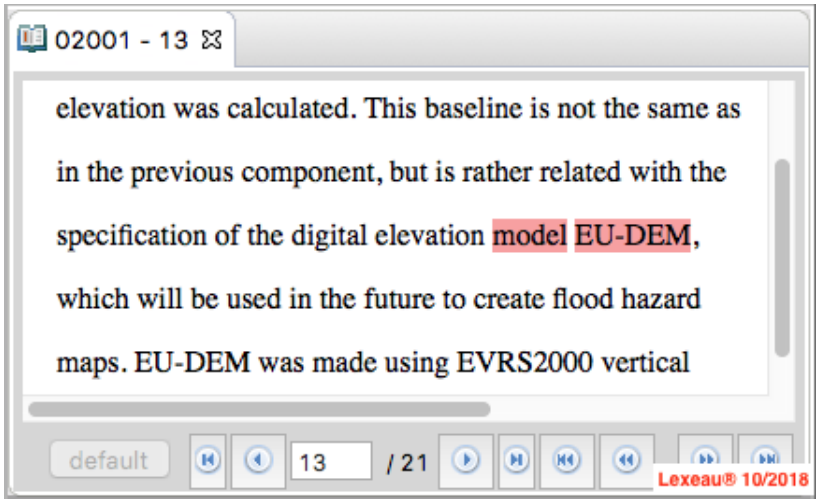


FIGURE 3.42 – Voisinage textuel du syntagme *model Eu-dem* (texte 02001)

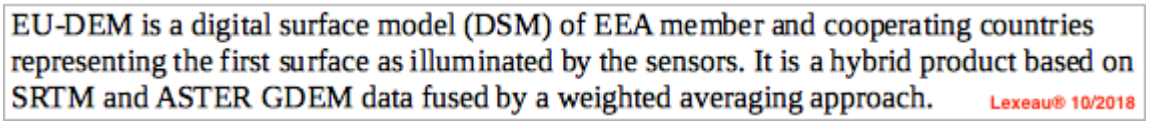


FIGURE 3.43 – Présentation du modèle numérique EU-DEM sur le site de l'AEE

L'intérêt de l'emploi du modèle de vagues SWAN est illustré par le voisinage textuel du syntagme *Swan wave model* présenté figure 3.44.

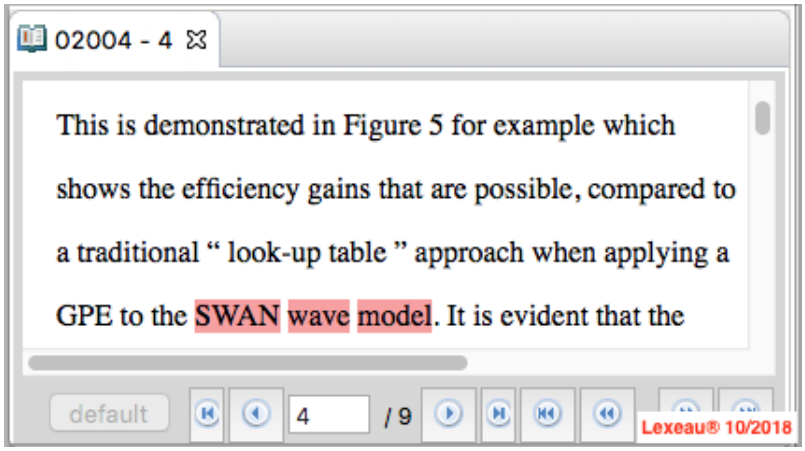


FIGURE 3.44 – Voisinage textuel du syntagme *Swan wave model* (texte 02004)

La présentation de ce modèle sur le site de l'Université technologique de Delft [8] est reprise dans la figure 3.45 .

Le voisinage textuel du syntagme *Pot models* présenté figure 3.46 donne le libellé complet de l'acronyme (*Peak Over Threshold*), ce qui renvoie à une famille de modèles basés sur une

8. http ://swanmodel.sourceforge.net/

SWAN is a third-generation wave model, developed at Delft University of Technology, that computes random, short-crested wind-generated waves in coastal regions and inland waters. For more information about SWAN, see a short overview of model features. This list reflects on the scientific relevance of the development of SWAN. The current version of SWAN is **41.20**. The release notes can be found here.

Lexeau® 10/2018

FIGURE 3.45 – Présentation du modèle de vagues SWAN sur le site de l'Université de Delft

théorie du dépassement des valeurs extrêmes en calcul des probabilités. On ne peut considérer POT comme un nom de modèle à introduire dans le lexique.

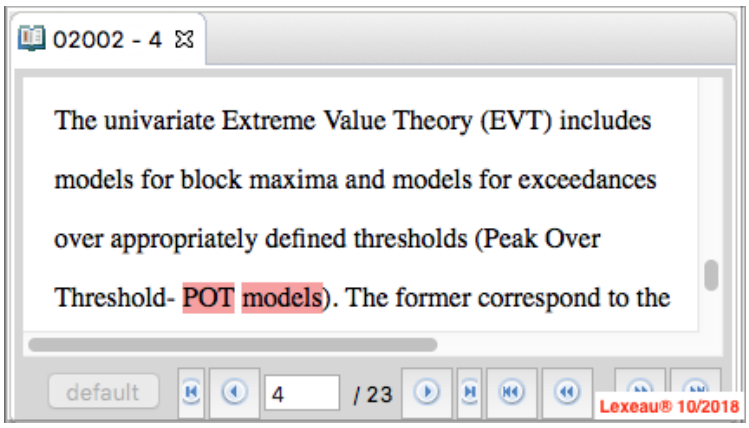


FIGURE 3.46 – Voisinage textuel du syntagme *Pot models* (texte 02004)

Le voisinage textuel du syntagme *tpx08 tide model* est présenté figure 3.47.

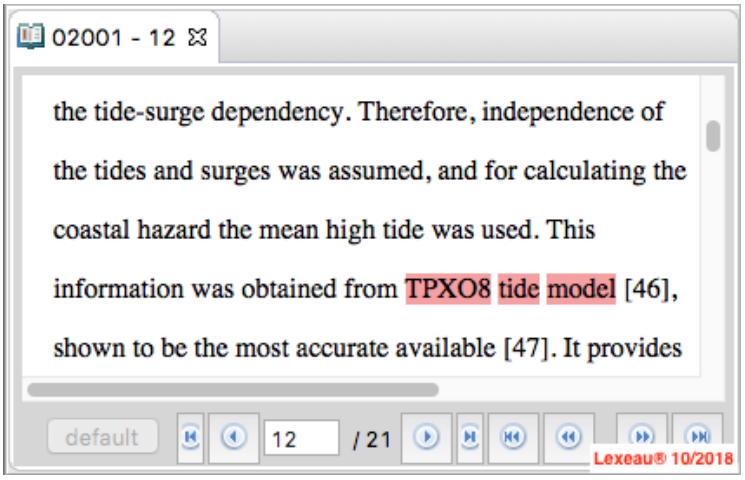


FIGURE 3.47 – Voisinage textuel du syntagme *tpx08 tide model* (texte 02001)

Notre recherche sur internet à partir de l'acronyme du modèle a débouché sur une présentation de la série des modèles de marées sur le site de l'université de l'état de l'Orégon [9], dont la

9. http ://volkov.oce.orst.edu/tides/global.html

figure 3.48 donne un extrait, et sur l'article de Gary D. Egbert et Svetlana Y. Erofeeva, article de référence sur le logiciel OTIS développé par le Collège des Sciences Océaniques et Atmosphériques de cette université, dont la figure 3.49 présente l'en-tête et le résumé, extraits du document téléchargé au format *pdf*[10].

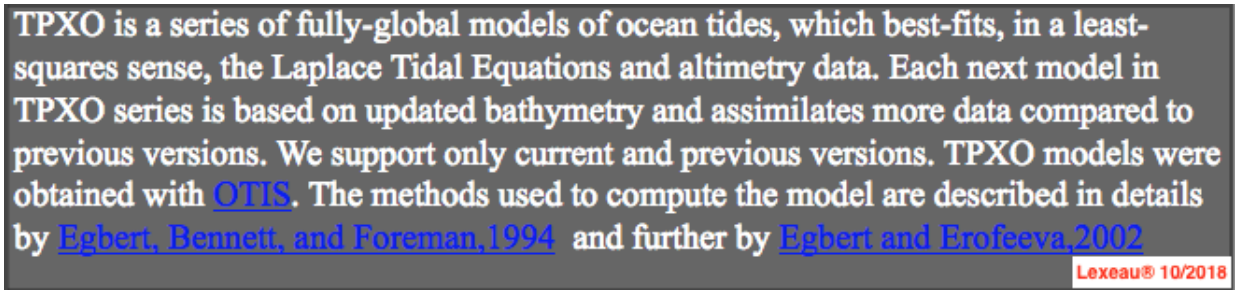
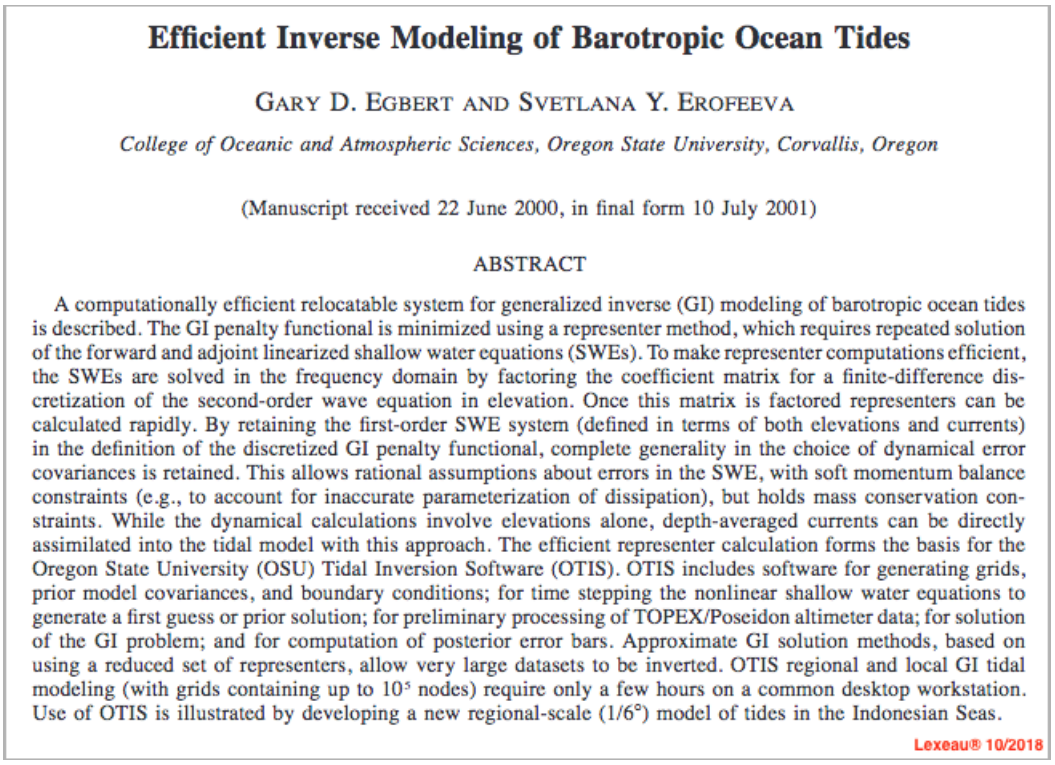


FIGURE 3.48 – Présentation de la série des modèles de marées TPXO (extrait)



Efficient Inverse Modeling of Barotropic Ocean Tides

GARY D. EGBERT AND SVETLANA Y. EROFEEVA

College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, Oregon

(Manuscript received 22 June 2000, in final form 10 July 2001)

ABSTRACT

A computationally efficient relocatable system for generalized inverse (GI) modeling of barotropic ocean tides is described. The GI penalty functional is minimized using a representer method, which requires repeated solution of the forward and adjoint linearized shallow water equations (SWEs). To make representer computations efficient, the SWEs are solved in the frequency domain by factoring the coefficient matrix for a finite-difference discretization of the second-order wave equation in elevation. Once this matrix is factored representers can be calculated rapidly. By retaining the first-order SWE system (defined in terms of both elevations and currents) in the definition of the discretized GI penalty functional, complete generality in the choice of dynamical error covariances is retained. This allows rational assumptions about errors in the SWE, with soft momentum balance constraints (e.g., to account for inaccurate parameterization of dissipation), but holds mass conservation constraints. While the dynamical calculations involve elevations alone, depth-averaged currents can be directly assimilated into the tidal model with this approach. The efficient representer calculation forms the basis for the Oregon State University (OSU) Tidal Inversion Software (OTIS). OTIS includes software for generating grids, prior model covariances, and boundary conditions; for time stepping the nonlinear shallow water equations to generate a first guess or prior solution; for preliminary processing of TOPEX/Poseidon altimeter data; for solution of the GI problem; and for computation of posterior error bars. Approximate GI solution methods, based on using a reduced set of representers, allow very large datasets to be inverted. OTIS regional and local GI tidal modeling (with grids containing up to 10^5 nodes) require only a few hours on a common desktop workstation. Use of OTIS is illustrated by developing a new regional-scale (1/6°) model of tides in the Indonesian Seas.

FIGURE 3.49 – Résumé de l'article de présentation du logiciel OTIS (10/07/2001)

Au terme de cette recherche qui va bien au delà des textes où les modèles sont mentionnés, il est possible de faire le choix des noms propres à introduire dans le lexique en éliminant des noms propres artificiels sélectionnés par la textométrie. Le tableau de la figure 3.50 présente sur fond vert les noms propres des modèles à lexicaliser.

10. https ://doi.org/10.1175/1520-0426(2002)019<0183 :EIMOBO>2.0.CO ;2.

FLR-02, partition par texte	T	02001	02002	02003	02004	Nb de	Nb de	Numéro
Syntagmes nominaux de la base « model »	22 317	6 612	7 550	4 166	3 989	textes	lemmes	requête
Noms propres des modèles à lexicaliser	F							
Agency model	1	0	0	0	1	2	2	2
ARPEGE–Climat model	3	0	0	3	0	1	2	2
Delft3D model	3	3	0	0	0	1	2	5
EURO-CORDEX model	1	1	0	0	0	1	2	2
Hydraulic model	1	0	0	0	1	1	2	2
ISBA model	1	0	0	1	0	1	2	2
Mesoscale Model	1	0	1	0	0	1	2	2
model **CFFlood**	1	1	0	0	0	1	2	3
model **EU-DEM**	1	1	0	0	0	1	2	3
model **SWAN**	1	0	1	0	0	1	2	3
Pot model	1	0	1	0	0	1	2	2
SWAN model	2	0	0	0	2	1	2	2
Ccr impact model	2	0	0	2	0	1	3	7
Ccr's impact model	1	0	0	1	0	1	3	7
Estuaries Hydraulic model	1	0	0	0	1	1	3	7
Exe Estuary model	1	0	0	0	1	1	3	7
Mesoscale Model version	1	0	0	0	1	1	3	4
SWAN wave model	1	0	0	0	1	1	3	7
TPXO8 tide model	1	1	0	0	0	1	3	7
in-house flood hazard model	1	0	0	1	0	1	4	12
in-house surface run-off model	1	0	0	1	0	1	4	12
Teign Estuaries Hydraulic models	1	0	0	0	1	1	4	8
								Lexeau® 11/2018

FIGURE 3.50 – Les noms propres des modèles à lexicaliser (FLR-02)

Nous avons recherché dans les textes du corpus FLR-1 une mention des modèles à lexicaliser tirés du corpus FLR-02. Les résultats sont présentés dans le tableau de la figure 3.51. Les noms

FLR-1 : mention des modèles de FLR-02	FLR-1	Textes du corpus FLR-1											Nb de textes
		01001	01007	01010	01013	02001	02002	02003	02004	03006	03015	04002	
SWAN	59	1	23	2	0	0	1	0	8	3	4	17	8
Mesoscale	2	0	0	0	1	0	1	0	0	0	0	0	2
ARPEGE	3	0	0	0	0	0	2	1	0	0	0	0	1
CFFlood	1	0	0	0	0	1	0	0	0	0	0	0	1
Delft3D	3	0	0	0	0	3	0	0	0	0	0	0	1
EU-DEM	3	0	0	0	0	3	0	0	0	0	0	0	1
EURO-CORDEX	7	0	0	0	0	7	0	0	0	0	0	0	1
ISBA	4	0	0	0	0	0	0	4	0	0	0	0	1
TPXO8	1	0	0	0	0	1	0	0	0	0	0	0	1
Total des fréquences	83	1	23	2	1	15	4	5	8	3	4	17	
												Lexeau® 11/2018	

FIGURE 3.51 – Partage des modèles à lexicaliser du corpus FLR-02 dans le corpus FLR-1

des modèles sont classés dans l'ordre décroissant du nombre de textes concernés. En dehors du modèle SWAN et du modèle Mesoscale, les modèles numériques utilisés pour les recherches du thème 02 sont propres à ce thème.

Il ressort de la recherche des noms propres et des adjectifs en collocation avec la base nominale *model* que tous les noms propres détectés par textométrie dans le corpus FLR-02 ne sont pas qualifiés pour être lexicalisés en les rattachant à l'unité lexicale « Modèle/*Model*. Certaines

appellations des modèles sont génériques. Elles peuvent être retenues en entrée du lexique s'il s'agit d'une classe de modèles bien identifiée, par exemple *wave model* et *tide model*, et doivent être écartées lorsque le nom du modèle est trop vague (*Hydraulic model, Peak over threshold model*) ou renvoie à une entité géographique (*Exe Estuary, Teign Estuary*) ou un organisme (*Environment Agency*, CCR [11], in-house). Le choix passe par des requêtes sur plus de deux mots en cooccurrence, par la lecture du voisinage des syntagmes issus des requêtes et le plus souvent être complétée par des recherches sur internet sont souvent fructueuses. Les cooccurrences d'un nom propre et d'un nom commun jouent un grand rôle dans le lexique de l'eau car toutes les instances des classes d'entités typées du domaine font appel à des noms propres associés aux types, quel que soit leur libellé dans l'ontologie du domaine (la Loire, le fleuve Sénégal, la région Nouvelle Aquitaine, etc.). La classe des noms propres lexicalisés sert de classe d'accueil des noms propres dans le lexique, avec un lien sur la classe des exemples d'emploi édités dans le logiciel de textométrie.

Le graphe de la figure illustre, sur un exemple, l'appariement d'un nom propre récurrent (Npr SWAN) avec un nom propre typé (Npt SWAN) issu d'une entité typée du domaine (Modèle SWAN), appariement qui débouche sur un nom propre lexicalisé (Nl Modèle SWAN), rattaché à une unité lexicale existante (1-TS Modèle) transposée du nom de la classe source du nom propre typé (Modèle).

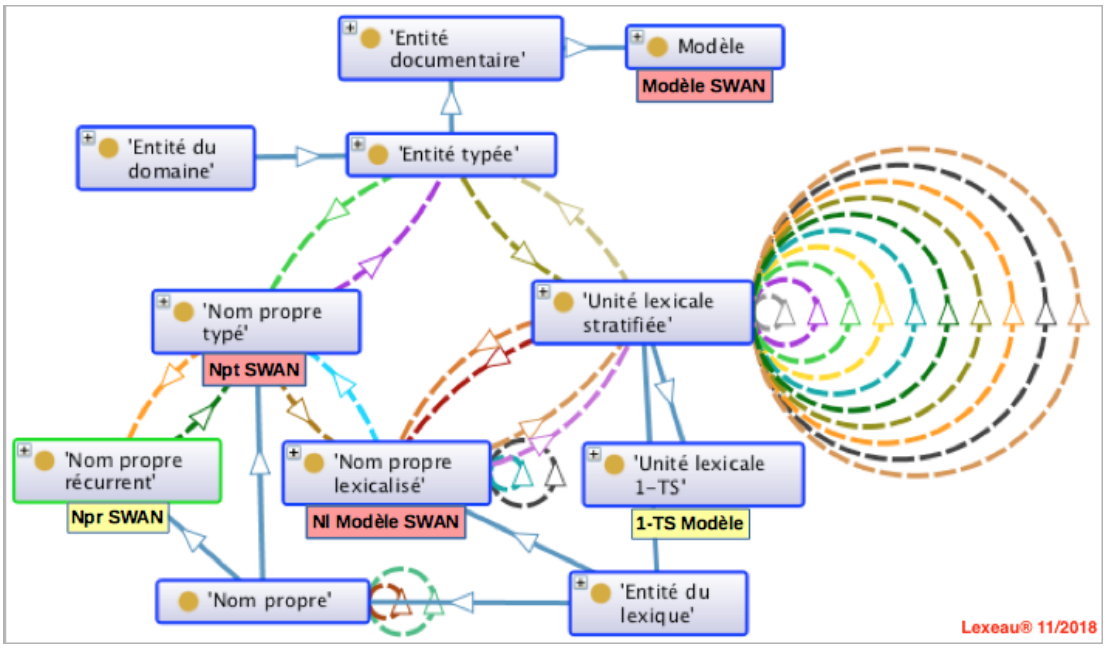


FIGURE 3.52 – Lexicalisation du nom propre récurrent SWAN

Ce chapitre a montré comment la recherche de cooccurrences dans un corpus de texte sur une base nominale a fait émerger une sélection de noms, de syntagmes et de noms propres récurrents. Ils ont permis d'enrichir le lexique d'unités lexicales, de collocations et de noms propres lexicalisés à partir de nouvelles instances et de nouvelles classes dans l'ontologie du domaine.

11. Caisse centrale de réassurance

Chapitre 4

Données, textes et discours

Le soutitre du dictionnaire de l'Académie de l'Eau — Un lexique bilingue de textes et des données sur l'eau — met en parallèle les textes et les données sur l'eau, lesquelles renvoient à première vue aux données qualitatives et quantitatives des usages de l'eau. Ces « données » peuvent être entendues de façon plus large, comme étant ce que l'analyse des textes du domaine fait ressortir des discours sur l'eau. Quels qu'il soient, les textes disent du « rapport à l'eau » ce que les locuteurs veulent faire partager par leur auditoire. La reformulation des problématiques du partage de la ressource, de la baisse de la biodiversité et de l'augmentation des risques liés à l'eau génère des objectifs économiques et sociétaux confrontés, non sans contradictions, l'injonction d'un développement durable et équitable dans un contexte de réchauffement climatique et de mondialisation de l'économie. L'écart entre les objectifs et le « ressenti » quotidien se traduit sur le plan politique et sociétal. Il apparait aussi dans le langage, avec des ambiguïtés sur le sens de mots qui affectent la compréhension des discours et l'intercompréhension entre ceux qui les tiennent ou les entendent. Cette situation s'applique dans le domaine de l'eau, au sein du *nexus* de l'eau, de l'énergie et de l'alimentation et sert de fil conducteur à l'élaboration d'un lexique stratifié. Après les apports de la linguistique de corpus à cette élaboration présentés dans les chapitres 2 et 3, ce chapitre propose une « métrique » graphique pour rendre compte de l'agencement des mots en locutions récurrentes dans les textes, sur un thème donné. Le thème des inondations a été retenu pour tester cette métrique sur les articles du journal britannique *The Guardian* et la directive européenne sur les inondations.

4.1 Recherche des mots les plus fréquents des deux corpus

Le corpus de cinq années d'articles du journal britannique *The Guardian* contenant le mot *flooding* a été présenté dans la section 2.4.3. Parmi les 975 articles extraits de la base FACTIVA, 798 articles ont été importés dans TXM pour former un corpus de textes, après avoir éliminé les articles non pertinents. Un premier travail dans TXM a permis d'extraire les 5 lemmes les plus fréquents sur une partition du corpus par tranche annuelle, du 22/02/2011 au 27/02/2016. Les résultats sont présentés dans la figure 4.1 en regroupant les *pos* avérés des emplois du lemme comme nom commun ou comme nom propre au singulier ou au pluriel.

Journal The Guardian		Partition par tranches annuelles				
Corpus sur 5 ans (2011-2016)	T	2011-2012	2012-2013	2013-2014	2014-2015	2015-2016
Les 5 lemmes nominaux les plus fréquents	507 680	55 559	71 700	121 893	62 248	196 280
	F					
flood_nn/nns/np	2 476	195	351	671	177	1 082
flooding_nn/np	1 881	177	296	482	266	660
people_nns/np/nps	1 393	181	181	327	136	568
rain_nn/nn/np	1 316	147	267	237	294	371
water_nn/nns/np	1 524	161	218	421	118	606
Total des fréquences	8 590	861	1 313	2 138	991	3 287
						Lexeau® 11/2018

FIGURE 4.1 – Fréquences de 5 premiers lemmes du *Guardian* par tranches annuelles

On observe globalement des pics d'emploi des lemmes dans les tranches 2013-2014 et 2015-2016, qui se retrouvent sur les 5 lemmes, sauf pour le lemme *rain* en 2013-2014 dont l'emploi culmine en 2012-2013. La place occupée par les submersions marines peut expliquer ce résultat. Les pics en 2015-2016 sont supérieurs à tous les autres.

Le même travail a été réalisé sur le texte de la directive européenne sur les inondations, nettement moins volumineux (fig. 4.2)

Directive « Inondations »	
Les 5 premiers lemmes	F
flood_NN/NNS	142
risk_NN/NNS	92
management_NN	71
river_NN/NNS	41
basin_NN/NNS	37
	Lexeau® 11/2018

FIGURE 4.2 – Fréquences de 5 premiers lemmes de la directive européenne sur les inondations

On observe l'apparition des lemmes *risk*, *management*, *river* et *basin* parmi les 5 premiers lemmes présents dans la directive européenne, alors qu'ils sont remplacés par *flooding*, *people*, *rain* et *water* dans les articles du *Guardian*. La presse rend plus compte du phénomène de l'inondation que de sa gestion, ce qui est normal car elle s'adresse à ceux qui les subissent, alors que la directive s'adresse à ceux qui doivent les prévoir et les prévenir, avec une prise de conscience du risque et de sa gestion par bassin. L'emploi par la presse de *flooding* (en dehors de la forme verbale) montre que c'est une bonne clef de sélection des articles. Il n'apparait pas en tête de la sélection de la directive mais seulement six fois, ce qui donne une indication sur l'emploi du mot au quotidien et dans un discours technico-administratif.

4.2 Recherche des cooccurrences immédiates à gauche et à droite

L'étape suivante a consisté à rechercher les mots en cooccurrence immédiate, à gauche et à droite des mots les plus fréquents du corpus, choisis comme lemmes pivots. La sélection des cooccurrents est faite sur la base de l'indice de spécificité de la cooccurrence, qui exprime l'écart

du positionnement des cooccurrents par rapport à un positionnement purement aléatoire du pivot et du cooccurrent dans le texte. Un seuil de 20 pour l'indice de spécificité a été retenu pour cette première sélection, à la place du seuil de 2 fixé par défaut. Le tableau 4.1 donne la fréquence et la cofréquence des lemmes placés immédiatement à droite du lemme nominal *flood* dans les articles du Guardian. Les lemmes cooccurrents sont classés par valeur décroissante de l'indice de spécificité.

TABLE 4.1 – Les 14 premiers lemmes après 'flood' dans les articles du Guardian

Cooccurrent	Fréquence	Cofréquence	Indice	Distance moyenne
defence	617	377	1000	,0
warning	659	213	1000	,0
plain	99	81	184	,0
risk	750	124	169	,0
protection	127	63	121	,0
alert	153	61	110	,0
water	1 496	106	104	,0
victim	82	32	57	,0
barrier	113	27	42	,0
insurance	284	29	33	,0
damage	393	29	29	,0
management	102	19	27	,0
relief	88	18	26	,0
prevention	26	13	25	,0

Nous avons recherché dans les articles les lemmes suivis du lemme nominal *flood*, en les triant par *pos*. Avec le même seuil de spécificité, on obtient trois adjectifs, une préposition et des nombres cardinaux exprimés en chiffres (par exemple *62 flood warnings*).

TABLE 4.2 – Les cinq premiers lemmes suivis de 'flood' dans les articles du Guardian

Cooccurrent	Fréquence	Cofréquence	Indice	Distance moyenne
severe_JJ	368	68	96	,0
on_IN	3 884	119	75	,0
@card@_CD	9 286	94	21	,0
new_JJ	448	24	21	,0
flash_JJ	145	17	21	,0

Il est possible de poursuivre en boucle la recherche de cooccurrences à la droite et à gauche d'un lemme. Nous l'avons fait pour les cooccurrences à droite de flood et de @card@, placés en position de pivot dans la requête, en neutralisant les recherches à gauche. Les expressions récurrentes à la droite de *flood* sont présentées dans le tableau 4.3 avec leur fréquence.

Les éléments recueillis avec les requêtes mises en oeuvre par itération permettent d'élaborer une présentation graphique des résultats sur un choix limité de lemmes.

TABLE 4.3 – Expressions récurrentes du *Guardian* à droite de *flood* et de *@card@*

Expression	Fréquence
flood defence spending	32
flood defence scheme(s)	29
flood warning(s) in place	26
@card@ flood warnings	52
@card@ flood alerts	27

4.3 Un graphe des locutions construites par cooccurrence

L'image graphique des réseaux de cooccurrents obtenus par textométrie repose sur le choix d'un lemme pivot et sur les trois données numériques des lemmes représentés, ce qui permet de dimensionner les éléments du graphique en utilisant les valeurs obtenues pour :

- la fréquence du lemme dans le corpus, pour dimensionner son cadre d'écriture sur une taille de police qui augmente avec cette fréquence ;
- la cofréquence, pour choisir de la même façon la taille de la police d'écriture du lemme dans son cadre ;
- l'indice de spécificité de la cooccurrence, pour choisir de la même façon l'épaisseur du trait qui va relier les cadre des lemmes en cooccurrence.

La taille de la police utilisée pour écrire le lemme ou dimensionner son cadre augmente avec sa cofréquence ou sa fréquence, selon une échelle que nous avons déterminée de façon empirique. Elles est présentée dans le tableau 4.4.

TABLE 4.4 – Échelle de taille de la police

Cofréquence/Fréquence	Taille de la police
>10 000	60
6 001-10 000	48
4 001-6 000	40
2 001-4 000	36
1 501-2 000	32
1 001-1 500	28
751-1 000	26
501-750	24
351-500	22
201-350	20
101-200	18
51-100	16
25-50	14
10-24	12
3-9	10
<3	8

Il en est de même pour l'épaisseur du trait de liaison entre deux lemmes en fonction de l'indice de spécificité de la cooccurrence (tableau 4.5). Dans les réseaux de la figure 4.3 les locutions

TABLE 4.5 – Échelle de l'épaisseur du trait

Indice	Épaisseur (cm)
<3	0,05
3-10	0,10
11-25	0,12
26-50	0,14
51-100	0,16
101-200	0,18
201-350	0,20
351-500	0,22
501-750	0,24
751-1000	0,28
>1000	0,32

construites de part et d'autre de *flood* sont disposées l'une sous l'autre par fréquence décroissante, avec une restitution graphiquement de la fréquence et de la cofréquence des lemmes qui les composent. Le principe de la construction du graphique est d'attacher horizontalement

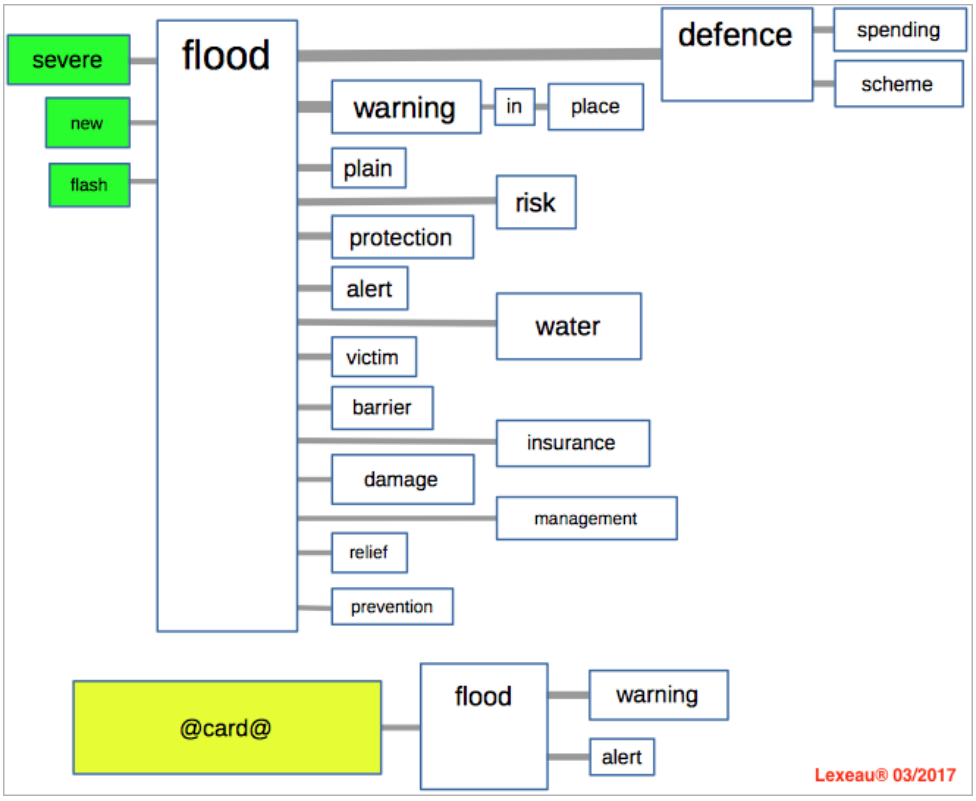


FIGURE 4.3 – Les deux réseaux des cooccurrences de *flood* dans les articles du Guardian

les cadres des lemmes qui figurent dans les locutions issues de la recherche des cooccurrents immédiats, à gauche et à droite du lemme « pivot » puis des les lemmes suivants, du même côté. Le cadre d'un lemme doté de plusieurs cooccurrents à droite ou à gauche est étiré en hauteur pour attacher les locutions les unes au dessus des autres, dans l'ordre décroissant de l'épaisseur du lien qui les relie au lemme « pivot ». Deux réseaux distincts ont été construits pour les lemmes *flood* et *@c@rd@*, compte tenu de la place occupée par les chiffres à gauche de *flood* dans les textes. La recherche des cooccurrents à gauche de *flood* a été limité aux adjectifs et aux nombres cardinaux. Les adjectifs sont reportés dans leur cadre sur fond vert et les chiffres sur fond vert jaune. Les locutions sont à lire d'un seul côté du lemme pivot.

Un travail analogue a été réalisé sur le texte anglais de la directive de 2007 sur la prévention des inondations. Le tableau 4.6 présente les cooccurrences en chaine à droite des lemmes nominaux *flood* et *river* pour un seuil de l'indice de spécificité de 15. Le tableau 4.7 présente les adjectifs

TABLE 4.6 – Cooccurrences en chaîne à droite de *flood* et *river* dans la directive « Inondations »

Lemme nominal	Cooccurrent nominal à droite	Fréquence	Cofréquence	Indice
flood	risk	92	80	139
risk	management	71	46	76
management	plan	46	39	73
risk	map	33	13	16
risk	assessment	22	11	15
flood	hazard	16	15	25
hazard	map	33	14	31
river	basin	37	37	94
basin	district	24	24	59

cooccurrents à gauche des lemmes nominaux *flood* et *river* pour un seuil de spécificité de 17. Le tableau 4.8 présente un tableau des locutions associées à *flood*, avec leur fréquence.

TABLE 4.7 – Adjectifs cooccurrents à gauche de *flood* et *river*, directive « Inondations »

Lemme nominal	Cooccurrent nominal à gauche	Fréquence	Cofréquence	Indice
flood	preliminary_JJ	10	10	17
river	international_JJ	17	12	23

TABLE 4.8 – Collocations avec 'flood' et 'river' (directive 'Inondations')

Expression	Fréquence
flood risk management plan(s)	35
flood risk map(s)	12
preliminary flood risk assessment(s)	10
flood hazard maps	14
river basin district(s)	24

Les réseaux de cooccurrence en chaine à gauche et à droite des lemmes nominaux *flood* et *river* dans la directive 'Inondations' sont présentés figure 4.4, avec deux adjectifs à gauche.

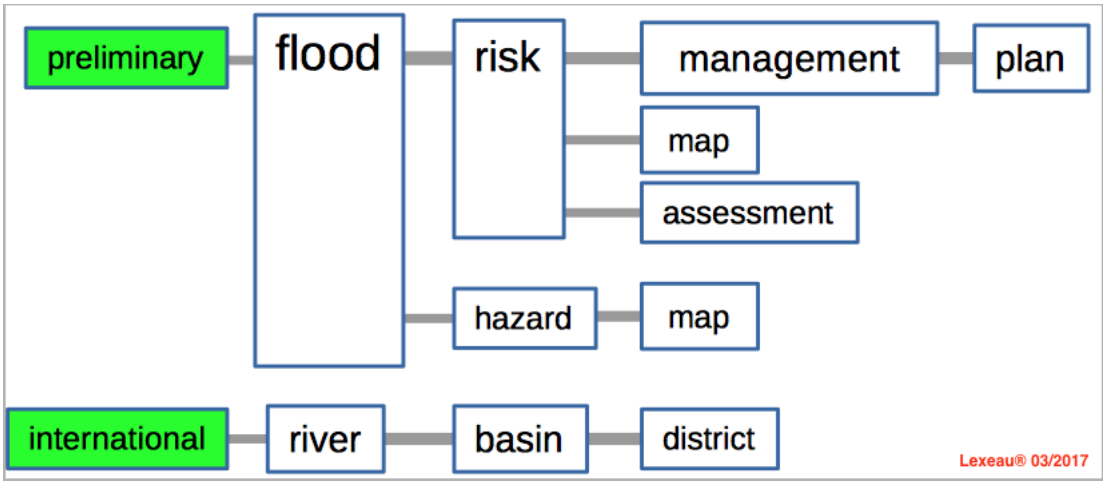


FIGURE 4.4 – Réseau des cooccurrents de *flood* et *river* dans la directive « Inondations »

En comparant les figures 4.3 et 4.4, on observe que le nombre de cooccurrents à droite de *flood* dans la directive est plus limité que dans les articles du *Guardian*, avec les mêmes règles de sélection. Les articles du *Guardian* s'attachent à une description plus large des inondations et de leur contexte et aux préoccupations des populations concernées. La question des assurances y est fréquemment évoquée, alors que le mot n'apparait pas dans la directive à ce niveau de sélection du lexique. La directive est « normative » en termes de vocabulaire, avec l'opposition sémantique clairement affichée entre *risk* et *hazard*. Le mot *risk*, qui peut se traduire par « risque », évoque l'exposition des personnes et des biens aux conséquences d'un événement exceptionnel, avec une expression comme *flood risk assessment*, que l'on peut traduire par « évaluation du risque d'inondation ».

Le mot *hazard*, évoque un aléa, un danger, avec l'expression *hazard map*. L'opposition de *risk* et *hazard* est moins marquée en français, où *hazard mapping* est traduit par *cartographie des risques*. Les mots *risk* et *hazard* relèvent de la strate discursive technico-administrative. Le mot *risk* figure seul dans les articles du *Guardian*. Il relève de la strate discursive courante, où *hazard* aurait peut-être sa place en le traduisant par « danger ».

4.4 Conclusion du chapitre 4

Ce chapitre a ouvert une perspective de développement d'un module graphique associé aux analyses en cooccurrence immédiate à droite et à gauche d'un lemme pivot réalisées avec le logiciel TXM. Ce module rendrait les résultats de la textométrie plus lisibles et valoriserait un logiciel de qualité qui mérite d'être développé au delà de son utilisation dans un contexte académique, notamment dans le cadre du projet LEXEAU. Pour atteindre un tel objectif, la réalisation d'un module informatisé qui facilite le paramétrage et la construction des réseaux testée ici « à la main » est suggérée, sur la base d'un cahier des charges préparé dans le prolongement de ces tests avec l'équipe de développement du logiciel.

Chapitre 5

Une sémantique lexicale revisitée

La sémantique lexicale mise en oeuvre dans le lexique est basée sur la stratification des unités lexicales, dotées chacune d'une définition et d'un libellé différent pour les unités lexicales de la même strate. Les strates discursives sont présentées et discutées dans la section 5.1 ainsi que l'attribution d'un numéro d'ordre pour structurer les liens entre des unités lexicales de strates différentes. L'état d'avancement des trois ontologies du dispositif est présenté dans la section 5.2. Le statut et les fonctions des des unités lexicales sont discutés dans la section 5.3 à la lumière des travaux d'Anna Wierzbicka à l'université de Canberra et d'Igor Mel'čuk à l'université de Montréal. La nature bilingue du lexique et sa mise en oeuvre sont discutés dans la section 5.4, avec le rôle et la structure des unités discursives placées en entrée, à l'instar des lexèmes définis par M. Lynne Murphy dans son manuel de sémantique lexicale [Murphy, 2010].

5.1 Strates discursives et façons de parler

Cette section consacrée aux strates discursives a pour but de montrer comment la notion de strate discursive s'est imposée sous une forme adaptée à ce travail sur le lexique de l'eau. Au départ, le langage technique et scientifique s'oppose au langage quotidien. Cette opposition classique et massive ne résout pas le problème. Quelles sont les positions respectives du journaliste et du juriste ? Le langage du juriste n'est pas un langage technique ou scientifique en ce sens qu'il est le produit des sociétés humaines et non la description d'un « objet du monde » indépendant de sa propre finalité. Il est possible de faire une recherche scientifique sur l'évolution du droit mais cela ne vous dispensera pas de payer votre contravention pour un excès de vitesse en invoquant le droit romain, avant l'arrivée des voitures automobiles. La position du journaliste comme « médiateur » au quotidien de la science et de la technique ou du droit, sans en être un spécialiste, l'amène à utiliser le langage courant pour se faire comprendre. Il contribue à enrichir ce langage une fois instruit de la signification de nouveaux mots auprès des spécialistes, dont une bonne partie sont dans la sphère administrative chargée de mettre en oeuvre les lois et les règlements dans des domaines qui sont souvent techniques.

Un journaliste qui enquête sur l'application du code de l'environnement aux rejets d'eau chaude par les centrales nucléaires aura un discours en langage courant, différent du discours

de l'administration chargée d'appliquer ce droit, qui s'appuiera sur des textes votés par le parlement et leurs décrets d'application, différent aussi du discours de l'ingénieur qui calcule les conditions de refroidissement du coeur de la centrale en tenant compte des contraintes réglementaires sur des rejets d'eau chaude dans la mer ou dans un fleuve. Les mots du journaliste ne vont pas renvoyer aux mentions que l'on peut trouver dans un dictionnaire de langue de telle discipline ou tel secteur d'activité pour les acceptions qui relèvent d'un langage spécialisé, ils seront plus globalement associés à un discours technico-scientifique ou technico-administratif dont il cherche à rendre compte. Les trois strates discursive qui apparaissent à cette occasion dépassent l'opposition binaire entre le langage technique et scientifique et le langage courant des dictionnaires de langue.

Confronté au discours poétique, religieux, moral ou éthique, voire politique, le journaliste aura affaire à un discours dans laquelle le discours scientifique n'a pas sa place. Il n'est pas possible de faire entrer ce type de discours dans la strate technico-administrative car il vise d'abord la sphère individuelle et les émotions, par une incitation à aller dans tel ou tel sens, à prendre telle ou telle décision. Ces discours et le mots et expressions qui les caractérisent relèveraient donc d'une quatrième strate discursive.

Une expérience professionnelle marquante est présentée dans la section 5.1.1 avec des développements plus récents qui ont amené à retenir quatre strates discursives dans le domaine de l'eau. La section 5.1.2 présente les quatre strates et justifie leur numéro d'ordre. Le traitement conceptuel et lexical des aléas et de la vigilance administrative sur les risques naturels et celui des catastrophes, avec parmi les noms propres la catastrophe de la rupture du barrage de Malpasset sont présentés dans cette section pour illustrer l'inter-connexion entre les strates discursives. Le modèle de stratification des documents à la source des textes du lexique est présenté dans le section 5.1.3.

5.1.1 Une expérience professionnelle marquante

J'ai pris conscience de l'existence des strates discursives à travers une expérience professionnelle qui m'a marqué, avant même qu'un nom soit donné au concept par Henri Portine, lorsque je l'ai rencontré pour la première fois. Cette prise de conscience a son origine dans une vive frustration professionnelle, lorsque je me suis rendu compte que je n'arrivais pas à me faire comprendre sur l'intérêt de mettre un prélèvement d'eau en rapport avec le flux d'eau qu'il génère et le devenir de ce flux. Que devient l'eau après avoir été prélevée ? Un de mes interlocuteurs, pourtant diplômé d'une grande école scientifique, ne comprenait pas la question. Pour lui, il suffisait d'administrer la police de l'eau et la perception des redevances en attribuant au prélèvement envisagé un volume et un débit maximum, prévisionnel ou mesuré, puis en délivrant au pétitionnaire une autorisation de prélèvement pour la campagne d'irrigation à venir ou au redevable un titre de perception de la redevance. Son discours était résolument hostile à l'idée d'un modèle hydraulique des écoulements d'eau auxquels l'administration avait affaire, dans une succession de « mouvements d'eau » et de flux d'origine anthropique générés en amont et récupérés en aval. Il faut dire que le modèle présenté dans la section 1.3 n'était pas vraiment au point à l'époque. Il y avait seulement l'idée qu'il pouvait en exister un pour maîtriser la connaissance des flux d'eau anthropiques, sur lesquels les agences de l'eau et les services de la police de l'eau en département n'arrivaient pas à s'accorder pour une bonne partie de leurs dossiers.

J'ai été frappé que ceux qui devraient dépasser cette opposition binaire, dont les cadres d'une administration de formation scientifique, n'acceptent pas facilement de reconnaître qu'ils se sont éloignés de la langue courante, non pas en tant que scientifiques mais parce qu'ils sont tributaires d'acceptions définies dans des textes réglementaires et les lois qu'il leur faut appliquer. Ces acceptions les tiennent éloignés des propriétés des « objets du monde » auxquels ils ont affaire, que le scientifique s'efforce de décrire dans une autre strate discursive.

Un phénomène analogue de rejet se fait jour lorsque deux structures administratives ont des missions distinctes sur les mêmes objets, avec des propriétés qui ne se recoupent pas, ce qui est normal. C'est le cas des prélèvements d'eau pour la police de l'eau et pour l'agence de l'eau. Pour certains, cette évidence ne se discute pas. Elle ne doit ni ne peut être élucidée. Ce serait ébranler des systèmes d'informations complexes, construits au fil du temps de part et d'autre pour remplir au mieux les missions imparties. La crainte, que l'on peut comprendre, est de perdre la maîtrise de ces systèmes au profit d'une super administration qui s'emparerait de données individuelles enrichies, avec de nouveaux enjeux dans la distribution du pouvoir administratif, ce qui relève effectivement de la commission Informatique et Liberté.

La quatrième strate discursive est apparue plus tard, lors d'une réunion avec un « politique », familier du domaine de l'eau, à propos du projet LEXEAU. Ce qu'il nous a dit sur sa vision du projet — il était chargé à l'époque d'un rapport sur l'eau pour un ministre en exercice — nous a fait comprendre que son discours n'entrait pas dans une des trois strates discursives. L'agencement et le choix de certains mots ne relevaient pas du langage courant. Le discours était construit pour inciter l'auditoire à aller dans telle ou telle direction, à prendre telle ou telle décision, et non pour un rappel au règlement ou à la loi, ou pour présenter le fruit d'un travail scientifique. Il fallait donc une quatrième strate, baptisée « décisionnelle incitative » (DI).

5.1.2 Quatre strates discursives structurées

Nous entendons par strate discursive un type de discours, une façon de parler, adaptés à un auditoire pour qu'il comprenne ce qui est dit et l'apprécie dans la mesure où cela répond à son attente.

Un scientifique s'adressera à ses collègues dans une **strate discursive technico-scientifique**, en se mettant autant que possible à la portée de l'auditoire que forment ses pairs pour présenter le résultat de ses recherches et leur fil conducteur. Il évitera des facilités de langage (des « pirouettes », de la « poudre aux yeux »), ce qui ne l'empêchera pas d'employer des métaphores pour se faire comprendre d'un auditoire moins spécialisé qu'il ne l'est sur le sujet traité. Le locuteur est engagé dans un contrat avec son auditoire, qui sera en droit d'en contester le respect en faisant ressortir le flou de certains concepts, le manque d'informations pour pouvoir répéter les expériences dont il présente les résultats, le parti pris sans fondement explicite sur certains points, etc. Les textes de cette strate sont issus d'articles publiés dans des revues scientifiques. Ces revues font généralement appel à des pairs avant de décider de la publication des articles (y compris des compte rendus d'ouvrages), ce qui leur vaut une forme de label dont les auteurs peuvent se prévaloir. Le *publish or perish* du monde anglo-saxon s'applique de plus en plus en langue anglaise dans le monde, ce qui ne manque pas de générer des difficultés de compréhension, voire de qualité dans les travaux qui sous-tendent de

telles publications. Les publications dans d'autres langues n'échappent pas à ce risque quand elles sont rédigés dans une langue seconde. Les doctorants prennent conscience très tôt de cet impératif dont l'institution scientifique a fait une règle. Dans le domaine de l'eau, l'éventail des thèmes et des disciplines à caractère scientifique est extrêmement large, ce qui pourrait inciter à découper la strate discursive en sous-strates autonomes. Ceci ferait perdre de sa valeur au concept qui permet de juger, sur le même thème, de l'emploi des mots et de leur agencement en opposant le sens qu'ils prennent dans d'autres strates discursives, et de croiser l'emploi des mêmes mots dans plusieurs thèmes pour voir s'ils prennent des sens différents. Le sens d'un mot ou d'une expression dans cette strate discursive est en effet réputé unique, en tout cas défini sans ambiguïté sur le thème et dans une discipline qui l'a vu naître. Il ne viendra à l'idée de personne de contester qu'un « anneau » est bien un « objet mathématique » parfaitement défini et d'en promouvoir l'emploi dans une autre discipline.

Les conditions de réception d'un discours dans la **strate technico-administrative** sont comparable d'une strate discursive à une autre à ceci ceci près que le locuteur ne s'adresse pas à son auditoire dans un langage scientifique. Il se s'agit pas de démontrer la réalité d'un « fait de nature » jusqu'alors ignoré, ou d'en expliquer le fonctionnement, comme la photosynthèse ou l'eutrophisation d'un milieu aquatique. Dans le contrat implicite passé avec l'auditoire, qui peut comprendre des éléments ayant une formation scientifique et technique poussée, il s'agit de l'amener à comprendre une règle ou d'une procédure administrative décrite dans un texte et de répondre aux modalités pratiques de leur application. Nous ne sommes pas ici dans une spéculation pour savoir si telle ou telle proposition de loi va être invalidée par le conseil constitutionnel, ou dans l'attente du résultat du recours déposé devant une juridiction d'appel. Les textes sont établis, y compris les jugements au contentieux, il n'y a plus qu'à les appliquer. C'est le travail de l'administration publique, mais aussi des entreprises privées ou publiques qui obtiennent des contrats de fournitures et de services, des particuliers et des organismes qui sollicitent une autorisation administrative, souscrivent un contrat d'assurance ou s'acquittent d'une redevance. Le champ de cette strate dans le domaine de l'eau est extrêmement vaste. Les liens avec les « objets » de la strate technico-scientifique sont fréquent, et peuvent soulever des problèmes délicats, par exemple la délimitation d'une « zone humide », faute de cartographie officielle dans le référentiel hydrographique national, ou celle d'une « nappe d'accompagnement », non répertoriée dans le référentiel hydrogéologique en cours de remaniement. Une bonne partie des textes qui relèvent de cette strate sont destinés aux natifs de la langue dans laquelle ils sont rédigés. Les textes rédigés en plusieurs langues comme les directives de la communauté européennes ne sont appliquées dans un pays que lorsqu'ils ont été transposés en droit national. Pour élaborer un lexique bilingue dans le domaine de l'eau, le recours au texte des directives à l'origine des lois françaises qui opèrent cette transposition est indispensable, afin de rapprocher des définitions et des exemples d'emploi des entrées du lexique dans cette strate. Il faut aussi distinguer entre les mesures qui relèvent du droit national inchangé et celles qui correspondent à la transposition de la législation européenne. Les sites internet des législations nationales et européenne comme « Legifrance » [1], pour la France, « GOV.UK » [2], pour le Royaume Uni, et « Eur-Lex » [3], pour l'Union européenne seront les auxiliaires indispensable de la recherche des textes du discours technico-administratif sur l'eau. Là encore, on évitera d'en extraire des entrées lexicales ayant plusieurs sens avec le même libellé.

1. https ://www.legifrance.gouv.fr/
2. https ://www.gov.uk/
3. https ://eur-lex.europa.eu/homepage.html ?locale=fr

On pourrait s'étonner de voir figurer dans le lexique des entrées issues du langage courant dans la **strate discursive courante**. Il s'agit là d'un élément essentiel de l'élaboration du lexique, dont l'ambition est de rapprocher, sans les déformer, des mots et des expressions qui visent la même réalité vue sous un angle différent, par les scientifiques, les « administratifs » et les media, lesquels reflètent et alimentent les échanges au quotidien sur la problématique de l'eau. L'objectif du lexique n'est pas de remplacer dans l'esprit des non spécialistes, qui constituent l'immense majorité de la population, le sens des mots et des expressions qui leurs sont familiers par des significations plus exactes sur le plan administratif ou scientifique. Il s'agit de faire comprendre à tout le monde, spécialistes compris, que ces mots et ces expressions peuvent avoir d'autres significations dans d'autres strates discursives, sans avoir à rougir du sens qui leur est donné dans la langue courante ni à vitupérer contre cette langue. Le procès des journalistes « ignorants » des acceptions propres aux univers restreints des administrations et des laboratoires serait encore plus absurde, écartant toute tentative de compréhension par les locuteurs de ces univers de la façon dont les « choses » du domaine de l'eau sont perçues au quotidien par des gens de tous les niveaux sociaux dont ce n'est pas le métier. Ceci procède du dépouillement des journaux quotidiens, y compris régionaux, lorsque l'actualité du domaine fait la une, ce qui est fréquent avec des intempéries qui entrainent parfois des situations de « catastrophe naturelle », et des rubriques connexes comme le droit à l'eau, le prix de l'eau, les restrictions dans l'usage de l'eau ou le réchauffement climatique. Les bases de données de la presse sont irremplaçables pour capter les préoccupations communes sur l'eau et les mots qui les expriment, ce que savent très bien faire les journalistes. On pourrait imaginer des contrats avec les organes de presse pour nourrir la strate discursive courante de façon à la mettre en regard des autres strates discursives et mieux comprendre les liens que les mots et les expressions partagées induisent entre les strates. Cette démarche pourrait intéresser les spécialistes et les scientifiques soucieux de voir la portée pratique de leur message transposé au quotidien. Le développement d'émissions à caractère scientifique destinées au grand public (« La tête au carré », par exemple) enrichit cette strate discursive. Les décideurs, y compris les décideurs financiers, ne sont pas forcément plus avertis que « l'homme de la rue » sur les questions du domaine. Ils sont eux-mêmes ou par personne interposée des lecteurs assidus de la presse quotidienne, de l'actualité radio-télévisée et des émissions « grand public ». Le secteur des *media* dans les deux langues du lexique est une mine de mots et d'expressions qui peuvent être confrontés au discours scientifique. Ce discours l'inspire plus qu'on ne l'imagine dans la recherche d'une signification stable des mots et des expressions du langage courant. Il est important de rétablir des liens entre les deux. Le pari de cette approche de la langue courante dans un domaine où le langage spécialisé domine est que l'acception d'une locution employée au quotidien soit reliée de façon stable à des acceptions plus sophistiquées et moins répandues.

Le dernier type de discours qui correspond pour nous à la **strate discursive décisionnelle incitative** est d'abord un discours qui n'entre pas dans les trois catégories du discours scientifique, administratif ou courant. Ce type de discours est marqué par la volonté du locuteur de faire adhérer l'auditoire à une vision du monde en faisant jouer des ressorts qui ne relèvent pas d'une description scientifique ou d'une contrainte administrative mais plutôt d'un ressenti marqué par une émotion ou une conviction personnelle qui ne s'exprime pas dans la langue pratiquée au quotidien. Le champ d'application de ce type de discours est très large. C'est celui de de la religion, de l'éthique, de la morale, de l'art, d'une « cause » ou d'une idéologie particulière. Ce champ intervient dans le domaine de l'eau de différentes façons, avec

des mots et des expressions qui lui sont propres et qui renvoient à des objets concrets (l'eau du baptême, les ablutions religieuses, le tableau des nénuphares de Monet, « La Mer » de Debussy) ou plus abstraits (la « fureur » des éléments, comme image de la colère divine). On aurait tort de négliger l'impact de ce discours au prétexte qu'il ne relève pas d'une science ou d'une loi humaine. Il traduit un mode d'observation du réel qui est très ancien et ne cesse pas de se renouveler, de s'actualiser, notamment dans l'art et la littérature ou la poésie, sur des questions proches du domaine de la loi (les droits de l'homme) ou des prises de conscience qui débouchent sur des investigations scientifiques (la pollution des océans). Le caractère militant de ce type de discours, qui peut aller jusqu'à des extrêmes meurtrières, ne doit pas faire obstacle à une investigation scientifique des mots et des expressions employés. La difficulté tient au fait qu'il est pratiqué dans des lieux très divers, à des fins parfois dissimulées, y compris sur les réseaux sociaux, qui sont une des sources à exploiter. Comme avec les autres strates, la question est de savoir si des concepts propres au domaine de l'eau peuvent émerger de cette investigation et comment ils s'articulent entre eux. Une approche factuelle des théologies et des idéologies du moment dans leur « rapport à l'eau » devrait permettre de comprendre les écarts de sens entre le langage courant et celui qui est pratiqué dans cette strate, écart qui peuvent jouer aussi avec les deux autres strates.

La structuration des strates discursives est selon nous un élément de leur « productivité » linguistique. La structure hiérarchique, qui consiste à les placer dans un certain ordre, reflète l'influence qu'une strate peut avoir sur une autre strate en y faisant émerger des mots ou des expressions de libellé identique ou proche, avec une signification et un emploi que se démarquent de la signification et de l'emploi dans la strate source. L'ordre des strates implique que cette influence ne joue que dans cet ordre, qui se traduit dans notre cas par un numéro d'ordre de 1 à 4. Nous avons placé la **strate discursive technico-scientifique** en première position, en considérant que le discours scientifique repose sur des concepts réputés stables, tant qu'ils ne sont pas remis en cause par les scientifiques eux-mêmes, qui les font sortir du domaine scientifique ou les considèrent comme obsolètes, ce qui se produit dans la durée, à une échéance difficile à évaluer à un moment donné, sinon *a posteriori*. Des concepts plus anciens peuvent être intégrés dans de nouvelles constructions conceptuelles qui ne les remettent pas en cause dans leur domaine d'application. La réalité de l'espace-temps de la théorie de la relativité a été démontrée par de nombreuses expériences, cela n'empêche pas les trains d'arriver à l'heure et les aiguillages de fonctionner en automatique dans notre univers terrestre à trois dimensions. Cette position implique que l'influence de la strate discursive peut jouer sur les trois autres strates, mais la réciproque n'est pas possible. Nous avons attribué la deuxième place à la **strate discursive décisionnelle incitative** considérant, sur l'exemple des religions, que celles-ci ont plutôt tendance à inspirer les lois humaines, qui fondent l'action administrative, que l'inverse. On peut soutenir que les religions sont des créations humaines, il n'empêche qu'elles semblent bien avoir inspiré dans l'histoire un discours administratif qui s'est transformé, avec des mots et des expressions nouvelles. La loi humaine d'aujourd'hui n'est pas celle de Jupiter, dont il reste pourtant une trace avec l'adjectif « jupitérien ». Lorsque la religion n'est pas encore passée à l'état de mythe, le passage d'une strate à l'autre tente de se faire dans l'autre sens, dans la contestation du discours scientifique, avec la création du monde en sept jours ou la venue proche de l'apocalypse dans certains milieux religieux aux États Unis, par exemple. Les scientifiques qui voient là de nouveaux concepts dans leur domaine propre ne semblent pas soutenus par leurs pairs. La « loi islamique » est un exemple problématique, pour un esprit rationnel, de fusion des lois humaines et divines. C'est une nouvelle version des tables

de la loi qui a une traduction très concrète dans les peines encourues si elle est transgressée. La **strate discursive technico-administrative** ne peut prétendre inspirer une strate discursive qui rend compte de la loi divine. Elle occupe la troisième position dans l'ordre des strates. La **strate discursive courante** se place en quatrième position. Cette position la rend disponible pour des unités lexicales satellites d'unités pivot dans les autres strates, dans une définition qui s'en démarque, et pour des concepts qui méritent de figurer dans l'ontologie du domaine, ce qui exclu une bonne partie des concepts du quotidien qui n'ont aucun rôle spécifique dans le domaine de l'eau. Dans notre esprit, la notion de strate discursive n'implique aucun jugement de valeur sur le contenu du discours de telle ou telle strate. Elle rend compte du phénomène de stratification des discours sur l'eau que nous avons étudié en français et en anglais.

Les strates discursives du lexique sont présentées avec leur préfixe dans la figure 5.18 en tant qu'entités du discours.

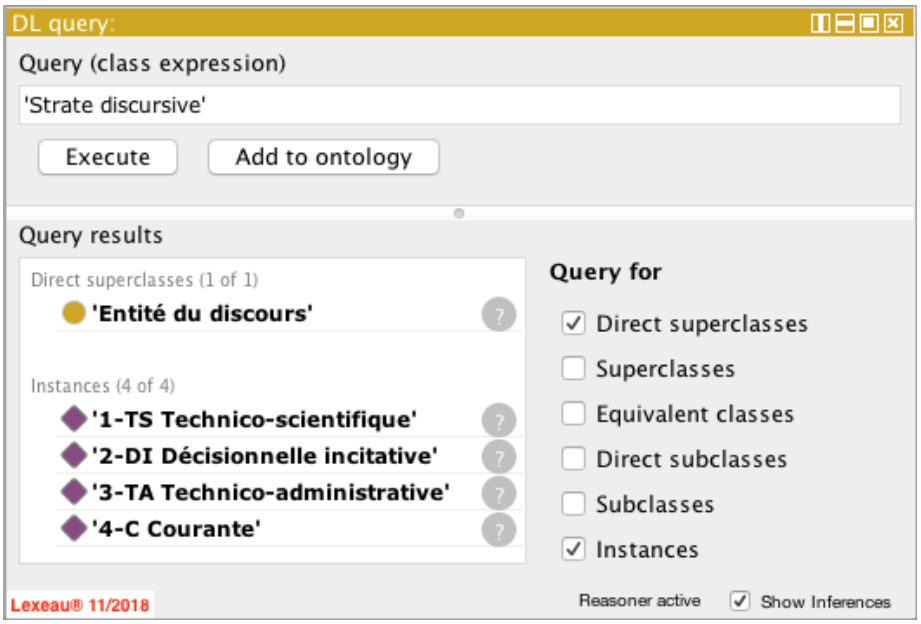


FIGURE 5.1 – Les quatre strates discursives en tant qu'entités du discours

L'ordre incorporé dans leur préfixe marque un lien structurel entre les strates discursives par l'intermédiaire des unités lexicales pivot et satellite que l'on y trouve. Un concept du domaine ne peut être transposé que dans une unité lexicale pivot. Le langage quotidien présente des caractéristiques inverses avec le phénomène de polysémie, qui marque son adaptabilité à la très grande diversité des situations de parole. Nous avons donc un double niveau, celui de l'ontologie du domaine, des discours scientifiques et techniques ou administratifs, et celui du « lexique quotidien » et des discours auxquels il donne lieu. Ce phénomène est exploité par le lexique en ce sens que les unités lexicales pivot peuvent avoir plusieurs unités lexicales satellites dans d'autres strates discursives avec un libellé identique ou proche (hors préfixe) et une définition qui leur est propre. Nous avons présenté la transposition du concept de prélèvement d'eau dans le lexique avec trois unités lexicales satellites dont deux dans la strate discursive administrative de libellés différents (cf. fig. 1.22). Les définitions des unités lexicales satellites de cet exemple sont présentées en annexe dans les figures A.1 à A.4.

5.1.2.1 Aléas et vigilance : ajuster les concepts pour harmoniser les discours

Pour rendre compte de la vigilance mise en oeuvre recommandée aux populations et aux autorités administratives concernées par des phénomènes climatiques extrêmes et des crues, nous avons établi une liste d'aléas incluant ceux qui en sont l'objet en France et en Angleterre (avec le *weather warning* et le *flood warning*), qui débouche sur les classes de la figure 5.2.

FIGURE 5.2 – Les 20 classes d'aléas prises en compte dans l'ontologie du domaine

Les dix phénomènes extrêmes suivis en France par Météo France et le SCHAPI [4] (avalanche, canicule, crue de rivière, grand froid, inondation de rivière, neige-verglas, orages, pluie inondation, vagues-submersion, vent violent) ont été complétés par dix aléas (brouillard, chute de neige, grêle, pluie, onde de tempête, rupture de barrage, sécheresse, submersion de barrage, verglas). Certains aléas de cette 2ème liste recoupent ceux de la liste précédente car ils sont considérés pour eux-mêmes, notamment le verglas qui fait l'objet d'une alerte spécifique en Angleterre pour la sécurité routière, avec le brouillard. Cette liste de 20 aléas n'est pas une nomenclature, mais un compromis entre la pratique des services de vigilance ou d'alerte (*warnings*) et la prise en compte d'évènements exceptionnels qui peuvent être à l'origine d'aléas déjà suivis, comme la rupture d'un barrage ou une onde de tempête (*storm surge*) exceptionnelle en zone côtière, combinée avec un fort coefficient de marée. Nous avons inclus la grêle, qui figure dans certaines polices d'assurance des cultures, et la sécheresse, pour son impact en agriculture.

Le graphe relationnel dans l'ontologie du domaine des vigilances recommandées en matière de risques naturels et des aléas correspondants est présenté dans la figure 5.3. Chaque vigilance

4. Service central d'hydrométéorologie et d'appui à la prévision des inondations

est datée, localisée et dotée d'un niveau de vigilance.

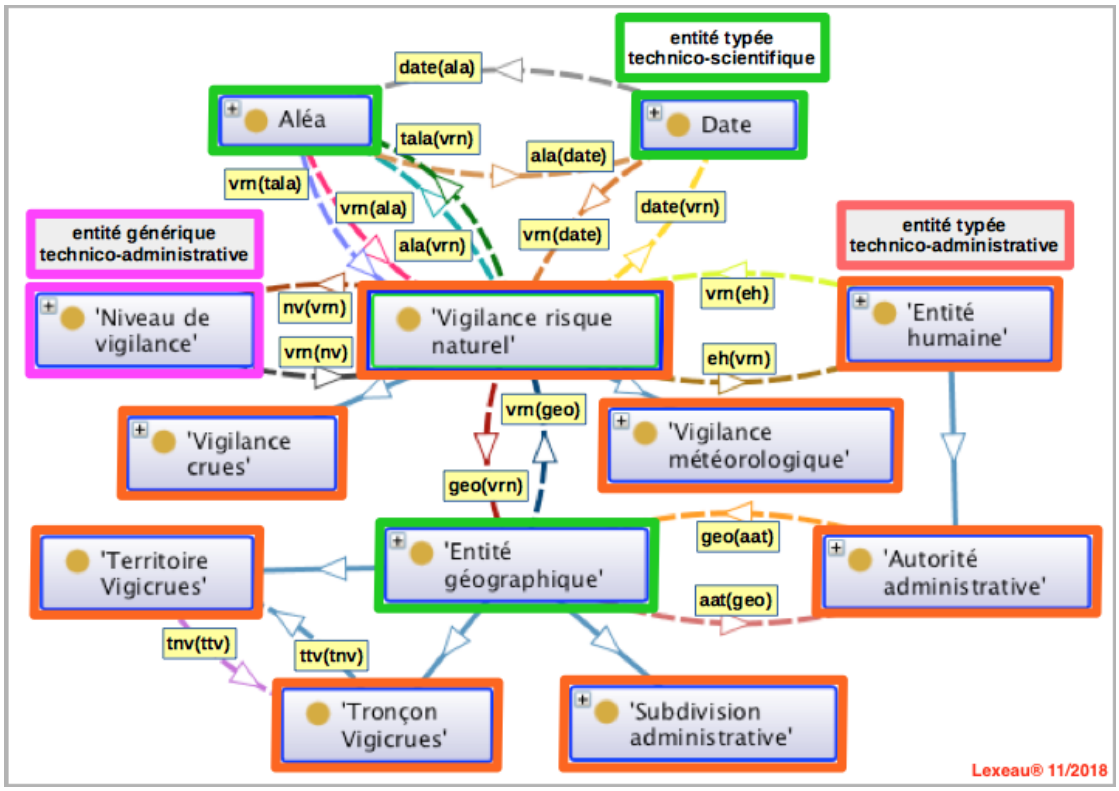


FIGURE 5.3 – Graphe des « vigilances » datées et localisées, par type d'aléa et par niveau

Les vigilances requises à l'égard d'un risque naturel sont enregistrées dans l'ontologie par type d'aléa (relation « tala(vrn) »). Chaque aléa, en tant que phénomène concret, est relié à une ou plusieurs vigilances (relation « vrn(ala) »), en fonction du déroulement du phénomène dans l'espace et dans le temps. Les phénomènes qui relèvent de la vigilance météorologique sont ceux de la première liste, à l'exception des crues de rivière, qui relèvent d'une vigilance baptisée « Vigicrues ». Chaque vigilance renvoie à une entité géographique (un territoire) et à l'autorité administrative habilitée à agir sur ce territoire pour la prévention des risques naturels. En France, c'est le préfet du département ou le maire de la commune qui prend éventuellement des mesures préventives (évacuation, etc.) et organise les secours. Chaque vigilance datée et localisée correspond à un niveau de vigilance, symbolisé par une couleur qui marque la dangerosité du phénomène (jaune, orange, rouge) et qui est assorti de conseils sur la conduite à tenir.

Dans la figure 5.3, un choix de la strate discursive des éléments transposés dans le lexique est proposé pour chaque classe (contour coloré). Ce choix renvoie en réalité au concept. Il porte sur le nom de la classe, pour une classe d'entités typées ou génériques. Les noms des classes d'aléas sont transposés dans la strate technico-scientifique. Il en est de même pour la classe des dates et celle des entité géographiques. Les autres classes de la figure induisent une transposition dans la strate discursive technico-administrative, avec des objets de connaissance qui n'ont pas cours dans les deux langues, comme par exemple les « territoires vigicrues » et les « tronçons vigicrues » qui n'ont pas cours en anglais, à notre connaissance. On serait tenté de qualifier d'entités technico-scientifiques les sous-classes d'une classe ayant cette « qualité ». Ce n'est

manifestement pas le cas. Nous considérons le concept d'entité géographique comme technico-scientifique car une entité géographique peut être définie dans un discours technico-scientifique, par exemple un cours d'eau ou un aquifère. Cette « qualité » conférée à la super-classe pour la transposition de son nom dans le lexique n'engage pas le choix de la strate discursive des hyponymes du nom. Les instances de la classe du niveau de vigilance relèvent de la strate discursive technico-administrative. Le choix de plusieurs couleurs pour qualifier le niveau de vigilance dont il faut faire preuve en face d'un risque naturel n'est pas inscrit comme un « fait de nature » dans les neurones de l'utilisateur du site de Météo France ou de Vigicrues. Il est appris et mémorisé comme le code de la route, tel qu'il a été imaginé et transmis par le service qui gère le site. La différence entre strates discursives apparait ainsi selon que le concept correspondant à l'élément de l'ontologie transposé est universel ou propre à la langue dans laquelle il est exprimé, à travers le contenu culturel restitué.

Le graphe relationnel de la figure 5.4 complète celui de la figure 5.3.

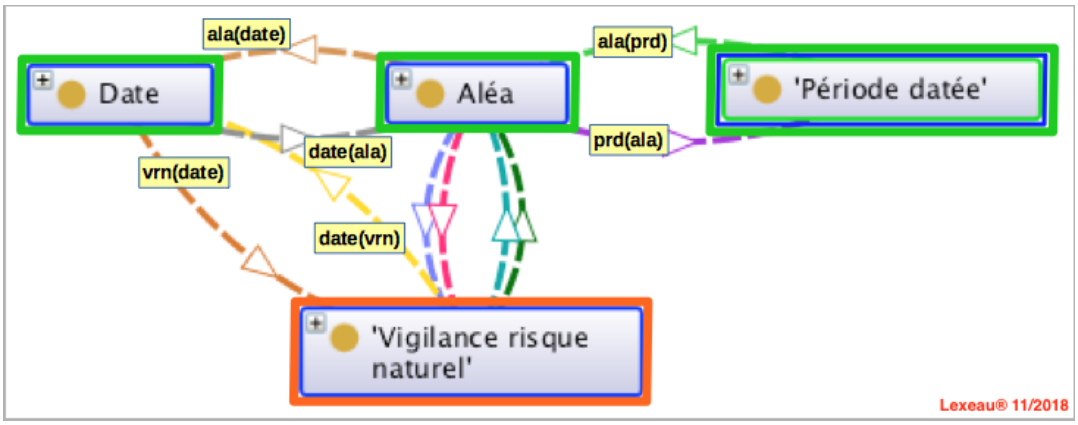


FIGURE 5.4 – Graphe relationnel de l'inscription de l'aléa dans le temps

Les aléas sont inscrits de deux façons dans la temporalité, par une date de référence, celle du pic de l'aléa, s'il existe (hauteur maximum de la crue, de l'inondation) et par la période d'occurrence du phénomène, pour une durée brève (une pluie de 2 heures, de 10 minutes) ou plus longue (une sécheresse de 6 mois, un épisode pluvieux de 4 jours). Nous avons fait figurer la classe des vigilances sur ce graphe pour rappeler qu'elles sont datées, ce qui permet de rassembler les vigilances associées à un phénomène exceptionnel sur sa période d'occurrence et de faire un bilan de l'action de prévention des services de l'état pour ce phénomène, auxquelles on pourrait ajouter les actions engagées à ce titre par les autorités administratives, à partir de leur modélisation conceptuelle et d'un enrichissement du lexique.

La définition des aléas dans la strate discursive technico-scientifique est reportée en annexe (figure A.5). Cet exemple soulève la question de la pertinence de notre distinction entre aléas et vigilance en termes de strate discursive. Après tout, la vigilance consiste à mettre en oeuvre des relevés climatiques ou hydrologiques et des modèles de prévision, construits de façon scientifique, pour alerter les populations sur les risques auxquels elles s'exposent en ne tenant pas compte de ces alertes. Le discours de la vigilance serait-il purement scientifique, et la « mauvaise volonté » du côté de l'auditoire ? Nous répondrons que le discours de la vigilance est tenu dans un contexte administratif qui varie d'un pays à l'autre, pour des aléas qui ne sont pas les

mêmes d'un pays à l'autre mais qui relèvent bien d'un discours scientifique. La vigilance n'est pas un « fait de nature ». Il s'agit de diminuer le risque d'accident. L'introduction des deux strates discursives permet de ne pas mélanger la mise au point et l'évaluation d'un dispositif mis en place par l'homme dans un discours technico-administratif et la connaissance scientifique des phénomènes et des événements qui ont conduit à cette mise en place. C'est pourquoi nous avons ajusté la liste des aléas sur ceux qui sont identifiés par les administrations française et britannique. Le graphe relationnel de la figure 5.5 présente cette mise en correspondance conceptuelle.

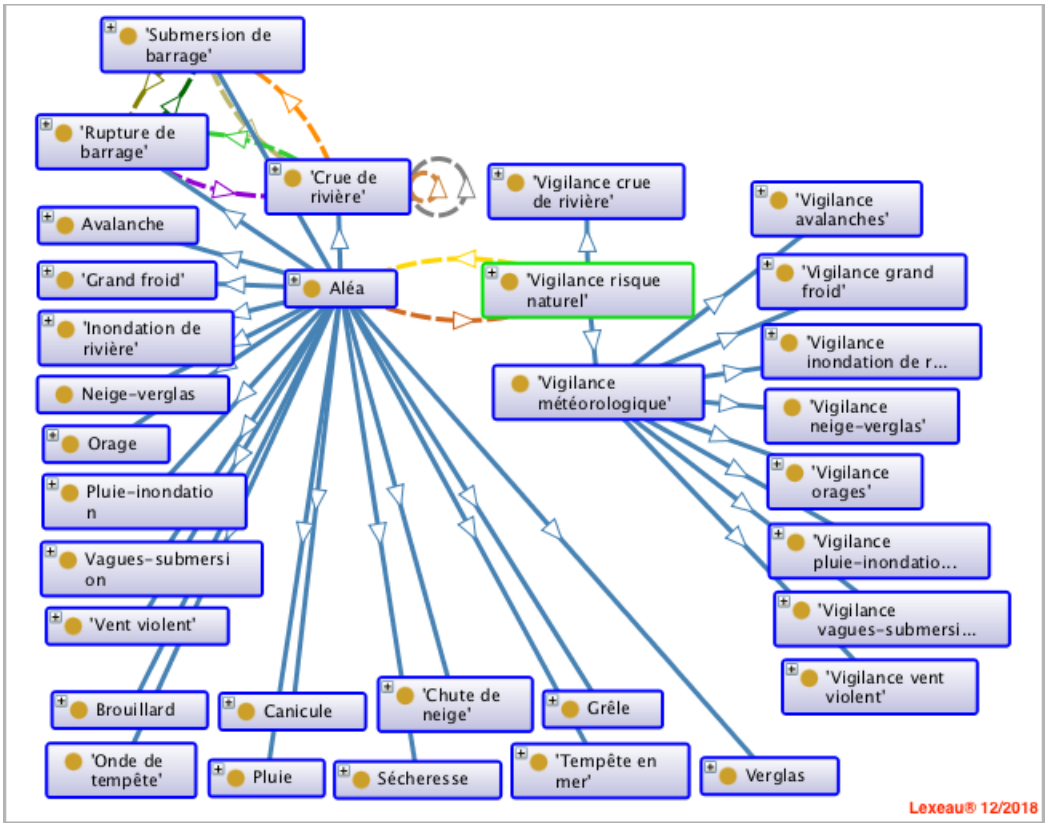


FIGURE 5.5 – Graphe relationnel des aléas et de la vigilance en matière de risques naturels

5.1.2.2 Transposition de l'ajustement conceptuel des aléas et de la vigilance

La figure 5.5 est à l'image de la transposition des concepts dans le lexique, avec, par exemple, la vigilance en matière d'inondations provoquées par une forte pluie sur un petit territoire (« 3-TA Vigilance Pluie-inondation ») en relation avec l'aléa « 1-TS Pluie inondation/ *1-TS Flash flood* ». La figure 5.6 présente une carte de vigilance publiée sur le site de Météo France [5] et les tableaux de vigilance dans les départements du Morbihan (fig. 5.7) et du Finistère (5.8), en vigilance jaune pour le risque « pluie-inondation ».

5. URL : http ://vigilance.meteofrance.com/index.html#

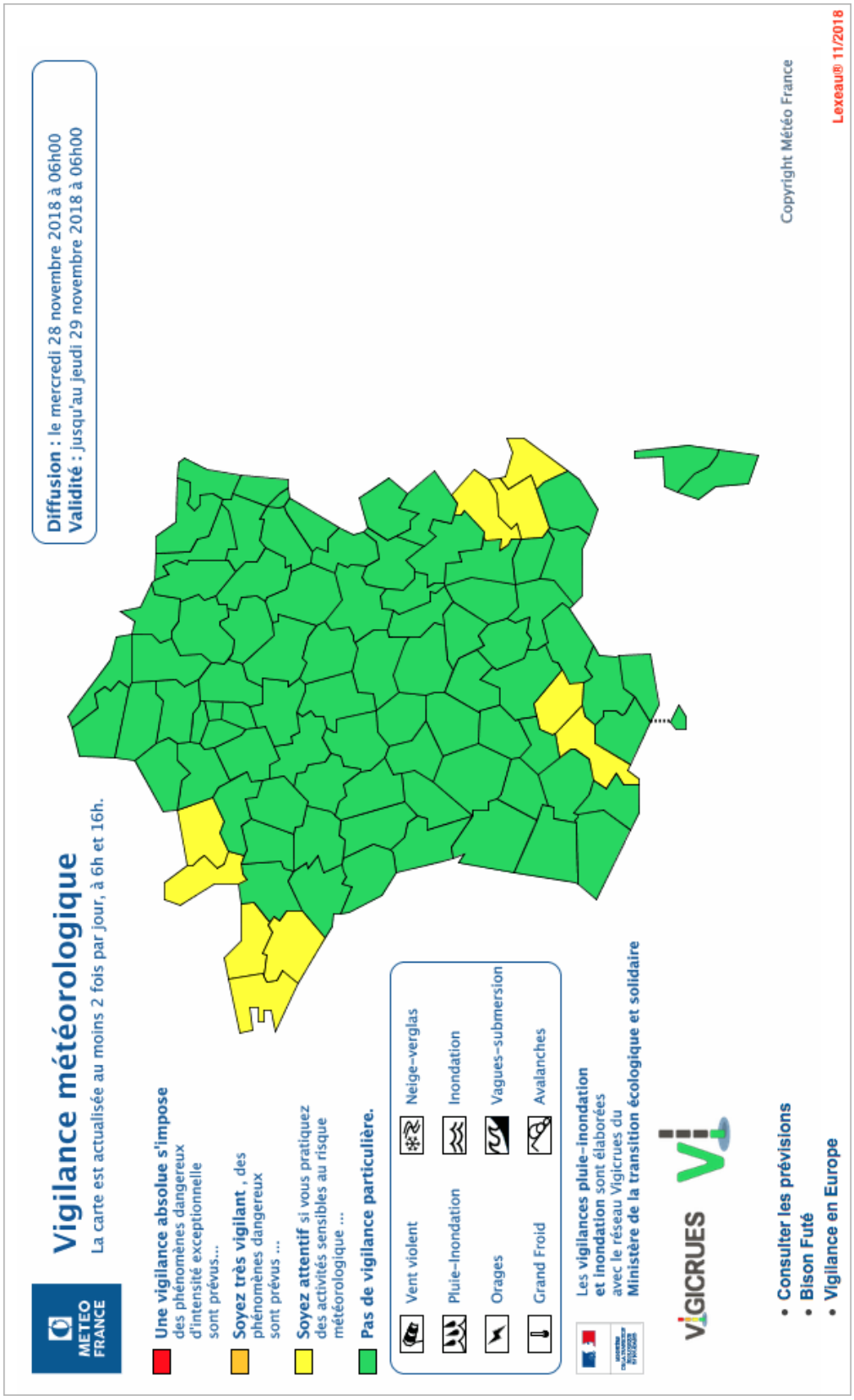


FIGURE 5.6 – Une carte de la vigilance météorologique départementale bi-quotidienne

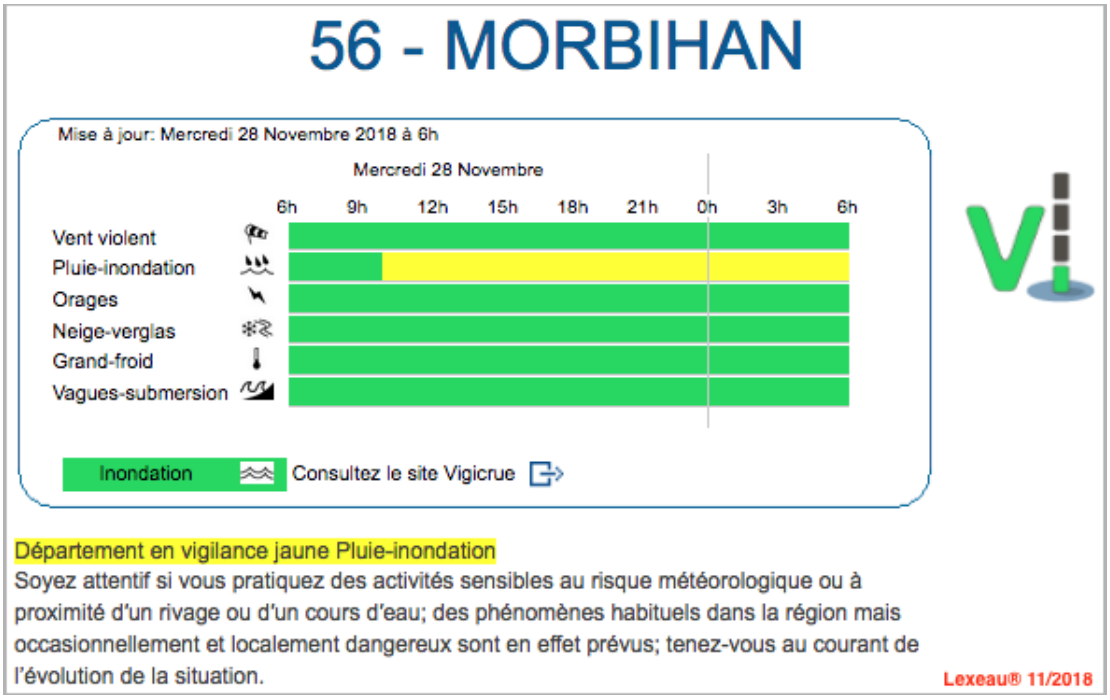


FIGURE 5.7 – Vigilance météorologique pour le département du Morbihan (extrait)

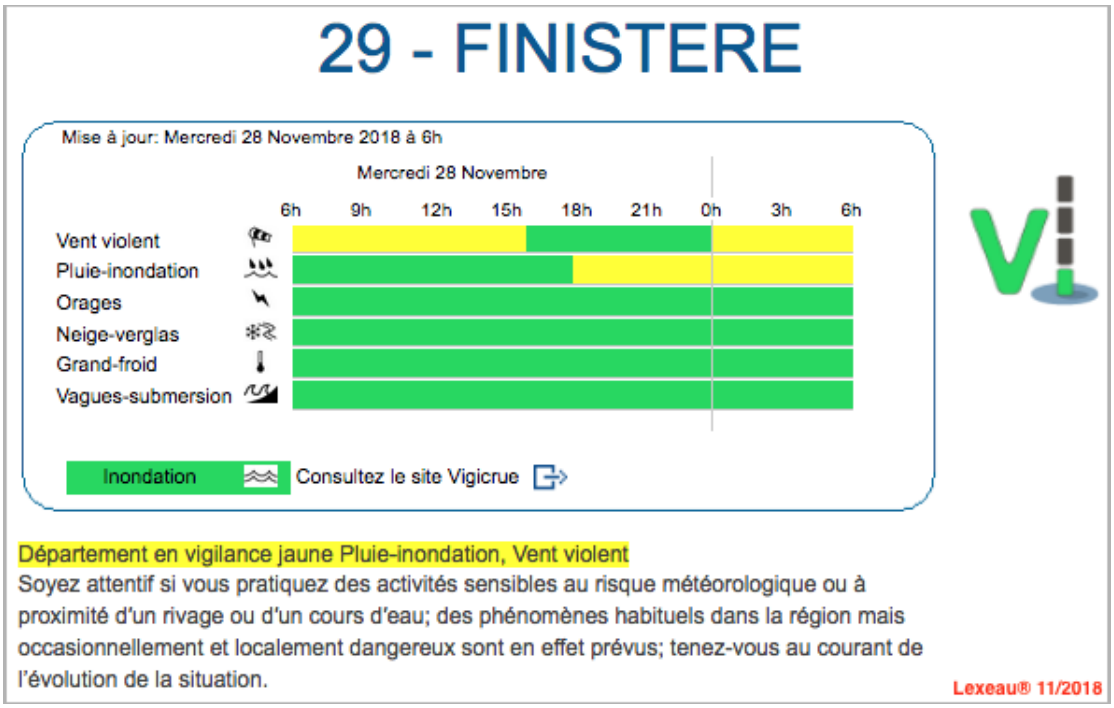


FIGURE 5.8 – Vigilance météorologique pour le département du Finistère (extrait)

Ce lien entre unités lexicales pivot n'est pas de même nature que le lien entre unités lexicales pivot et satellite, évoqué dans la section 1.5.2. Nous y reviendrons dans la section 5.2.3.

5.1.2.3 De nouveaux liens entre les crues et les données de vigilance

La vigilance « crues » se présente sur le site du SCHAPI sous la forme d'une carte de France avec les principaux cours d'eau.

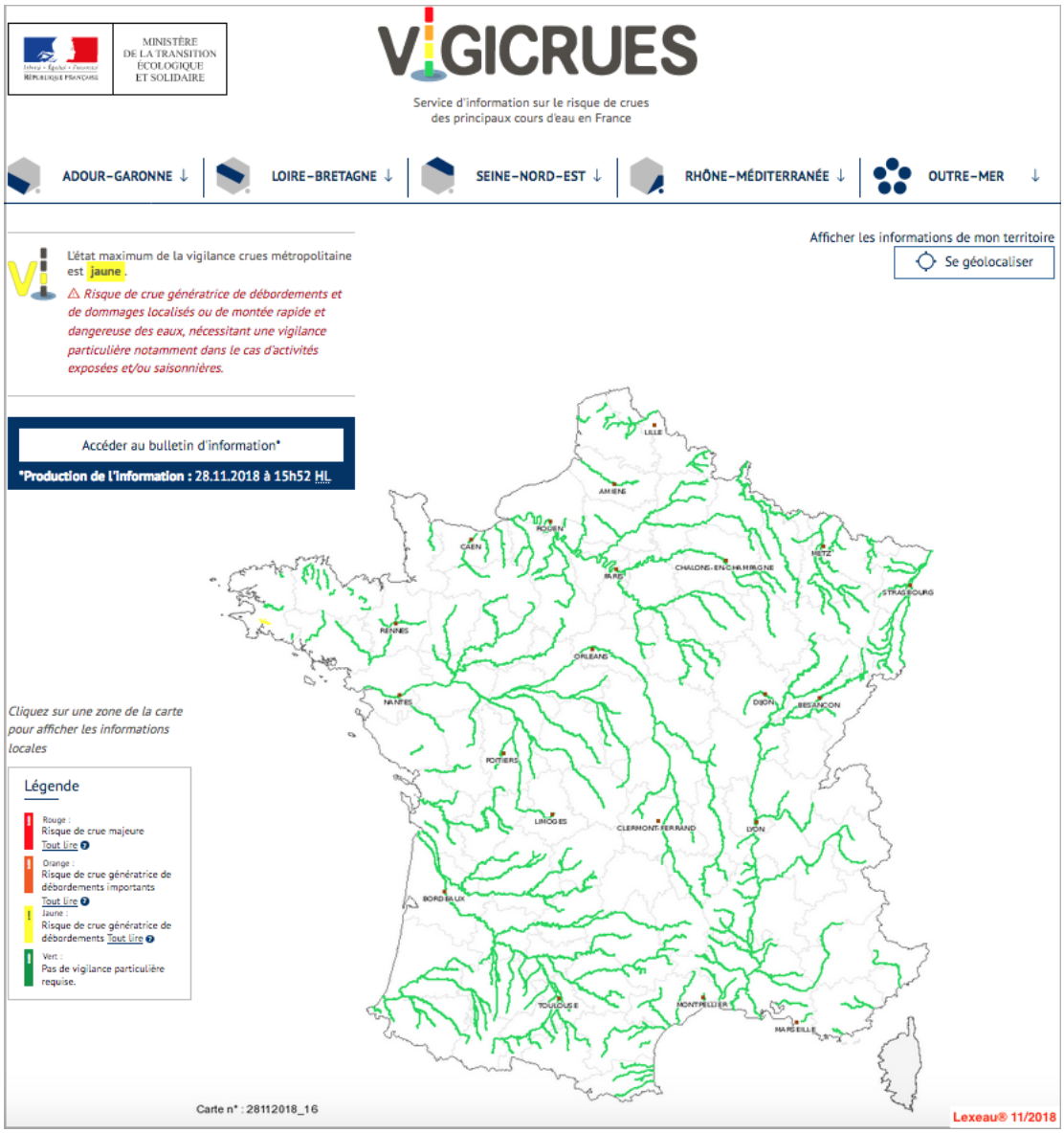


FIGURE 5.9 – Un carte de la vigilance « crues » pour le 28/11/2018

Les « tronçons vigicrues » portent la couleur du niveau de vigilance à la date d'émission de la carte. Les tronçons sont regroupés par « territoire vigicrues ». Ces territoires découpent les grands bassins hydrographiques des agences de l'eau. La vigilance est assurée dans chaque territoire par un service de protection des crues (SCP) rattaché au SCHAPI, soit un total de 22 centres de services pour 21 territoires. Chaque vigilance porte sur un « tronçon vigicrues ». La carte du territoire « Vilaine-Côtiers Bretons » est présenté figure 5.10.

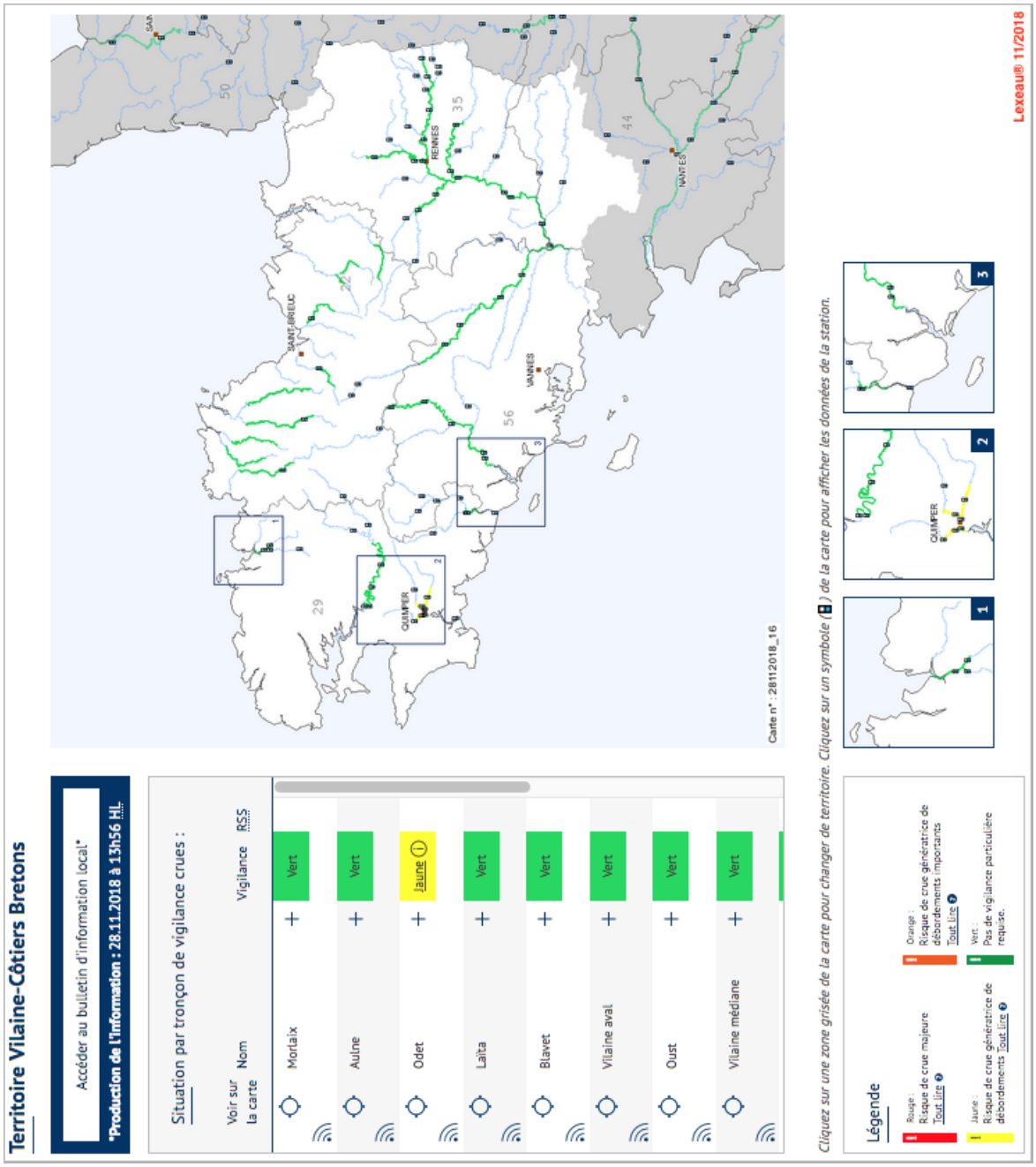


FIGURE 5.10 – Un carte de la vigilance « crues » pour le territoire « Vilaine-Côtiers Bretons »

Ces nouveaux objets de connaissance ont été introduits dans l'ontologie du domaine de façon à figurer dans le lexique. La localisation de la vigilance du risque d'inondation de rivière par tronçons vigicrues permet d'établir un modèle du débordement des tronçons en vigilance rouge dans la zone inondée enregistrée par photo aérienne ou des relevés de terrain comme un objet géographique témoin de l'inondation attestée à une date de référence et sur une période donnée. Le modèle relationnel correspondant est présenté dans la figure 5.11. Dans ce schéma, le lien entre la crue et l'inondation sont très indirects. La crue en rivière peut être reliée plus

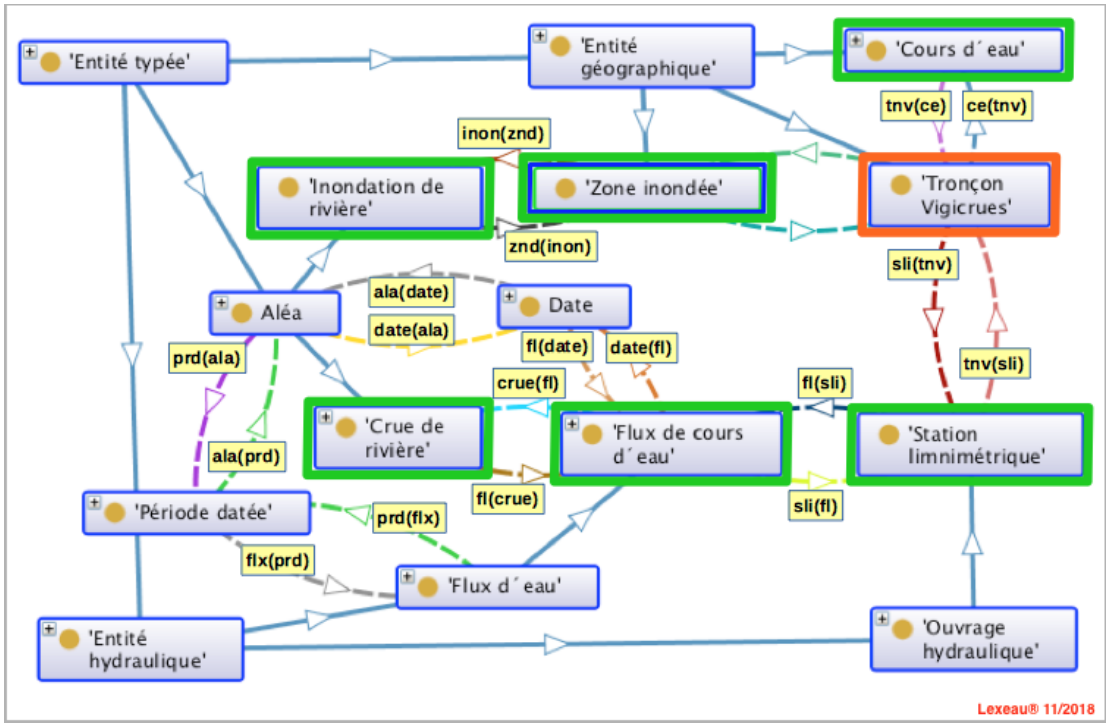


FIGURE 5.11 – Modélisation des zones inondées par débordement des tronçons vigicrues

directement à l'inondation qu'elle provoque par les tronçons hydrographiques du référentiel hydrographique français qui ont débordé au moment de son passage (fig. 5.12).

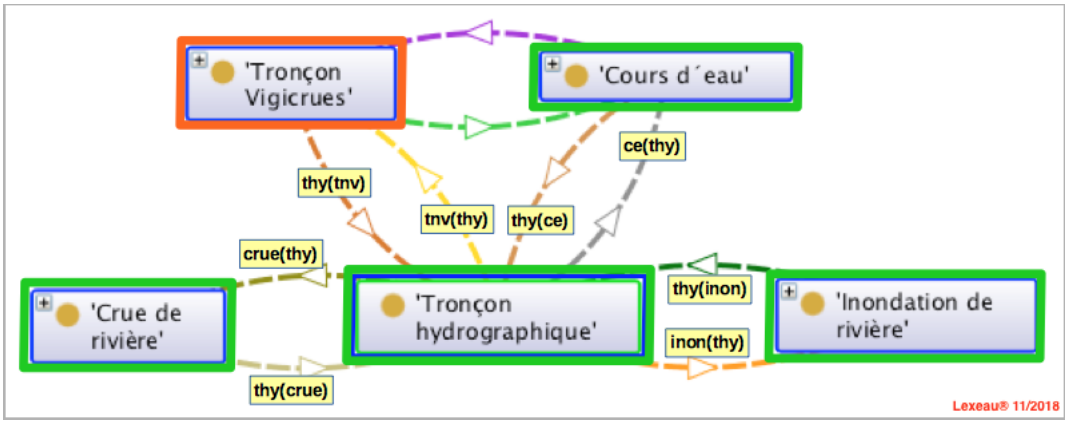


FIGURE 5.12 – Inondation de rivière par débordement de ses tronçons hydrographiques

Des liens de causalité factuelle peuvent être établis entre aléas. Nous allons voir successivement la relation entres les pluies sur un bassin versant et les crues en rivière à son exutoire, puis le cas d'une crue provoquée par la rupture d'un barrage de retenue d'eau et celui de la submersion d'un barrage par une crue qui entraine ou non sa rupture.

La crue à l'exutoire d'un bassin versant est le résultat de pluies intenses et continues sur

ce bassin sur une durée qui dépasse le « temps de concentration » du bassin, c'est à dire, schématiquement, le temps qu'il faut à une goutte d'eau tombée sur le point le plus éloigné de l'exutoire pour y parvenir. Le modèle de cette relation pluie-débit est présenté dans la figure 5.13. Il permet à l'hydrologue qui a fait le choix du temps de concentration de rapprocher

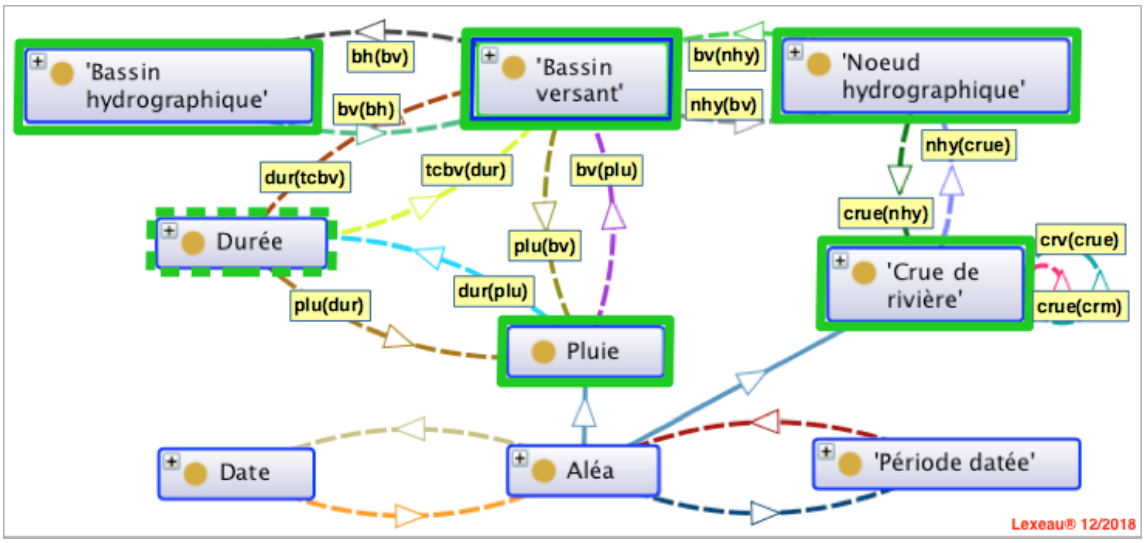


FIGURE 5.13 – Graphe du rapprochement des pluies et des crues sur un bassin versant

la hauteur d'eau tombée sur cet intervalle de temps du débit maximum de la crue observée à l'exutoire. Il obtient ainsi la période de retour d'un débit de crue donné en fonction de la période de retour de la pluie qui a provoqué la crue.

Comme nous l'avons déjà souligné, encore faut-il que l'évolution du climat sur la période d'observation des pluies, généralement plus longue que celle des crues, ce qui fait l'intérêt de la méthode, ne vienne pas mettre en défaut la courbe d'ajustement des pluies pour obtenir des hauteurs d'eau sur une période de retour qui dépasse la période d'observations.

Pour prendre un exemple « vécu », nous avons pu reconstituer vers 1975 une pluie de 140 mm en 2 heures sur un petit bassin versant dont c'est le temps de concentration, et ce à partir du débit maximum à l'exutoire calculé à partir des laisses du passage de la crue, sans submersion, dans le déversoir du petit barrage situé à l'exutoire du bassin. Cette pluie a été considérée il y a quarante ans comme déca-millénaire (durée de retour 10 000 ans), en application de la loi de Gumbel. Quelle est la durée de retour de cette pluie aujourd'hui ? Faut-il, par précaution, lui affecter une durée de retour de 1000 ans, voire 100 ans ? Sur quelles bases ? On voit tout l'intérêt d'une coopération des hydrologues pour apporter de nouvelles données et discuter de la validité des lois d'ajustement des hauteurs de pluie sur de longues périodes et du calcul des débits de crue correspondants avec un modèle simplifié de la relation « pluie-débit ».

5.1.2.4 Les catastrophes dans la strate discursive courante

Pour rendre compte des conséquences de la rupture d'un barrage, considérée comme un aléa, même si les causes de la rupture peuvent être établies *a posteriori*, nous avons introduit le concept de catastrophe (*disaster*), en lien avec les aléas impliqués dans la catastrophe. La

transposition du nom de cette nouvelle classe est effectuée dans la strate discursive courante. Le graphe relationnel de la figure 5.14 présente les liens entre les aléas concernés dans le cas de la rupture d'un barrage. Il n'y a pas de lien direct entre une crue de rivière et une inondation de rivière, faute de les situer dans le temps et dans l'espace. L'inondation de rivière figure néanmoins sur le graphe, pour rappeler les circonstances d'une catastrophe résultant de la rupture d'un barrage et de la crue que cette rupture a entrainée. La submersion d'un barrage ayant entrainé sa rupture figure également sur le graphe, ainsi que la crue à l'origine de cette rupture. Les relations de causalité ont été établies dans l'ontologie du domaine entre entités

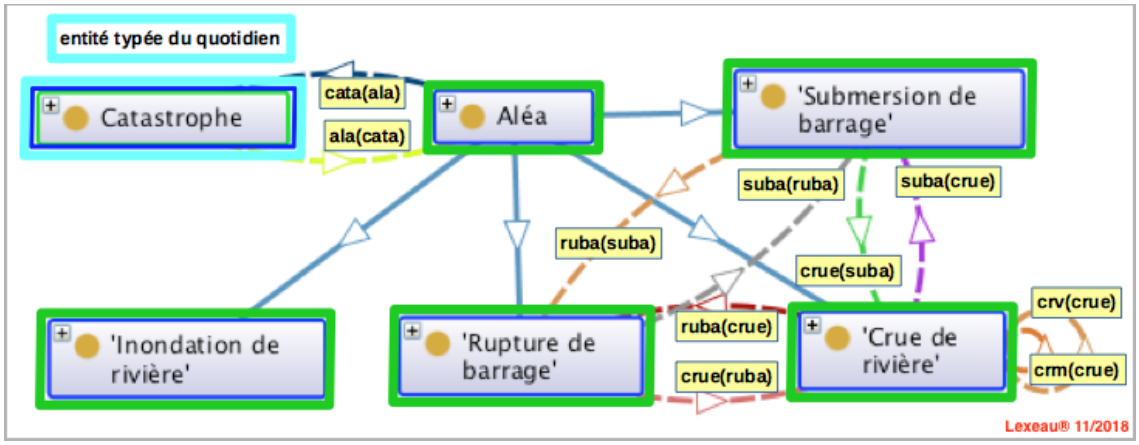


FIGURE 5.14 – Graphe relationnel d'une catastrophe résultant de la rupture d'un barrage

typées, lesquelles renvoient, dans cet exemple, à la strate discursive technico-scientifique (cf. les encadrement en vert des classes dans les figures 5.11 à 5.14).

La chronologie des événements est à considérer dans le triangle des relations conceptuelles entre la rupture d'un barrage, la crue d'une rivière et la submersion d'un barrage. Dans ce schéma, la crue d'une rivière peut être le résultat de la rupture d'un barrage sur laquelle il est construit. Elle peut se propager à l'aval dans une autre rivière. La crue d'une rivière qui alimente la retenue d'un barrage peut provoquer sa submersion si le déversoir de crue du barrage n'est pas suffisamment dimensionné. La submersion du barrage peut provoquer sa rupture, laquelle provoquera une autre crue sur la même rivière. La submersion du barrage peut provenir aussi d'un effondrement de terrain massif dans la retenue, ce qui génère une crue en aval du barrage. La catastrophe résulte de la rupture du barrage ou de sa submersion sans rupture par suite de l'inondation des zones habitées en aval. Elle portera le nom de la rupture du barrage. Dans le cas où l'inondation des zones habitées est provoquée par la crue de la rivière sans rupture ou submersion de barrage, la catastrophe reprendra le nom de la crue de la rivière. Le nom propre de la catastrophe dépend de l'enchaînement des évènements.

Après la rupture du barrage de Malpasset la question de la responsabilité humaine (une autre forme de causalité qui a des caractéristiques modales) a été tranchée par un arrêt du conseil d'état du 28/05/1971 dont un extrait est présenté dans la figure 5.15 comme exemple d'emploi du nom propre lexicalisé « Rupture du barrage de Malpasset », emploi rattaché à l'unité lexicale « 3-TA Rupture de barrage » satellite de l'unité pivot « 1-TS Rupture de barrage ». L'autre emploi présenté dans la figure 5.16 est tiré de Wikipedia [6]. Il est rattaché

6. https ://fr.wikipedia.org/wiki/Barrage_de_Malpasset#Les_causes_naturelles

à l'unité lexicale pivot. L'exemple d'emploi du nom propre de la catastrophe, dans la même source, est rattaché à l'unité lexicale « 4-C Catastrophe ».

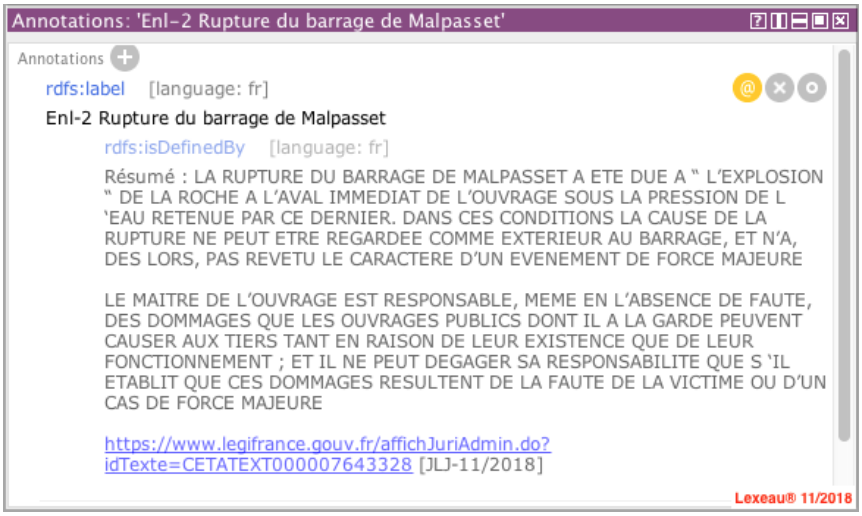


FIGURE 5.15 – Extrait de l'arrêt du Conseil d'État sur la rupture du barrage de Malpasset

FIGURE 5.16 – La rupture du barrage de Malpasset (extrait de Wikipedia)

FIGURE 5.17 – La catastrophe de la rupture du barrage de Malpasset (extrait de Wikipedia)

5.1.3 La stratification discursive des documents source

La stratification discursive des documents sources contribue au choix de la strate discursive d'une entrée lexicale tirée d'un texte ou pour valider ce choix lorsque cette entrée est issue de l'ontologie du domaine. Les documents source sont à l'origine des définitions et des exemples d'emploi du lexique. Les textes du lexique ont pour but d'éclairer le lecteur sur le sens d'un mot ou d'une expression dans sa strate discursive. Il n'y a pas lieu de stratifier les textes du lexique car ce serait redondant. Il en est de même pour les gloses qui portent sur plusieurs entrées du lexique. Elles évoquent des liens entre des entités de l'ontologie du domaine, laquelle n'obéit pas à un principe de stratification mais d'articulation des concepts entre eux. Stratifier « en amont » les corpus textuels serait contre-productif. Ce sont des regroupements de textes sur un thème et/ou dans une discipline. La liberté du regroupement permet de faire varier l'origine des documents source pour mettre en lumière les différences de sens et l'agencement des mots dans des strates discursives différentes. Une stratification des documents source est par contre porteuse de sens. Le graphe relationnel de la figure 5.18 présente le résultat de ce choix.

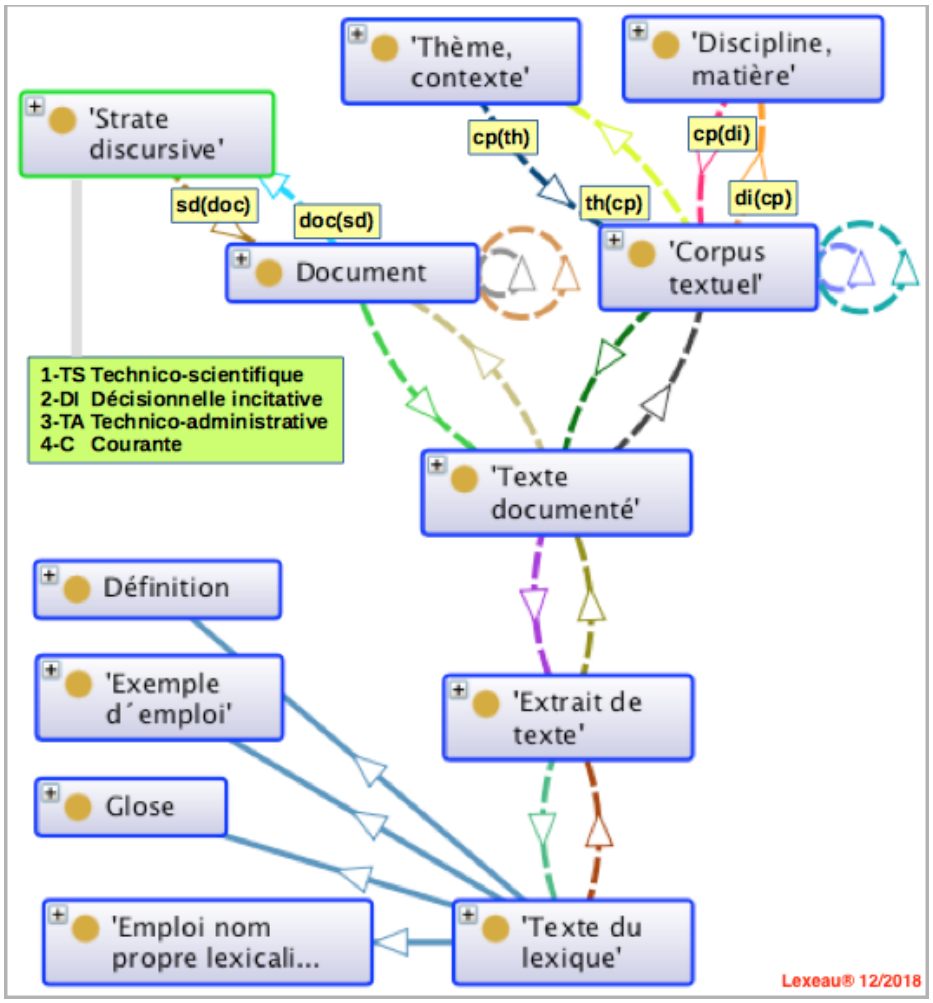


FIGURE 5.18 – Graphe relationnel des strates discursives

5.2 Concepts et unités lexicales

Dans l'ontologie du domaine, nous avons fait appel à des concepts que l'on serait tenté de qualifier de « scientifiques » — noms de classes typées ou génériques, instances de classes génériques, propriétés nommées relationnelles ou « à valeur » — avant de les transposer en unités lexicales pivot puis de déployer éventuellement des unités lexicales satellites dans d'autres strates discursives. Est-ce à dire qu'il n'y a pas d'autres concepts à l'oeuvre dans des échanges au quotidien, sur votre « grande surface » favorite, par exemple (Auchan ? Carrefour ? Leclerc ?). On retrouve ici une classe d'objets incontestable. Nous avons introduit les « PAPI » (Programme d'actions de prévention contre les inondations) comme classe de documents administratifs, dont l'élaboration est à la charge de certaines collectivités territoriales, pour en faire un inventaire et analyser les cadres successifs donnés par l'administration, en lien avec d'autres documents administratifs (les « TRI », territoires à risque important d'inondation, les « PGRI, plans de gestion des risques d'inondation, etc.), pour répondre aux injonctions de la directive européennes sur les inondations de 2007 transposée entre temps en droit français. Le concept administratif a sa place dans l'ontologie du domaine, tout comme l'acronyme a sa place comme unité lexicale pivot dans la strate technico-administrative du lexique.

Certains concepts du quotidien ont leur place dans l'ontologie du domaine, comme le concept de « catastrophe » (sans plus de précisions), dont nous avons vu les liens avec le concept d'aléa décliné en classes d'aléas, comme par exemple la rupture d'un barrage. Les concepts introduits dans l'ontologie du domaine ont pour caractéristique d'être reliés à d'autres concepts, de fonctionner en réseaux, ce qui les différencie des « idées » platoniciennes isolées (le Beau, le Bien). Les unités lexicales, à partir de la deuxième strate discursive (2DI Décisionnelle incitative), sont directement ou indirectement ancrées sur des concepts, en tant qu'unité lexicale pivot ou satellite, ce qui n'empêche pas des relations lexicales d'un type plus classique (hyperonymie/hyponymie, holonymie/méronymie) entre unités lexicales de la même strate discursive ou de strates discursives différentes.

Certains concepts dont la presse, la radio et la télévision nous parlent tous les jours comme le réchauffement climatique ou la transition écologique, peuvent être introduits par défaut dans l'ontologie du domaine de l'eau en tant que concepts du quotidien, dans la classe générique des thèmes du domaine et transposés dans la strate discursive courante. S'ils sont introduits comme concepts scientifiques, voire juridiques, et transposés dans les strates discursives technico-scientifique et technico-administrative, ramenant l'unité lexicale de la strate courante au statut d'unité lexicale satellite, il est souhaitable que des publications scientifiques ou des dispositions législatives et réglementaires (quotas individuels de consommation annuelle d'énergie fossile pour tel ou tel usage de la ressource, par exemple) donnent une épaisseur sémantique à ces unités lexicales et permettent de les définir, avec des exemples d'emploi dans les textes, en parallèle avec d'autres textes, d'autres définitions et d'autres exemples d'emploi dans les *media*. Pour le lexique comme pour l'ontologie du domaine, la réponse aux questions que l'on peut se poser est dans les textes, ce qui vaut aussi pour les concepts scientifiques, ce qu'un « scientisme » platonicien pourrait nous inciter à oublier.

Avant de faire le point sur l'état d'avancement des trois ontologies issues de la partition de la base de connaissance proposée dans la figure 1.1, la figure 5.19 présente un graphe relationnel aux frontières des trois ontologies, matérialisées par un trait discontinu, avec les relations internes et externes des classes situées à la frontière.

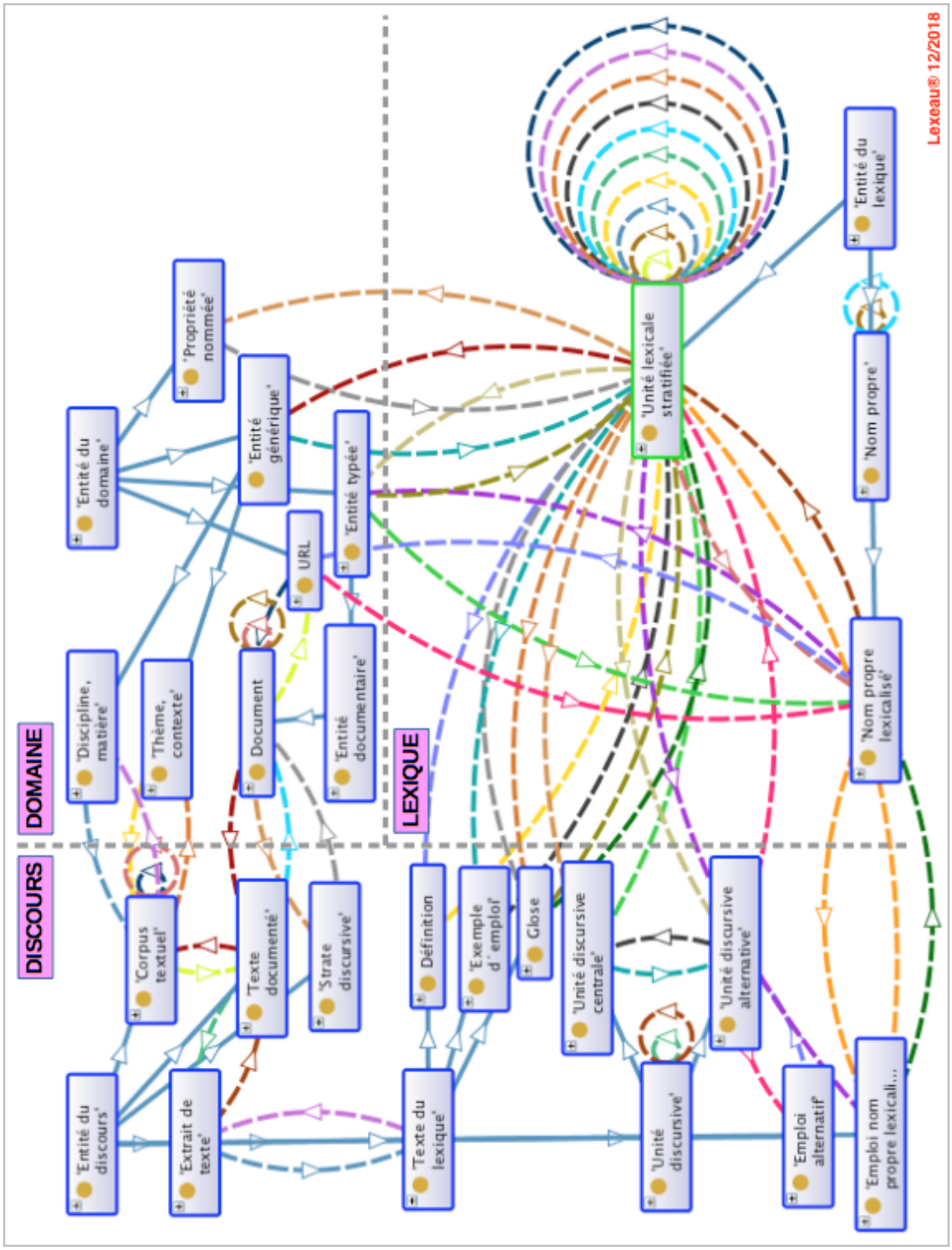


FIGURE 5.19 – Graphe relationnel aux frontières des trois ontologies

5.2.1 L'ontologie du domaine

L'ontologie du domaine vise *a priori* tous les concepts du domaine de l'eau, sous la forme du nom d'une classe, de l'instance d'une classe ou d'une propriété, et les objets concrets auxquels ces concepts font référence. L'objectif de la construction progressive des classes de l'ontologie est de mettre les concepts les plus généraux en haut des arbres de classe pour que de nouveaux concepts trouvent leur place dans l'un des arbres. Le nom d'une classe instanciée rend compte du regroupement d'objets concrets dans cette classe mais le regroupement de ces concepts permet de construire des arbres de classes sur des classes instanciées. L'apparition de nouveaux concepts peut entraîner des changements du niveau des classes dans leur arbre, voire leur contenu, lorsque l'instance d'une classe devient une nouvelle classe. Dans les classes de l'ontologie du domaine, nous avons distingué entre les objets et les entités humaines du monde matériel, que nous avons appelés des entités « typées », et les objets immatériels, comme les modalités d'une action ou une appellation générique (du blé), que nous avons appelés des entités « génériques ». La figure 5.20 présente le graphe des relations des classes primaires du domaine dans le dispositif.

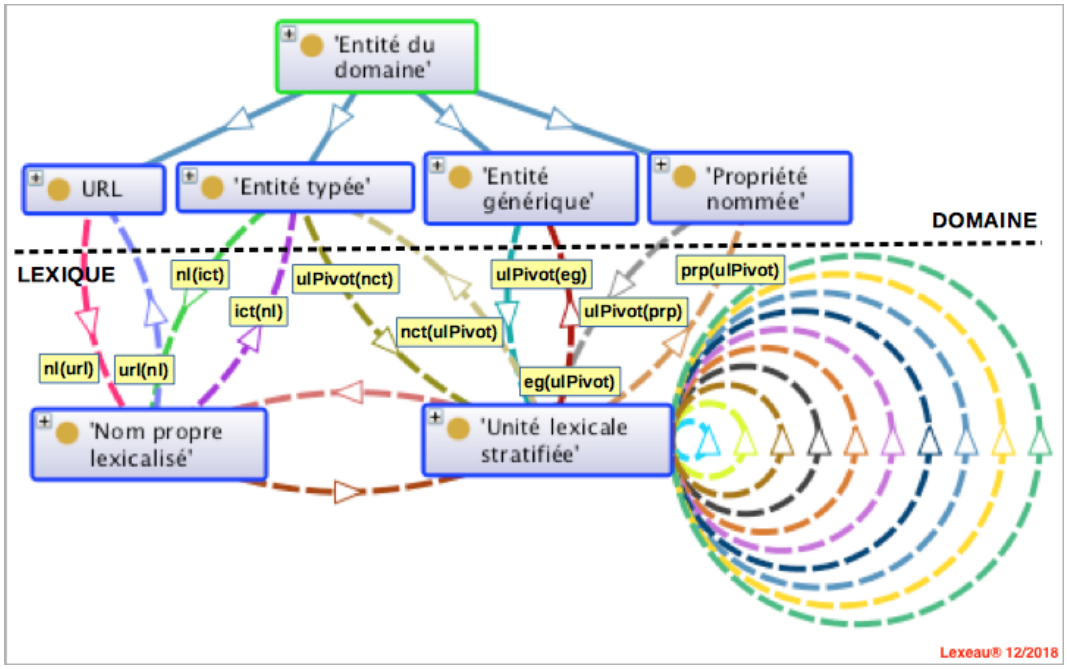


FIGURE 5.20 – Graphe relationnel des classes primaires du domaine

5.2.1.1 Les thèmes de l'ontologie du domaine

Les figures 5.21 et 5.22 présentent en deux vues l'arbre de la classe « Thème, contexte » qui est une classe d'entités génériques. Les classes non instanciées sont encadrées en vert et leurs sous-classes en orange clair. Le choix des arbres conceptuels reste arbitraire, de nature différente de celui des arbres d'« objets du monde » (les entités typées). Il joue sur le référencement des corpus sur un ou plusieurs thèmes plus ou moins spécifique, suivant son contenu (éventuellement un seul texte documenté).

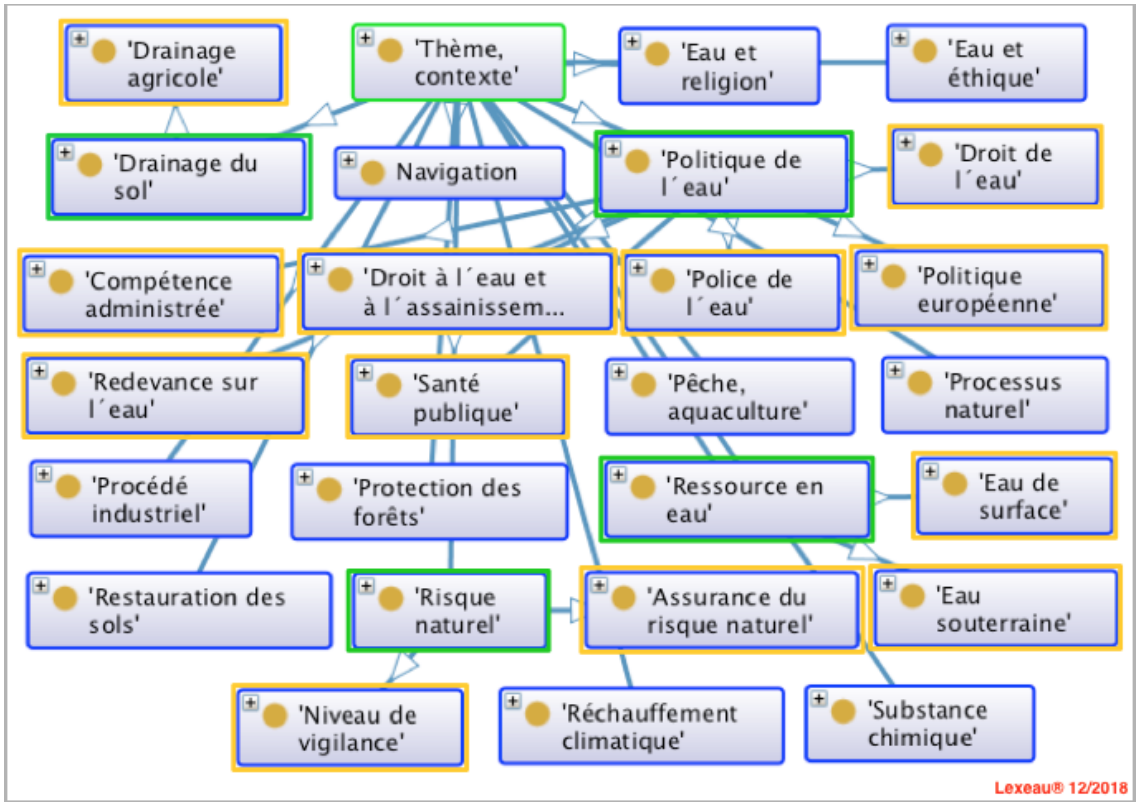


FIGURE 5.21 – Classement des thèmes du domaine (hors usages de la ressource)

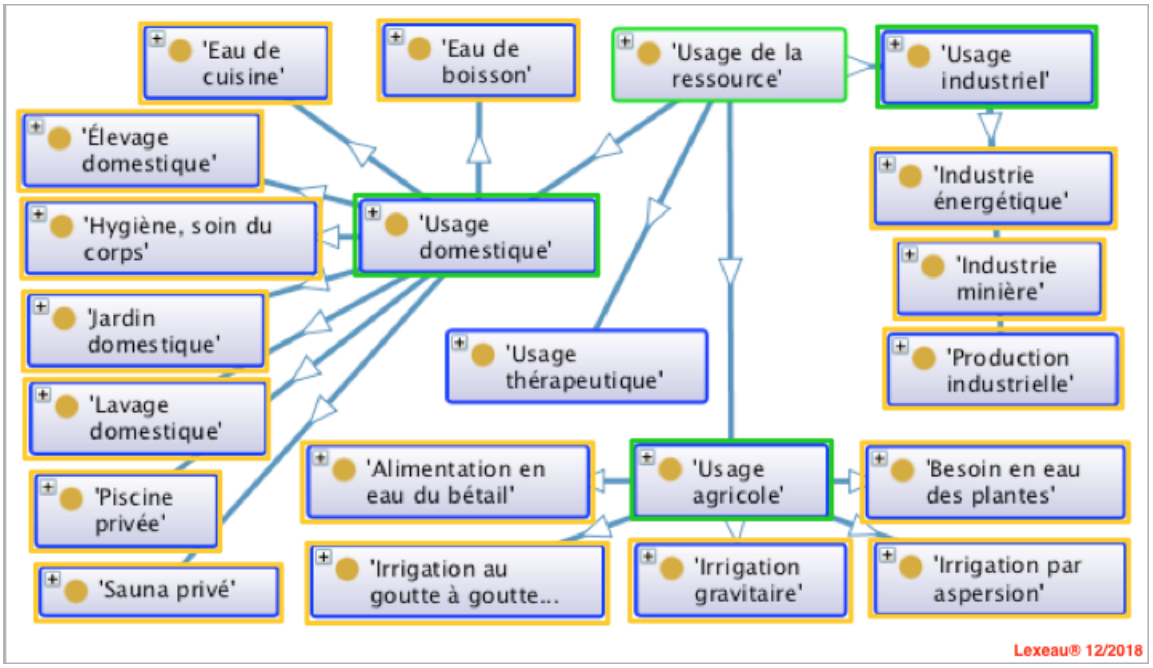


FIGURE 5.22 – Les classes du thème « Usage de la ressource »

Les thèmes sont détaillés dans les figures 5.23 à 5.32.

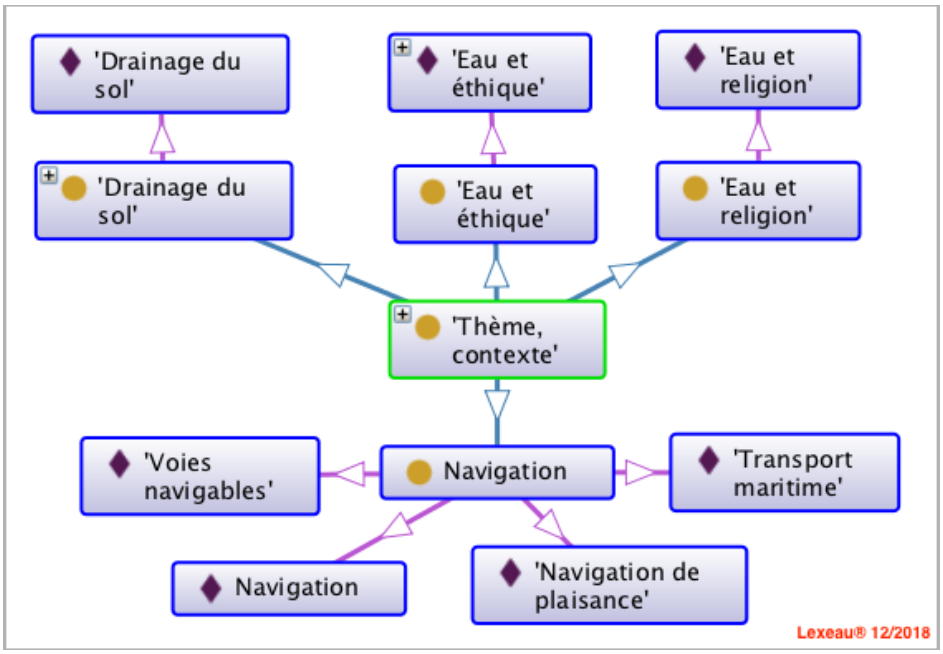


FIGURE 5.23 – Les thèmes, de « Drainage du sol » à « Navigation » (hors cultures agricoles)

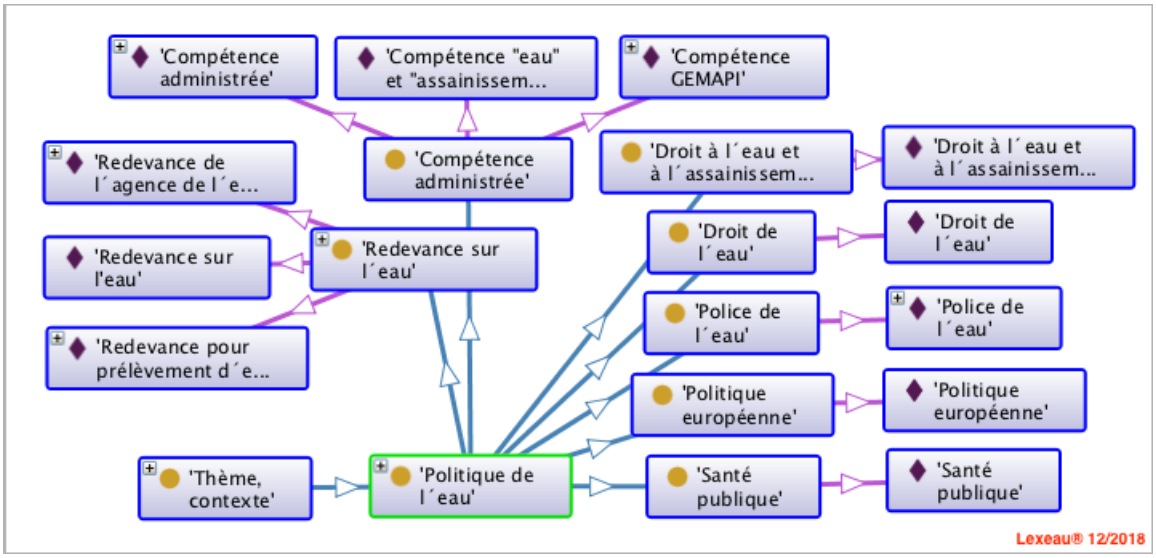


FIGURE 5.24 – Les thèmes de la classe « Politique de l'eau »

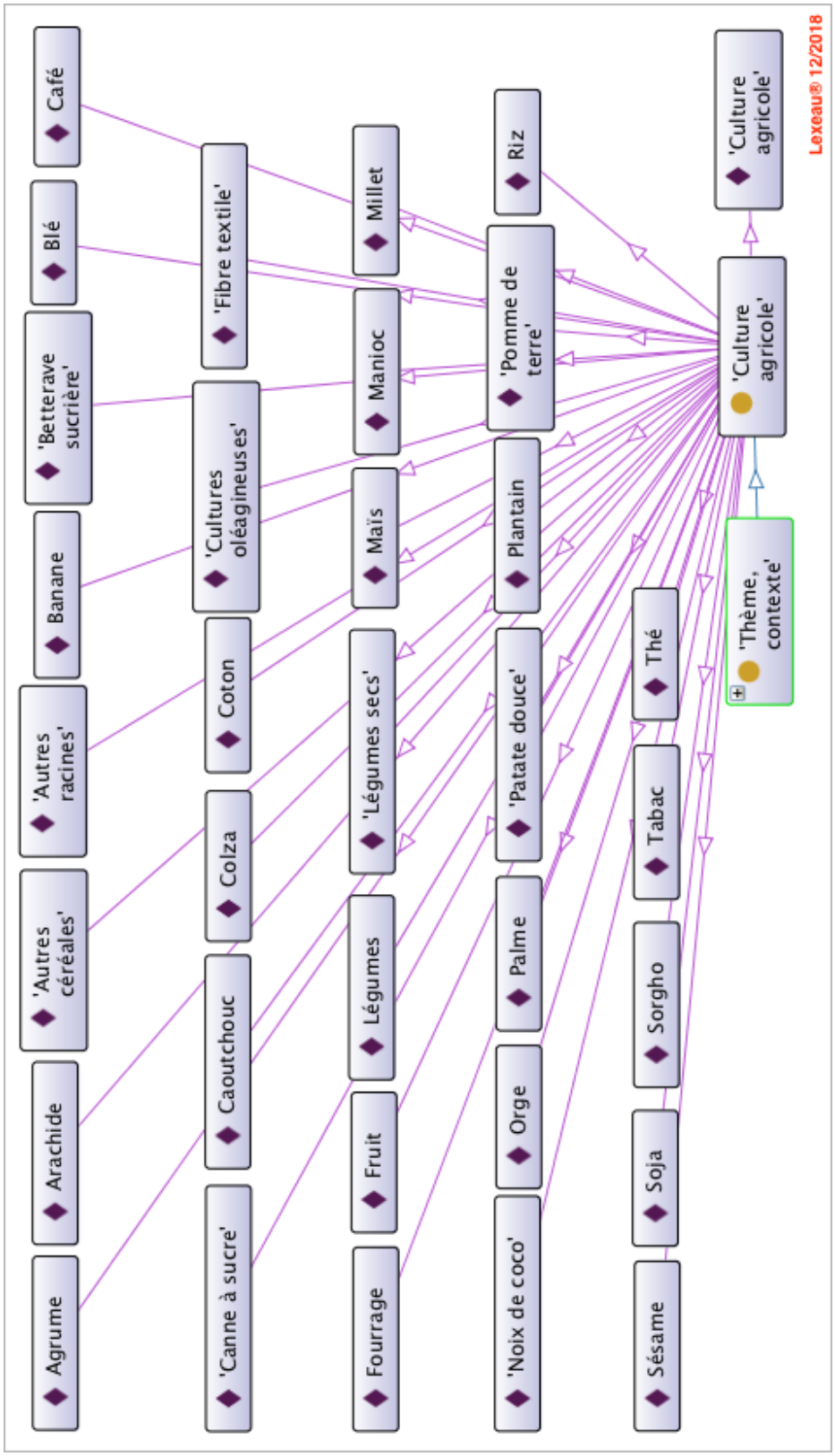


FIGURE 5.25 – Les cultures du thème « Culture agricole »

Les classes « Processus naturel » et « Procédé industriel » regroupent des thèmes transversaux (fig. 5.26). Le dessalement de l'eau de mer et le recyclage des eaux usées ont été classés parmi les procédés industriels alors qu'ils relèvent d'un nouvel usage de la ressource (eau de mer et eau usée), dans la classe « Usage de la ressource », après la mise au point des instances des deux sous-classes à créer. Les thèmes des classes « Pêche aquaculture », « Protection des forêts » et « Réchauffement climatique » sont présentés dans la figure 5.27 La thématique de

FIGURE 5.26 – Les thèmes des classes « Processus naturel » et « Procédé industriel »

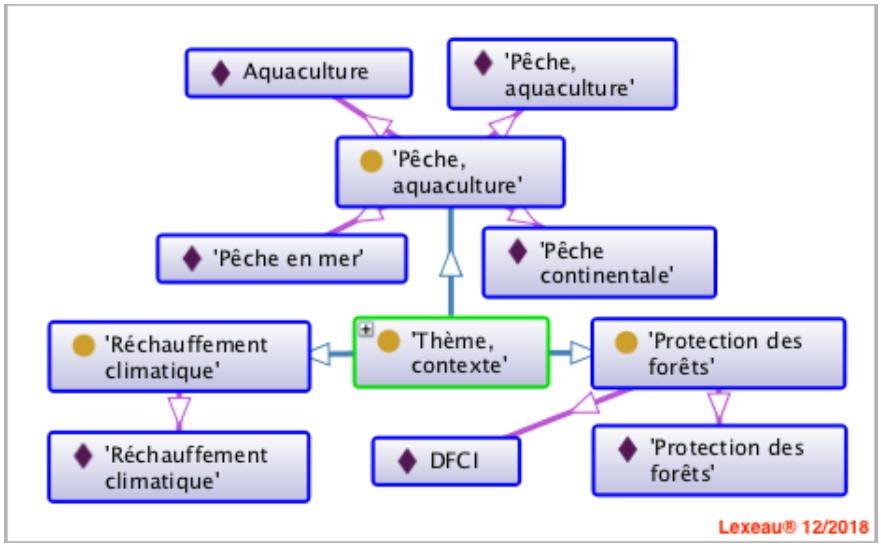


FIGURE 5.27 – Les thèmes des classes « Pêche aquaculture » à « Réchauffement climatique »

la ressource en eau, de la restauration des sols, du risque naturel et des substances chimiques est présentée dans la figure 5.28. L'acronyme RTM (Restauration des terrains en montagne) est ici dans l'attente d'une appellation plus générale dans les deux langues. L'ébauche du thème « Substance chimique » devra être revue à partir des substances identifiées dans les législations européennes et nationales, en créant éventuellement des sous-classes. Les thèmes de l'usage de la ressource sont présentés dans les figures 5.29 à 5.32.

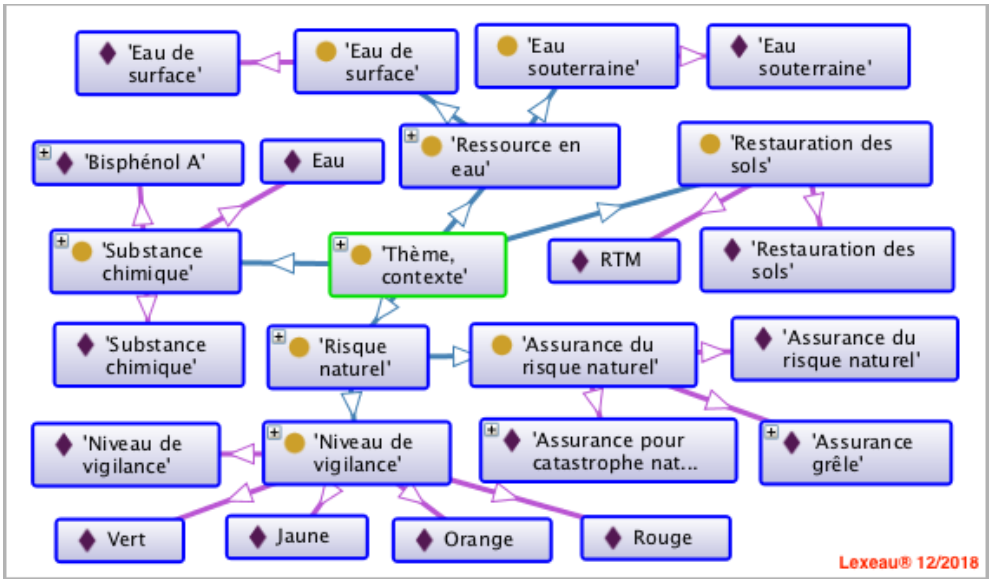


FIGURE 5.28 – Les thèmes des classes « Ressource en eau » à « Substance chimique »

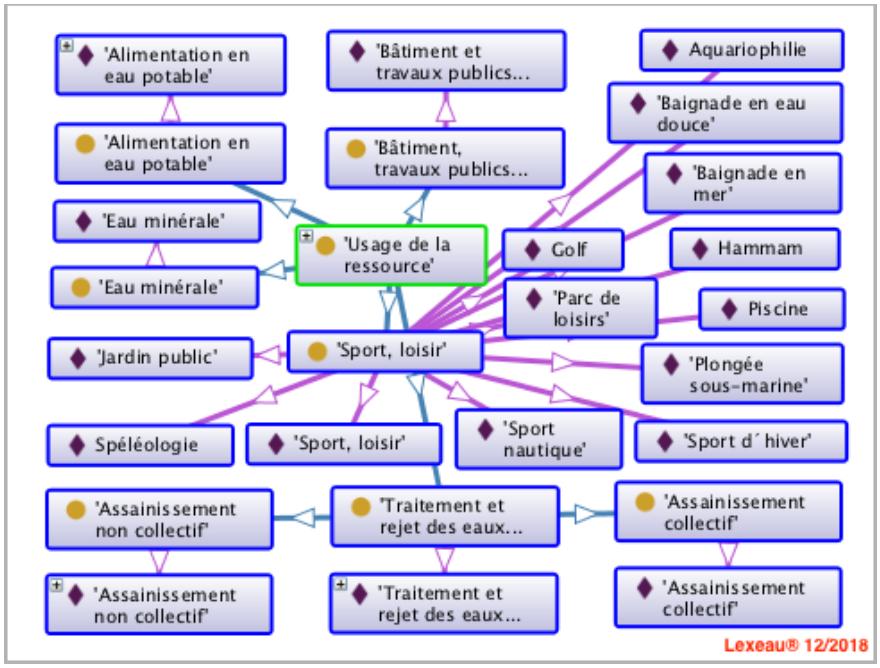


FIGURE 5.29 – Les thèmes de l'usage de la ressource, de l'AEP au traitement des eaux usées

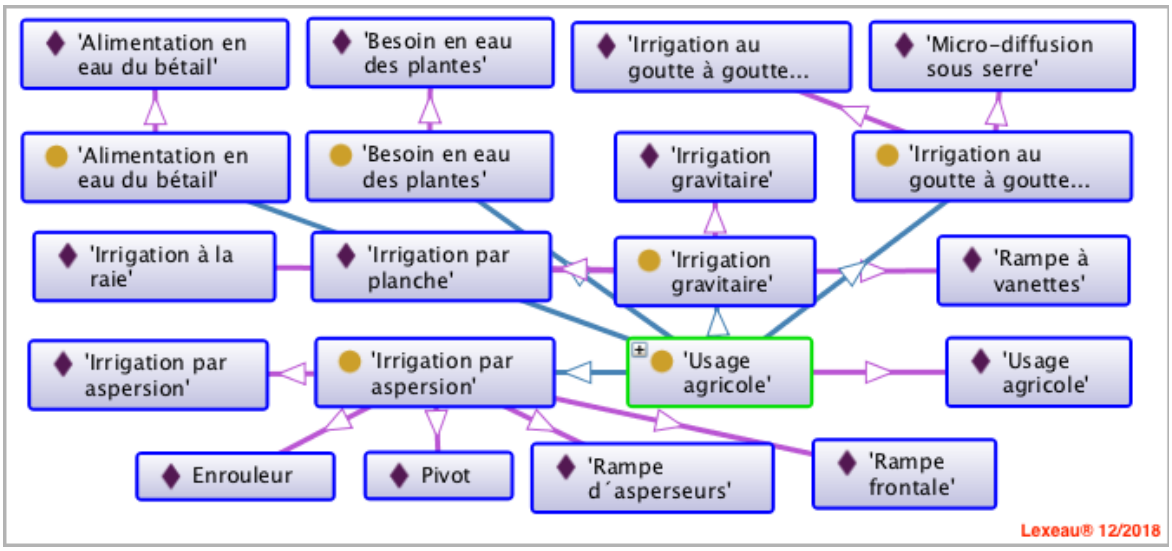


FIGURE 5.30 – Les thèmes de l'usage agricole de la ressource

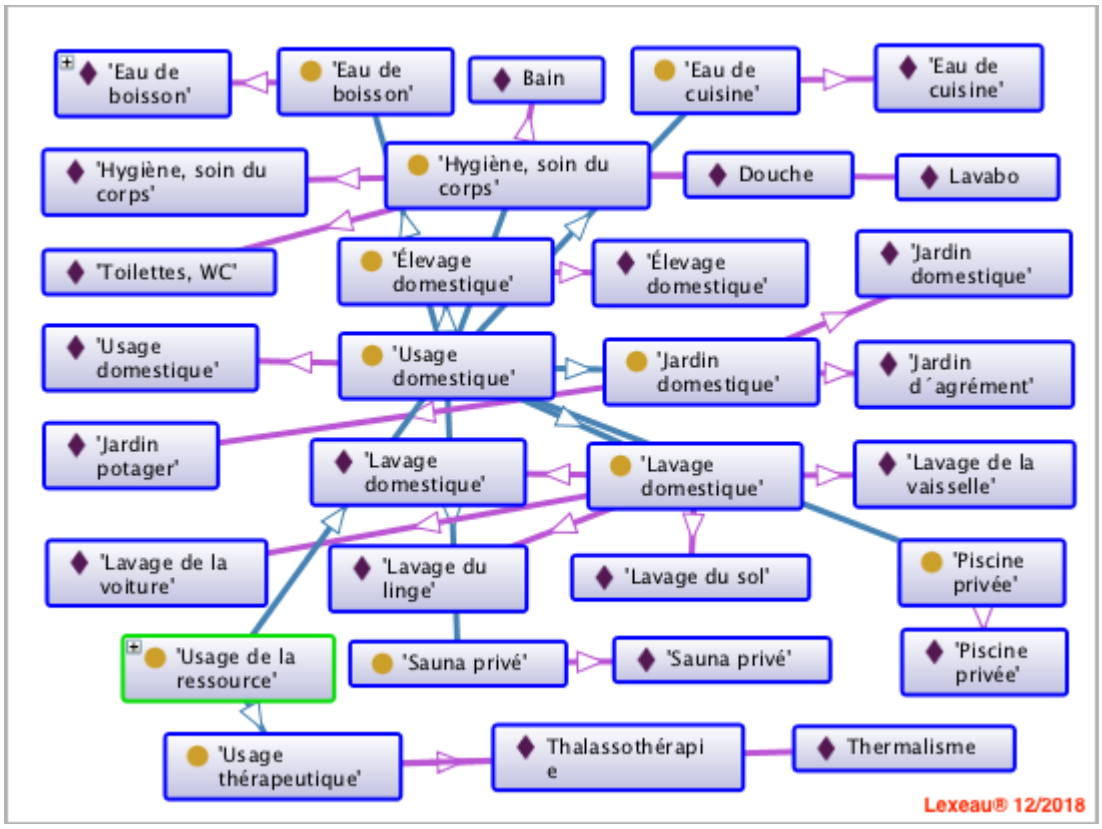


FIGURE 5.31 – Les thèmes de l'usage domestique et de l'usage thérapeutique de la ressource

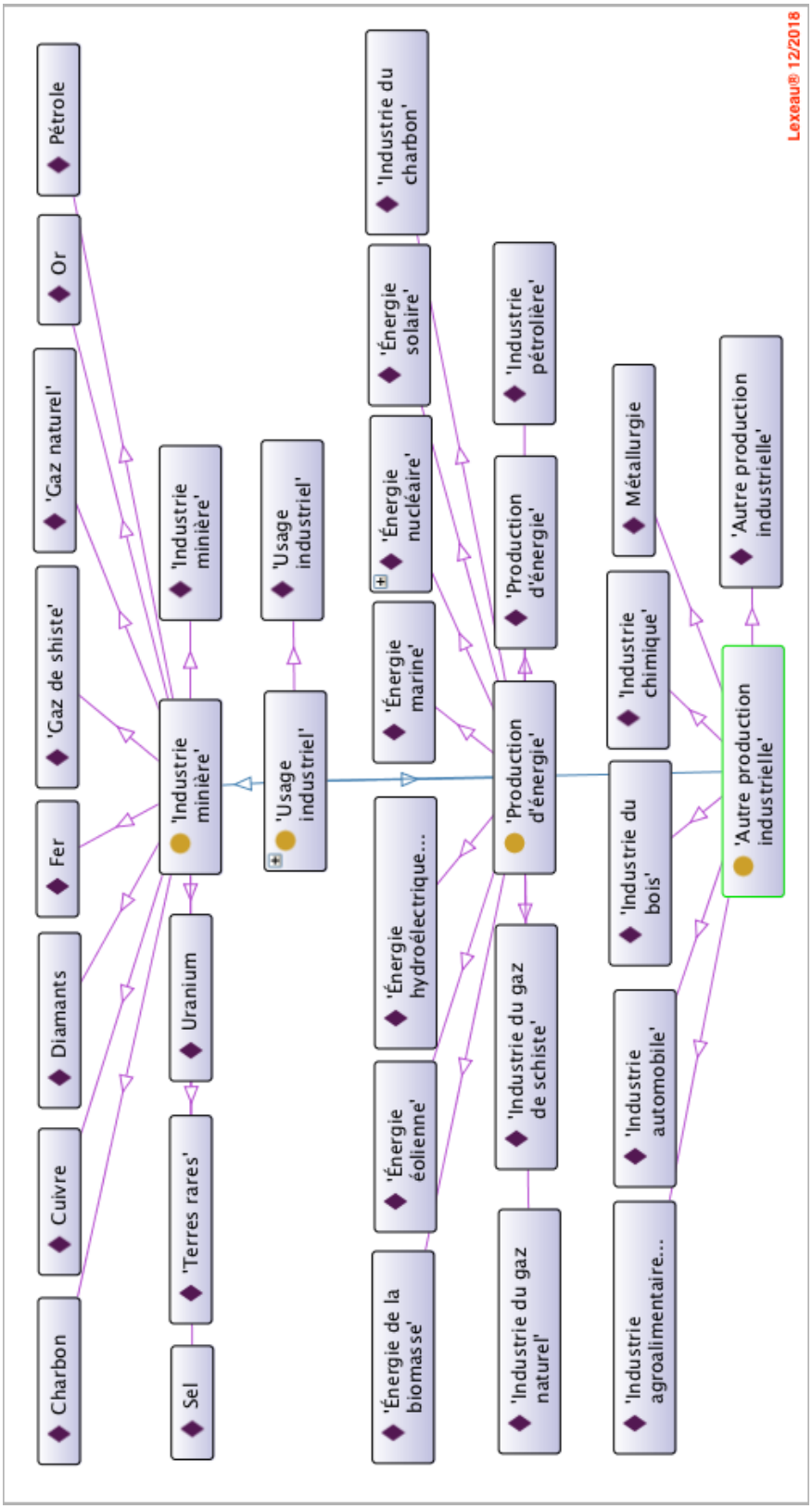


FIGURE 5.32 – Les thèmes de l'usage industriel de la ressource

5.2.1.2 Les disciplines de l'ontologie du domaine

Les disciplines ou « matières », au sens d'un cursus d'études supérieures, permettent de référencer un corpus textuel tiré des documents du domaine, à côté des thèmes, et de façon indépendante. Les disciplines apparaissent en tant qu'entités typées dans la figure 5.33 avec les relations de transposition dans le lexique comme unités lexicales stratifiées, dont nous avons fait figurer les 4 classes. Leurs liens avec des documents du domaine s'établissent à travers les textes documentés des corpus textuels qui leurs sont rattachés. Nous avons reporté la classe des URL sur la figure 5.33 pour rappeler que les documents, dont le titre est repris comme titre du texte documenté, peuvent être associés à une ressource internet, une URL [7], avec la possibilité d'un téléchargement au format *pdf*. Ce lien avec la classe des URL du domaine existe par ailleurs pour les noms propres lexicalisés issus des entités typées, que sont aussi les documents. On veillera donc à ce que les URL soient les mêmes si le titre du document est transposé dans le lexique comme nom propre lexicalisé.

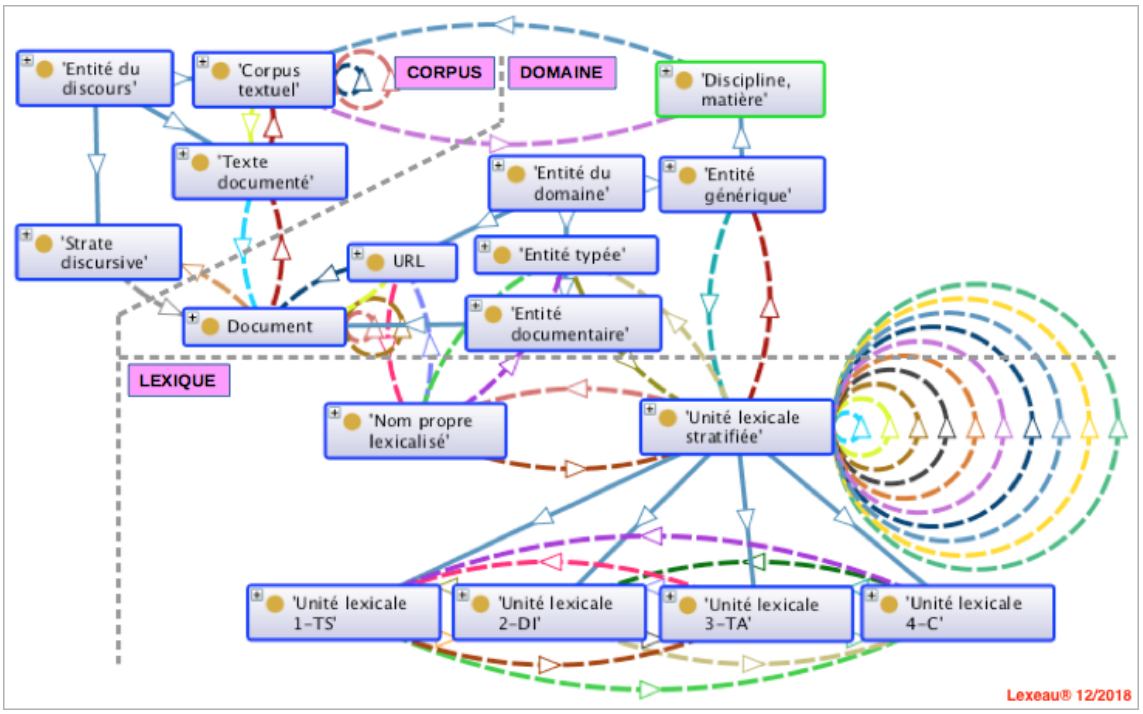


FIGURE 5.33 – Disciplines et documents du domaine en relation avec le lexique et le corpus

La nature de la discipline détermine la strate discursive dans laquelle elle est transposée dans le lexique comme unité lexicale pivot, sans exclure la possibilité théorique d'y figurer aussi comme unité lexicale satellite dans une autre strate discursive. La figure 5.34 présente une première liste de disciplines. La plupart des disciplines relèvent d'un discours sur l'eau dont la visée est scientifique, y compris en sciences humaines (anthropologie, droit, économie, géographie, histoire, linguistique, science politique, sociologie, urbanisme). Les disciplines non scientifiques de la liste ont été rattachées à la strate discursive « décisionnelle incitative » (littérature, poésie, musique, peinture, dessin, philosophie, théologie).

7. Acronyme francisé pour *Uniform resource locator*

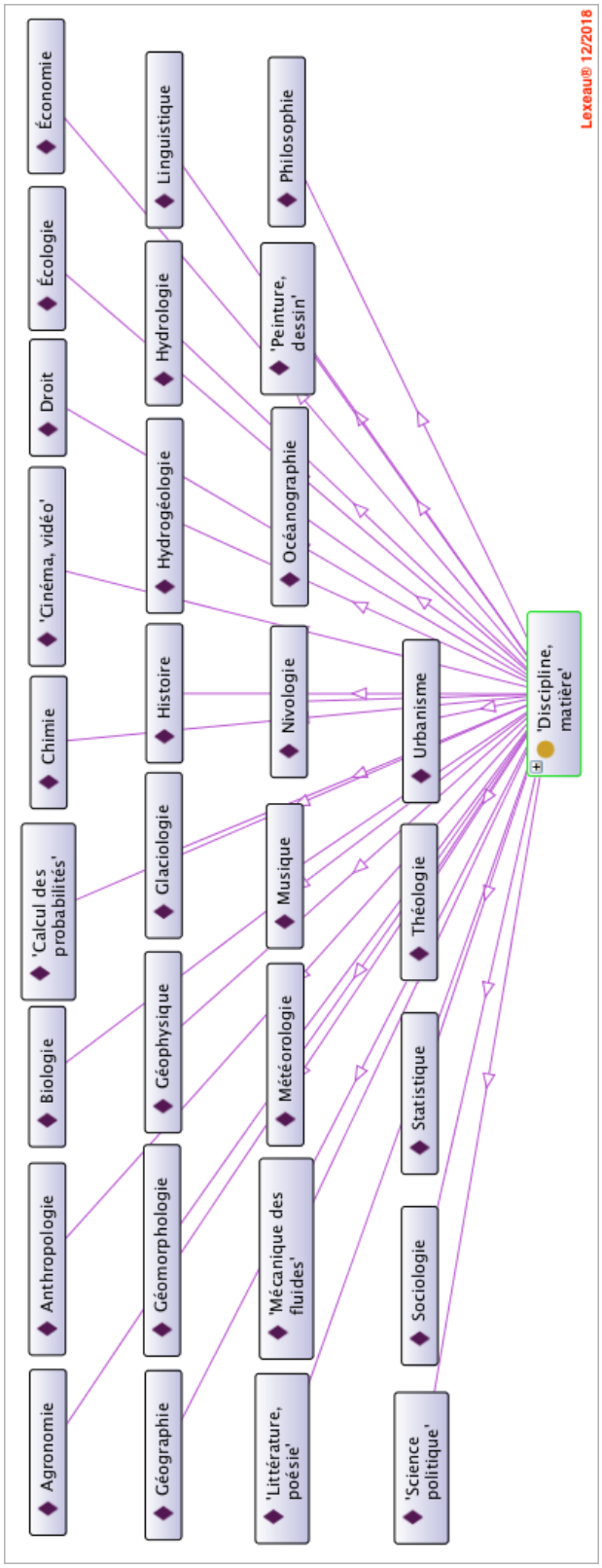


FIGURE 5.34 – Les disciplines de l'ontologie du domaine

5.2.2 Les entités typées de l'ontologie du domaine

Les entités typées du domaine sont transposées dans le lexique comme unités lexicales, pour les noms des classes ajoutés à l'identique comme individus dans les classes nommées, et comme noms propres lexicalisés, pour les instances des classes, en dehors de leurs noms. La figure 5.35 présente le graphe des relations internes et externes des entités typées primaires. On y retrouve les relations entre les catastrophes, les aléas, la vigilance sur les risques naturels et les entités géographiques concernées, présentés plus en détail dans les figures 5.3 et 5.14 de la section 5.1. On y trouve aussi le lien entre les rencontres organisées (congrès, colloques, etc.) et les organisateurs de ces rencontres.

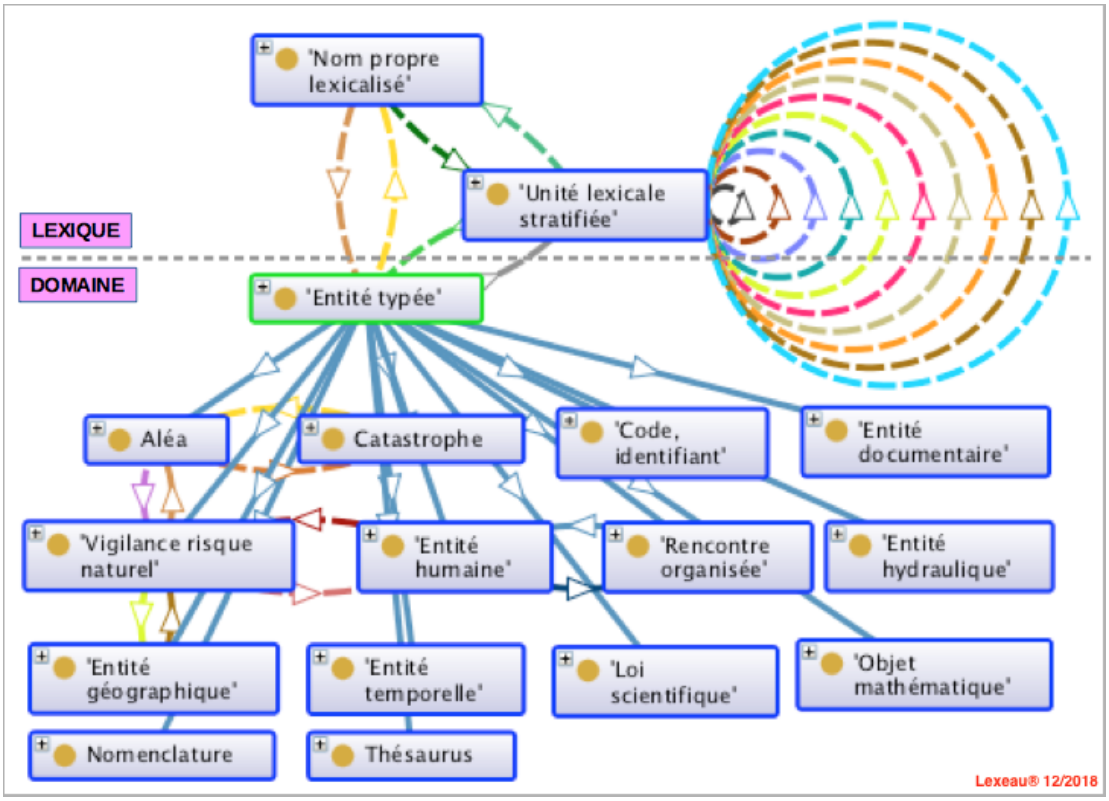


FIGURE 5.35 – Graphe relationnel des entités typées primaires de l'ontologie du domaine

Le choix des entités typées primaires a été guidé par la recherche d'une couverture aussi large que possible des objets concrets et des entités humaines du domaine de l'eau, avec le souci de les localiser dans l'espace et dans le temps dans une ontologie « située » déjà évoquée. Nous allons présenter un certain nombre de graphes relationnels de ces classes et sous-classes.

5.2.2.1 Une classe de codes et d'identifiants

Les codes et identifiants ont été créés au fur et à mesure des besoins d'identification des objets et des personnes, ou des entités humaines faute de personnalité juridique, comme le montre la figure 5.36.

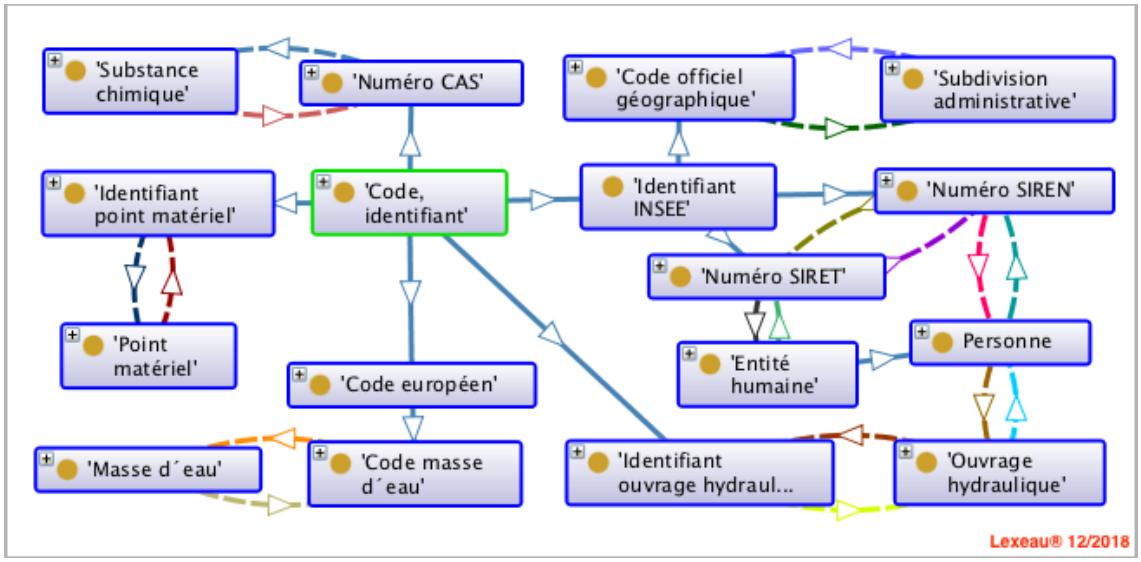


FIGURE 5.36 – Graphe relationnel des codes et identifiants

Ces codes visent certaines entités typées, mais aussi des entités génériques, comme par exemple les substances chimiques, avec le numéro CAS dont la définition tirée de Wikipedia a été reprise dans le lexique (fig. 5.37).

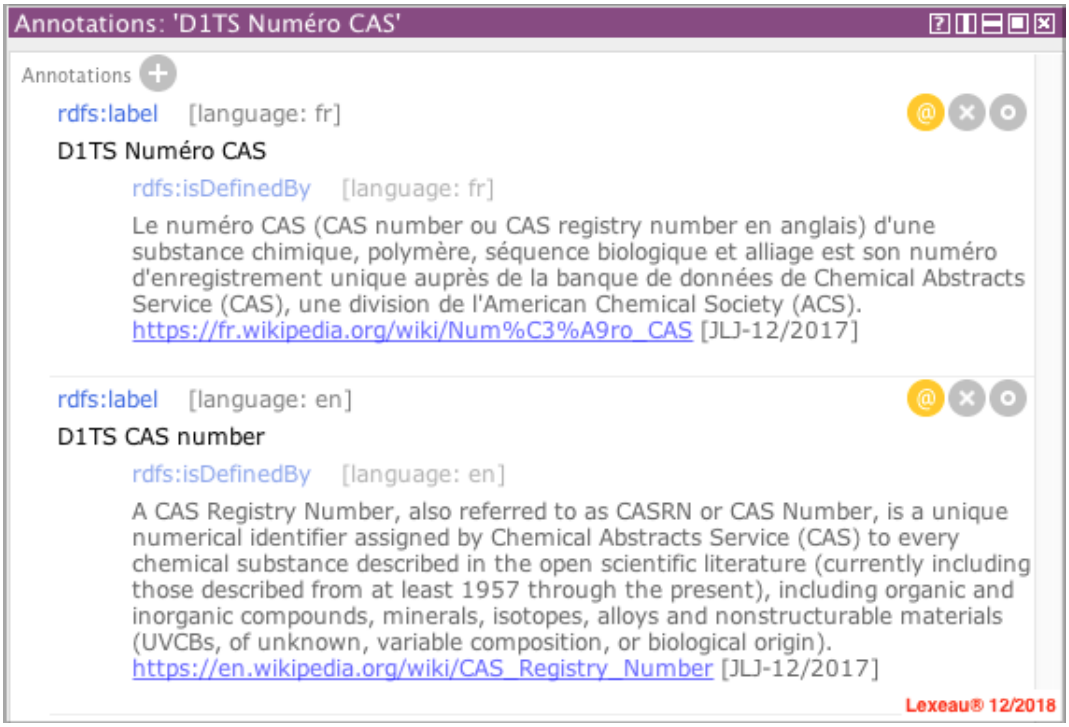


FIGURE 5.37 – Définition bilingue du numéro CAS tirée de Wikipedia

5.2.2.2 Un référentiel temporel en construction

Le référentiel temporel créé avec la classe des entités temporelles permet de mémoriser et réemployer des dates et des périodes. Le graphe relationnel est présenté dans la figure 5.38. La

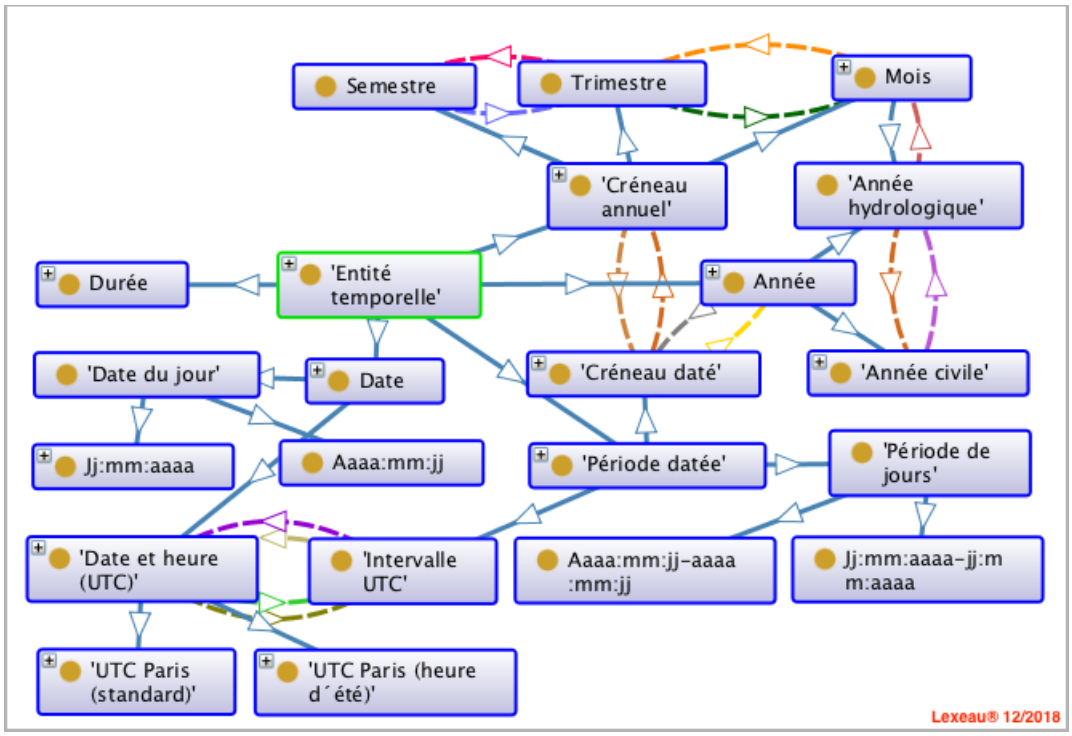


FIGURE 5.38 – Graphe relationnel du référentiel temporel en construction

classe « Date et heure » regroupe des « temps universels coordonnés » (UTC), définis dans le lexique à partir d'un article de Wikipedia (fig. 5.39).

FIGURE 5.39 – Définition bilingue du temps universel coordonné, tirée de Wikipedia

Les figures 5.40 et 5.41 présentent les formats d'écriture des UTC pour Paris, en standard et en heure d'été, extraits d'un article publié sous le titre *W3C XML Schema Part 2 : Datatypes Second Edition W3C Recommendation 28 October 2004* sur le site du *World Wide Web Consortium*.

FIGURE 5.40 – Format d'écriture d'une date et heure en UTC pour Paris (standard)

FIGURE 5.41 – Format d'écriture d'une date et heure en UTC pour Paris (heure d'été)

Le référentiel proposé est un référentiel en construction. Il peut être développé sur d'autres créneaux dans l'année (bimestre, quadrimestre, saison, semaine, décade) et sur des durées conventionnelles regroupées en sous-classes de durées en minutes, heures et jours. Il pourrait également être développé sur d'autres calendriers que le calendrier Julien.

5.2.2.3 Un référentiel géographique administratif et scientifique

Le référentiel géographique du domaine est à la fois administratif et scientifique, compte tenu de l'importance du discours administratif dans le domaine de l'eau, qui découle des lois sur l'eau 1992 et 2006 et de la transposition en droit français de la directive cadre sur l'eau de 2000, avec ses directives « filles ». La question de l'eau souterraine n'est pas vue de la même manière par les hydrogéologues, avec la notion d'entité hydrogéologique, et dans la directive cadre, avec les masses d'eau, désormais mises en oeuvre dans les systèmes d'information et dans les mesures prises par les agences de l'eau pour en améliorer l'état ou pour éviter qu'il ne se dégrade. La notion de nappe d'accompagnement qui intervient dans les procédures d'autorisation de prélèvement issues de la loi de 1992 n'a pas été reconnue pour l'instant comme une entité hydrogéologique avec une définition scientifique. La figure 5.42 présente le graphe relationnel des entités géographiques en l'état. Les relations entre les entités hydrogéologiques réparties sur trois niveaux sont issues des définitions qu'en donne le SANDRE [8] dans une fiche technique [9] Ces définitions sont reprises dans la figure 5.43.

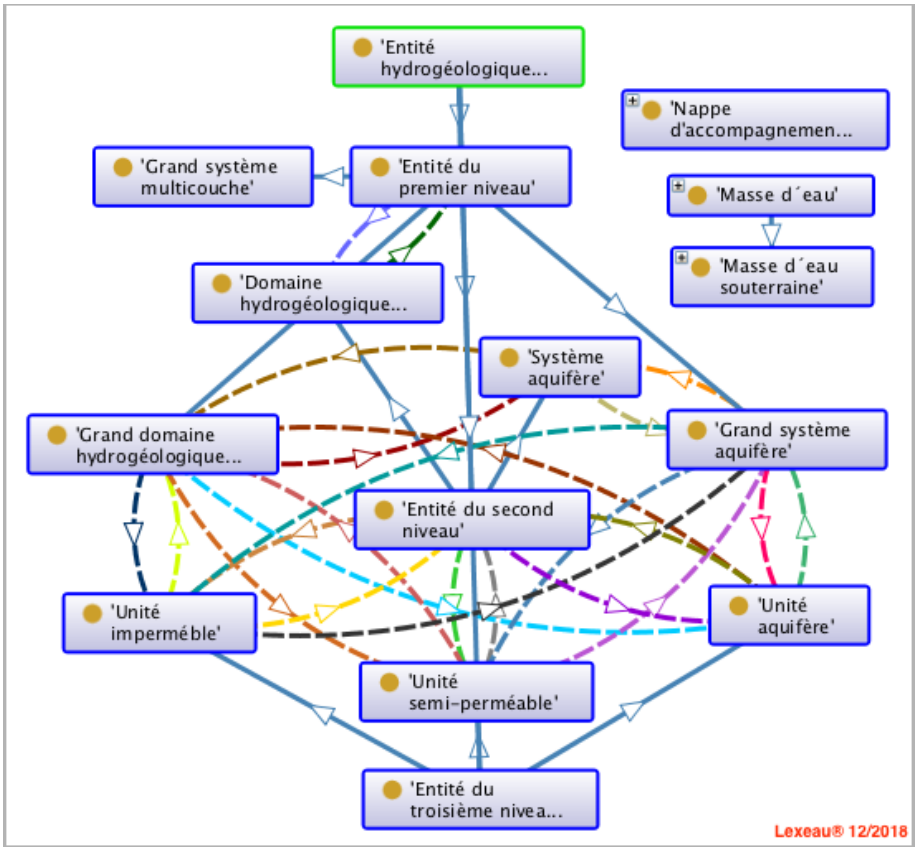


FIGURE 5.42 – Graphe relationnel des entités géographiques de l'eau souterraine

Dans un article publié sur son site [10], le BRGM annonce la sortie de la version 2 de la BDLISA (Base de Données des LImites des Systèmes Aquifères) avec des précisions importantes.

8. Service d'administration nationale des données et référentiels sur l'eau
9. http ://www.sandre.eaufrance.fr/ ?urn=urn :sandre :donnees :86 : : : : : :referentiel :3.1 :html
10. http ://www.brgm.fr/projet/referentiel-hydrogeologique-francais-bdlisa

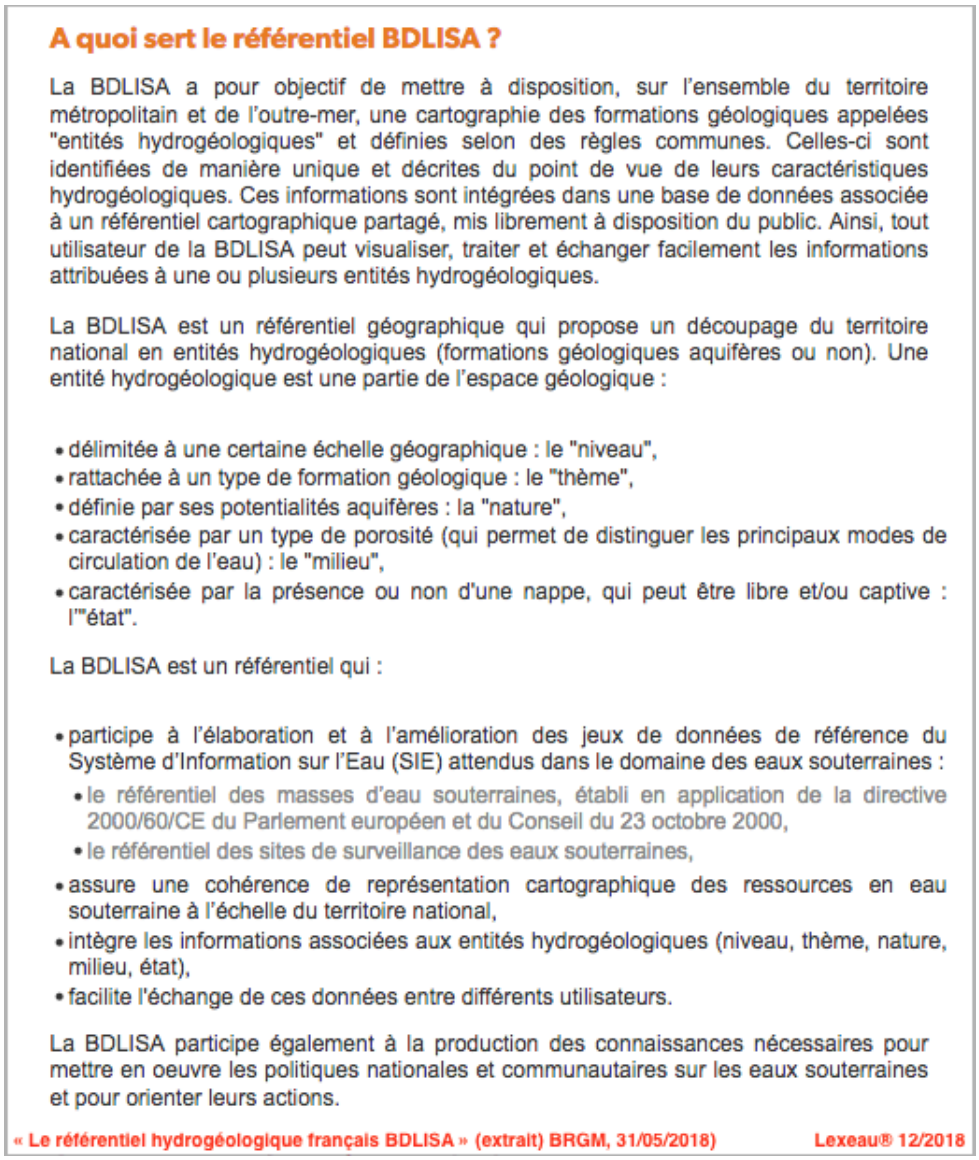
GSA	Grand système aquifère	Le grand système aquifère est un système physique composé d'une ou plusieurs unités aquifères, globalement en liaison hydraulique et qui est circonscrit par des limites lithostratigraphiques et/ou structurales. Le grand système aquifère est une entité de premier niveau.
GDH	Grand domaine hydrogéologique	Le grand domaine hydrogéologique est un système physique peu ou pas aquifère. Il peut contenir des unités aquifères mais sans grande extension latérale et isolées dans le massif imperméable. Le grand domaine hydrogéologique est une entité de premier niveau.
GSM	Grand système multicouche	Superposition du Grand système aquifère et du Grand système hydrogéologique. Le Grand système multicouche est une entité de premier niveau.
SA	Système aquifère	Un système aquifère est une entité hydrogéologique aquifère issue d'une subdivision verticale ou horizontale d'un grand système aquifère ou d'un grand domaine hydrogéologique. La subdivision s'effectue sur au moins l'un des critères suivants : lithologie, structurale, stratigraphie, piézométrie, géochimique, hydraulique. La constitution des systèmes est issue de la connaissance à un instant donné du milieu souterrain. Le système aquifère est une entité de second niveau.
DH	Domaine hydrogéologique	Un domaine hydrogéologique est une entité hydrogéologique peu aquifère issue d'une subdivision verticale ou horizontale d'un grand domaine hydrogéologique ou d'un grand système. La subdivision s'effectue sur, au moins l'un des critères suivants : lithologie,-structurale-stratigraphie-piézométrie-géochimique-hydraulique. Le domaine hydrogéologique est une entité du second niveau.
UA	Unité aquifère	L'unité aquifère est un système physique élémentaire présentant des conditions hydrodynamiques homogènes, suffisamment conductrice pour permettre la circulation de l'eau souterraine. Une unité aquifère est une entité hydrogéologique de niveau local présentant une perméabilité moyenne réputée supérieure à 10-6 m/s présentant des ressources en eau suffisante pour être exploitée. L'unité aquifère est une entité du 3ème niveau et elle correspond à la description la plus fine des entités hydrogéologiques pour le référentiel national. Ce concept résulte du découpage des domaines hydrogéologiques et des systèmes aquifères (éventuellement directement des grands domaines et des grands systèmes aquifères).
USP	Unité semi-perméable	Une unité semi-perméable est une entité hydrogéologique de niveau local présentant une perméabilité moyenne réputée comprise entre 10-9 m/s et 10-6 m/s et/ou présentant des ressources en eau mais de productivité insuffisante pour être exploitées. L'unité semi-perméable est une entité du 3ème niveau et elle correspond à la description la plus fine des entités hydrogéologiques pour le référentiel national. Ce concept résulte du découpage des domaines hydrogéologique et des systèmes aquifères (éventuellement directement des grands domaines et des grands systèmes aquifères).
UI	Unité imperméable	L'unité imperméable est un système physique élémentaire présentant des faibles circulations d'eau. Une unité imperméable est une entité hydrogéologique présentant une perméabilité moyenne réputée inférieure à 10-9 m/s.« Qualifie un milieu théoriquement impénétrable et non traversable par un fluide et en pratique ne laissant passer aucun flux significatif sous un gradient de potentiel hydraulique donné" » [Dictionnaire Hydrogéologique Français] L'unité imperméable est une entité du 3ème niveau et elle correspond à la description la plus fine des entités hydrogéologiques pour le référentiel national. Ce concept résulte du découpage des domaines hydrogéologiques et des systèmes aquifères (éventuellement directement des grands domaines et des grands systèmes aquifères).

Extrait de la fiche 86 de la nomenclature du SANDRE « Nature de l'entité hydrogéologique »

Lexeau® 12/2018

FIGURE 5.43 – Nature des entités hydrogéologiques de la BD LISA

L'extrait de cet article reproduit dans la figure 5.44 permet de comprendre la méthode utilisée pour disposer d'un référentiel qui puisse évoluer dans le temps, en distinguant les trois niveaux de représentation géographique, repris dans le graphe relationnel. L'utilisation des masses d'eau dans la politique de l'eau se fait au niveau local, ce qui correspond au troisième niveau de la BDLISA. Le lien des masses d'eau souterraines devra donc s'établir avec les unités aquifères, avec les ajustements nécessaires et en intégrant des nappes d'accompagnement, ce qui doit permettre d'harmoniser le discours scientifique et le discours administratif sur l'exploitation des eaux souterraines et le renouvellement de la ressource.



A quoi sert le référentiel BDLISA ?

La BDLISA a pour objectif de mettre à disposition, sur l'ensemble du territoire métropolitain et de l'outre-mer, une cartographie des formations géologiques appelées "entités hydrogéologiques" et définies selon des règles communes. Celles-ci sont identifiées de manière unique et décrites du point de vue de leurs caractéristiques hydrogéologiques. Ces informations sont intégrées dans une base de données associée à un référentiel cartographique partagé, mis librement à disposition du public. Ainsi, tout utilisateur de la BDLISA peut visualiser, traiter et échanger facilement les informations attribuées à une ou plusieurs entités hydrogéologiques.

La BDLISA est un référentiel géographique qui propose un découpage du territoire national en entités hydrogéologiques (formations géologiques aquifères ou non). Une entité hydrogéologique est une partie de l'espace géologique :

- délimitée à une certaine échelle géographique : le "niveau",
- rattachée à un type de formation géologique : le "thème",
- définie par ses potentialités aquifères : la "nature",
- caractérisée par un type de porosité (qui permet de distinguer les principaux modes de circulation de l'eau) : le "milieu",
- caractérisée par la présence ou non d'une nappe, qui peut être libre et/ou captive : l'"état".

La BDLISA est un référentiel qui :

- participe à l'élaboration et à l'amélioration des jeux de données de référence du Système d'Information sur l'Eau (SIE) attendus dans le domaine des eaux souterraines :
  - le référentiel des masses d'eau souterraines, établi en application de la directive 2000/60/CE du Parlement européen et du Conseil du 23 octobre 2000,
  - le référentiel des sites de surveillance des eaux souterraines,
- assure une cohérence de représentation cartographique des ressources en eau souterraine à l'échelle du territoire national,
- intègre les informations associées aux entités hydrogéologiques (niveau, thème, nature, milieu, état),
- facilite l'échange de ces données entre différents utilisateurs.

La BDLISA participe également à la production des connaissances nécessaires pour mettre en oeuvre les politiques nationales et communautaires sur les eaux souterraines et pour orienter leurs actions.

« Le référentiel hydrogéologique français BDLISA » (extrait) BRGM, 31/05/2018) Lexeau® 12/2018

FIGURE 5.44 – Présentation de la BDLISA sur le site du BRGM (extrait)

Le graphe relationnel des entités géographiques est présenté dans la figure 5.45. Les masses d'eau rivière, lac, de transition et côtières, qui sont des masses d'eau de surface, et les entités hydrogéologiques en dehors des unités aquifères n'ont pas été reportées.

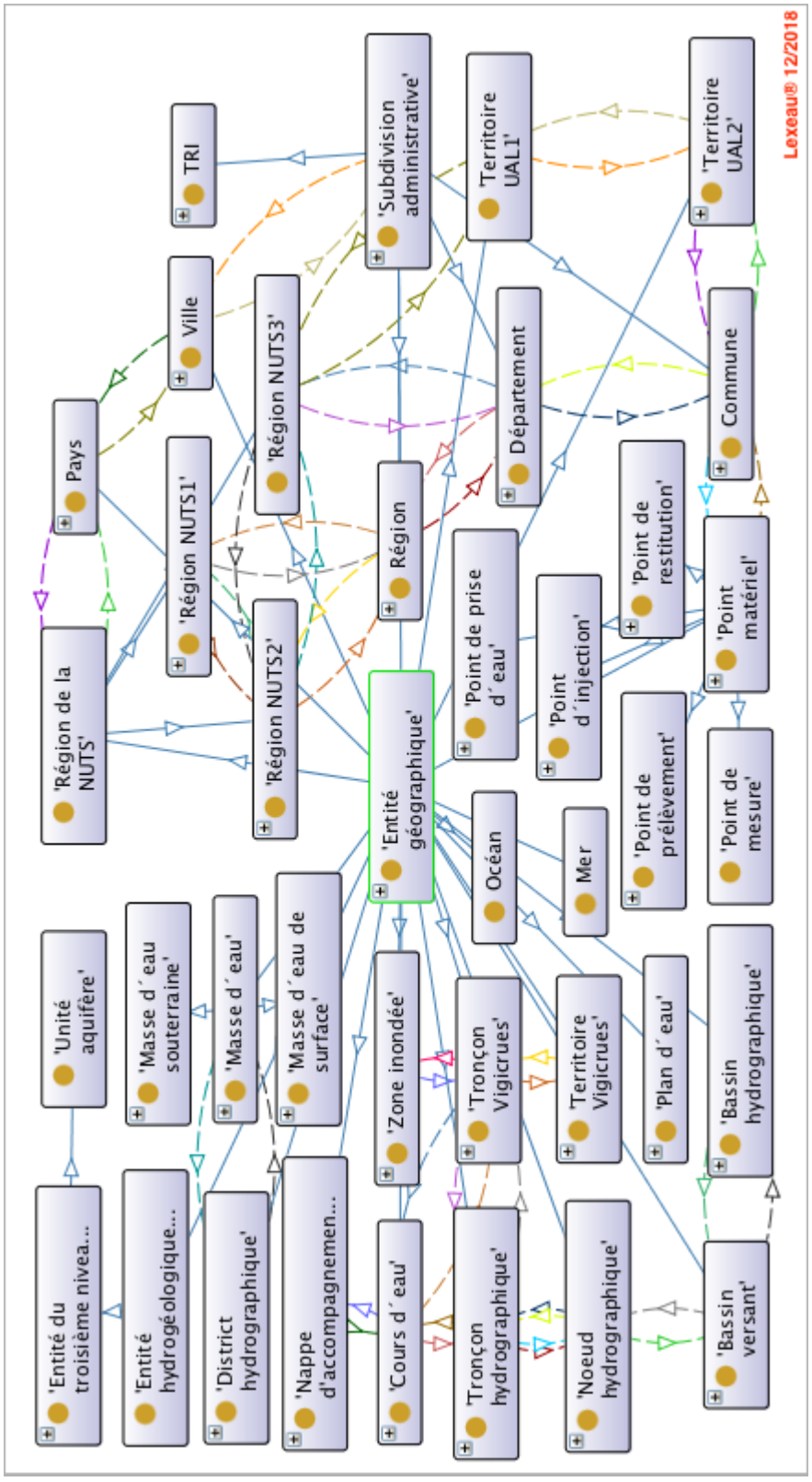


FIGURE 5.45 – Graphe relationnel des entités géographiques

Les régions de la NUTS (nomenclatures des unités territoriales statistiques) correspondent à un découpage économique du territoire européen, établi en accord avec les autorités nationales sur la base des subdivisions administratives, pour publier des statistiques locales élaborées par EUROSTAT (l'office des statistiques de l'union européenne) et localiser les crédits d'intervention des fonds européens. Les NUTS 3 correspondent aux départements. Les territoires UAL1 et UAL2 (unités administratives locales de niveau 1 et 2) correspondent à un découpage plus fin, avec un lien des UAL1 sur les NUTS 3. Le niveau UAL2 correspond aux communes. Les TRI sont des territoires à risque important d'inondation qui regroupent des communes entières. Ils ont été délimités après la transposition en droit français de la directive inondation de 2007. Les pays introduits comme entités géographiques comprennent les pays membres de l'union européenne. Les villes ont été introduites par pays, avec un lien pour la France en tant comme chef-lieu d'une subdivision administrative. La figure 5.46 tirée de Wikipedia [11] présente la carte des bassins DCE et des districts hydrographiques pour la France métropolitaine.

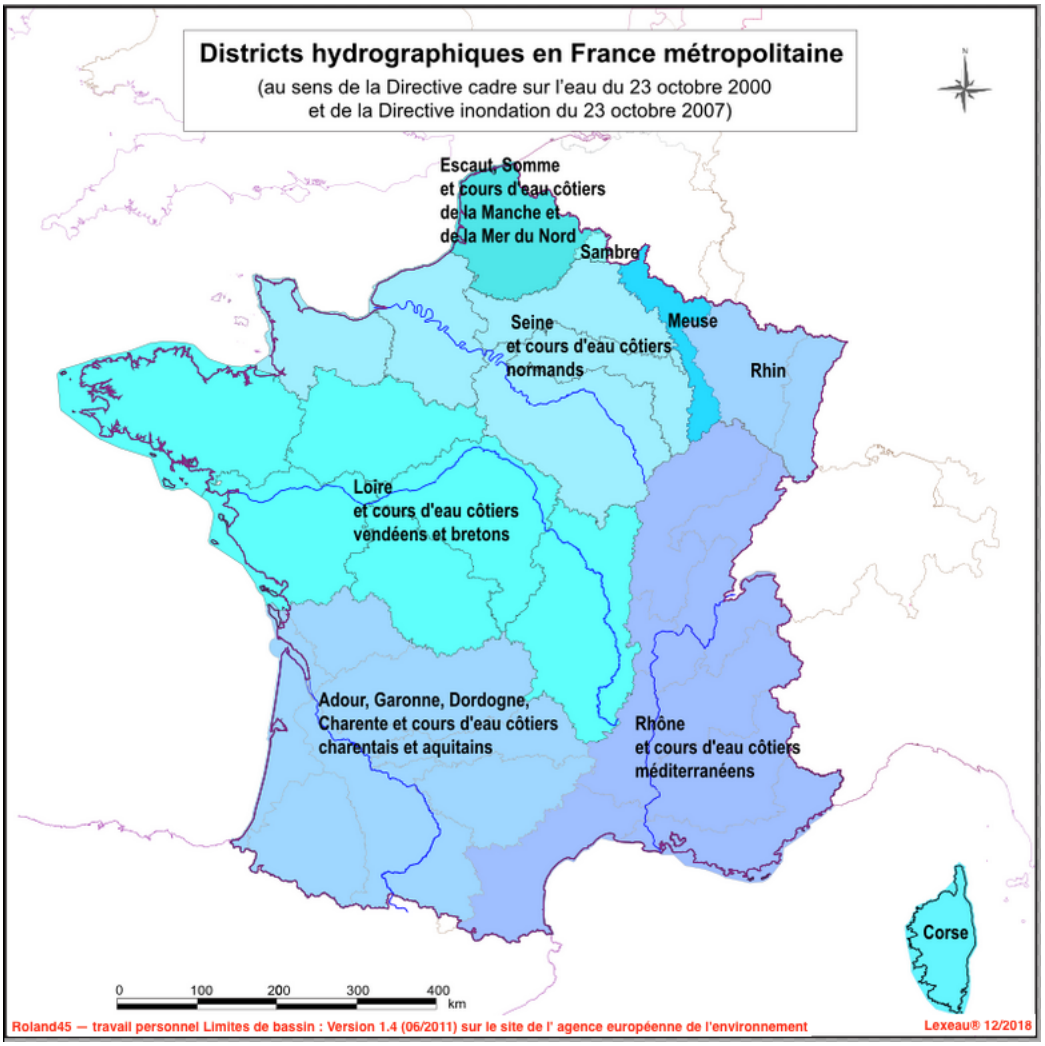


FIGURE 5.46 – Carte des 9 districts hydrographiques pour la France métropolitaine

11. https ://fr.wikipedia.org/wiki/District_hydrographique

5.2.3 L'ontologie du lexique

Les entités du lexique sont présentées dans la figure 5.47 avec leurs relations en interne. Nous avons vu que des unités récurrentes sélectionnées comme noms (substantifs) ou comme

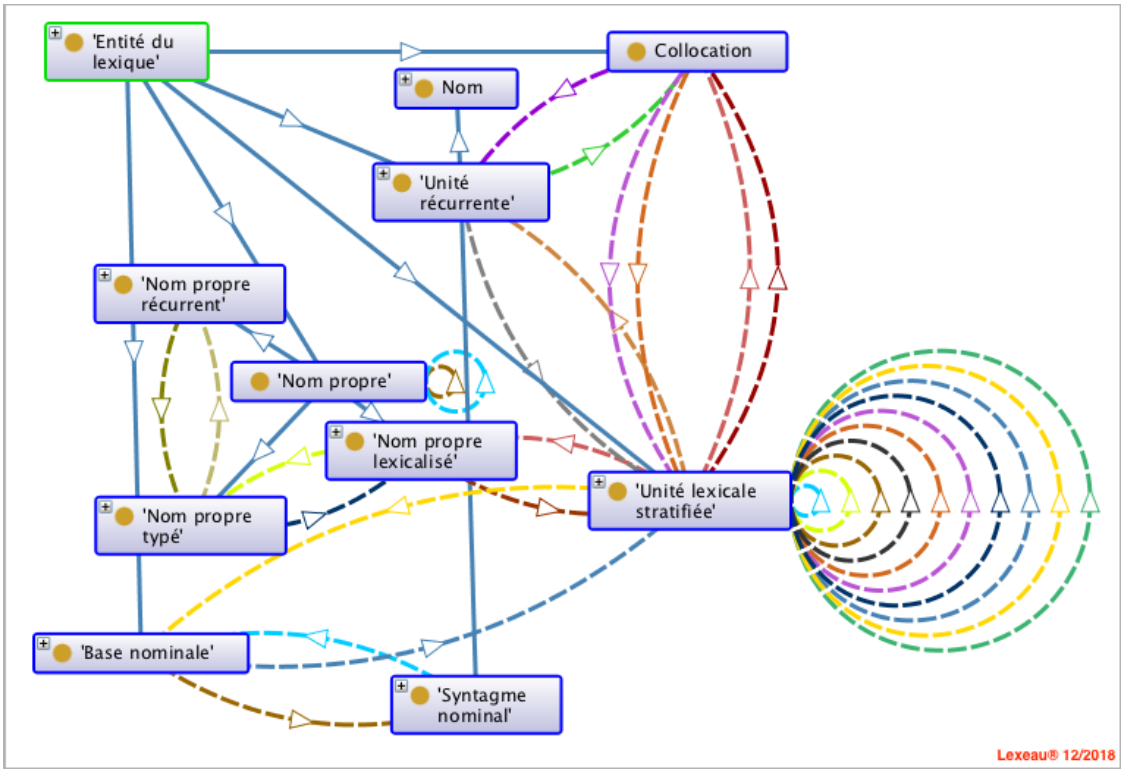


FIGURE 5.47 – Relations internes des entités du lexique

syntagmes nominaux par l'analyse des textes peuvent être introduites dans le lexique. Certains syntagmes nominaux peuvent être considérés comme des collocations, en tant que combinaison d'unités lexicales qui exprime une idée précise. Ils sont alors associés à des unités lexicales, en tant que base de la collocation ou collocatif, et seront intégrés dans l'article de l'unité.

La classe des unités récurrentes est donc une classe de transit utilisée pour finaliser les travaux de textométrie. Le double lien entre la classe des collocations et celle des entités lexicales s'explique par la possibilité, pour une collocation, d'avoir un collocatif et sa base dans le lexique. Lorsqu'un syntagme nominal est introduit dans le lexique, sa base nominale, si elle n'y figure pas, est conservée sa classe et rattachée à l'unité lexicale.

Les noms propres associés au lexique sont des noms propres lexicalisés qui renvoient à des objets concrets. Ils ont pour origine l'instance d'une classe d'entités typées dont le nom de classe a été transposé dans le lexique comme unité lexicale pivot, ce qui explique le premier lien entre les deux classes du lexique. Le second lien tient à l'emploi du nom propre, qui peut ou non correspondre au sens d'une unité lexicale satellite de l'unité lexicale pivot dont il est issu. Nous en avons eu un exemple avec le prélèvement d'eau effectué par EDF pour alimenter la centrale nucléaire de Golfech en 2016, intitulé « Prélèvement EDF Golfech 2016, entendu comme un prélèvement d'eau soumis à redevance, dans le système d'information source géré

par l'agence de l'eau Adour-Garonne. Il est de ce fait rattaché aux deux unités lexicales « 1-TS Prélèvement d'eau » et « 3-TA Prélèvement d'eau soumis à redevance ».

Les noms propres récurrents issus de l'analyse textométrique construits sur un ou plusieurs noms propres typés, sur lesquels sont déjà construits des noms propres lexicalisés, peuvent leur être aoutés comme référents d'une unité lexicale existante, voire à créer, à l'occasion d'une mise à jour de l'ontologie du domaine, sous réserve d'un contrôle des instances typées qui n'auraient pas été transposées.

Les figure 5.48 et 5.49 présentent les relations externes des entités du lexique avec celles du domaine et du discours. Les URL donnent accès à des pages web en français et en anglais auxquelles renvoient les noms propres lexicalisés. Les propriétés nommées transposées dans le lexique sont des propriétés relationnelles (l'auteur d'un document, l'opérateur d'un prélèvement d'eau) ou « à valeur » (le volume d'un flux d'eau). La première lettre en entrée du lexique sera notée en minuscule pour distinguer ce type d'entrée.

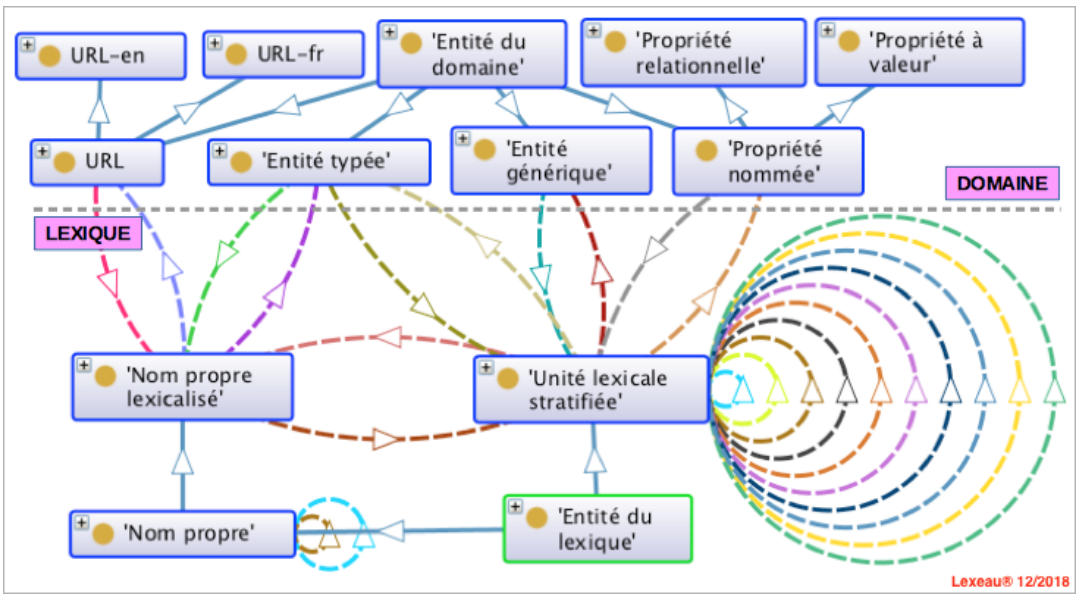


FIGURE 5.48 – Relations externes des entités du lexique avec celles du domaine

5.2.4 L'ontologie du discours

Les entités du discours donnent accès au lexique à travers les unités discursives. Elles fournissent un contenu aux articles du lexique (définitions et exemples d'emploi des unités lexicales) et des exemples d'emploi des noms propres lexicalisés, sous la forme de textes et d'éléments hors-texte (cf. fig. 5.48). Ces éléments sont extraits des textes de documents référencés et/ou accessibles sur le web. Les étiquettes morphosyntaxiques sont présentées en annexe (section B.1). Elles permettent d'étiqueter les unités discursives. Elles correspondent aux *pos* de l'analyse textométrique des textes lemmatisés, où un lemme et une étiquette sont attribués à chaque mot du texte.

Les textes documentés sont regroupés en corpus avant de les soumettre à des analyses texto-

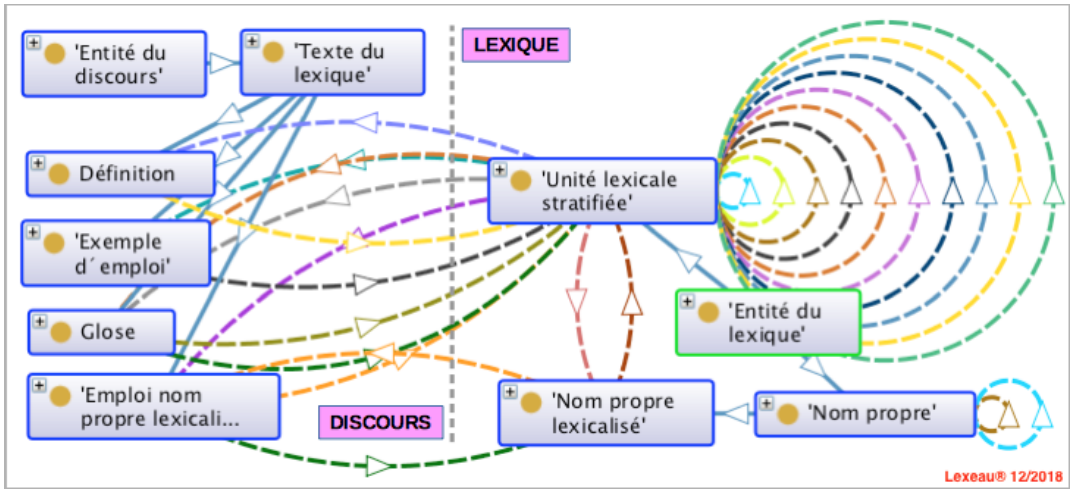


FIGURE 5.49 – Relations des entités du lexique avec les textes du lexique

métriques. La partition d'un corpus consiste à répartir ses textes en sous-corpus jointifs (sans recouvrement). Dans une partition textuelle fine, chaque texte correspond à un sous-corpus. Dans les relations entre les entités corpus et du domaine, on retrouve le rattachement des cor-

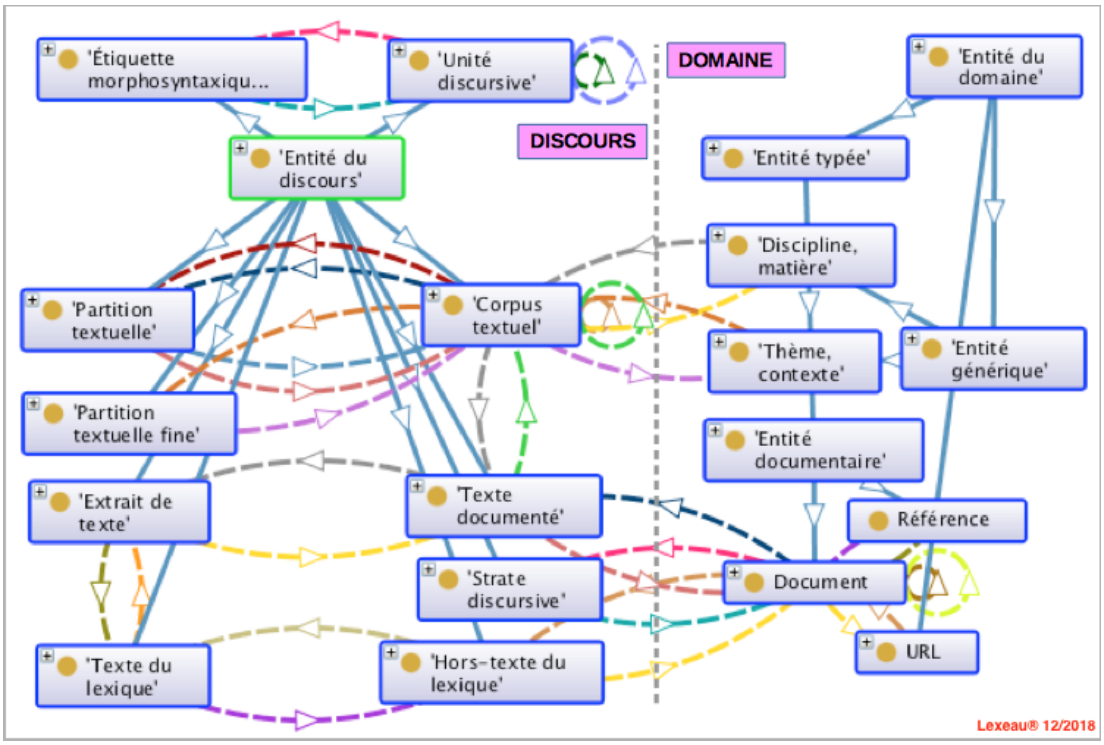


FIGURE 5.50 – Relations entre les entités primaires du discours et avec les entités du domaine

pus à un thème et/ou une discipline, le rattachement à un document des textes documentés et des éléments hors texte destinés à figurer dans le lexique et le rattachement des documents à une strate discursive.

La figure 5.51 présente le graphe relationnel qui résulte de l'extraction dans les corpus, par analyse textométrique, de noms (substantifs), de syntagmes nominaux et de noms propres récurrents. Ces éléments lexicaux sont transposés dans les classes correspondantes du lexique dans un processus d'extraction-sélection qui n'apparait pas dans ce graphe, où ne figurent pas les seuils de fréquence utilisés pour extraire les éléments récurrents dans les corpus et d'autres critères de sélection des substantifs et des syntagmes nominaux après extraction. Il s'agit d'un constat *a posteriori* des résultats d'une textométrie « dirigée » qui demande une présentation plus détaillée de son déroulement.

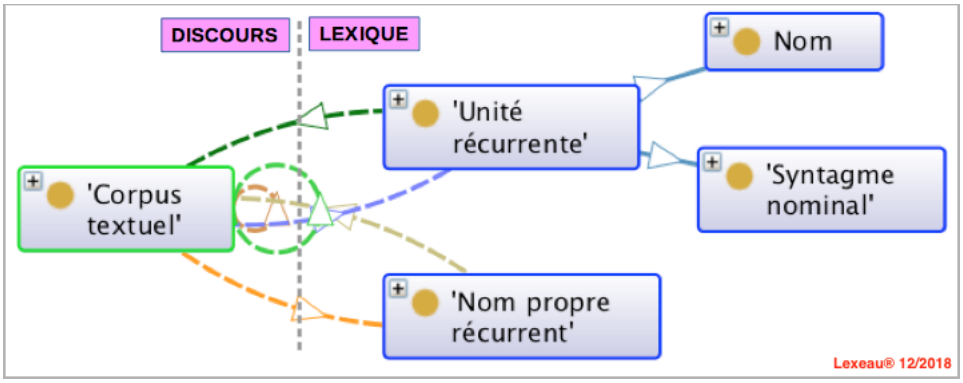


FIGURE 5.51 – Graphe de l'extraction d'éléments lexicaux récurrents des corpus textuels

5.3 Statut et fonctions des unités lexicales

Les lexèmes sont les points d'arrivée de l'élaboration de notre lexique à partir des concepts. Ils constituent, en linguistique, un point d'ancrage des unités lexicales. Dans le dictionnaire du langage de Franck Neveu [Neveu, 2010], pages 176, on lit :

> Les lexèmes sont des morphèmes lexicaux. Ils assurent la spécificité sémantique d'un mot, par distinction avec les grammèmes, (ou morphèmes grammaticaux), qui ont pour fonction de marquer les rapports morpho-syntaxiques et sémantiques entre les constituant de l'énoncé. Les lexèmes forment une classe ouverte, susceptible de s'enrichir par la formation d'unités nouvelles.

Dans la suite de l'article, l'auteur note qu'il existe d'autres approches de la notion, selon lesquelles « le terme de lexème désigne une entité abstraite susceptible d'être formée d'un ou de plusieurs morphèmes. ».

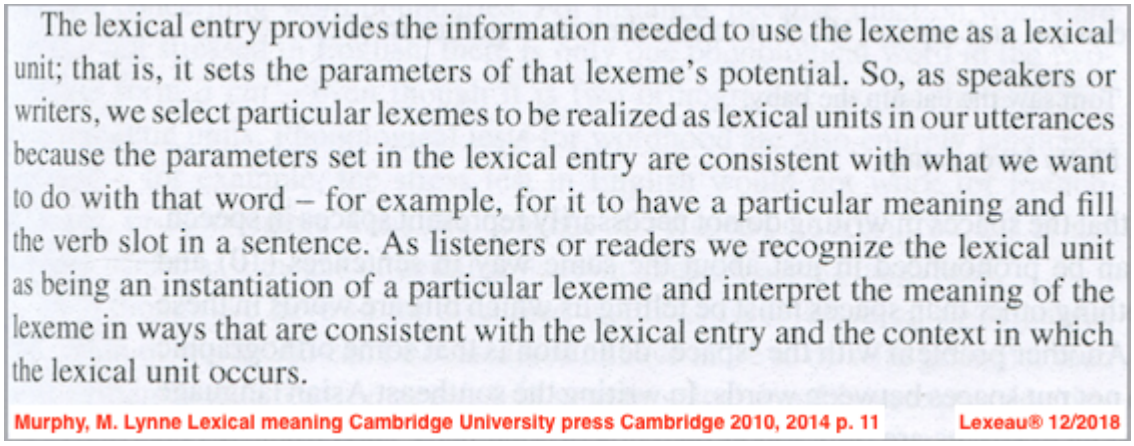
Pour donner un statut aux lexèmes qui soit indépendant des considérations morphologiques, nous utiliserons l'approche de M. Lynn Murphy [Murphy, 2010] dans son ouvrage intitulé *Lexical meaning* [Murphy, 2010], ce que l'on peut traduire littéralement par « signification lexicale ». Parmi les définitions pages 3 et 4 de cet ouvrage, le lexique (*lexicon*) est vu comme le vocabu-laire d'une langue — le *lexis* — ou encore comme le vocabulaire dont dispose l'usager d'une langue.

L'auteur laisse explicitement de côté le lexique en tant que dictionnaire de la langue (*dictionary*), notamment de la langue classique (*classical language*).

Les lexème (*lexemes*) sont définis comme des abstractions des mots utilisés dans la langue :

Lexemes are, essentially, abstractions of actual words that occur in the real language use

Quand nous utilisons un mot dans une phrase, le lexème n'est pas dans la phrase, on y trouve l'instanciation particulière du lexème (« *a particular instantiation (i.e. instance of use) of that lexeme* »). Ces instanciations sont des unités lexicales (« *lexical units* ») Les lexèmes sont instanciés dans les mots d'une phrase pour former des unités lexicales (*lexical units*). Ceci permet d'associer un lexème à une gamme de sens (« *a range of meanings* »). La fonction que l'auteur assigne à une entrée du lexique (« *a lexical entry* ») est définie dans un passage de son ouvrage (p. 11) reproduit dans la figure 5.52.


The lexical entry provides the information needed to use the lexeme as a lexical unit; that is, it sets the parameters of that lexeme's potential. So, as speakers or writers, we select particular lexemes to be realized as lexical units in our utterances because the parameters set in the lexical entry are consistent with what we want to do with that word – for example, for it to have a particular meaning and fill the verb slot in a sentence. As listeners or readers we recognize the lexical unit as being an instantiation of a particular lexeme and interpret the meaning of the lexeme in ways that are consistent with the lexical entry and the context in which the lexical unit occurs.

Murphy, M. Lynne Lexical meaning Cambridge University press Cambridge 2010, 2014 p. 11 Lexeau® 12/2018

FIGURE 5.52 – Fonction d'une entrée du lexique selon M. Lynne Murphy

Dans notre lexique, cette fonction des entrées du lexique (avant stratification) est assurée par les unités discursives. Les lexicographes les appellent parfois les « vedettes » du lexique. Nous reviendrons sur la structure de cette classe pour prendre en compte les libellés alternatifs, à propos de la nature bilingue du lexique (section 5.4.3).

La limitation des actes langagiers dans le domaine de l'eau entraine une limitation du nombre d'acceptions d'un mot ou d'une expression employé dans le domaine par rapport à ce que l'on peut trouver dans le lexique de la langue. Le statut conféré à l'unité lexicale comme instanciation d'un lexème permet de rattacher chaque unité lexicale à l'unité discursive qui correspond au lexème instancié. Ce rattachement s'opère sur les deux libellés de l'unité, en français et en anglais. Le choix d'un lexique stratifié fait que les unités lexicales sont monosémiques dans leur strate discursive. Ces unités sont souvent composées de plusieurs mots. Les unités introduites dans le lexique sont pour l'instant des substantifs ou des syntagmes nominaux. Les syntagmes verbaux interviennent dans les relations entre les entités du domaine et les gloses des unités lexicales mais ne font pas partie du lexique. Les adjectifs du lexique sont intégrés dans les syntagmes nominaux.

Les classes d'entités typées de l'ontologie transposées dans le lexique constituent le plus grand nombre des unités lexicales. Ces unités gardent la trace de leur classe d'origine (une « indivision » n'est pas une « personne morale », elle n'est pas dotée de la personnalité juridique, attribut de la personne morale). Les propriétés relationnelles comme « auteur », « opérateur », « usager » et « propriétaire » sont transposées dans le lexique comme des substantifs, avec une minuscule en début de mot pour marquer le fait qu'ils portent dans l'ontologie du

domaine le nom de la relation entre les réalisations de « types » différents de cette ontologie (X auteur d'un livre L, Y opérateur d'un mouvement d'eau M, Z usager d'un flux anthropique F, T propriétaire d'un ouvrage O). Le premier terme de la relation est ici une entité humaine. Les propriétés « à valeur » comme « volume d'eau (m3) » et « débit maximum (m3/h) » sont également transposées dans le lexique avec une lettre minuscule au début de l'unité. Ces unités lexicales sont des attributs « à valeur » d'autres unités lexicales (le volume d'eau d'un flux anthropique en m3, son débit maximum en m3/h). Les entités génériques transposées dans le lexique comme « évapotranspiration », « eutrophisation », « hydrologie » sont rattachées à des unités lexicales plus générales du même type (l'évapotranspiration et l'eutrophisation sont des processus naturels, l'hydrologie est une discipline). Ainsi, les unités lexicales n'ont pas la même fonction en discours, en tant que concepts édités dans l'ontologie du domaine et transposés dans le lexique, selon leur place et leur fonction (nom ou instance d'une classe d'entités génériques, nom d'une classe d'entités typées, propriété relationnelle, propriété à valeur). Cette règle est indépendante de la strate discursive et s'applique aux entités lexicales satellites comme aux entités lexicales pivot.

Les unités lexicales prennent, comme les unités discursives, la forme type d'un *lemme*, c'est à dire, selon le dictionnaire des sciences du langage de Franck Neveu :

> La forme type d'un mot, adoptée par convention, permettant le regroupement sous une même adresse lexicale (ou entrée), des diverses réalisations flexionnelles que ce mot peut connaître en discours.

La forme grammaticale du lemme correspond à l'étiquette morphosyntaxique de rattachement de l'unité discursive.

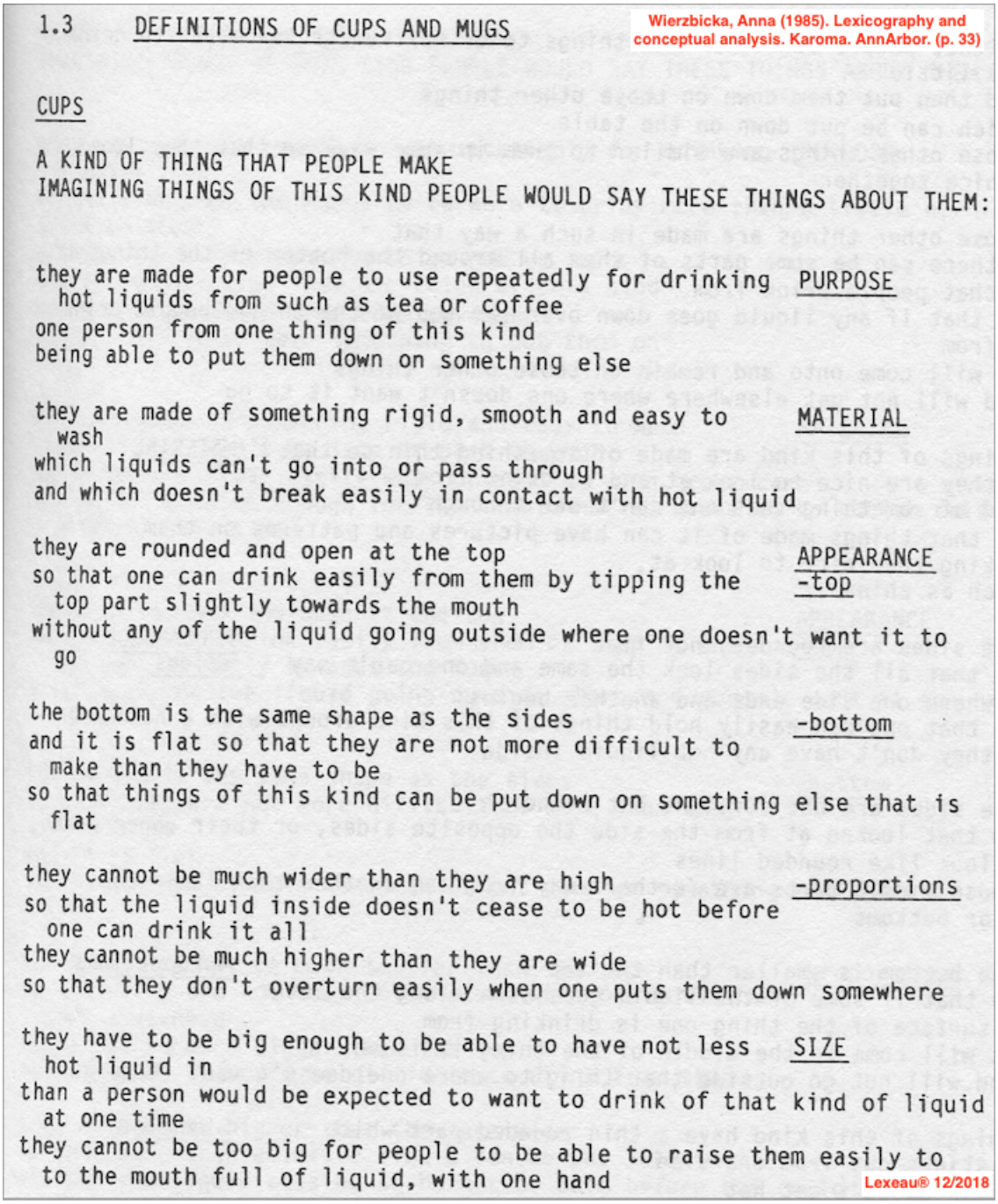
5.3.1 Statut et fonctions des unités lexicales chez Anna Wierzbicka

Les unités lexicales semblent assimilées par Anna Wierzbicka à des concepts insérés dans un réseau de concepts. La lexicographie développée dans son ouvrage intitulé *Lexicography and conceptual analysis* [Wierzbicka, 1985] est inspirée des relations conceptuelles que l'unité lexicale entretien avec d'autres unités. Nous reprenons ci-dessous l'extrait d'une citation de Humboldt [Humboldt, 1903] traduite par Robert Miller [Miller, 1968] reproduite en introduc-tion de l'ouvrage, page 15 :

> *For language does not represent objects, but rather the concepts which, in the process of speech, have been formed by the mind independent of those objects.*

Dans son analyse comparative des tasses à café et des *mugs* (chopes), la définition des tasses à café dont nous présentons un extrait dans la figure 5.53 est une glose où sont passées en revue les objectifs poursuivis par les concepteurs de l'objet à certaines fins que l'on peut qualifier de « fonctionnalités » : *purpose*/utilisation, *material*/matériau, *appearance*/aspect, *size*/taille. Dans le langage de l'ontologie du lexique, la définition d'une tasse à café est composée d'un ensemble de propriétés « à valeur textuelle » correspondant chacune à une « fonctionnalité » de l'objet fini. À la différence d'Anna Wierzbicka, nous faisons la distinction entre un concept, son édition dans l'ontologie du domaine qui lui donne une forme et sa transposition dans le lexique qui en fait une unité lexicale. Les fonctionnalités de la tasse de café, dans l'exemple de définition qu'elle donne de l'unité lexicale, nous semblent assez éloignées des fonctions que peuvent remplir les unités lexicales en tant que « produit » direct ou indirect de l'ontologie

du domaine. Il est vrai que son « domaine » est celui du langage courant, où elle cherche à établir des définitions universelles dans la forme et dans le fond, pour peu que les mêmes fonctionnalités soient recherchées pour l'objet défini quelle que soit la langue.



1.3 DEFINITIONS OF CUPS AND MUGS

Wierzbicka, Anna (1985). Lexicography and conceptual analysis. Karoma. AnnArbor. (p. 33)

CUPS

A KIND OF THING THAT PEOPLE MAKE
IMAGINING THINGS OF THIS KIND PEOPLE WOULD SAY THESE THINGS ABOUT THEM:

they are made for people to use repeatedly for drinking hot liquids from such as tea or coffee — PURPOSE
one person from one thing of this kind
being able to put them down on something else

they are made of something rigid, smooth and easy to wash — MATERIAL
which liquids can't go into or pass through
and which doesn't break easily in contact with hot liquid

they are rounded and open at the top — APPEARANCE
so that one can drink easily from them by tipping the top part slightly towards the mouth — -top
without any of the liquid going outside where one doesn't want it to go

the bottom is the same shape as the sides — -bottom
and it is flat so that they are not more difficult to make than they have to be
so that things of this kind can be put down on something else that is flat

they cannot be much wider than they are high — -proportions
so that the liquid inside doesn't cease to be hot before one can drink it all
they cannot be much higher than they are wide
so that they don't overturn easily when one puts them down somewhere

they have to be big enough to be able to have not less hot liquid in — SIZE
than a person would be expected to want to drink of that kind of liquid at one time
they cannot be too big for people to be able to raise them easily to to the mouth full of liquid, with one hand

Lexeau® 12/2018

FIGURE 5.53 – Définition d'une tasse à café à partir de sa conception (A. Wierzbicka)

Notre point de vue, dans un domaine du discours plus limité, est que les concepts ne sont pas forcément les mêmes dans toutes les langues, ce qui est également son point de vue affirmé

avec force par Anna Wierzbicka dans son ouvrage intitulé *Semantics, Culture and Cognition* [Wierzbicka, 1992] et sa critique de l'impérialisme d'une langue qui en aurait le monopole. Nous pensons qu'une ontologie conceptuelle bilingue doit néanmoins pouvoir intégrer ce fait, comme nous avons tenté de le faire pour ajuster les concepts d'aléas aux pratiques françaises et britanniques en matière de vigilance météorologique et d'annonce des crues, et ce en travaillant sur plusieurs strates discursives.

Dans la suite de ses travaux, avec la théorie des « primitives »/*primes*, Anna Wierzbicka a développé un métalangage sémantique naturel (*natural semantic metalanguage* ou NSM), pour expliquer les unités lexicales du langage courant à partir d'un nombre limité de primitives, formes lexicalisées de concepts universels que l'on trouve dans toutes les langues. Pour éviter toute circularité dans les définitions, les primitives ne sont pas définies en propre et sont les seules à figurer dans la définition de l'unité lexicale, en des termes proches de concepts que tout le monde peut comprendre dans sa propre langue. Cette méthode évite les biais de l'ethno-centrisme dans les définitions des anthropologues. Elle ouvre la voie d'une réelle intercompréhension entre des locuteurs de différentes langues et de différentes cultures sur le vocabulaire de la vie courante. Ce métalangage est présenté par Lynn Murphy (ouvrage cité) en s'appuyant sur des ouvrages d'Anna Wierbizcka ([Wierzbicka, 1972] et [Wierzbicka, 1996]) et de Cliff Goddard [Goddard, 1998]. L'explication de *happy*, présentée page 70 de l'ouvrage de Lynn Murphy, est reprise ici dans le tableau 5.1, sur la base d'une liste des primitives dans la version anglaise du métalangage, reprise dans le tableau 5.2. Les primitives sont notées en petites majuscules. Elles renvoient à 16 groupes de concepts.

TABLE 5.1 – An explication of *happy* with the NSM metalanguage

X feels *happy* =	sometimes someone thinks something like this something good happened to me I wanted this I don't want other things now because of this, someone feels something good X feels like this

Nous avons relevé des groupes de concepts qui peuvent être rapprochés de certaines entités de l'ontologie du domaine, par exemple SOMEWHERE, proche des entités géographiques et TIME, proche des entités temporelles, avec un écart important entre l'ontologie du domaine, qui compte actuellement 382 classes d'objets, et le nombre limité de concepts du métalangage de la NSM lexicalisés au quotidien dans toutes les langues (63, dans l'ouvrage cité de M. Goddard), et ce pour construire une explication du vocabulaire courant des locuteurs qui leur permette de se comprendre d'une langue à l'autre. Comme le rappelle Anna Wierzbicka p. 8 et 9 d'un autre ouvrage intitulé *Semantics, Culture, and Cognition, Universal Human Concepts in Culture-Specific Configurations* [Wierzbicka, 1992], cette idée a été défendue par Leibnitz avec son *alphabet des pensées humaines*. Elle cite en anglais l'extrait d'une version française de son oeuvre [Leibnitz, 1903] que nous avons retraduit vers le français :

> Bien que le nombre d'idées qui peuvent être conçues soit infini, il est possible que le nombre de celles qui peuvent être conçues pour elles-mêmes soit très petit, parce qu'un nombre infini de choses peut être exprimé en combinant très peu d'éléments. Au contraire, cela

TABLE 5.2 – English exponents of NSM primes [Goddard, 1998]

PRIMES	Group of primes
I, YOU, SOMEONE, SOMETHING/THING, PEOPLE, BODY	substantives
KIND, PART	relational substantives
THIS, THE SAME, OTHER/ELSE	determiners
ONE, TWO, SOME, ALL, MUCH/MANY	quantifiers
GOOD, BAD	evaluators
BIG, SMALL	descriptors
THINK, KNOW, WANT, FEEL, SEE, HEAR	mental predicates
SAY, WORDS, TRUE	speech
DO, HAPPEN, MOVE, TOUCH	action, event, movement, contact
BE (SOMEWHERE), THERE IS, HAVE, BE (SOMEONE/SOMETHING)	location, existence, possession, specification
LIVE, DIE	life and death
WHEN/TIME, NOW, BEFORE, AFTER, A LONG TIME, A SHORT TIME, FOR SOME TIME, MOMENT	time
WHERE/PLACE, HERE, ABOVE, BELOW, FAR, NEAR, SIDE, INSIDE	space
NOT, MAYBE, CAN, BECAUSE, IF	logical concepts
VERY, MORE	intensifier, augmentor
LIKE/WAY	similarity

> est non seulement possible mais probable, parce que la nature tend naturellement à réaliser beaucoup avec très peu, en opérant de la manière la plus simple ... L'alphabet des pensées humaines est le catalogue de ces concepts qui peuvent être compris pour eux-mêmes et dont la combinaison forme toutes nos idées.

Le caractère opérationnel de la théorie des primitives ne fait pas de doute dans le champ du langage courant. À l'image des concepts humains universels (*universal human concepts*) que nous trouvons dans notre propre tête (*in our own minds*), les concepts qui nous intéressent dans le domaine de l'eau relèvent de discours diversifiés qui portent sur des *objets du monde* vus sous un angle technique, scientifique, juridique ou administratif, voire religieux, éthique, moral ou idéologique. En dehors des concepts scientifiques, ces concepts peuvent varier d'une langue à l'autre et d'une culture à l'autre. L'introduction de quatre strates discursives permet de les qualifier d'une façon grossière, supposée indépendante de la langue et de la culture, ce qui livre une clef d'intercompréhension entre les acteurs du domaine. La notion de concept qui domine les travaux de recherche initiés par Anna Wierzbicka dans la construction d'un lexique de la langue courante basé sur des concepts universels domine aussi, d'une autre façon, la construction d'un lexique conceptualisé du domaine de l'eau.

5.3.2 Statut et fonctions des unités lexicales chez Igor Mel'čuk

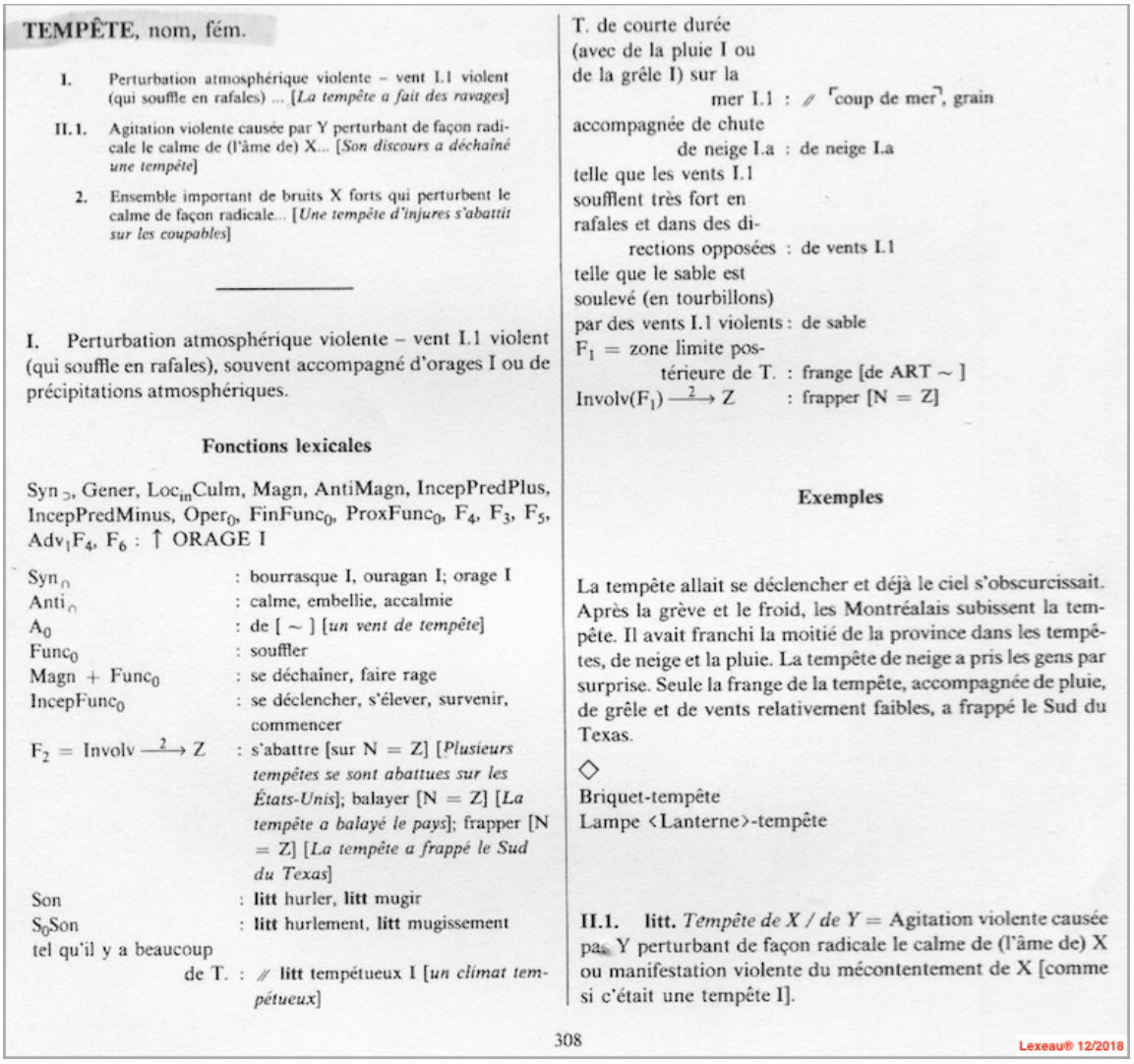
Les réflexions qui suivent sont inspirées de la lecture du Dictionnaire « explicatif et combinatoire du français contemporain », publié en 4 tomes de 1984 à 1999 dans le cadre de recherches lexico-sémantiques conduites par l'OLST [12] et le GRESLET [13] à l'université de Montréal sur cette période [Mel'čuk *et al.*, 1984, Mel'čuk *et al.*, 1988, Mel'čuk *et al.*, 1992, Mel'čuk *et al.*, 1999]. Les recherches ont donné lieu à la production d'articles théoriques et d'articles sur les lexies du dictionnaire pour 508 vocables. Les figures 5.54 et reproduisent l'article consacré au vocable « Tempête » dans le DEC II. Les 3 lexies « Tempête I », « Tempête II.1 » et « Tempête II.2 » sont numérotées sur un arbre d'acceptions du vocable organisées par niveau. Les définitions renvoient à d'autres lexies. Dans le cas d'un système d'actants (X et Y ou X), le tableau du régime des actants est suivi d'une liste de combinaisons. Les fonctions lexicales qui s'appliquent à la lexie selon une nomenclature codée sont indiquées ensuite, avant des exemples d'emploi. Selon le modèle dictionnairique mis en oeuvre, l'article doit couvrir tous les emplois possibles de la lexie, en lien avec les autres lexies, ce qui conduit les auteurs du dictionnaire à travailler par « champ lexical », qui est ici le champ des phénomènes atmosphériques, avec 12 vocables traités dans les tomes 1 et 2 du dictionnaire (brouillard, brume, grêle, grelon, neige, neiger, orage, pleuvoir, pluie, tempête et vent). La présentation du DEC sur le site de l'OLST [14] complète cet exemple :

> Le Dictionnaire explicatif et combinatoire du français contemporain (DECFC) est le produit d'une recherche qui se consacre à la description formelle du lexique français, selon une approche sémantique. Il est élaboré à partir des principes de la lexicographie explicative et combinatoire — voir Mel'c˘uk, Clas, Polguère (1995) — postulée par la théorie linguistique Sens-Texte. Selon cette théorie, un DEC doit fournir, de façon systématique et rigoureuse, les informations permettant à un locuteur de construire toutes les expressions linguistiques correctes de n'importe quelle pensée et ce, dans n'importe quel contexte : c'est un dictionnaire de production.

La lexie correspond ici à la réalisation d'un vocable qu'il est possible de définir dans un réseau hiérarchisé de lexies du vocable, lequel pourrait être comparé à une unité discursive. Elle a donc un statut analogue au statut de l'unité lexicale du lexique. Les choses se compliquent en termes de fonction, ou de fonctionnalité. Le système des fonctions lexicales tel qu'il est illustré dans l'article consacré au vocable « Tempête » et présenté dans le manuel intitulé « Introduction à la lexicologie explicative et combinatoire » [Mel'čuk *et al.*, 1995] ne correspond pas aux « fonctions » des concepts à l'origine des unités lexicales. Il ne correspond pas non plus à la traduction en glose principale des relations d'objet introduites dans l'ontologie du domaine que nous proposons de mettre en oeuvre pour les unités lexicales pivot du lexique, et ce pour rendre compte des propriétés des concepts source de ces unités.

Les verbes utilisés dans ces gloses, comme par exemple le verbe « générer » dans la relation entre un prélèvement d'eau et un flux d'eau anthropique, peuvent-ils être considérés comme la réalisation d'une fonction plus générale qui produirait, entre autre, un flux à partir d'un prélèvement ? Autant la notion nous parait productive dans l'analyse du langage courant, encore que probablement très complexe à mettre en oeuvre et surtout à assimiler par les

12. Observatoire de linguistique Sens-Texte
13. Groupe de recherhce en sémantique, lexicologie et terminologie
14. http ://olst.ling.umontreal.ca/ ?page_id=56#Vocables



TEMPÊTE, nom, fém.

1. Perturbation atmosphérique violente – vent I.1 violent (qui souffle en rafales) ... [*La tempête a fait des ravages*]

II.1. Agitation violente causée par Y perturbant de façon radicale le calme de (l'âme de) X... [*Son discours a déchaîné une tempête*]

2. Ensemble important de bruits X forts qui perturbent le calme de façon radicale... [*Une tempête d'injures s'abattit sur les coupables*]

I. Perturbation atmosphérique violente – vent I.1 violent (qui souffle en rafales), souvent accompagné d'orages I ou de précipitations atmosphériques.

Fonctions lexicales

$Syn_{\supset}$, Gener, $Loc_{in}Culm$, Magn, AntiMagn, IncepPredPlus, IncepPredMinus, $Oper_0$, $FinFunc_0$, $ProxFunc_0$, F_4, F_3, F_5, Adv_1F_4, F_6 : ↑ ORAGE I

$Syn_{\cap}$	: bourrasque I, ouragan I; orage I
$Anti_{\cap}$	: calme, embellie, accalmie
A_0	: de [~] [*un vent de tempête*]
$Func_0$	: souffler
Magn + $Func_0$	: se déchaîner, faire rage
$IncepFunc_0$	: se déclencher, s'élever, survenir, commencer
F_2 = Involv $\xrightarrow{2}$ Z	: s'abattre [sur N = Z] [*Plusieurs tempêtes se sont abattues sur les États-Unis*]; balayer [N = Z] [*La tempête a balayé le pays*]; frapper [N = Z] [*La tempête a frappé le Sud du Texas*]
Son	: **litt** hurler, **litt** mugir
S_0Son	: **litt** hurlement, **litt** mugissement
tel qu'il y a beaucoup de T.	: // **litt** tempétueux I [*un climat tempétueux*]
T. de courte durée (avec de la pluie I ou de la grêle I) sur la mer I.1	: // ⌜coup de mer⌝, grain
accompagnée de chute de neige I.a	: de neige I.a
telle que les vents I.1 soufflent très fort en rafales et dans des directions opposées	: de vents I.1
telle que le sable est soulevé (en tourbillons) par des vents I.1 violents	: de sable
F_1 = zone limite postérieure de T.	: frange [de ART ~]
Involv(F_1) $\xrightarrow{2}$ Z	: frapper [N = Z]

Exemples

La tempête allait se déclencher et déjà le ciel s'obscurcissait. Après la grève et le froid, les Montréalais subissent la tempête. Il avait franchi la moitié de la province dans les tempêtes, de neige et la pluie. La tempête de neige a pris les gens par surprise. Seule la frange de la tempête, accompagnée de pluie, de grêle et de vents relativement faibles, a frappé le Sud du Texas.

◇

Briquet-tempête

Lampe ‹Lanterne›-tempête

II.1. **litt.** *Tempête de X / de Y* = Agitation violente causée par Y perturbant de façon radicale le calme de (l'âme de) X ou manifestation violente du mécontentement de X [comme si c'était une tempête I].

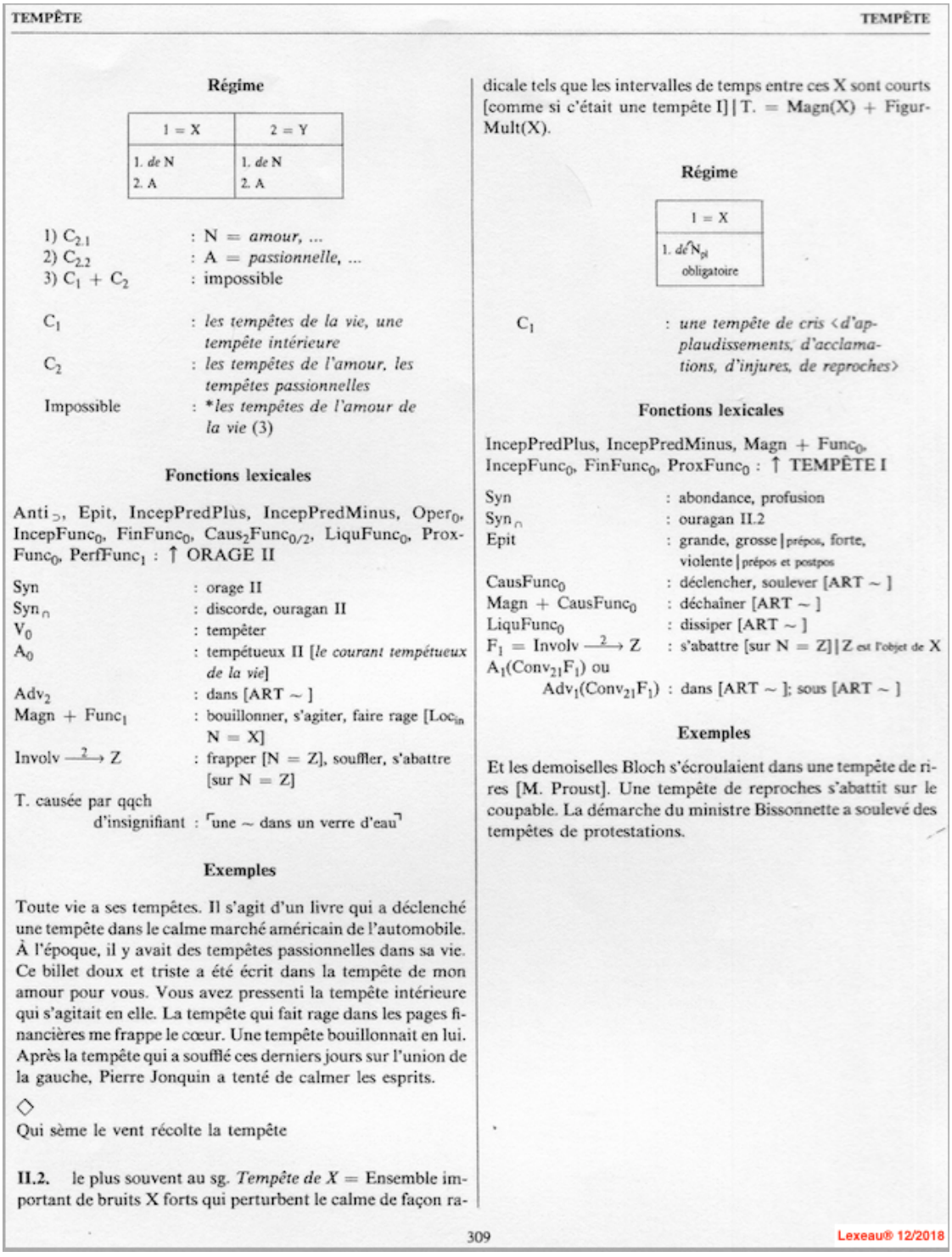
308

Lexeau® 12/2018

FIGURE 5.54 – Article du DEC sur le vocable « Tempête » (début)

lexicographes, autant la relation d'objet nous parait relativement simple pour construire un modèle relationnel entre les concepts d'un domaine bien délimité comme celui de l'eau, quitte à revenir sur les relations de l'ontologie et les gloses du lexique pour en faire une analyse plus systématique sous l'angle de fonctions lexicales.

Nous en avons retenu du modèle dictionnairique du DEC le caractère rigoureux de la composition des articles et l'idée centrale de l'individualisation des lexies par un code accolé au vocable. Dans le cas des unités de notre lexique, l'encodage de la strate discursive de rattachement est effectué dans le préfixe de l'unité. Le recours à des libellés différents pour les unités lexicales d'une même strate permet d'éviter leur numérotation. Le rôle du vocable dans le DEC est assuré par l'unité discursive de rattachement de l'unité lexicale. On trouvera dans la thèse de doctorat d'Anne-Laure Jousse dirigée par Sylvain Kahane et Alain Polguère et soutenue à l'université Paris Diderot, une présentation très complète des ressources et modèles lexicaux relationnels (DiCo, FrameNet, WordNet, etc.) et plusieurs chapitres sur les fonctions



TEMPÊTE TEMPÊTE

Régime

1 = X	2 = Y
1. *de* N 2. A	1. *de* N 2. A

1) $C_{2.1}$: N = *amour, ...*
2) $C_{2.2}$: A = *passionnelle, ...*
3) $C_1 + C_2$: impossible

C_1 : *les tempêtes de la vie, une tempête intérieure*
C_2 : *les tempêtes de l'amour, les tempêtes passionnelles*
Impossible : **les tempêtes de l'amour de la vie* (3)

Fonctions lexicales

$Anti_{\supset}$, Epit, IncepPredPlus, IncepPredMinus, $Oper_0$, $IncepFunc_0$, $FinFunc_0$, $Caus_2Func_{0/2}$, $LiquFunc_0$, $ProxFunc_0$, $PerfFunc_1$: ↑ ORAGE II

Syn : orage II
$Syn_{\cap}$: discorde, ouragan II
V_0 : tempêter
A_0 : tempétueux II [*le courant tempétueux de la vie*]
Adv_2 : dans [ART ~]
Magn + $Func_1$: bouillonner, s'agiter, faire rage [Loc_{in} N = X]
Involv $\xrightarrow{2}$ Z : frapper [N = Z], souffler, s'abattre [sur N = Z]
T. causée par qqch d'insignifiant : ⌜une ~ dans un verre d'eau⌝

Exemples

Toute vie a ses tempêtes. Il s'agit d'un livre qui a déclenché une tempête dans le calme marché américain de l'automobile. À l'époque, il y avait des tempêtes passionnelles dans sa vie. Ce billet doux et triste a été écrit dans la tempête de mon amour pour vous. Vous avez pressenti la tempête intérieure qui s'agitait en elle. La tempête qui fait rage dans les pages financières me frappe le cœur. Une tempête bouillonnait en lui. Après la tempête qui a soufflé ces derniers jours sur l'union de la gauche, Pierre Jonquin a tenté de calmer les esprits.

◇

Qui sème le vent récolte la tempête

II.2. le plus souvent au sg. *Tempête de X* = Ensemble important de bruits X forts qui perturbent le calme de façon radicale tels que les intervalles de temps entre ces X sont courts [comme si c'était une tempête I] | T. = Magn(X) + FigurMult(X).

Régime

1 = X
1. *de* N_{pl} obligatoire

C_1 : *une tempête de cris ‹d'applaudissements, d'acclamations, d'injures, de reproches›*

Fonctions lexicales

IncepPredPlus, IncepPredMinus, Magn + $Func_0$, $IncepFunc_0$, $FinFunc_0$, $ProxFunc_0$: ↑ TEMPÊTE I

Syn : abondance, profusion
$Syn_{\cap}$: ouragan II.2
Epit : grande, grosse | prépos, forte, violente | prépos et postpos
$CausFunc_0$: déclencher, soulever [ART ~]
Magn + $CausFunc_0$: déchaîner [ART ~]
$LiquFunc_0$: dissiper [ART ~]
F_1 = Involv $\xrightarrow{2}$ Z : s'abattre [sur N = Z] | Z est l'objet de X
$A_1(Conv_{21}F_1)$ ou $Adv_1(Conv_{21}F_1)$: dans [ART ~]; sous [ART ~]

Exemples

Et les demoiselles Bloch s'écroulaient dans une tempête de rires [M. Proust]. Une tempête de reproches s'abattit sur le coupable. La démarche du ministre Bissonnette a soulevé des tempêtes de protestations.

309 Lexeau® 12/2018

FIGURE 5.55 – Article du DEC sur le vocable « Tempête » (fin)

lexicales, la normalisation des fonctions non standard, l'élaboration d'un système de fonctions et les applications dérivées de ce système [Jousse, 2010].

5.4 Un lexique bilingue

Cette section a pour but de préciser la gestion du bilinguisme du lexique. Ce bilinguisme ne consiste pas à traduire en anglais des libellés des unités lexicales écrits en français, et inversement, car le lexique doit pouvoir fonctionner de façon autonome dans chaque langue, à partir de son lien avec une ontologie du domaine où les concepts sont édités dans la même langue. Ceci écarte l'usage du lexique comme d'une « terminologie » du domaine de l'eau, comme nous allons le préciser dans une première partie de cette section. Le caractère bilingue du lexique n'obère en rien sa capacité à gérer dans chaque langue les relations lexicales propres à la langue ou des libellés alternatifs d'unités lexicales, comme nous le verrons sur des exemples dans les sections 5.4.2 et 5.4.3. On retrouve ici la problématique de la traduction : une unité lexicale dans une langue L1 fonctionne en réseau au sein de cette langue et l'unité lexicale correspondante dans la langue L2 fonctionne, elle aussi, en réseau mais au sein de L2. Il n'y a donc pas correspondance mais corrélation entre deux unités lexicales munies de leurs réseaux. C'est pourquoi la correspondance entre deux unités lexicales de deux langues différentes passe par l'étude de chacune d'elles. Stricto sensu, on ne devrait pas mettre deux unités lexicales en correspondance mais deux réseaux lexicaux. La traduction d'une unité lexicale de L1 par une unité lexicale de L2 est donc un « effet de raccourci » de cette corrélation de deux réseaux.

5.4.1 Lexique et terminologie

La notion de terminologie domine dans le monde de l'industrie et, comme le souligne Loïc Depecker, ([Depecker, 2003], p. 7), chez « les ingénieurs, scientifiques, ouvriers et techniciens [...] naturellement amenés à manier des termes et à développer des systèmes de concepts ». Dans cet essai, présenté en Sorbonne pour l'habilitation à diriger des recherches, l'auteur revisite son expérience de chargé de mission pour la terminologie dans les services du Premier ministre puis au ministère de la culture, de 1980 à 1998, et son travail de recherche universitaire axé sur la pratique et la normalisation de l'activité terminologique, en inscrivant sa réflexion dans le champ de la linguistique. Son approche normative des « termes » nous parait tout à fait compréhensible dans une communauté appelée à des échanges quotidiens entres collègues, pour l'organisation d'une production industrielle répartie sur plusieurs pays, dans plusieurs langues, avec les échanges commerciaux que cela implique. La question est de savoir qui décide de la norme, du « bon » terme à employer. Cette question est d'autant plus facile à traiter que le domaine ou l'auditoire concernés sont restreints. Elle peut être abordée en termes de définition, en se limitant au français. Dans le domaine de l'eau, on peut citer le glossaire accessible en ligne sur le site Eaufrance [15] et celui de l'agence de l'eau Adour-Garonne, accessible sur le site du système d'information du bassin Adour-Garonne (SIE Adour-Garonne [16]).

Pour aborder la traduction en plusieurs langues, il existe des lexiques multilingues publiés ou accessibles sur abonnement comme de lexique de l'eau, en 6 langues (français, anglais, allemand, italien, espagnol, portugais), sans définitions, [Johanet, 2001]. La question sous-jacente de la terminologie est aussi le choix d'un terme « officiel » à substituer à l'emploi d'un terme d'une autre langue, notamment l'anglais, dont le sens n'est compris que par les

15. http ://www.data.eaufrance.fr/glossaire
16. http ://adour-garonne.eaufrance.fr/component/option,com_rd_glossary/Itemid,20/lang,/

spécialistes et qui entre dans l'usage courant sans définition claire ou avec plusieurs acceptions. L'exemple en est le travail de la délégation générale à la langue française qui a débouché sur la publication en 1994 d'un dictionnaire des termes officiels de la langue française [générale à la langue française, 1994]. Cette approche semble avoir rencontré ses limites au niveau institutionnel. On peut enfin citer les travaux de l'AFNOR et ceux de l'ISO qui réunissent des spécialistes en groupes de travail avant la publication de normes payantes où la question de la traduction est traitée au fond par le groupe de travail.

Dans notre cas, nous avons considéré que le nom d'une classe d'objets concrets pouvait être considéré comme un concept et transposé dans le lexique dans une strate discursive et une seule, en fonction du discours qui accompagne l'emploi de cette entrée, de l'auditoire de ce discours et des documents qui l'accompagnent. Ce travail est mené dans une langue donnée, ce qui induit des différences, voire des absences d'une langue à l'autre dans la strate discursive technico-administrative, dans la mesure où les concepts qui relèvent de cette strate sont le reflet de la culture et de la vie institutionnelle des pays où cette langue est pratiquée comme langue première. Lorsqu'il s'agit d'une langue seconde, il peut être difficile ou tout simplement impossible de remonter au concept source dans la langue première, ce qui peut invalider la construction d'une ontologie dans une langue seconde, comme la critique du *globish* (pour *global english*) le fait bien comprendre. C'est néanmoins l'exercice auquel nous nous sommes livrés tout au long de ce travail pour évaluer la difficulté d'élaborer un lexique de l'eau bilingue, condition première à remplir pour qu'un tel lexique remplisse son office d'outil d'intercompréhension entre les acteurs du domaine, et pas seulement de la « traduction » en « bon » français d'un terme à la mode en *globish*. Nous sommes loin d'une conceptualisation « officielle » des termes du domaine, sinon, encore une fois, dans la strate technico-administrative, avec la directive cadre sur l'eau et ses directives filles, comme la directive sur les inondations. La directive cadre a introduit en 2000 dans sa version française une acception nouvelle de l'expression « masse d'eau » en la conceptualisant comme une entité géographique. L'expression n'est pas employée dans ce sens dans le langage courant et les *media* n'ont pas cherché à en relayer l'apparition, pour des raisons qui nous paraissent tout à fait recevables. Nous pouvons témoigner de la forte réticence des milieux administratif et scientifique du domaine de l'eau — un ancien directeur d'administration centrale et un hydrogéologue directeur d'une école d'ingénieurs — à employer cette expression et surtout à en comprendre la portée dans la politique de l'eau en France. Le principal grief est qu'il s'agit d'une traduction inappropriée de l'anglais *water body*, l'expression française n'ayant pas la même résonance géographique qu'elle a en anglais. Nous sommes là devant un phénomène de rejet de type « professionnel » d'un nouveau concept dans la même strate ou dans une autre strate discursive, comme nous l'avons vu avec la BDLISA. Aucun arrêté ministériel ne pourra y remédier en accélérant la lente assimilation du concept dans la langue, y compris dans la langue courante, si le concept s'avère véritablement opérationnel, ce qui n'est pas le cas actuellement.

5.4.2 Bilinguisme et relations des unités lexicales

Pour prendre un autre exemple de la problématique du lexique, le concept de crue est bien distinct du concept d'inondation, en français, ce qui n'est pas le cas en anglais, où l'emploi des mots *flood* et *flooding* domine pour évoquer les inondations, sans qu'il y ait l'équivalent pour évoquer les crues. Pour distinguer entre les crues et les inondations dans l'ontologie du domaine,

nous avons créé deux classes d'objets concrets différents, en proposant respectivement *High river flow* et *Flood* comme noms des classes dans la version anglaise, ce qui reste à valider en langue native. L'objectif de l'ontologie est de ne pas mélanger la crue de la Seine de 1905 à Paris et l'inondation consécutive de la ville. L'évocation de cet événement en anglais sera du genre *The 1905 flooding of Paris by the river Seine* plutôt qu'une traduction artificielle du type **The 1905 high river flow of the river Seine in Paris*. Chaque langue crée ses propres concepts pour exprimer la même chose sous des approches conceptuelles différentes.

Le traitement des inondations parmi d'autres aléas du domaine nous a conduit à écarter le concept d'inondation, trop vague, pour le remplacer par les trois concepts d'inondation de rivière, de pluie-inondation et de vagues-submersion qui permettent de regrouper en trois classes des phénomènes qui n'ont pas les mêmes caractéristiques et qui sont reconnus comme différents en tant qu'objets de la vigilance administrative. L'expression *flash flood* est plus parlante que « pluie-inondation » mais correspond à la même réalité, de même que « *river flood* », pour « inondation de rivière », et « *coastal submersion* », pour « vagues-submersion ». Les expressions du français administratif ne sont pas attestées dans les *media* et pour un usage au quotidien. Elles n'ont pas été introduites dans la strate discursive courante du lexique.

L'unité lexicale « 4-C Inondation/*4-C Flood* » a été retenue dans la strate discursive courante. Les variantes « Inondation/*Flooding* » et « Inondation/*Inundation* » figurent parmi les unités récurrentes comme noms récurrents alternatifs de cette unité lexicale. La figure 5.56

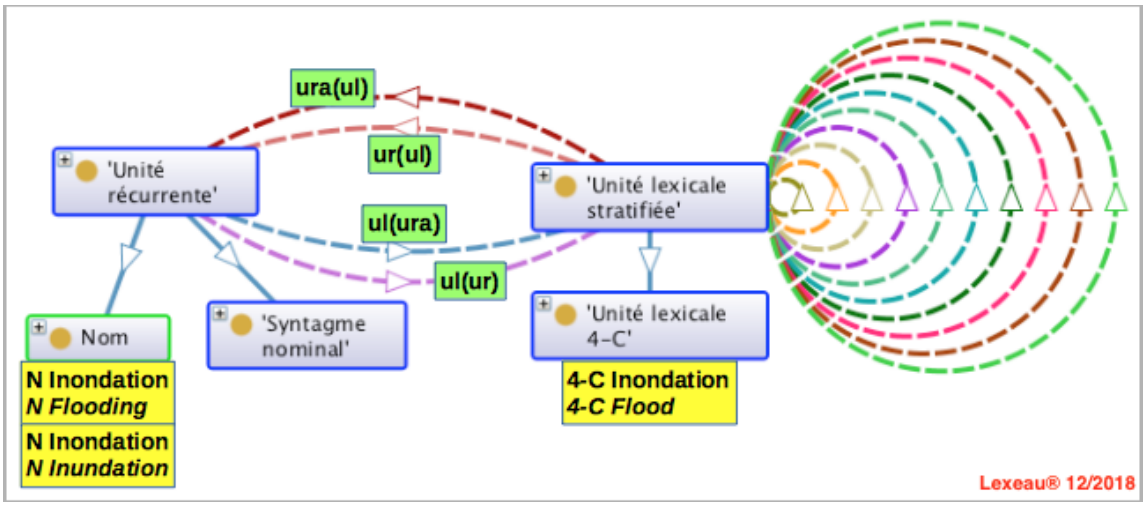


FIGURE 5.56 – Graphe relationnel des noms récurrents alternatifs de « 4-C Inondation »

Le rôle des unités récurrentes alternatives dans le modèle du lexique est différent de celui des unités discursives. Les unités discursives reprennent les libellés des unités lexicales stratifiées auxquelles elles sont reliées. Mais les unités lexicales peuvent être reliées à des unités récurrentes avec d'autres libellés dans la mesure où ces dernières répondent à la même définition, ce qui est le cas dans l'exemple de la figure 5.56.

La relation d'hyperonyme à hyponymes de « 4-C Inondation » avec les trois unités de la strate discursive technico-scientifique renvoie au discours scientifiques et au discours administratif de la vigilance. La figure 5.57 présente les relations lexicales correspondantes.

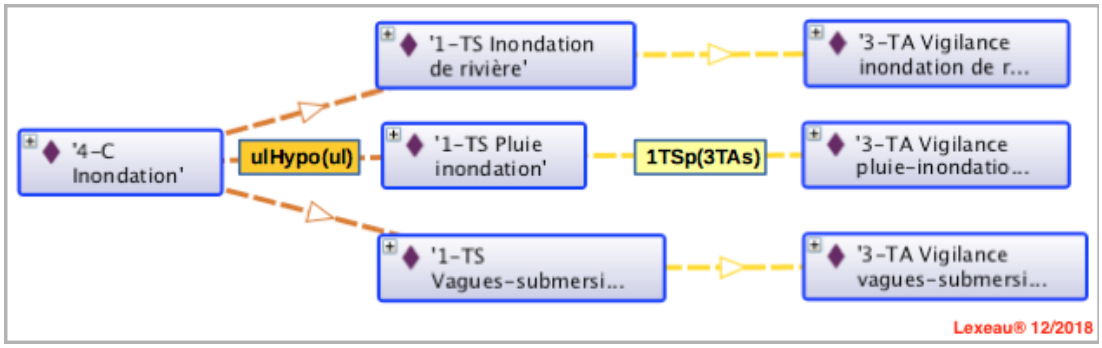


FIGURE 5.57 – Relations lexicales de l'entrée « 4-C Inondation »

5.4.3 Des unités discursive bilingues centrales ou alternatives

Les libellés d'une unité discursive sont déterminés dans les deux langues dès la transposition d'un concept édité en deux ou en une seule langue dans l'ontologie du domaine, après sa transposition dans le lexique comme unité lexicale pivot. Dans la suite de l'enrichissement du lexique en unités lexicales satellites ou par la mise en oeuvre de relations lexicales « classiques » comme la relation d'hyperonyme à hyponyme, des changements de libellés peuvent se produire dans une seule langue entre unités lexicales associées. Ces changements sont gérés au niveau des unités discursives en distinguant entre deux types d'unités, les unités discursives centrales, tributaire du lien du lexique avec l'ontologie du domaine, et les unités discursives alternatives qui permettent de gérer les variantes de libellé dans une seul langue. Ceci apparait dans le cas des prélèvements d'eau. Les deux unités discursives bilingues sont présentées dans les figures 5.58 et 5.59. Les libellés des deux unités lexicales de la strate discursive technico-administrative (uniquement en français) ne sont pas repris dans l'unité discursive à laquelle elles sont rattachées, ce qui apparait dans la figure 5.58. Ces libellés ne correspondent actuellement à aucun usage attesté dans des textes publiés.

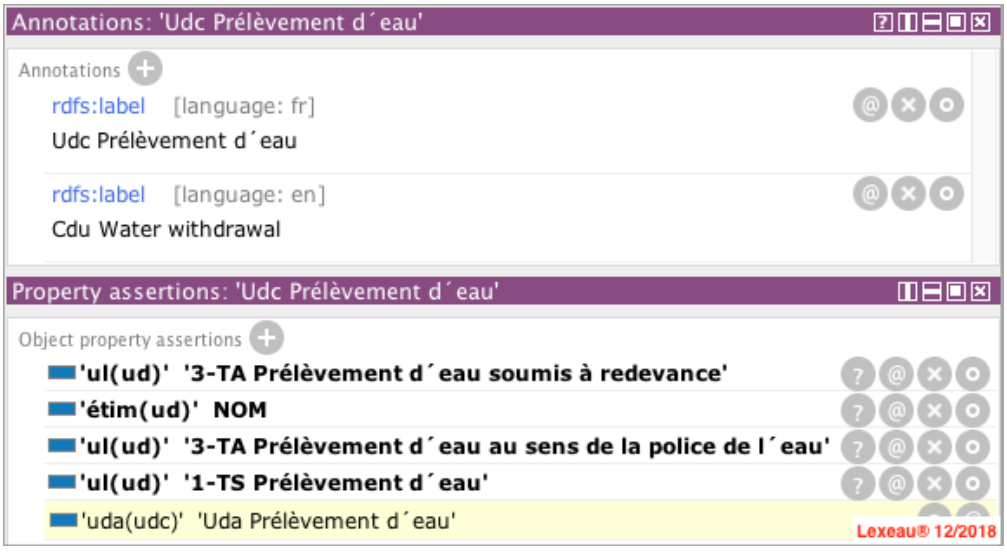


FIGURE 5.58 – L'unité discursive centrale « Udc Prélèvement d'eau »

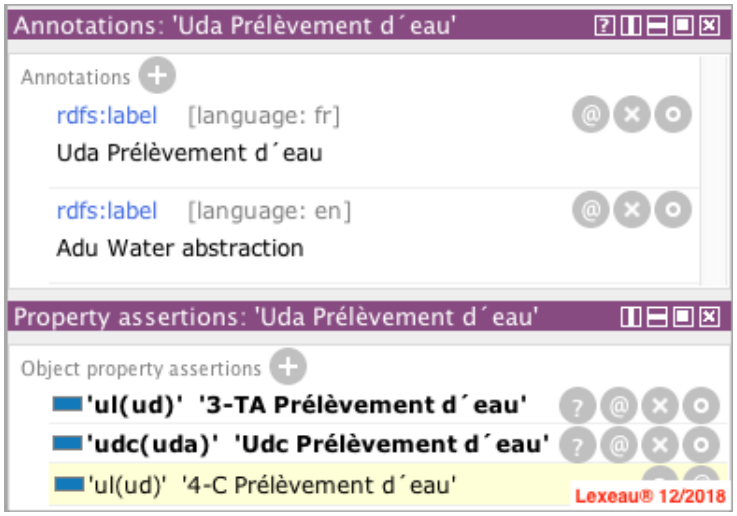


FIGURE 5.59 – L'unité discursive alternative « Uda Prélèvement d'eau »

L'unité discursive alternative bilingue donne accès aux unités lexicales « 3-TA Prélèvement d'eau » et « 4-C Prélèvement d'eau » présentées dans les figures 5.60 et 5.61. En français, l'unité lexicale « 3-TA Prélèvement d'eau » est l'hyperonyme des deux autres unités lexicales de la même strate, ce qui apparait dans les propriétés de la figure 5.60.

Cette unité lexicale n'est pas définie dans la version française du lexique en plus de ses hyponymes pour éviter une redondance entre les définitions et une ambiguïté dans l'emploi de l'expression. L'unité bilingue a été imposée dans le lexique par l'appellation officielle (*water abstraction*) qui est reprise dans la définition présentée dans la figure 5.62, définition tirée d'un texte accessible sur le site du gouvernement britannique. Elle garde néanmoins la qualité d'unité lexicale satellite de l'unité pivot « 1-TS Prélèvement d'eau », ce qui apparait dans les propriétés de la figure 5.60. On voit que les relations d'unité pivot à unité satellite peuvent s'établir sans que les deux libellés coïncident dans les deux langues.

Cet exemple illustre la gestion « a minima » des synonymes. La relation lexicale de synonymie n'est pas introduite dans le lexique compte tenu des difficultés à la définir. Les notions d'unité discursive centrale et alternative et d'unité récurrente permettent de gérer des variantes de libellé à partir du choix des libellés effectués dans l'ontologie du domaines et dans les unités lexicales stratifiées dotées d'une définition. Ces choix restent arbitraires et peuvent être modifiés à l'occasion d'une révision, à la condition de faire suivre les modifications, au delà de l'ontologie et des unités lexicales, dans les classes d'unités discursives et récurrentes. Ces choix doivent être validés en langue native.

Dans notre exemple, le libellé en anglais de *water withdrawal* a été **choisi par nous** en traduction du concept scientifique de « prélèvement d'eau » introduit dans l'ontologie du domaine sous la forme d'une classe d'objets bilingues. Ce choix a été conservé pour transposer le nom de la classe dans l'unité lexicale « 1-TS Prélèvement d'eau ». Le choix du libellé *water abstraction* dans la strate discursive courante nous a paru approprié dans la mesure où il s'est imposé dans la strate discursive administrative dans la version anglaise (sans être défini dans la version française). Ceci a débouché sur la définition bilingue de l'unité lexicale « 4-C Prélèvement d'eau » de la figure 5.63 en laissant de côté la variante *water withdrawal* qui nous

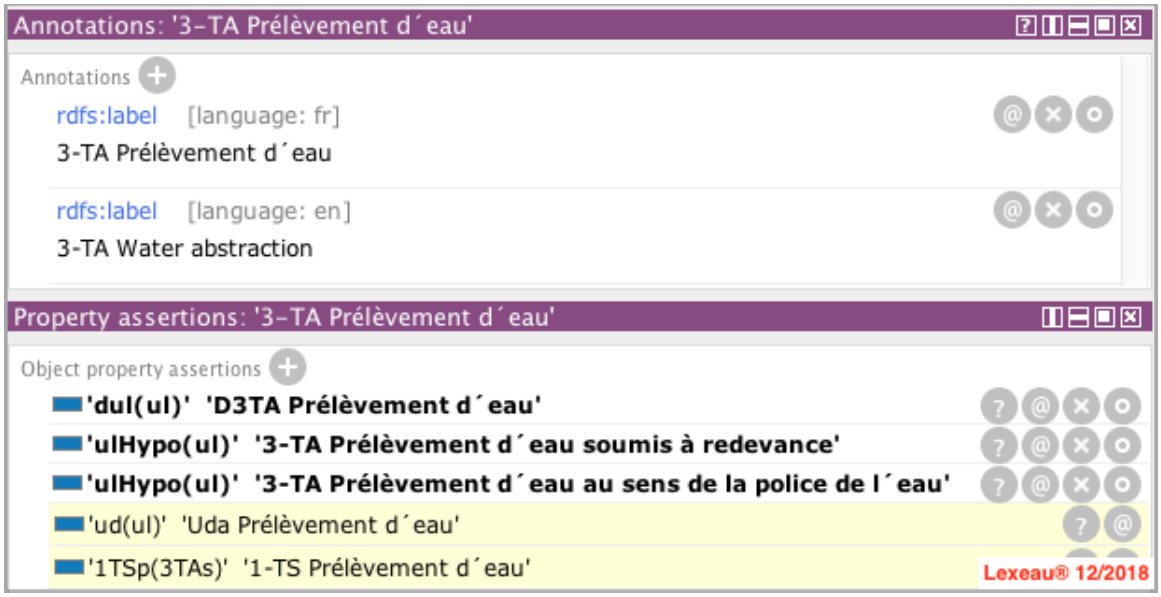


FIGURE 5.60 – L'unité lexicale « 3-TA Prélèvement d'eau »

FIGURE 5.61 – L'unité lexicale « 4-C Prélèvement d'eau »

a paru moins utilisée.

Le problème ne s'est pas posé de la même manière avec les variantes *flooding* et *inundation* de l'anglais *flood*, associé au français « inondation » parce que ces libellés ne figurent pas parmi ceux des unités lexicales. La solution a été de les intégrer comme noms récurrents alternatif parmi les unités récurrentes du lexique (cf. fig. 5.56).

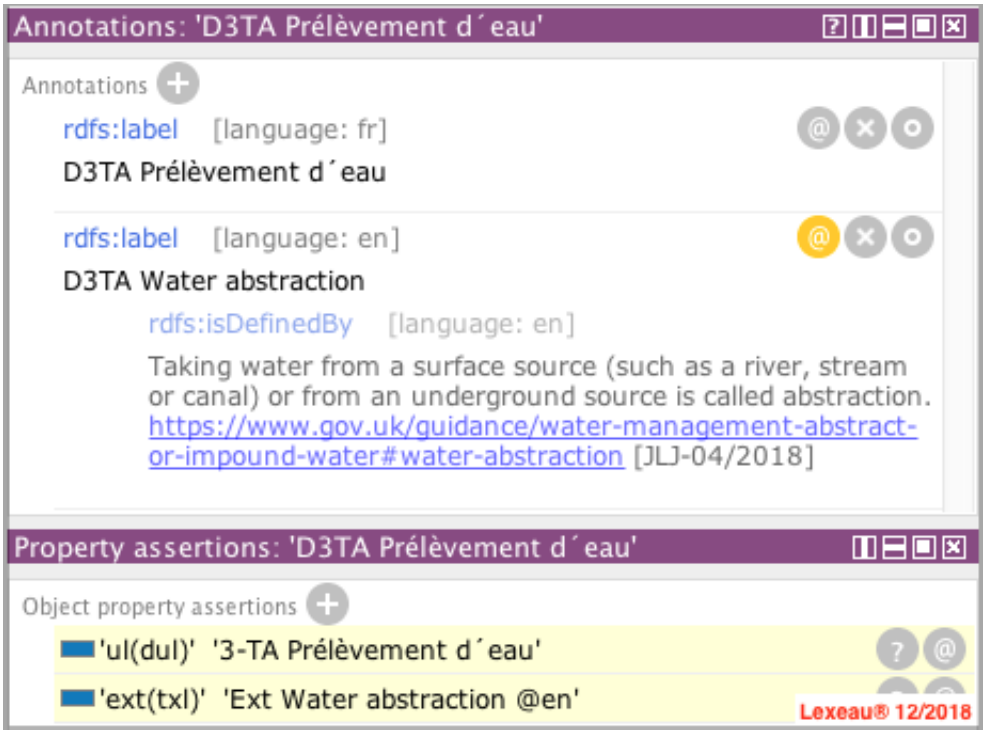


FIGURE 5.62 – Définition de l'unité lexicale « 3-TA Prélèvement d'eau »

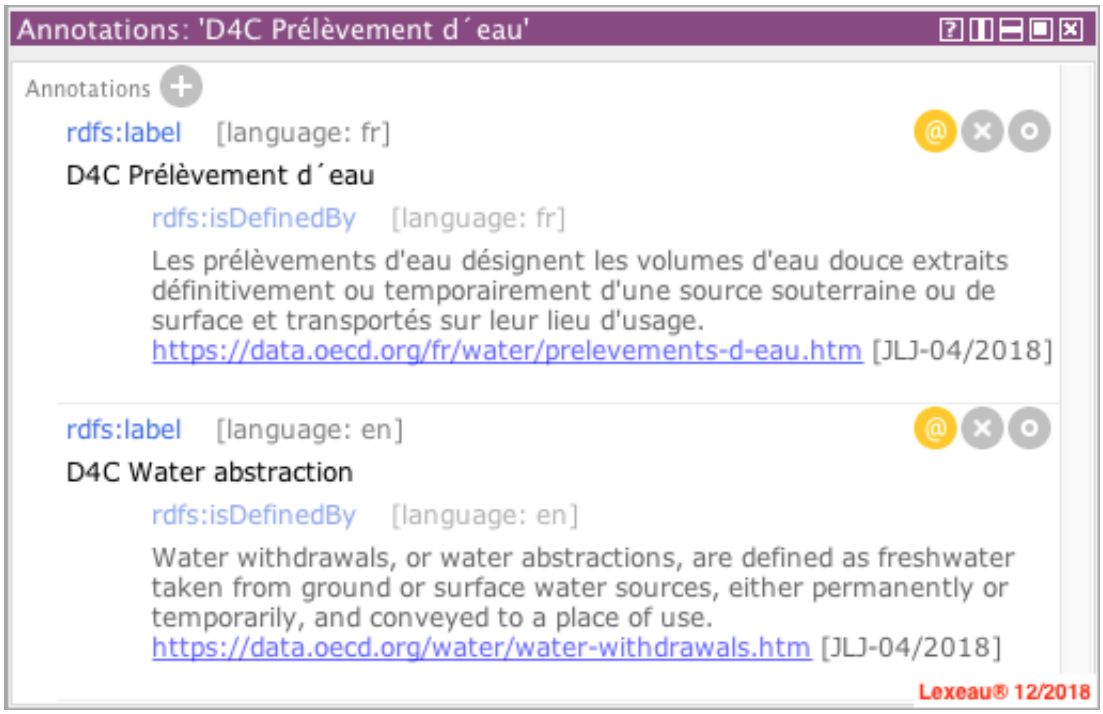


FIGURE 5.63 – Définition de l'unité lexicale « 4-C Prélèvement d'eau »

Conclusion

Ce travail a montré qu'il est possible de construire un lexique de l'eau stratifié qui mette en correspondance les mots et les expressions employés dans les discours du domaine de l'eau adaptés à des auditoires différents sur des problématiques communes, comme l'usage quantitatif de la ressource et la prévention contre les risques naturels.

L'étude de la problématique linguistique en sémantique lexicale comme en dictionnairique a débouché sur une stratification du lexique proprement dit en quatre strates discursives, avec des unités lexicales monosémiques issues en partie des concepts édités dans une ontologie des « objets » de connaissance du domaine de l'eau. Le modèle fait la distinction entre les unités lexicales « pivot », issues des concepts, et les unités lexicales « satellites », associées aux unités lexicales pivot dans d'autres strates discursives avec les mêmes libellés ou des libellés proches. Les unités discursives non stratifiées reprennent les libellés des unités lexicales. Elles forment, avec les unités récurrentes associées aux unités lexicales avec d'autres libellés alternatifs, les entrées traditionnelles dans un dictionnaire qui permettent dans notre cas d'accéder aux unités lexicales stratifiées. Les noms propres issus des instances des classes d'entités typées de l'ontologie du domaine sont associés au lexique sous forme de noms propres « lexicalisés », avec le nom de leur classe d'origine transposé dans le lexique.

Les textes du lexique sont des définitions, des exemples d'emploi et des gloses des unités lexicales, ainsi que des exemples d'emploi des noms propres lexicalisés. Les gloses sont rédigées pour les unités lexicales « pivot » issues de l'ontologie du domaine, sur la base du graphe relationnel des concepts proches du concept source. Elles permettent de mettre en réseaux les unités lexicales sur une base sémantique. En dehors des gloses, les textes du lexique sont extraits des documents enregistrés dans l'ontologie du domaine et/ou de pages référencées sur l'internet.

Les textes des documents du domaine peuvent être regroupés en corpus par thème et/ou par discipline. Les thèmes du domaine ont fait l'objet d'une nomenclature hiérarchisée pour faciliter la constitution des corpus. Les disciplines et les « matières » mises à contribution dans la connaissance du domaine sont regroupées dans une liste ouverte, non hiérarchisée.

L'enrichissement du lexique a été testé sur deux démarches, destinées à se rencontrer pour faire du lexique un produit des textes et de l'ontologie du domaine. La première démarche, qui part de l'ontologie du domaine, est inspirée de la façon dont les spécialistes d'un domaine, sur un plan technique, scientifique, juridique ou administratif, tirent leur vocabulaire des concepts qu'ils utilisent dans leur travail quotidien. La deuxième démarche part d'un corpus de textes sur un thème donné pour en extraire, par textométrie, des substantifs et des syntagmes nominaux récurrents qui sont intégrés dans les strates discursives du lexique, avec, si nécessaire,

un enrichissement de l'ontologie pour articuler les nouvelles unités lexicales pivot et satellites. Dans les tests que nous avons effectués, les textes du domaine ont été les arbitres de l'enrichissement du lexique, y compris pour des concepts du langage courant qui ont été intégrés dans l'ontologie à cause de leurs liens avec des concepts scientifiques et administratifs.

Les tests effectués ont permis de développer un modèle robuste de réalisation d'un lexique stratifié du domaine de l'eau. Des essais en vraie grandeur du dispositif de consultation et d'enrichissement sont nécessaires, avec la mise en œuvre informatisée des logiciels utilisés « à la main » dans notre travail et une interface appropriée sur un portail internet testé par des utilisateurs et des contributeurs. Nous entrons là dans le domaine du projet LEXEAU®. Nous espérons qu'il permettra d'associer des linguistes intéressés par des recherches liées à des applications opérationnelles à la réalisation d'un projet pilote, confié à un groupe de projet, qui pourrait être lancé en 2019 avec un financement limité.

Avis au lecteur : Les annexes sont présentées dans une version provisoire au 15/12/2018, très incomplète. Elles devraient être finalisées avant la présentation de la thèse.

Annexe A

Annexes du chapitre 1

A.1 Exemples de définition d'unités lexicales

A.1.1 Définition des prélèvements d'eau

La figure A.1 présente la définition de l'unité lexicale pivot « 1-TS Prélèvement d'eau ».

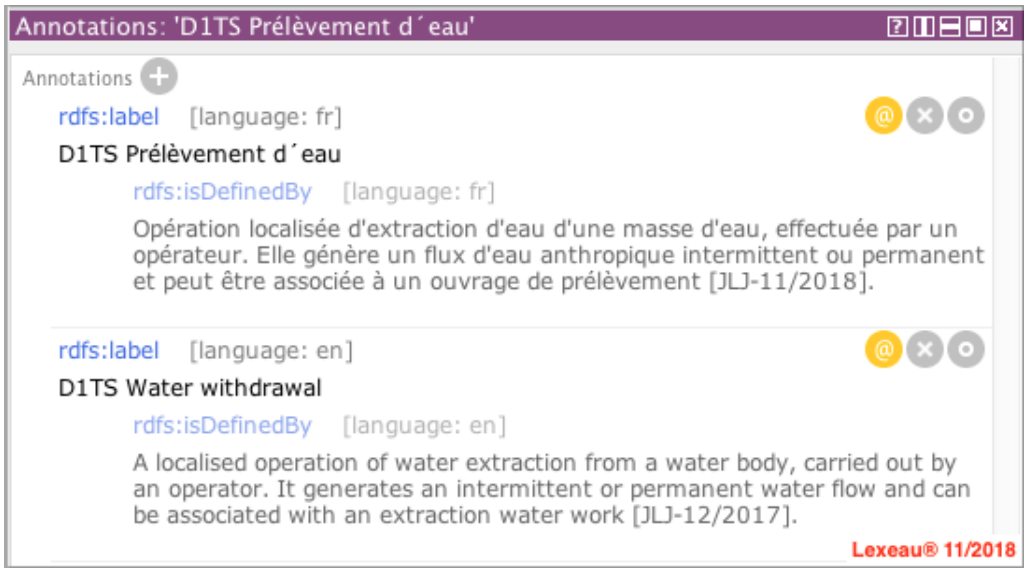


FIGURE A.1 – Définition bilingue de l'unité lexicale pivot « 1-TS Prélèvement d'eau »

Les définitions des prélèvements d'eau au sens de la police de l'eau et de ceux qui sont soumis à redevance sont présentées dans les figures A.2 et A.3. Elles ne sont valables qu'en français.

La définition bilingue des prélèvements d'eau dans le langage courant, en tant que volume de l'eau prélevée, est celle que propose un site de l'OCDE en lui associant des données par habitant et par pays (fig. A.4).

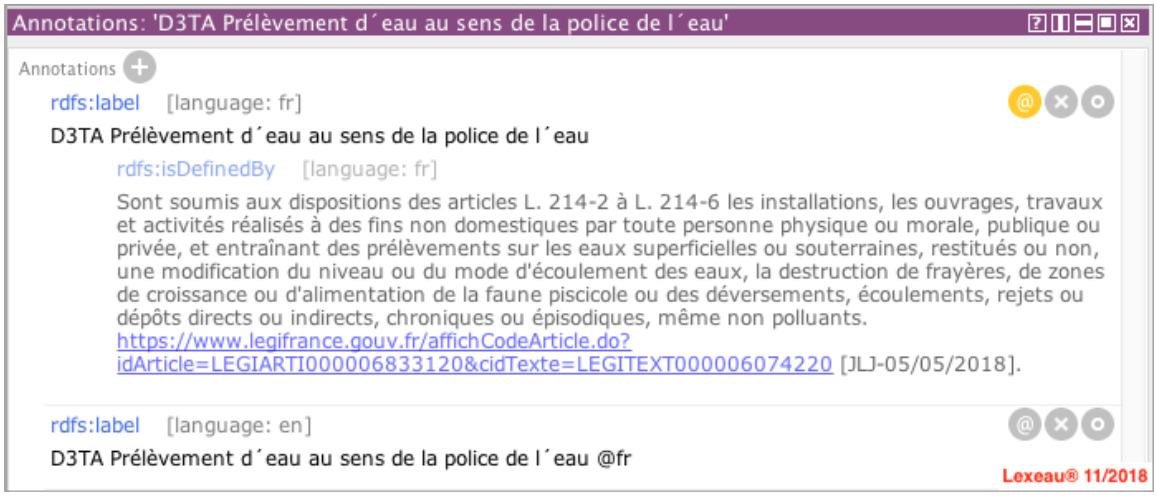


FIGURE A.2 – Définition des prélèvements d'eau au sens de la police de l'eau

FIGURE A.3 – Définition des prélèvements d'eau soumis à redevance

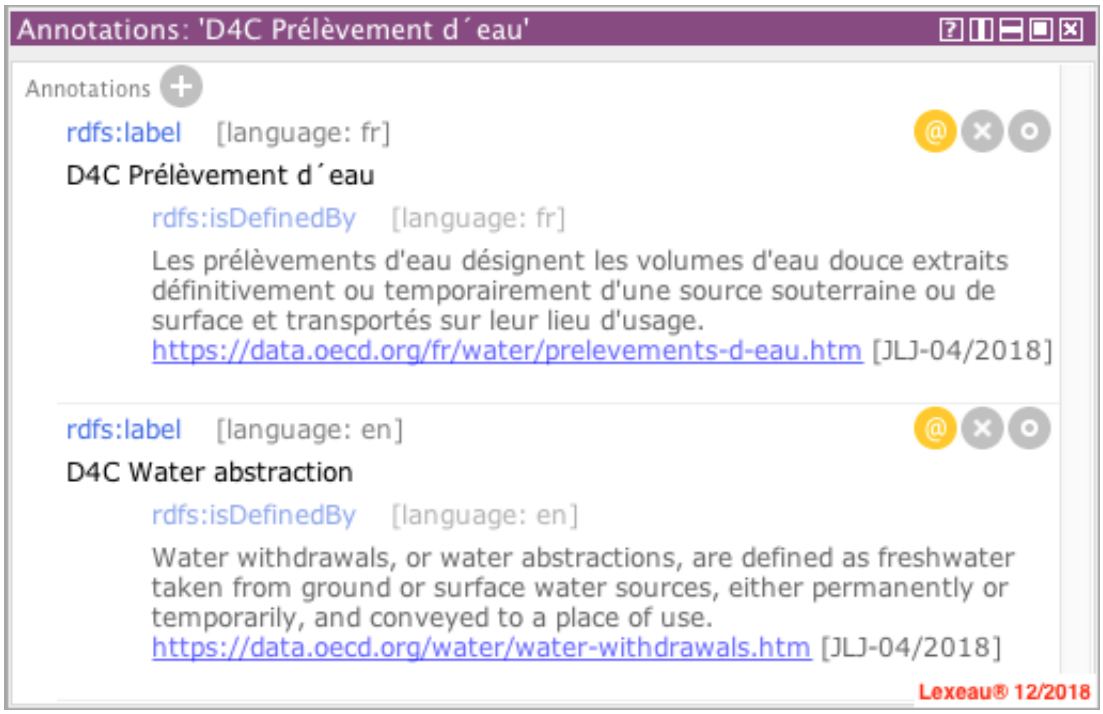


FIGURE A.4 – Définition des prélèvements d'eau selon l'OCDE

A.1.1.1 Exemple de définition des aléas

La figure A.5 présente la définition de l'entrée « Aléa/*Hazard* » du lexique dans la strate discursive technico-scientifique, tirée de Wikipédia [1].

FIGURE A.5 – Définition des aléas en termes de prévention de risques naturels

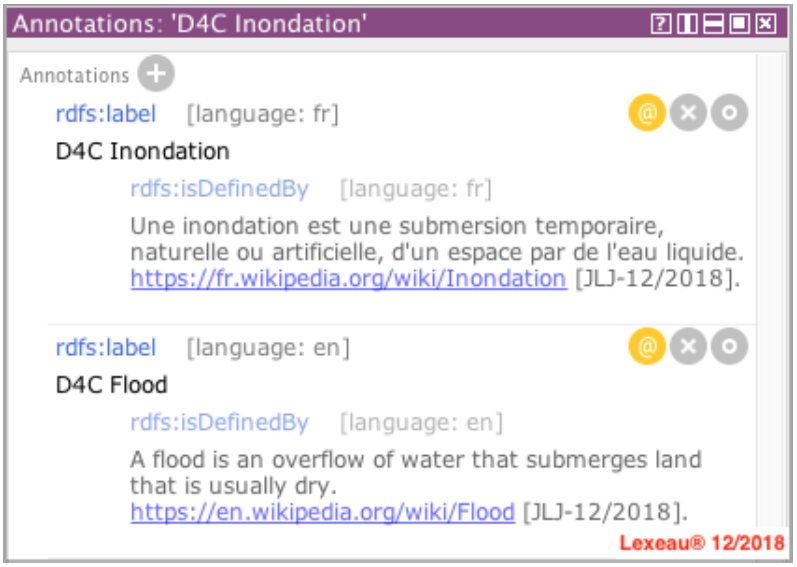


FIGURE A.6 – Définition des inondations dans la strate discursive courante

1. https ://fr.wikipedia.org/wiki/Al%C3%A9a

Annexe B

Annexes du chapitre 2

B.1 Étiquettes morphosyntaxiques

B.1.1 Étiquettes morphosyntaxiques du français

TABLE B.1 – Les étiquettes morphosyntaxiques du français (format d'import *odt* dans TXM)

abr	abréviation
ad	adjectif
adv	adverbe
det :art	article
det :pos	pronom possessif (ma,ta...)
int	interjection
kon	conjonction
nam	nom propre
nom	nom
num	numeral
pro	pronom
pro :dem	pronom démonstratif
pro :ind	pronom indéfini
pro :per	pronom personnel
pro :pos	pronom possessif (mien, tien...)
pro :rel	pronom relatif
prp	préposition
prp :det	préposition plus article (au, du, aux, des)
pun	ponctuation
pun :cit	ponctuation de citation
sent	balise de phrase
sym	symbole
ver :cond	verbe au conditionnel
ver :futu	verbe au futur
ver :impe	verbe à l'impératif
ver :impf	verbe à l'imparfait
ver :infi	verbe à l'infinitf
ver :pper	verbe au participe passé
ver :ppre	verbe au participe présent
ver :pres	verbe au présent
ver :simp	verbe au passé simple
ver :subi	verbe au subjonctif imparfait
ver :subp	verbe au subjonctif présent

B.1.2 Étiquettes morphosyntaxiques de l'anglais

Les figures B.1 et B.2 sont des copies d'écran du tableau accessible sur le site de l'Université d'État du Washington [1]

POS Tag	Description	Example
CC	coordinating conjunction	and, but, or, &
CD	cardinal number	1, three
DT	determiner	the
EX	existential there	*there is*
FW	foreign word	d'œuvre
IN	preposition/subord. conj.	in,of,like,after,whether
IN/that	complementizer	that
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	(1),
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NP	proper noun, singular	John
NPS	proper noun, plural	Vikings
PDT	predeterminer	*both the boys*
POS	possessive ending	friend's
PP	personal pronoun	I, he, it
PP$	possessive pronoun	my, his
RB	adverb	however, usually, here, not
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give up
SENT	end punctuation	?, !, .
SYM	symbol	@, +, *, ^, |, =
TO	*to*	*to go, to him*
UH	interjection	uhhuhhuhh
		Lexeau® 2018

FIGURE B.1 – Étiquettes morphosyntaxiques de l'anglais (début)

1. https://courses.washington.edu/hypertxt/csar-v02/penntable.html

POS Tag	Description	Example
VB	verb be, base form	be
VBD	verb be, past	was|were
VBG	verb be, gerund/participle	being
VBN	verb be, past participle	been
VBZ	verb be, pres, 3rd p. sing	is
VBP	verb be, pres non-3rd p.	am|are
VD	verb do, base form	do
VDD	verb do, past	did
VDG	verb do gerund/participle	doing
VDN	verb do, past participle	done
VDZ	verb do, pres, 3rd per.sing	does
VDP	verb do, pres, non-3rd per.	do
VH	verb have, base form	have
VHD	verb have, past	had
VHG	verb have, gerund/participle	having
VHN	verb have, past participle	had
VHZ	verb have, pres 3rd per.sing	has
VHP	verb have, pres non-3rd per.	have
VV	verb, base form	take
VVD	verb, past tense	took
VVG	verb, gerund/participle	taking
VVN	verb, past participle	taken
VVP	verb, present, non-3rd p.	take
VVZ	verb, present 3d p. sing.	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when
:	general joiner	;, -, --
$	currency symbol	$, £
		Lexeau® 10/2018

FIGURE B.2 – Étiquettes morphosyntaxiques de l'anglais (fin)

B.2 Annexes du chapitre 3

B.3 Recherche des modèles mathématiques (Floodrisk 2016)

FLR-02, partition par texte	T	02001	02002	02003	02004	Nb de	Nb de	Numéro
Base nominale « model »	22 317	6 612	7 550	4 166	3 989	textes	lemmes	requête
117 syntagmes étiquetés (début)	F							
model_nn	89	45	5	23	16	4	1	1
model_nns	36	5	16	13	2	4	1	1
model_np	2	0	2	0	0	1	1	1
6-hour_jj model_nn	1	1	0	0	0	1	2	5
agency_np model_nn	1	0	0	0	1	1	2	2
arpege–climat_np model_nn	3	0	0	3	0	1	2	2
big_jjr model_nn	1	1	0	0	0	1	2	5
bivariate_jj model_nns	1	0	1	0	0	1	2	5
candidate_nn model_nns	1	0	1	0	0	1	2	2
circulation_nn model_nn	2	2	0	0	0	1	2	2
climate_nn model_nn	2	1	0	1	0	2	2	2
climate_nn model_nns	2	1	1	0	0	2	2	2
climate_nn model_np	1	0	1	0	0	1	2	2
climatic_jj model_nn	1	0	0	1	0	1	2	5
cnrm-cm5_jj model_nn	1	1	0	0	0	1	2	5
damage_nn model_nn	2	0	0	2	0	1	2	2
delft3d_jj model_nn	3	3	0	0	0	1	2	5
digital_jj model_nn	1	0	0	1	0	1	2	5
disaster_nn model_nns	1	0	0	1	0	1	2	2
elevation_nn model_nn	1	1	0	0	0	1	2	2
estuary_np model_nn	1	0	0	0	1	1	2	2
euro-cordex_np model_nn	1	1	0	0	0	1	2	2
external_jj model_nn	1	1	0	0	0	1	2	5
fitted_jj model_nns	1	0	1	0	0	1	2	5
full_jj model_nn	1	1	0	0	0	1	2	5
global_jj model_nns	1	1	0	0	0	1	2	5
good_jj model_nn	1	1	0	0	0	1	2	5
hazard_nn model_nn	3	0	0	3	0	1	2	2
hazard_nn model_nns	1	0	0	1	0	1	2	2
high-resolution_nn model_nn	1	1	0	0	0	1	2	2
hydraulic_np model_nns	1	0	0	0	1	1	2	2
hydrodynamic_jj model_nn	5	5	0	0	0	1	2	5
iii_np model_nn	1	0	0	0	1	1	2	2
impact_nn model_nn	1	0	0	1	0	1	2	2
impact_nn model_nns	7	0	0	7	0	1	2	2
isba_np model_nn	1	0	0	1	0	1	2	2
mesoscale_np model_np	1	0	1	0	0	1	2	2
model_nn cfflood_np	1	1	0	0	0	1	2	3
model_nn control_nn	1	1	0	0	0	1	2	3
model_nn domain_nn	1	0	0	0	1	1	2	3

Lexeau® 10/2018

FIGURE B.3 – Recherche des noms modèles mathématiques (1/3, corpus FLR-02)

FLR-02, partition par texte	T	02001	02002	02003	02004	Nb de	Nb de	Numéro
Base nominale « model »	22 317	6 612	7 550	4 166	3 989	textes	lemmes	requête
117 syntagmes étiquetés (suite 1)	F							
model_nn eu-dem_np	1	1	0	0	0	1	2	3
model_nn grid_nn	1	0	1	0	0	1	2	3
model_nn output_nn	1	1	0	0	0	1	2	3
model_nn run_nns	2	1	0	0	1	2	2	3
model_nn set-up_nn	1	1	0	0	0	1	2	3
model_nn set-up_nns	1	1	0	0	0	1	2	3
model_nn simulation_nns	1	1	0	0	0	1	2	3
model_nn space_nn	1	0	0	0	1	1	2	3
model_nn swan_np	1	0	1	0	0	1	2	3
model_nn validation_nn	1	1	0	0	0	1	2	3
model_nns result_nns	2	0	2	0	0	1	2	3
model_np version_nn	1	0	1	0	0	1	2	3
multivariate_jj model_nns	1	0	1	0	0	1	2	5
new_jj model_nn	1	0	0	0	1	1	2	5
numerical_jj model_nns	1	1	0	0	0	1	2	5
other_jj model_nns	2	2	0	0	0	1	2	5
overflow_nn model_nn	2	0	0	2	0	1	2	2
polynomial_jj model_nns	1	0	1	0	0	1	2	5
pot_np model_nns	1	0	1	0	0	1	2	2
run-off_nn model_nn	2	0	0	2	0	1	2	2
several_jj model_nns	1	0	0	1	0	1	2	5
surge_nn model_nn	1	0	1	0	0	1	2	2
swan_np model_nn	2	0	0	0	2	1	2	2
terrain_nn model_nn	1	1	0	0	0	1	2	2
tide_nn model_nn	2	2	0	0	0	1	2	2
transformation_nn model_nn	1	0	0	0	1	1	2	2
trend_nn model_nns	1	0	1	0	0	1	2	2
twodimensional_jj model_nn	1	0	1	0	0	1	2	5
value_nn model_nns	3	0	3	0	0	1	2	2
wave_nn model_nn	3	0	1	0	2	2	2	2
wind-driven_jj model_nn	1	1	0	0	0	1	2	5
1d_jj wave_nn model_nn	1	0	0	0	1	1	3	11
2d_jj wave_nn transformation_nn model_nn	1	0	0	0	1	1	3	12
6-hour_jj model_nn output_nn	1	1	0	0	0	1	3	6
big_jjr model_nn cfflood_np	1	1	0	0	0	1	3	6
ccr_np impact_nn model_nns	2	0	0	2	0	1	3	7
ccr's_np impact_nn model_nns	1	0	0	1	0	1	3	7
climate_nn model_nn control_nn	1	1	0	0	0	1	3	4
dedicated_jj tide_nn model_nn	1	1	0	0	0	1	3	11
different_jj candidate_nn model_nns	1	0	1	0	0	1	3	11
							Lexeau® 10/2018	

FIGURE B.4 – Recherche des noms des modèles mathématiques (2/3, corpus FLR-02)

FLR-02, partition par texte	T	02001	02002	02003	02004	Nb de	Nb de	Numéro
Base nominale « model »	22 317	6 612	7 550	4 166	3 989	textes	lemmes	requête
117 syntagmes étiquetés (fin)	F							
digital_jj elevation_nn model_nn	1	1	0	0	0	1	3	11
digital_jj terrain_nn model_nn	1	1	0	0	0	1	3	11
drought_jj hazard_nn model_nn	1	0	0	1	0	1	3	11
drought_nn impact_nn model_nn	1	0	0	1	0	1	3	7
drought_nn impact_nn model_nns	1	0	0	1	0	1	3	7
elevation_nn model_nn eu-dem_np	1	1	0	0	0	1	3	4
environment_np agency_np model_nn	1	0	0	0	1	1	3	7
estuaries_np hydraulic_np model_nns	1	0	0	0	1	1	3	7
exe_np estuary_np model_nn	1	0	0	0	1	1	3	7
extreme_jj value_nn model_nns	3	0	3	0	0	1	3	11
flood_nn hazard_nn model_nn	1	0	0	1	0	1	3	7
flood_np hazard_nn model_nn	1	0	0	1	0	1	3	7
full_jj model_nn run_nns	1	1	0	0	0	1	3	6
general_jj circulation_nn model_nn	1	1	0	0	0	1	3	11
hydrodynamic_jj model_nn set-up_nn	1	1	0	0	0	1	3	6
insurance_nn impact_nn model_nns	1	0	0	1	0	1	3	7
large-scale_jj climate_nn model_nn	1	0	0	1	0	1	3	11
mesoscale_np model_np version_nn	1	0	1	0	0	1	3	4
natural_jj disaster_nn model_nns	1	0	0	1	0	1	3	11
own_jj impact_nn model_nns	1	0	0	1	0	1	3	11
regional_jj circulation_nn model_nn	1	1	0	0	0	1	3	11
regional_jj climate_nn model_nns	1	0	1	0	0	1	3	11
regional_jj climate_nn model_np	1	0	1	0	0	1	3	11
river_nn overflow_nn model_nn	2	0	0	2	0	1	3	7
storm_nn surge_nn model_nn	1	0	1	0	0	1	3	7
surface_nn run-off_nn model_nn	1	0	0	1	0	1	3	7
surge_nn model_nn grid_nn	1	0	1	0	0	1	3	4
swan_np wave_nn model_nn	1	0	0	0	1	1	3	7
tpxo8_np tide_nn model_nn	1	1	0	0	0	1	3	7
value_nn model_nns result_nns	1	0	1	0	0	1	3	4
wave_nn model_nn swan_np	1	0	1	0	0	1	3	4
wave_nn transformation_nn model_nn	1	0	0	0	1	1	3	7
wavewatch_np iii_np model_nn	1	0	0	0	1	1	3	7
in-house_jj flood_nn hazard_nn model_nn	1	0	0	1	0	1	4	12
in-house_jj surface_nn run-off_nn model_nn	1	0	0	1	0	1	4	12
storm_nn surge_nn model_nn grid_nn	1	0	1	0	0	1	4	8
teign_np estuaries_np hydraulic_np model_nns	1	0	0	0	1	1	4	10
							Lexeau® 10/2018	

FIGURE B.5 – Recherche des noms des modèles mathématiques (3/3, corpus FLR-02)

Annexe C

Notations de l'ontologie du domaine (*à finaliser*)

C.1 Préfixes des instances de certaines classes

Table C.1 – Glossaire des préfixes en français et en anglais

Préfixe	Forme développée	*Prefix*	*Full form*
1TS/1-TS	Strate technico-scientifique	*1TS/1-TS*	*Technico-scientific stratum*
2DI/2-DI	Strate décisionnelle incitative	*2DI/2-DI*	*Decisional incitative stratum*
3TA/3-TA	Strate technico-administrative	*3TA/3-TA*	*Technico-administrative stratum*
4C/4-C	Strate courante	*4C/4-C*	*Current stratum*
D1TS	Définition unité lexicale 1TS	*D1TS*	*1TS lexical unit definition*
D2DI	Définition unité lexicale 2DI	*D1TS*	*2DI lexical unit definition*
D3TA	Définition unité lexicale 3TA	*D1TS*	*3TA lexical unit definition*
D4C	Définition unité lexicale 4C	*D1TS*	*4C lexical unit definition*
Uda	Unité discursive alternative	*Adu*	*Alternative discursive unit*
Udc	Unité discursive centrale	*Cdu*	*Central discursive unit*

C.2 Acronymes des noms de classes

TABLE C.2 – Glossaire des acronymes en français et en anglais

Abréviation	Forme développée	*Abbreviation*	*Full form*
cvo	Compteur du volume d'eau	*vom*	*Water volume meter*
hu	Entité humaine	*hu*	*Human entity*
flx	Flux	*flw*	*Flow*
ijn	Injection d'eau	*ijn*	*Water injection*
mav	Mouvement d'eau aval	*dom*	*Downstream water movement*
mam	Mouvement d'eau amont	*upm*	*Upstream water movement*
me	Masse d'eau	*wb*	*Water body*
mve	Mouvement d'eau	*wmv*	*Water movement*
ovg	Ouvrage	*wrk*	*Work*
pri	Prise d'eau	*wtp*	*Water tapping*
prl	Prélèvement d'eau	*waw*	*Water withdrawal*
prp	Propriété dans 'Protégé'	*prp*	*Property in 'Protégé'*
rse	Réseau d'eau	*wnw*	*Water network*
ul	Unité lexicale	*lu*	*Lexical unit*

Annexe D

Propriétés

Les propriétés (*properties*) introduites dans la base de connaissance sont entendues au sens du logiciel 'Protégé', de l'Université de Stanford, éditeur de l'ontologie de cette base dans ses trois composantes (domaine de connaissance de l'eau, lexique de l'eau, discours sur l'eau).

TABLE D.1: Glossaire des propriétés

Propriété *Property*	Forme développée *Full form*	Prop. inverse *Inverse prop.*
1TSp(2DIs) *1TSp(2DIs)*	Unité lexicale pivot 1TS de l'unité lexicale satellite 2DI *1TS pivot lexical unit of the 2DI satellite lexical unit*	2DIs(1TSp) *2DIs(1TSp)*
1TSp(3TAs) *1TSp(3TAs)*	Unité lexicale pivot 1TS de l'unité lexicale satellite 3TA *1TS pivot lexical unit of the 3TA satellite lexical unit*	3TAs(1TSp) *3TAs(1TSp)*
1TSp(4Cs) *1TSp(4Cs)*	Unité lexicale pivot 1TS de l'unité lexicale satellite 4C *1TS pivot lexical unit of the 4C satellite lexical unit*	4Cs(1TSp) *4Cs(1TSp)*
2DIp(3TAs) *2DIp(3TAs)*	Unité lexicale pivot 2DI de l'unité lexicale satellite 3TA *2DI pivot lexical unit of the 3TA satellite lexical unit*	3TAs(2DIp) *3TAs(2DIp)*
2DIp(4Cs) *2DIp(4Cs)*	Unité lexicale pivot 2DI de l'unité lexicale satellite 4C *2DI pivot lexical unit of the 4C satellite lexical unit*2DI	4Cs(2DIp) *4Cs(2DIp)*
2DIs(1TSp) *2DIs(1TSp)*	Unité lexicale satellite 4C de l'unité lexicale pivot 1TS *2DI satellite lexical unit of the 1TS pivot lexical unit*	1TSp(2DIs) *1TSp(2DIs)*
3TAp(4Cs) *3TAp(4Cs)*	Unité lexicale pivot 3TA de l'unité lexicale satellite 4C *3TAp pivot lexical unit of the 4C satellite lexical unit*	4Cs(3TAp) *4Cs(3TAp)*
3TAs(1TSp) *3TAs(1TSp)*	Unité lexicale satellite 3TA de l'unité lexicale pivot 1TS *3TA satellite lexical unit of the 1TS pivot lexical unit*	1TSp(3TAs) *1TSp(3TAs)*
3TAs(2DIp) *3TAs(2DIp)*	Unité lexicale satellite 3TA de l'unité lexicale pivot 2DI *3TA satellite lexical unit of the 1TS pivot lexical unit*	2DIp(3TA) *2DIp(3TAs)*
4Cs(1TSp) *4Cs(1TSp)*	Unité lexicale satellite 4C de l'unité lexicale pivot 1TS *4C satellite lexical unit of the 1TS pivot lexical unit*	1TSp(4Cs) *1TSp(4Cs)*
4Cs(2DIp) *4Cs(2DIp)*	Unité lexicale satellite 4C de l'unité lexicale pivot 2DI *4C satellite lexical unit of the 2DI pivot lexical unit*	2DIp(4Cs) *2DIp(4Cs)*
4Cs(3TAp) *4Cs(3TAp)*	Unité lexicale satellite 4C de l'unité lexicale pivot 3TA *4C satellite lexical unit of the 3TA pivot lexical unit*	3TAp(4Cs) *3TAp(4Cs)*

TABLE D.1: Glossaire des propriétés (suite)

Propriété *Property*	Forme développée *Full form*	Prop. inverse *Inverse prop.*
aaaa(date) *yyyy(date)*	Année de la date *Year of the date*	date(aaaa) *date(yyyy)*
aléaP(cata) *pHzd(cata)*	Aléa primaire associé à la catastrophe *Primary hazard associated with the catastrophe*	cata(aléaP) *cata(pHzd)*
arti(dic) *arti(dic)*	Article du dictionnaire ou du lexique *Article of the dictionary or the lexicon*	dic(arti) *dic(arti)*
arti(prq) *arti(prc))*	Article du périodique *Article of the periodical*	prq(arti) *prc(arti)*
auto(pe) pe(auto)	Autorisation du plan d'eau Plan d'eau autorisé	*@fr auto(pe)* *@fr pe(auto)*
auto(prl) prl(auto)	Autorisation du prélèvement d'eau Prélèvement d'eau autorisé	*@fr auto(prl)* *@fr prl(auto)*
bge(pe) pe(bge)	Barrage de retenue du plan d'eau Plan d'eau retenu par le barrage	*dam(bw)* *bw(dam)*
bge(rte) rte(bge)	Barrage associé à la retenue d'eau Retenue d'eau associée au barrage	*dam(imp)* *imp(dam)*
cata(aléaP) *catapHzd)*	Catastrophe associée à l'aléa primaire *Catastrophe associated with the primary hazard*	aléaP(cata) *pHzd(cata)*
cvo(ovg) ovg(cvo)	Compteur du volume d'eau écoulé qui équipe l'ouvrage Ouvrage équipé du compteur du volume d'eau écoulé	*vom(wrk)* *wrk(vom)*
date(aaaa) *date(yyyy)*	Date de l'année *Date of the year*	aaaa(date) *yyyy(date)*
dic(arti) *dic(arti)*	Dictionnaire ou lexique de l'article *Dictionary or lexicon of the article*	arti(dic)) *dic(arti)*
flx(mam) mam(flx)	Flux d'eau généré par le mouvement d'eau amont (prl/pri) Mouvement d'eau amont (prl/pri) qui génère le flux d'eau	*flw(upm)* *upm(flw)*
flx(mav) mav(flx)	Flux d'eau récupéré par le mouvement d'eau aval (resti/ijn) Mouvement d'eau aval (resti/ijn) qui récupère le flux d'eau	*flw(dom)* *dom(flw)*
huOp(mve) mve(huOp)	Entité humaine qui opère le mouvement d'eau Mouvement d'eau opéré par l'entité humaine	*huOp(wmv)* *wmv(huOp)*
huUf(flx) flx(huUf)	Entité humaine usager final du flux anthropique Flux anthropique dont l'entité humaine est l'usager final	*huUf(flw)* *flw(huUf)*
ijn(ovg) ovg(ijn)	Injection d'eau effectuée avec l'ouvrage d'injection d'eau Ouvrage d'injection d'eau utilisé pour l'injection d'eau	*ijn(wrk)* *wrk(ijn)*
ijn(prl)Ovg prl(ijn)Ovg	Ouvrage d'injection raccordé à l'ouvrage de prélèvement Ouvrage de prélèvement raccordé à l'ouvrage d'injection	*ijn(waw)Wrk* *waw(ijn)*Wrk
ijn(rse) rse(ijn)	Injection d'eau dans le réseau d'eau Réseau d'eau dans lequel est effectuée l'injection d'eau	*in(wnw)* *wnw(ijn)*
me(code) code(me)	Masse d'eau identifiée par son code européen Code européen de la masse d'eau	*wb(code)* *code(me)*
péri(flx) flx(péri)	Période d'écoulement du flux anthropique Flux anthropique écoulé sur la période	*peri(flw)* *flw(peri)*

Table D.1: Glossaire des propriétés (suite)

Propriété *Property*	Forme développée *Full form*	Prop. inverse *Inverse prop.*
prl(me)	Prélèvement d'eau effectué dans la masse d'eau	*waw(wb)*
me(prl)	Masse d'eau dans laquelle le prélèvement d'eau est effectué	*wb(waw)*
prl(ovg)	Prélèvement d'eau effectué avec l'ouvrage de prélèvement	*waw(wrk)*
ovg(prl)	Ouvrage de prélèvement utilisé pour le prélèvement d'eau	*wrk(waw)*
prq(arti) *prc(arti)*)	Périodique de l'article *Periodical of the article*	arti(prq) *arti(prc)*

Annexe E

Glossaire

Table des matières

Table des figures

Liste des tableaux

Bibliographie

[Depecker, 2003] Depecker, L. (2003). *Entre signe et concept, éléments de terminologie générale*. Presse Sorbonne Nouvelle.

[générale à la langue française, 1994] générale à la langue française, D. (1994). *Dictionnaire des termes ociels de la langue française*. Journal officiel de la République française, Paris.

[Goddard, 1998] Goddard, C. (1998). *Semantic analysis*. Oxford University Press, Oxford.

[Groenfeld, 2013] Groenfeld, D. (2013). *Water Ethics A Values Approach to Solving the Water Crisis*. Routledge, New York.

[Huddleston, 1988] Huddleston, R. (1988). *English grammar : an outline*. Cambridge University Press, Cambridge.

[Humboldt, 1903] HuMboldt, W. v. (1903). *Gesamelte Schriften*. B. Verlag, Berlin.

[Johanet, 2001] Johanet, V. (2001). *Lexique de l'eau*. Éditions Johanet.

[Jousse, 2010] Jousse, A.-L. (2010). *Modèle de structuration des relations lexicales fondé sur le formalisme des fonctions lexicales*. Thèse de doctorat, Université de Montréal et Université Paris Diderot (Paris 7), Paris.

[Kripke, 1980] Kripke, S. (1980). *Naming and necessity*. Harvard Unuversity Press, Cambridge (Mass.).

[Kripke, 1982] Kripke, S. (1982). *La logique des noms propres*. Les éditions de minuit, Paris.

[Leibnitz, 1903] Leibnitz, G. W. (1903). *Opuscules et fragments inédits de Leibnitz*. Ed. Louis Couturat, Paris.

[Linton, 2010] Linton, J. (2010). *What is water*. UBC Press, Vancouver.

[Mel'čuk *et al.*, 1988] Mel'čuk, I., Arbatchevsky-Jumarie, N., Dagenais, L., Elnitsky, L., Iordanskaja, L., Lefebvre, M.-N. et Mantha, S. (1988). *Dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques II*. Les Presses de l'Université de Montréal.

[Mel'čuk *et al.*, 1984] Mel'čuk, I., Arbatchevsky-Jumarie, N., Elnitsky, L., Iordanskaja, L. et Lessard, A. (1984). *Dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques I*. Les Presses de l'Université de Montréal.

[Mel'čuk *et al.*, 1992] Mel'čuk, I., Arbatchevsky-Jumarie, N., Iordanskaja, L. et Mantha, S. (1992). *Dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques III*. Les Presses de l'Université de Montréal, Montreal.

[Mel'čuk *et al.*, 1999] Mel'čuk, I., Arbatchevsky-Jumarie, N., Iordanskaja, L., Mantha, S. et Polguère, A. (1999). *Dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques IV*. Les Presses de l'Université de Montréal, Montreal.

[Mel'čuk *et al.*, 1995] Mel'čuk, I., Clas, A. et Polguère, A. (1995). *Intrduction à la lexicologie explicative et combinatoire*. Duculot, Louvain-la-Neuve.

[Miller, 1968] Miller, R. (1968). *The Linguistic Relativity Principle anf Humboldtian Ethnolinguistics*. Mouton, La Haye.

[Murphy, 2010] Murphy, M. L. (2010). *Lexical meaning*. Cambridge University Press, Cambridge.

[Neveu, 2010] Neveu, F. (2010). *Dictionnaire des sciences du langage*. Armand Colin.

[Russel, 1905] Russel, B. (1905). On denoting. *Mind*, 14:479–493.

[Wierzbicka, 1972] Wierzbicka, A. (1972). *Semantic primitives*. Athenäum, Frankfurt.

[Wierzbicka, 1985] Wierzbicka, A. (1985). *Lexicography and conceptual analysis*. Karoma, Ann Arbor.

[Wierzbicka, 1992] Wierzbicka, A. (1992). *Semantics, Culture and Cognition : universal human concepts in Culture-Specific Configurations*. Oxford University Press, Oxford.

[Wierzbicka, 1996] Wierzbicka, A. (1996). *Semantics : Primes and Universals*. Oxford University Press, Oxford.