Ontologie des noms propres

Introduction

Ce chapitre a pour objet de présenter dans un premier temps la notion d'ontologie, notamment à travers les nombreuses définitions trouvées dans la littérature scientifique. Il s'agit ensuite de détailler les caractéristiques d'une ontologie et l'intérêt d'utiliser une approche ontologique pour modéliser l'ensemble des connaissances relatives à un domaine particulier, comme le nôtre. Nous verrons aussi une méthodologie pour construire une ontologie, qui se base sur sept étapes.

Enfin, nous présenterons notre ontologie des noms propres.

4.1 Ontologie

La notion d'ontologie est apparue la première fois il y a environ 2 300 ans sous la Grèce antique en philosophie avec Aristote et même avec Platon. Les ontologies sont, depuis 1990, au cœur de nombreux travaux dans le domaine de l'organisation des connaissances. En Intelligence Artificielle, en Ingénierie des Connaissances, dans le Web Sémantique, dans le Traitement Automatique des Langues, etc., les approches ontologiques connaissent beaucoup de succès et apportent des solutions novatrices. Cela s'explique par le besoin et la recherche d'une modélisation du monde et du sens des mots qui soit accessible aussi bien par des humains que par des agents logiciels.

4.1.1 Définition d'une ontologie

Il n'est pas évident de définir précisément ce qu'est une ontologie. Il existe bien sûr de nombreuses définitions de la notion d'ontologie, mais nous allons présenter seulement quelques définitions que nous avons trouvées dans le domaine de la recherche en informatique et qui nous ont paru intéressantes. L'une d'entre elles, dans le domaine de l'intelligence artificielle, citée fréquemment, revient à [Gruber, 1993] :

Définition 1 *An ontology is a formal, explicit, specification of a shared conceptualization. (Une ontologie est une spécification formelle explicite d'une conceptualisation.)*

Construire une ontologie consiste dans un premier temps à mener un travail de conceptualisation, qui nécessite d'identifier les concepts du domaine à modéliser en se basant sur l'étude de corpus relatif à ce domaine. De nombreux autres travaux se sont basés sur cette définition. [Charlet et al., 2003] donnent la définition suivante :

Définition 2 *Une ontologie implique ou comprend une certaine vue du monde par rapport à un domaine donné. Cette vue est souvent conçue comme un ensemble de concepts —e.g. entités, attributs, processus —, leurs définitions et leurs interrelations. On appelle cela une conceptualisation.*
[...]
Une ontologie peut prendre différentes formes mais elle inclura nécessairement un vocabulaire de termes et une spécification de leur signification.
[...]
Une ontologie est une spécification rendant partiellement compte d'une conceptualisation.

Pour [Roche, 2005] une ontologie possède les caractéristiques suivantes :

Définition 3 *Définie pour un objectif donné et un domaine particulier, une ontologie est pour l'ingénierie des connaissances une représentation d'une modélisation d'un domaine partagée par une communauté d'acteurs. Objet informatique défini à l'aide d'un formalisme de représentation, elle se compose principalement d'un ensemble de concepts définis en compréhension, de relations et de propriétés logiques.*

Selon ces différentes définitions, toute ontologie doit au moins posséder les caractéristiques suivantes :

- des concepts : un concept peut être un objet concret ou abstrait, qui apparaît dans le domaine à modéliser.
- des propriétés : il s'agit de caractéristiques qui permettent de décrire plus précisément les concepts.
- des relations : les relations permettent relier les différents concepts de l'ontologie entre eux. Il existe de nombreuses relations : la relation de méronymie, la relation de synonymie, la relation de subsomption (*is-a*), etc.

Ces différentes définitions nous renseignent sur la notion d'ontologie dans un contexte informatique, mais elles ne nous donnent aucune méthodologie pour construire une ontologie relative à un domaine spécifique.

4.1.2 Méthodologie de construction d'ontologie

Il existe évidemment de nombreuses méthodologies qui permettent de développer des ontologies, mais aucune d'entre elles n'est admise ou reconnue par l'ensemble de la communauté scientifique.

Certaines méthodes relèvent parfois plus de l'intuition que de la rigueur scientifique. La plupart admettent qu'il est nécessaire d'identifier dans un premier temps les concepts et les relations. Mais on constate que, selon la méthodologie utilisée pour modéliser un même domaine, le résultat obtenu ne sera pas forcément le même [Mizoguchi, 2005], en raison des nombreux choix et critères que chaque ontologiste est amené à prendre au cours de cette première phase. La plupart des méthodes ne décrivent pas de manière précise les décisions à prendre ou les règles qu'il faut appliquer durant le processus de conceptualisation.

Nous allons présenter une méthodologie qui nous a semblé intéressante, celle proposée par [Noy and McGuinness, 2003].

Méthodologie de Noy et McGuinness

Selon [Noy and McGuinness, 2003], il est nécessaire au cours de la conception de l'ontologie de toujours se rappeler, particulièrement lorsque l'on est confronté à un problème, les trois règles de base suivantes :

1. Il n'y a pas qu'une seule façon correcte pour modéliser un domaine - il y a toujours des alternatives viables. La meilleure solution dépend presque toujours de l'application que vous voulez mettre en place et des évolutions que vous anticipez.
2. Le développement d'une ontologie est nécessairement un processus itératif.
3. Les concepts dans une ontologie doivent être très proches des objets (physiques ou logiques) et des relations dans votre domaine d'intérêt. Fort probablement, ce sont des noms (objets) ou verbes (relations) dans des phrases qui décrivent votre domaine.

Leur méthodologie de construction d'une ontologie relative à un domaine particulier repose sur une série de sept étapes.

Dans la première étape, il faut commencer par faire une description précise et détaillée du domaine sur lequel on va travailler afin de mieux percevoir ses limites, c'est-à-dire où il commence et où il s'arrête. Il faut aussi déterminer les applications que l'on souhaite faire de cette ontologie.

La deuxième étape consiste à rechercher dans des bibliothèques d'ontologies mises à disposition, par exemple sur Internet, s'il n'existe pas déjà une ontologie qui correspondrait à ses besoins. Si l'on n'a pas eu la chance de trouver son bonheur dans les travaux existants, il va falloir passer à l'étape suivante.

Dans la troisième étape, il faut lister les différents mots importants du domaine. Il ne faut surtout pas s'inquiéter si cette liste est extrêmement longue.

Dans la quatrième étape, on définit les différentes classes et on établit une hiérarchisation entre elles. On peut soit commencer par définir le concept le plus général pour finir par les concepts les plus spécialisés (méthode descendante, en anglais *top down*), soit appliquer la méthode inverse ascendante (*bottom up*), soit choisir une méthode mixte qui combine les deux précédentes.

L'étape cinq permet de décrire les classes plus précisément, en cherchant pour chacune ses propriétés ou attributs.

L'étape six consiste à définir la cardinalité et le type (chaîne, bouléen, etc.) associés à chaque attribut.

La dernière étape correspond au moment ou l'on pourra créer des instances (ou individus) de l'ontologie.

4.2 Typologie des noms propres

Dans cette partie, nous allons nous intéresser au domaine de la typologie des noms propres. Il s'agit maintenant de définir les différents concepts de notre typologie et les relations entre ces concepts sous la forme d'une ontologie.

Nous allons appliquer les quatre premières étapes[1] de la méthodologie de Noy et McGuinness. Pour décrire notre domaine, nous nous sommes basés sur les différentes typologies, utilisées dans le domaine de la linguistique et celles qui ont conduit à des systèmes de reconnaissance de noms propres, que nous avons décrits en détail au cours du premier chapitre. A partir de ces différents travaux, nous avons ensuite établi une liste de types de noms propres. Nous avons appliqué la méthode descendante pour définir et hiérarchiser nos différents concepts, que nous allons présenter dans cette partie. Ces différents concepts entretiennent entre eux une relation d'hyperonymie.

[1]Les étapes cinq et six seront présentées au chapitre 5 sur l'implémentation de notre modèle.

Cette typologie a pour racine le concept de nom propre, pour nœuds, des supertypes et pour feuilles, des types.

4.2.1 Les quatre premiers supertypes

Situés juste en dessous du concept de nom propre, les quatre premiers supertypes classent les noms propres suivant des traits syntaxo-sémantiques assez généraux. Ces traits peuvent facilement être reconnus par des systèmes d'extraction automatique de noms propres en se basant essentiellement sur le contexte linguistique apparaissant autour d'eux dans le texte.

Dans notre ontologie, nous avons distingué :

– les anthroponymes : trait humain
– les ergonymes : trait inanimé
– les pragmonymes : trait événement
– les toponymes : trait locatif

La figure 4.1 montre la représentation des différents concepts de supertype à l'aide du logiciel *Protégé 3.1*[2], permettant de créer des ontologies.

FIG. 4.1 – Les supertypes.

Les anthroponymes

Le supertype anthroponyme, comme le supertype toponyme, est un concept largement connu et communément admis dans le domaine de l'onomastique ou de l'étude des noms propres. Le trait humain est sans doute le trait le plus facile à percevoir et à reconnaître chez un nom propre. Les anthroponymes renvoient sur le plan sémantique à la notion de personne. Nous avons partagé le supertype anthroponyme en deux autres supertypes [Gross, 1995] : les anthroponymes individuels (*Lassie*, *George Orwell*, etc.) et les anthroponymes collectifs (*Mérovingiens*, *Organisation mondiale de la santé*, etc.). [Dubois, 1973], dans le *Dictionnaire de linguistique*, distingue les noms animés non humains, c'est-à-dire les animaux, et les noms animés, sous-catégorie dans laquelle il classe le trait humain. Cette distinction se fera au niveau des types célébrité et pseudo-anthroponyme (voir section 4.2.2).

[2] `http://protege.stanford.edu/`

Les toponymes

[Lepesant, 2000] définit les toponymes ainsi :

> *Les noms locatifs constituent une catégorie de noms d'objets dimensionnels, tels que leurs méronymes d'espace ont pour hyperonyme le mot lieu.*

Nous avons rassemblé sous le concept de toponyme tous les noms de lieu au sens général. Les toponymes regroupent diverses entités qui possèdent chacune une taille extrêmement variée. Cela peut aller du nom donné à une rue ou à un bâtiment, en passant par le nom d'une vaste zone géographique pouvant regrouper plusieurs pays, jusqu'à s'étendre au nom d'un ensemble contenant environ quelques millions de galaxies. Il est possible de diviser les toponymes en deux classes différentes : les toponymes naturels et les toponymes bâtis par les hommes.

Les systèmes de reconnaissance automatique de noms propres arrivent à extraire les toponymes dans un texte journalistique [Friburger, 2002], car la plupart du temps, ils apparaissent dans ces textes, accompagnés de preuve externe (*la* ***ville de*** *Tours*) ou de preuve interne (*le* ***Mont*** *Blanc*) [MacDonald, 1996].

Les ergonymes

Ergonyme (du grec *ergon* : travail, force) est un mot emprunté à [Bauer, 1985] :

> *Noms des installations créées par l'homme servant à la production, [...] noms de produits créés par et pour l'homme.*

Sous le type ergonyme, on peut retrouver des noms propres qui se rattachent soit au trait sémantique inanimé concret (*Coca-Cola*), soit au trait inanimé abstrait (*Alice au pays des merveilles*). Nous distinguons dans cette catégorie les ergonymes à caractère économique de ceux à caractère artistique.

Les pragmonymes

Les pragmonymes peuvent être définis comme des noms d'événements (comme *le 14 juillet*) ou de catastrophes naturelles (comme par exemple *Katrina*) ou non (comme par exemple *Tchernobyl*).

4.2.2 Type

Le type correspond à une classification plus détaillée que le supertype d'un nom propre. Cette classification est destinée principalement à la recherche d'information et à la traduction automatique. Pour associer un type à un nom propre, il faut souvent une intervention humaine. Dans le cadre de nos travaux, nous avons retenu au total 29 types que nous allons présenter dans cette partie. La figure 4.2 liste des exemples de noms propres classés en fonction de ces types.

Cependant, certaines distinctions sont difficiles à réaliser et peuvent sembler arbitraires. Nous avons donc décidé de créer deux autres supertypes :

- un supertype que nous appellerons Groupement et qui rassemble les anthroponymes collectifs correspondant à une association ou à une institution (politique, religieuse, culturelle, nationale, internationale, etc.). Ce supertype contient les types association, ensemble, entreprise, institution et organisation.
- un supertype que nous appellerons Territoire car il n'est parfois pas évident de faire une distinction entre les pays (au sens états indépendants) et les régions incluses ou non dans les pays. Ce supertype contient les types pays, région et supranational.

En cas de polysémie (voir section 3.1.2 page 55), comme par exemple pour le nom propre *Michelin* qui correspond à la fois à une célébrité et à une entreprise et pour le nom propre *Tempelhof* qui correspond à la fois à un faubourg de Berlin et à un de ses aéroports, nous avons décidé de créer deux noms propres conceptuels différents. Nous associerons à chacun de ces noms propres un unique type.

Rappelons aussi que les homonymes correspondent de même à des noms propres conceptuels différents même s'ils ont le même type. Par exemple, le nom propre *Vienne*, capitale de l'Autriche, sera lié au type ville, ses homonymes correspondant à une ville d'Isère et de Poitou-Charentes le type ville.

La figure 4.3 présente la hiérarchie des types correspondant à la relation d'hyperonymie primaire (voir section 4.3.2)

Types	**Exemples**
Association	*les Restaurants du cœur*, *l'Union chrétienne-démocrate* , etc.
Astronyme	*l'étoile Polaire*, *le Bélier*, *Pluton*, etc.
Catastrophe	*Erika*, *Tchernobyl*, *Katrina*, etc.
Célébrité	*Platon*, *Blanche-Neige*, *Antoine de Saint-Exupéry*, etc.
Dynastie	*Carolingien*, *Michelin*, *Ming*, etc.
Édifice	*le Colisée*, *le palais Bourbon*, *la Grande Muraille*, etc.
Ensemble	*Les Beatles*, *le cercle de Prague*, *De Stijl*, etc.
Entreprise	*Air France*, *Nestlé*, *DaimlerChrysler*, etc.
Ethnonyme	*Étrusque*, *Aztèque*, *Sabin*, etc.
Fête	*Noël*, *Halloween*, *la Pentecôte*, etc.
Géonyme	*les Alpes*, *le désert de Syrie*, *le Kilimandjaro*, etc.
Histoire	*le IIIe Reich*, *la bataille d'Austerlitz*, *le traité de Rome*, etc.
Hydronyme	*l'Amazone*, *le lac Léman*, *l'océan Pacifique*, etc.
Institution	*le Collège de France*, *Scotland Yard*, *l'institut Pasteur*, etc.
Manifestation	*le Tour de France*, *le Festival d'Avignon*, *la coupe Davis*, etc.
Météorologie	*l'anticyclone des Açores*, *El Niño*, *la Tramontane*, etc.
Objet	*le Saint-Graal*, *Durandal*, *la Toison d'or*, etc.
Œuvre	*l'Avare*, *les Demoiselles d'Avignon*, *la Vénus de Milo*, etc.
Organisation	*la Croix-Rouge*, *l'Organisation mondiale de la santé*, etc.
Patronyme	*Dupont*, *Durant*, etc.
Pays	*le Portugal*, *l'Australie*, *la République de Corée*, etc.
Prénom	*Louis*, *Jean*, *Pierre*, etc.
Produit	*Adidas*, *Ferrari 250 GTO*, *Coca-Cola*, etc.
Pseudo-anthroponyme	*Pégase*, *C-3PO*, *Donald*, etc.
Région	*l'Austrasie*, *le Tartare*, *la Californie*, etc.
Supranational	*les Antilles*, *l'Eurasie*, *les pays Baltes*, etc.
Vaisseau	*le Titanic*, *Apollo 11*, *Enterprise*
Ville	*Marseille*, *Nha Trang*, *Chiang Rai*, etc.
Voie	*la place Rouge*, *les Champs-Élysées*, *l'autoroute du Soleil*, etc.

FIG. 4.2 – Les types

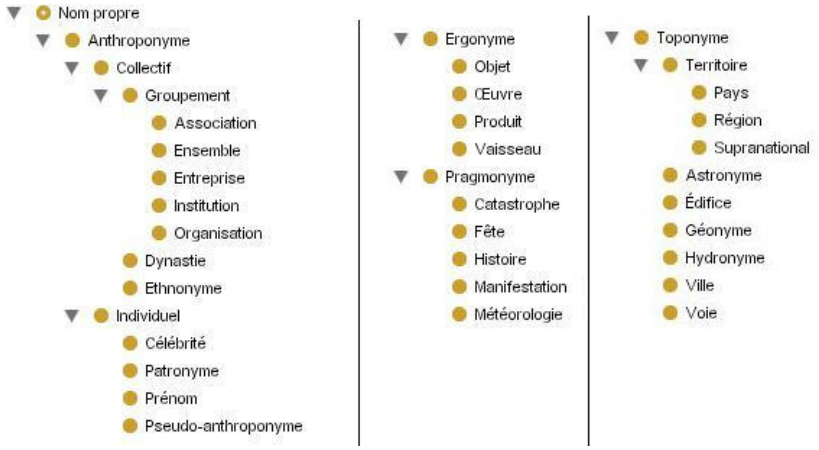

FIG. 4.3 – La hiérarchie des types.

Association

Le terme association désigne un groupe comprenant plusieurs personnes visant un but commun. Dans le type association nous avons regroupé les associations nationales, à caractère social et à but non lucratif, les partis politiques et les syndicats nationaux.

Astronyme

Les astronymes regroupent uniquement les noms propres relevant du domaine de l'astronomie. Il s'agit des noms que l'on attribue aux objets célestes, c'est-à-dire les planètes, les galaxies, les étoiles, les comètes, les constellations, etc.

Catastrophe

Les catastrophes rassemblent les noms de désastres ou tragédies entraînant le plus souvent la mort ou la destruction qui sont soit d'origine naturelle (les catastrophes climatiques comme les cyclones, ouragans, tempêtes, etc., les catastrophes sismiques, les éruptions volcaniques, etc.), soit d'origine humaine (les catastrophes industrielles, etc.).

Célébrité

Ce type regroupe les humains célèbres. Les célébrités constituent sans aucun doute une classe très vaste par rapport aux autres classes de notre typologie des noms propres et en perpétuelle expansion. Nous avons aussi regroupé dans cette classe les pseudonymes ou noms d'emprunt généralement utilisés par des artistes.

Les noms de célébrités apparaissent dans les textes sous des formes très diverses : un nom, un prénom, un prénom et un nom, etc.

Dynastie

La classe des dynasties correspond aux humains collectifs. La majorité des éléments de cette classe est constituée de noms de familles royales de divers pays ou empires, regroupant une succession de monarques qui ont marqué l'histoire. Nous avons aussi étendu cette classe aux noms de familles ayant un lien avec le pouvoir politique ou économique.

Édifice

Le type édifice correspond à des constructions humaines de toute sorte, telles que les bâtiments historiques ou officiels, les monuments, les châteaux, les ponts, les parcs, les musées, les bibliothèques, les théâtres, les bâtiments religieux (églises, basiliques, mosquées, temples, etc.), les aéroports, les prisons, les stades, les hôpitaux, les murs, etc.

Ensemble

Les ensembles sont essentiellement formés de noms de groupe de personnes relevant soit du domaine artistique, soit du domaine sportif.

Entreprise

Le type entreprise rassemble les sociétés industrielles, financières ou commerciales, qui peuvent être nationales ou multinationales.

Ethnonyme

Les ethnonymes sont des noms de peuples. Comme pour les gentilés, leur statut de nom propre est parfois contesté même s'ils prennent une majuscule initiale.

Fête

Le type fête comprend les noms des événements festifs et cycliques, qui ont pour but de rappeler certaines traditions ou faits historiques marquants.

Géonyme

La classe des géonymes est formée de noms donnés aux espaces géographiques naturels, tels que les déserts, les montagnes, les massifs montagneux, les glaciers, les plaines, les gouffres, les plateaux, les vallées, les volcans, les canyons, etc.

Histoire

La classe histoire comprend les événements qui ont marqué la mémoire des hommes. Il s'agit de traités ou accords signés entre différents pays, de batailles, de périodes historiques, de manifestations sociales, de révoltes, de crises, les ères, etc.

Hydronyme

Les hydronymes renvoient à des noms d'étendue d'eau. Il peut s'agir par exemple de rivières, de fleuves, d'étangs, de marais, de lacs, de mers, d'océans, de courants marins, de canaux, de sources, etc.

Institution

Le type institution regroupe :
- les instituts d'enseignement et de recherche
- les instituts religieux
- les fondations
- les instituts politiques

– les juridictions
– les instituts militaires
– les administrations
– etc.

Les noms d'institution sont généralement traduits d'une langue à l'autre.

Manifestation

Les manifestations regroupent toutes sortes d'activités ou d'événements sportifs ou culturels.

Météorologie

Il s'agit d'évènements météorologiques naturels et récurrents, tels que les vents, les phénomènes climatiques, etc.

Objet

Cette classe regroupe uniquement les noms d'objet qui sortent souvent de légendes, de la littérature ou qui relèvent du domaine religieux.

Œuvre

Nous avons regroupé dans cette classe toutes les formes d'œuvres artistiques. Il peut s'agir de sculptures, de livres, de tableaux de grands maîtres, de films, de pièces de théâtre ou d'opéra, de partitions de musique, etc.

Organisation

Le type organisation comprend seulement les organismes à caractère international et qui ne se rattachent pas à un gouvernement en particulier. Si l'organisation se rattache à un gouvernement, nous la classerons parmi les institutions.

Patronyme

Il s'agit de noms de famille. Les patronymes se traduisent rarement d'une langue à l'autre. Lorsqu'un nom de famille appartient à un système d'écriture différent, il est sujet à des translittérations ou à des transcriptions.

Pays

Nous avons rassemblé dans le type pays les noms d'états indépendants, des royaumes ou empires qui sont apparus au fil de l'histoire. Il peut aussi s'agir de noms de pays fictifs.

Prénom

Il s'agit de prénoms. Un même prénom peut se retrouver d'une langue à l'autre sous des formes différentes. C'est le cas du prénom français *Marie* qui donne :

– *Mary* en anglais
– *Maria* en espagnol
– etc.

Les prénoms évoluent d'une époque à l'autre (*Johan*, *Jehan*, *Jean*, etc.).

Produit

Il s'agit de noms de produits ou de marques qui ont été développés uniquement dans un but commercial. On peut trouver des noms de voitures, d'avions, de produits de consommation, de vêtements, d'outils, etc.

Pseudo-anthroponyme

Les pseudo-anthroponymes regroupent tous les noms propres qui peuvent être classés parmi les êtres vivants ou considérés comme tels et qui ne font pas partie de la catégorie des êtres humains. Le profil des entités que l'on retrouve dans cette classe est extrêmement varié. Il peut s'agir de noms donnés aux animaux (zoonymes), aux robots, aux machines, aux êtres venus d'une autre planète ou d'une autre dimension, etc.

Région

Les régions correspondent à un découpage, parfois administratif, d'un pays en plusieurs espaces géographiques de taille variable. Il peut s'agir par exemple de comté ou Land pour l'Allemagne, d'état pour les États-Unis, de canton pour la Suisse, de province pour le Canada, la Belgique et la Chine, etc.

Supranational

Le concept intitulé supranational est défini comme un regroupement de différents pays ou contenant des parties de différents pays.

Vaisseau

Cette catégorie regroupe les véhicules pouvant circuler sur l'eau (paquebots, navires de guerre, etc.) ou dans l'espace (fusées, stations spatiales, etc.). D'autres noms de véhicules peuvent rentrer dans cette catégorie. C'est le cas par exemple du nom propre *Batmobile*.

Ville

Nous avons regroupé sous le type ville les noms de villes et les noms de quartiers. Les quartiers correspondent à un découpage d'une ville, ayant parfois une densité de population bien supérieure à certaines villes. De plus, certains quartiers dans les grandes métropoles étaient autrefois des villes.

Voie

Le concept de voie est principalement formé de noms de rues, de places, de routes, d'autoroutes, etc.

4.3 Ontologie des noms propres

Les relations de notre ontologie des noms propres comprennent des relations qui ne dépendent pas de la langue vues au chapitre 3 et de la relation d'hyperonymie (voir section 4.3.2). Les concepts de notre ontologie comprennent les types, les supertypes, l'existence (voir section 4.3.1) et le nom propre conceptuel.

4.3.1 Existence

Le concept d'existence permet de préciser le domaine d'appartenance d'un nom propre. La majorité des noms propres appartiennent au domaine historique (*Mozart*, *le Danube*, *Paris*, etc.), qui se définit dans le Larousse 2004 comme :

> ***Historique** : [...] dont l'existence est considérée comme objectivement établie.*

D'autres relèvent plutôt du domaine de la croyance (*Zeus*, *Adam*, etc.) ou du domaine de la fiction (*Tintin*, *Utopie*, *Atlantis*, etc.).

La distinction entre des noms propres historiques et les autres s'avère utile pour la traduction, car ces derniers se traduisent d'une langue à l'autre. C'est le cas, par exemple, de Blanche-Neige qui devient :

- *Sneeuwwitje* en néerlandais
- *Biancaneve* en italien
- *Schneewitchen* en allemand
- *Snövit* en suèdois
- etc.

4.3.2 La relation d'hyperonymie

[Polguère, 2003] définit la relation d'hyperonymie de la façon suivante :

> *La lexie L_{hyper} est un hyperonyme de la lexie L_{hypo} si*
> *– le sens (L_{hyper}) est inclus dans le sens (L_{hypo})*
> *– et si (L_{hypo}) peut être considéré comme un cas particulier de (L_{hyper}).*
> *La lexie L_{hypo}, quant à elle, est appelée hyponyme de L_{hyper}.*

Hyperonyme secondaire	**Type**
Anthroponyme	Pays Région Supranational
Anthroponyme Ergonyme	Ville
Ergonyme	Édifice Voie
	Fête Histoire Manifestation
Ergonyme Toponyme	Association Ensemble Entreprise Institution Organisation
Toponyme	Vaisseau

FIG. 4.4 – Hyperonymie secondaire.

Les types et les supertypes de notre ontologie sont reliés par une relation d'hyperonymie (figure 4.3), que nous appellerons relation d'hyperonymie primaire.

Certains types peuvent être en relation d'hyperonymie secondaire avec d'autres supertypes (figure 4.4). Par exemple, les types association et organisation sont à la fois en relation d'hyperonymie primaire avec le supertype anthroponyme collectif et en relation d'hyperonymie secondaire avec les supertypes ergonyme et toponyme.

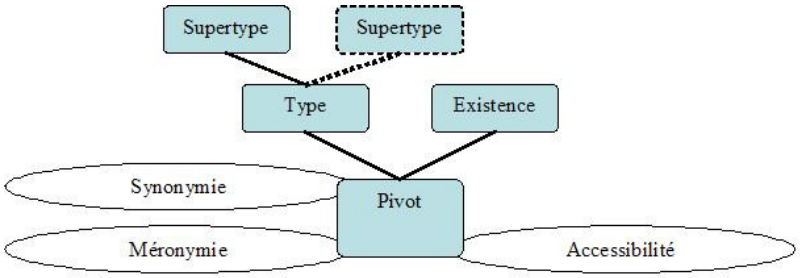

FIG. 4.5 – Ontologie des noms propres.

Chaque nom propre conceptuel (ou pivot) sera en relation d'hyperonymie avec un type et une existence (figure 4.5).

Conclusion

Nous avons présenté dans ce chapitre la modélisation de notre typologie et de notre partie qui ne dépend pas de la langue sous la forme d'une ontologie, composée de plusieurs concepts et relations d'hyperonymie permettant de hiérarchiser ces différents concepts.

Pour créer cette typologie des noms propres, nous nous sommes beaucoup inspiré des travaux de Thierry Grass, avec qui nous avons eu de nombreuses occasions de travailler durant ces trois années.

Les chapitres 3 et 4 nous permettent donc d'obtenir la définition complète de notre ontologie des noms propres (voir figure 4.5). Celle-ci comprend quarante concepts (le nom propre conceptuel, sept supertypes, vingt-neuf types et trois existences) et quatre relations (hyperonymie, synonymie, méronymie et accessibilité).

Maintenant que l'architecture de notre modèle a été définie, il s'agit d'implémenter ce modèle sur machine afin de pouvoir entrer des données. Les parties suivantes seront donc applicatives.

Troisième partie

Implémentation de Prolexbase

Chapitre 5

La base de données

Introduction

Ce chapitre est destiné à présenter l'implémentation des différents concepts et relations du domaine des noms propres sous la forme d'une base de données relationnelle, que nous avons appelée Prolexbase.

Pour créer notre base de données des noms propres, nous avons utilisé la méthode Merise. Dans la première partie, nous allons présenter quelques notions de base sur cette méthode. Ensuite, nous décrirons le modèle conceptuel de données et le modèle logique de données que nous avons mis en place pour les noms propres.

5.1 La méthode Merise

Développée en France en 1978, la méthode Merise (Méthode d'Étude et de Réalisation Informatique pour les Systèmes d'Entreprise) [Matheron, 1998] propose une démarche pour analyser et concevoir un système d'information. Dans cette partie, nous allons présenter brièvement deux étapes de cette méthode : le modèle conceptuel de données et le modèle logique de données.

5.1.1 Le modèle conceptuel de données

Le modèle conceptuel de données (MCD) permet de décrire les objets de la réalité et les dépendances ou associations entre ces objets. Un MCD aboutit à la création d'un schéma d'entité/association (E/A).

Les entités

Une entité est définie comme un objet concret ou abstrait du monde réel. Dans le modèle E/A, on représente une entité sous la forme d'un rectangle (figure 5.1) dans lequel on inscrit le nom de l'entité.

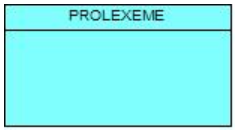

Fig. 5.1 – Représentation d'une entité.

Chaque entité peut posséder un ou plusieurs attributs, dont on devra préciser le type (Date, Entier, Booléen, Texte, etc.). Pour pouvoir identifier chaque occurrence d'une entité de manière unique, il faudra obligatoirement désigner parmi ses différents attributs un attribut ou un ensemble d'attributs qui jouera le rôle d'identifiant ou de clé primaire. Il arrive souvent que l'on rajoute un attribut fictif (un numéro) qui servira de clé primaire. Nous avons associé à chaque identifiant le type ID dans nos schémas E/A (figure 5.2).

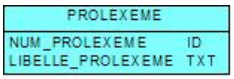

FIG. 5.2 – Représentation des attributs.

Les associations

Les associations sont des liens qui unissent les entités du modèle. Elles apparaissent dans un schéma E/A sous la forme d'un ovale (figure 5.3). On associe à chaque entité d'une association une cardinalité qui précise si une entité peut participer dans l'association zéro, une ou plusieurs fois.

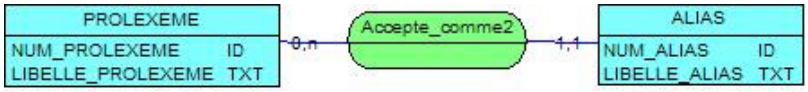

FIG. 5.3 – Représentation d'une association.

Dans la figure 5.3, les entités *PROLEXEME* et *ALIAS* sont reliées par l'association *Accepte_ comme2*. La cardinalité (0,n) indique qu'un prolexème accepte au minimum zéro alias et au maximum plusieurs alias. La cardinalité (1,1) précise qu'un alias correspond à un seul et unique prolexème.

5.1.2 Le modèle relationnel de données

Le modèle relationnel de données (MLD) correspond à une traduction des entités et des associations du MCD sous la forme de relations. Les principales règles de passage d'un MCD vers MLD sont les suivantes :

- Règle 1 : une entité du MCD se transforme en relation. Ses propriétés deviennent des attributs. La clé primaire de la relation sera représentée par l'identifiant.
- Règle 2 : soit R une association de type un-à-plusieurs reliant deux entités E1 et E2 (une occurrence de E1 peut être en relation avec au maximum une occurrence de E2 et une occurrence de E2 peut être en relation avec plusieurs occurrences de E1). R ne devient pas une relation. L'identifiant de E2 et les éventuelles propriétés de R sont rajoutés dans la relation E1.
- Règle 3 : soit R une association de type plusieurs-à-plusieurs reliant deux entités E1 et E2 (plusieurs occurrences de E1 peuvent être en relation avec plusieurs occurrences de E2 et plusieurs occurrences de E2 peuvent être en relation avec plusieurs occurrences de E1). R devient une relation et ses éventuelles propriétés seront des attributs. Les identifiants de E1 et E2 deviennent les clés primaires de R.

En appliquant ces règles sur la figure 5.3, nous obtenons le modèle relationnel suivant :

PROLEXEME (NUM_PROLEXEME, LIBELLE_PROLEXEME)
ALIAS (NUM_ALIAS, LIBELLE_ALIAS, NUM_PROLEXEME)

qu'il est possible de représenter sous la forme d'un schéma relationnel (figure 5.4).

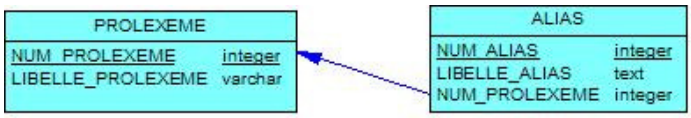

FIG. 5.4 – Schéma relationnel.

5.2 Modèle conceptuel de données

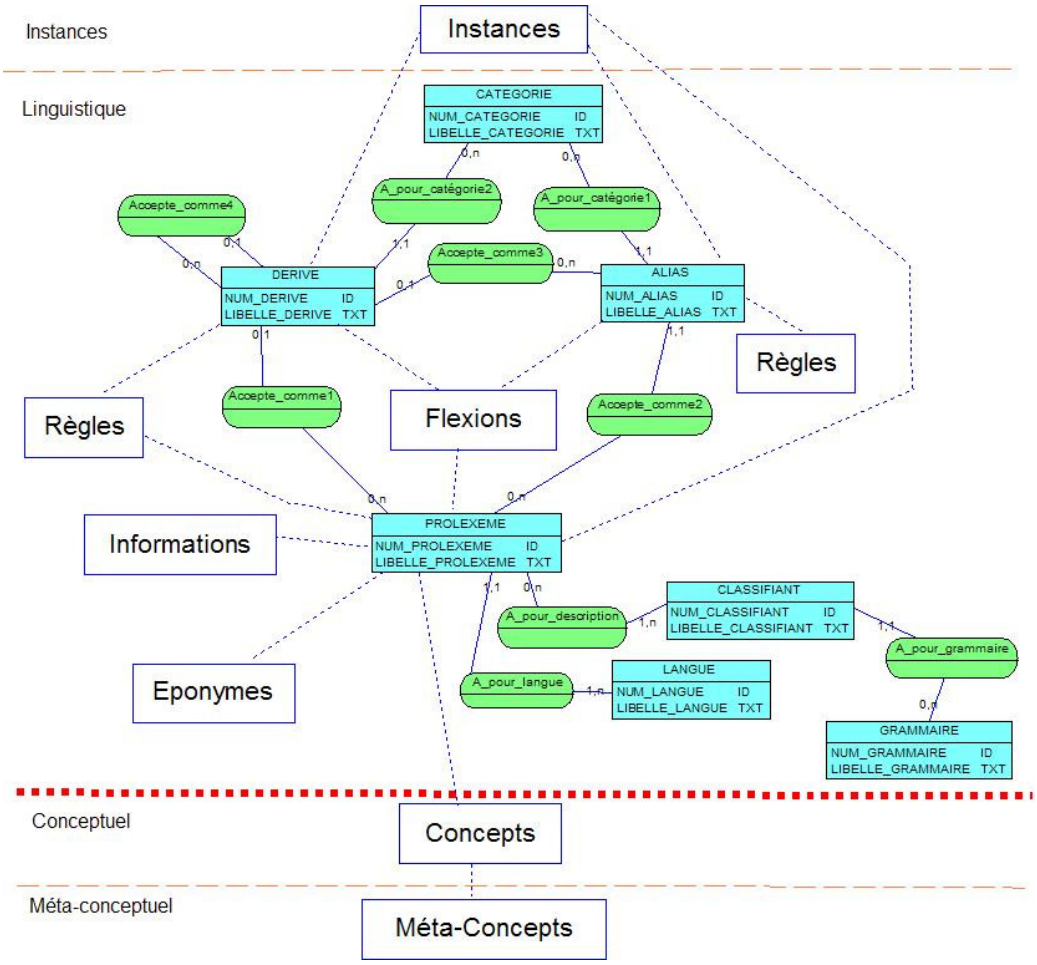

FIG. 5.5 – Le modèle conceptuel de données.

Nous avons établi notre MCD des noms propres à partir des différents concepts et relations définis dans le chapitre 2 et le chapitre 3. La figure 5.5 présente une version simplifiée de notre MCD (voir figure A.3 de l'annexe A page 151 pour un MCD complet).

Notre MCD peut être regroupé en quatre niveaux (méta-conceptuel, conceptuel, linguistique et instances) et comprend au total 28 entités et 41 associations.

Nous avons créé une entité pour chaque concept du domaine des noms propres (prolexème, alias, dérivé, etc.). L'entité *DERIVE* permet de stocker les dérivés de prolexème, d'alias ou d'autres dérivés (dans le cas du serbe). Nous avons associé à chaque alias, à travers la relation *A_pour_catégorie1*, une catégorie qui précise s'il s'agit d'une variante de caractères, d'une abréviation, d'acronymes ou sigles, d'une transcription, d'un synonyme diastratique ou d'un synonyme diatopique. Nous avons aussi associé à chaque dérivé, à travers la relation *A_pour_catégorie2*, une catégorie qui indique s'il s'agit d'un nom relationnel, d'un préfixe, d'un adjectif relationnel ou possessif. Les expansions classifiantes sont stockées dans l'entité *CLASSIFIANT* et chaque classifiant sera en relation avec une description (entité *GRAMMAIRE*) sous forme de grammaire locale, lien vers EuroWordNet ou Framenet (voir section 3.3.4). La relation *A_pour_langue* permet d'associer à chaque prolexème une langue.

5.2.1 Le niveau conceptuel

La partie conceptuelle (figure 5.6) est formée de quatre entités et cinq relations. L'entité *PIVOT* permet de stocker les noms propres conceptuels. La relation *Concept* associe à chaque nom propre conceptuel un ou plusieurs prolexèmes. On retrouve dans ce niveau la relation de méronymie, de synonymie et d'accessibilité.

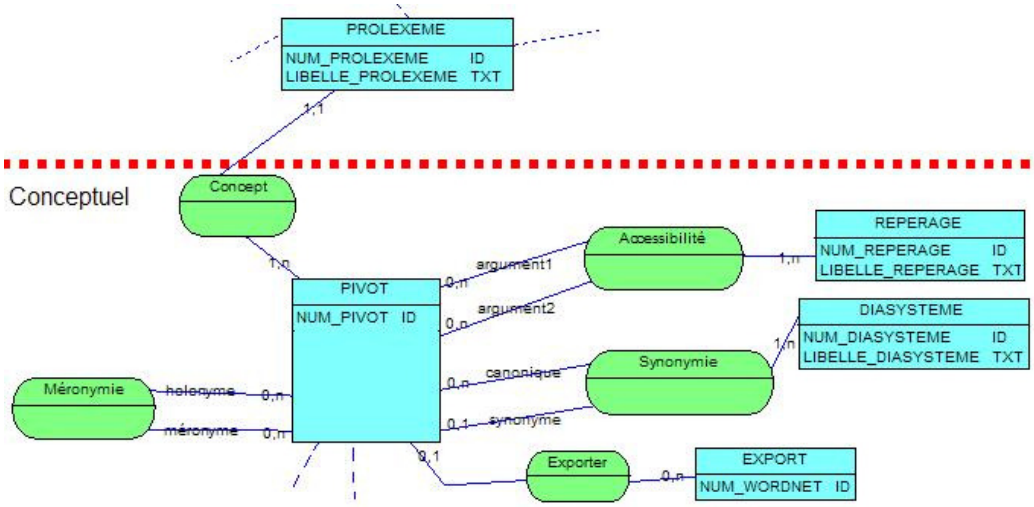

FIG. 5.6 – Le niveau conceptuel.

Un nom propre conceptuel sera en relation de synonymie avec un autre nom propre conceptuel suivant un diasystème (entité *DIASYSTEME*), qui peut être diachronique, diastratique ou diaphasique. La figure 5.7 présente la liste de repérages pour la relation d'accessibilité. Pour une relation de synonymie, nous avons imposé que chaque pivot peut être la forme canonique de plusieurs autres pivots et que chaque pivot peut être le synonyme d'une seule forme canonique.

L'entité *EXPORT* sert à relier les noms propres conceptuels de notre dictionnaire vers d'autres bases de données lexicales ou vers des encyclopédies. Des liens vers l'encyclopédie Wikipédia[1] et vers EuroWordNet ont été envisagés.

[1]Wikipédia est une encyclopédie gratuite accessible à l'adresse suivante : `http://www.wikipedia.org/`.

Repérage	**Exemple**
Capitale	Paris est la capitale de la France
Créateur	Auguste Rodin est le sculpteur du Penseur
Dirigeant non politique	Ray Norda est le patron de Novell
Dirigeant politique	Jacques Chirac est le président de la République française
Élève	Platon est l'élève de Socrate
Fondateur	Dardanos est le fondateur mythique de Troie
Héritier	Charles, prince de Galles, héritier du Royaume-Uni
Locataire	Jacques Chirac est le locataire de l'Elysée
Parent	Aaron est le frère de Moïse
Siège	Le Bureau Veritas a son siège à Paris
...	

FIG. 5.7 – Les repérages.

Le lien vers l'encyclopédie Wikipédia n'est pas conservé dans Prolexbase. Ce lien est généré dynamiquement sur le site de consultation en concaténant le code iso de la langue de consultation (fr, en, etc.), une url (wikipedia.org/wiki/Special:Search/) et le prolexème sélectionné par le visiteur. Pour le nom propre *France*, on produit ainsi le lien suivant :

http://fr.wikipedia.org/wiki/Special:Search/France

La génération automatique des liens vers l'encyclopédie Wikipédia présente un inconvénient. Tous les liens générés automatiquement n'ont pas été testés, l'interface de consultation peut par conséquent produire des liens qui n'existent pas, car certains articles ne sont pas présent dans cette encyclopédie, ou des liens vers un mauvais article. Pour éviter les liens incorrects, il faudrait vérifier manuellement chaque lien et les conserver dans la base de données. Il s'agit d'une tâche extrêmement longue. Par manque de temps, nous avons décidé de générer automatiquement les liens vers l'encyclopédie Wikipédia. Cette encyclopédie est en cours de développement : un lien incorrect aujourd'hui pourrait devenir correct le jour suivant.

Le lien vers la base lexicale EuroWordNet est conservé dans notre base de données grâce à l'entité *EXPORT*. Si le nom propre conceptuel existe dans EuroWordNet, son numéro ILI (Inter-Lingual-Index) apparaîtra dans l'entité *EXPORT*. Par exemple, on associera au nom propre *Paris* le numéro d'ILI *0558236n* (figure 5.8).

entity
location
region
area, country
center, middle, heart
seat
capital
national capital
Paris, City of Light, French capital, capital of France

FIG. 5.8 – Le nom propre *Paris* dans EuroWordNet.

5.2.2 Le niveau méta-conceptuel

La partie méta-conceptuelle (figure 5.9) comprend deux entités et quatre associations. L'entité *EXISTENCE* contient trois occurrences : historique, fictif et religieux. Nous avons regroupé les types et les supertypes dans une seule entité (*TYPE*), afin de pouvoir associer à un nom propre conceptuel un supertype, si l'on n'a pas d'information sur son type. Cela nous permet d'insérer dans notre dictionnaire des noms propres qui ont été trouvés par des systèmes de reconnaissance automatique de noms propres.

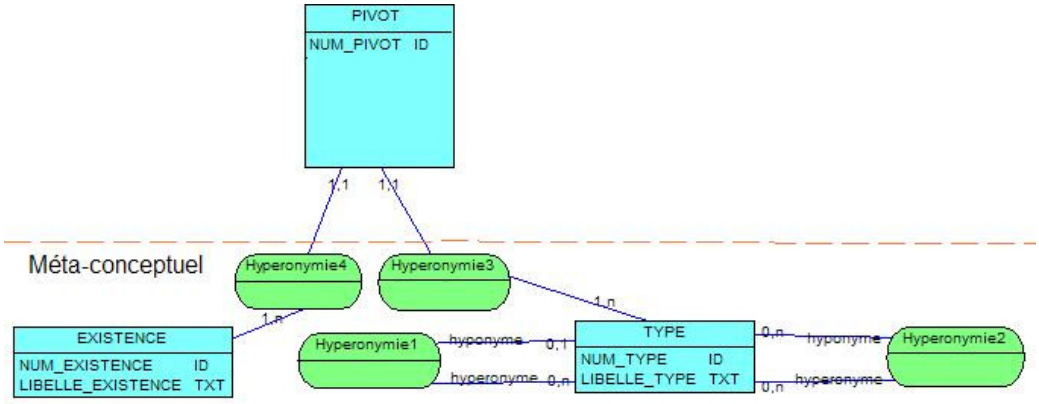

FIG. 5.9 – Le niveau méta-conceptuel.

5.2.3 L'éponymie

L'éponymie (figure 5.10) regroupe les entités *IDIOME*, *TERMINOLOGIE* et *ANTONOMASE*.

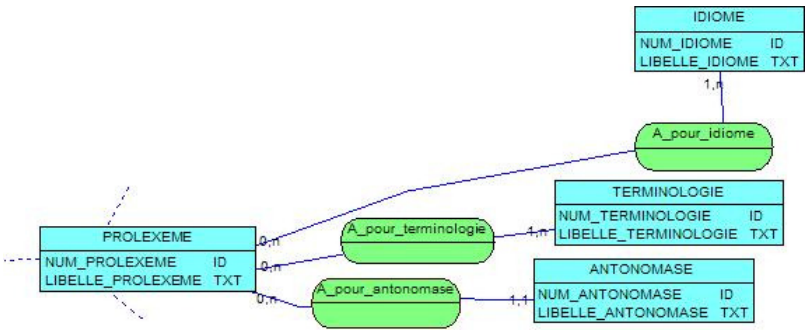

FIG. 5.10 – L'éponymie.

5.2.4 Les règles

L'entité *ALIASISATION* (figure 5.11) permet de stocker les règles de création d'alias à partir d'un prolexème. L'entité *DERIVATION* permet de stocker les règles de création de dérivés à partir d'un prolexème, d'un alias ou d'un dérivé.

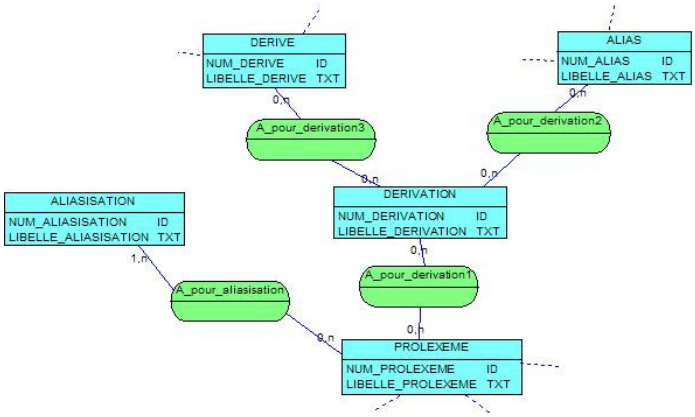

FIG. 5.11 – Les règles.

5.2.5 Les autres informations

Les informations supplémentaires (figure 5.12) sont formées de cinq entités et de cinq associations.

L'association *A_pour_statistique* permet d'associer à chaque prolexème des informations relatives à ses fréquences d'apparition (attribut *POIDS*) au sein d'un corpus donné (attribut *LIBELLE_STATISTIQUE*). Il peut s'agir, par exemple, d'étudier les fréquences d'apparition de noms propres sur quelques années d'un corpus journalistique. Certains noms propres apparaissant durant une année donnée pourront ne plus réapparaître quelques années plus tard. Cette étude statistique peut prendre en compte les différentes formes d'un même prolexème (ses alias, ses dérivés et leurs formes fléchies).

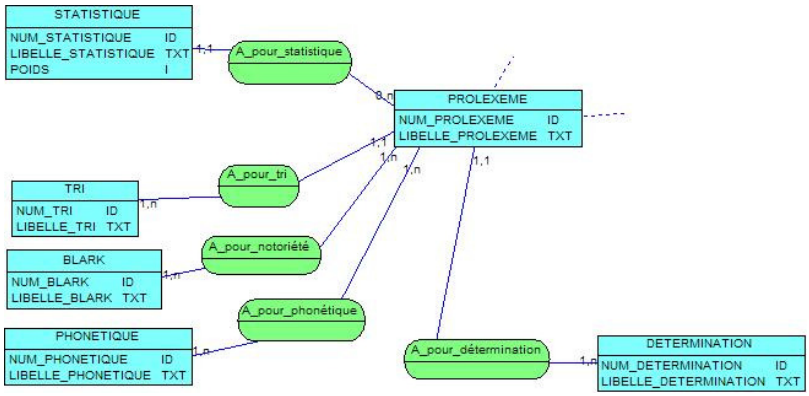

FIG. 5.12 – Les informations.

S'il est normal pour le simple mortel que nous sommes de ne pas posséder d'entrée dans les dictionnaires de noms propres, on peut parfois se demander pourquoi certains chanteurs ou chanteuses de variété française (*Johnny Hallyday*, etc.), actrices chinoises (*Michelle Yeoh*, etc.) ou autres célébrités ne figurent pas dans ces dictionnaires. On pourra aussi s'étonner que des villes telles que, par exemple, *Sainte-Enimie*, qui est le chef-lieu de canton de la *Lozère* comprenant à peine moins de 600 habitants et dont la majorité des Français ignore même l'existence, puisse apparaître dans le dictionnaire, alors que des villes de Russie, de Chine, et d'autres pays ayant une population nettement supérieure n'y figurent pas. Comme