UNIVERSITÉ LIBRE DE BRUXELLES
Faculté de Lettres, Traduction et Communication

RECONNAISSANCE D'ENTITÉS NOMMÉES À L'AIDE D'UNE BASE DE CONNAISSANCE

État de l'art, aperçus statistiques et apports de Cumuleo.be pour leur désambiguïsation dans des dépêches d'agence belges francophones

Ettore RIZZA

Mémoire présenté sous la direction de M. Seth VAN HOOLAND en vue de l'obtention du titre de Maître en Sciences et Technologies de l'Information et de la Communication

Année académique 2015–2016

Résumé

Ettore RIZZA

M-STICI

RECONNAISSANCE D'ENTITÉS NOMMÉES À L'AIDE D'UNE BASE DE CONNAISSANCE

État de l'art, aspects statistiques et apports de Cumuleo.be pour leur désambiguïsation dans des dépêches d'agence belges francophones

Année académique 2015-2016

Ce mémoire traite de la reconnaissance et de la désambiguïsation d'entités nommées dans un contexte journalistique. La première partie, théorique, retrace l'histoire du concept d'EN, dresse son état de l'art, ainsi que ceux du web sémantique. La seconde, plus pratique, se fonde sur l'annotation d'un corpus de 384 dépêches Belga. L'analyse statistique des noms de personnes, lieux et organisations contenus dans les textes révèle que le taux de ces entités est supérieur dans leur premier paragraphe. Une partie des entités dont aucun équivalent n'a pu être trouvé dans de grandes bases de connaissance, telles que DBpedia, a pu être réconciliée avec la base de données Cumuleo.

Mots clés : Extraction d'Information, Entités Nommées, *Linked Data*, *Web Scraping*, Annotation, Corpus, Dépêches, Agence Belga, Cumuleo.be.

Remerciements

Le présent mémoire n'aurait pu voir le jour sans l'appui, les conseils ou l'influence de quelques personnes.

Nous tenons ainsi à remercier M. Seth van Hooland, non seulement d'avoir accepté de le diriger, mais aussi de l'avoir inspiré par ses publications, ses cours et son enthousiasme communicatif.

Nous sommes également redevable envers M. Max De Wilde, dont les encouragements à l'heure de définir notre thématique ont été décisifs. Sa thèse de doctorat, consacrée à un sujet voisin du nôtre, a constitué pour nous un guide méthodologique précieux.

Ce mémoire n'aurait jamais abouti si nous avions suivi toutes les pistes évoquées dans notre travail préparatoire. Mme Isabelle Boydens nous a ramené à plus de raison ; nous lui en sommes reconnaissant.

Nous remercions également M. Christophe Van Gheluwe pour l'intérêt qu'il a accordé à nos recherches et pour nous avoir fourni une copie de la base de données Cumuleo.

Enfin, il nous est impossible d'exprimer ici notre gratitude à l'égard de Sophie Lejoly, sans laquelle ni ce travail, ni le reste de nos études, ni d'ailleurs rien n'aurait été possible.

Table des matières

Résumé i

Remerciements ii

Table des matières iii

Table des figures vi

Liste des tableaux viii

Liste des algorithmes viii

Introduction 1

Partie 1 : Fondements théoriques 5

1. Les entités nommées 6

1.1. Introduction 6
1.2. Historique 7
1.3. Définition 11

1.4. Typologies . . . 13
1.5. Méthodes de reconnaissance . . . 14
1.6. Évaluation des systèmes . . . 18
1.7. Quelques systèmes de NERC . . . 22
1.8. NERC : une tâche résolue ? . . . 24
1.9. La désambiguïsation sémantique . . . 26

2. Web de données et annotation sémantique 29
2.1. Du web... . . . 30
2.2. ...au web sémantique... . . . 31
2.3. ...puis au Linked Open Data . . . 39
2.4. Résolution d'entités . . . 42
2.5. Limites de Wikipedia . . . 49

Partie 2 : Étude de cas 52

3. Analyse du corpus : présentation et aspects méthodologiques 53
3.1. Présentation du corpus . . . 54
3.2. Récupération et traitement du corpus . . . 60
3.3. Traitement des résultats . . . 70

4. Analyse des résultats et *matching* avec Cumuleo 77
4.1. Corpus général . . . 78
4.2. Entités nommées . . . 80
4.3. Cumuleo . . . 87
4.4. Réconciliation avec le corpus . . . 90

Conclusion 94

Annexes 98

A. Scripts Python 99

B. Opérations dans Open Refine **106**

C. Exemple de requête SPARQL **108**

Bibliographie **110**

Table des figures

1.4.1 Hiérarchie étendue d'entités nommées, tirée de [Sekine 10] 14
1.5.1 Exemple de transducteur CasEN destiné à reconnaître les organisations politiques. Les boîtes grisées indiquent un lien vers un autre graphe ou vers un dictionnaire [Maurel 11] 16

2.2.1 La pile de standards du web sémantique, reprise de Wikipedia (user :Marobi1) 33
2.2.2 Représentation visuelle d'un graphe RDF 35
2.2.3 Représentation du graphe requête 37
2.3.1 Les cinq étoiles du web de données liées ouvertes, selon le W3C 41
2.3.2 *Linked Open Data Cloud* (détail), extrait de [Cyganiak 14] 42
2.4.1 Schéma général d'un système d'*Entity linking* 43
2.4.2 Infoboxe de Wikipedia et la page correspondante de DBpedia version française 46

3.1.1 La pyramide inversée selon [Ross 94], repris de [Labasse 12] 57
3.1.2 Aperçu d'une dépêche dans le système interne *Belga360* 58
3.2.1 Schéma du processus 60
3.2.2 Exemple de dépêche annotée au format XML 68
3.3.1 Choix du type de service dans Open Refine 74

3.3.2 Synthèse de la procédure dans Open Refine 75

4.2.1 Distribution des sous-classes d'entités 81
4.2.2 Nombre et taux moyens d'entités nommées par élément de la dépêche 82
4.2.3 Pourcentages par catégories d'entités 83
4.2.4 Répartition des entités non redondantes 84
4.2.5 Distribution des entités non reconnues, classées par sous-types . . . 86
4.3.1 Schéma simplifié de Cumuleo 89
4.3.2 Extrait de la déclaration de mandats 2015 telle que publiée en PDF au Moniteur belge 89
4.3.3 Répartition des mandataires par formation politique 90
4.4.1 Distribution des PER reconnus par Cumuleo seul 92

Liste des tableaux

1.1 Tableau de contingence des résultats d'une recherche documentaire, repris de [Boydens 15, p. 155] 20

2.1 Répartition de quelques types d'entités nommées dans DBpedia . . . 49

3.1 L'une des dépêches du corpus au format CSV (lignes tronquées) . . . 63
3.2 Hiérarchie d'entités issue de l'annotation 71

4.1 Longueur moyenne des articles du corpus en caractères, espaces compris 79
4.2 Répartition des articles par thématiques 80
4.3 Répartition entre les trois grandes classes d'entités 80
4.4 Exemples d'entités nommées uniques contenues dans les titres/*leads* 85
4.5 Nombre de fonctions distinctes après normalisation sommaire 92
4.6 Distribution des `ORG` reconnus par Cumuleo seul 93

Introduction

Cadre général

Depuis un quart de siècle, la croissance ininterrompue du web met à notre disposition une somme chaque jour plus considérable de connaissances, dispersées dans des milliards de pages [1]. Mais enfermées dans des documents rédigés à l'attention d'êtres humains, ces connaissances ne peuvent être traitées que superficiellement par une machine, incapable de comprendre le langage naturel. Certains types de mots, toutefois, peuvent être isolés et extraits automatiquement d'un texte. L'opération fait l'objet d'une discipline informatique à part entière nommée « extraction d'informations », dont la tâche principale consiste à remplir automatiquement des formulaires ou une base de données à partir de textes non structurés [Moens 06, p. 2]. Sur l'ensemble des informations susceptibles d'être extraites d'un texte, une catégorie particulière nourrit les recherches depuis deux décennies. Sous le terme « entités nommées » (*Named Entities*), forgé en 1996, des informaticiens ont regroupé un ensemble flou de concepts qui se rapprochent, sans se limiter à lui, de celui de « noms propres ». Il s'agit par exemple des noms de personnes, de lieux, d'organi-

1. Près de 4.8 milliards à la date du 4 août 2016, selon une estimation minimale du nombre de pages indexées par les moteurs de recherche Google, Bing et Yahoo Search, et basée sur les travaux de [Bosch 16] : http ://www.worldwidewebsize.com/. Ces chiffres ne tiennent compte que des pages accessibles à ces moteurs, à l'exclusion donc du « web profond » (*deep web*). Une étude un peu datée estimait la taille de ce dernier à 120 fois au moins celle du web visible [Bergman 01].

sations, mais aussi parfois les dates, voire les noms de composants chimiques mentionnés dans un article scientifique.

Enjeux

Sans entrer plus avant dans leur définition, qui sera abordée dans le chapitre 1, nous pouvons préciser que ces objets linguistiques jouent un rôle essentiel en traitement automatique des langues (TAL). Comme le souligne [Ehrmann 08, pp. 25-28], pouvoir reconnaître les entités nommées dans un texte, et leur attribuer une catégorie, facilite d'autres tâches de TAL : l'analyse syntaxique, la résolution de coréférence, la désambiguïsation lexicale, la traduction automatique... Cette reconnaissance constitue également un maillon essentiel dans des applications particulières, que ce soit en matière d'extraction d'information[2] ou de veille informationnelle, dans les tâches de questions-réponses ou dans l'anonymisation de documents.

Plus près de notre point de vue, qui est celui des humanités numériques, les entités nommées peuvent également jouer un rôle en classification de documents. Comme le montre [Friburger 02], confirmant une intuition répandue, « les noms propres possèdent une qualité "classificatrice" plus importante que les autres mots d'un texte journalistique ». Pour le dire autrement, savoir de *qui* parle un texte permet de mieux deviner de *quoi* il parle.

Un autre enjeu essentiel est celui de la *résolution* ou de la *désambiguïsation* de ces entités nommées : une fois le segment de texte « Charles Michel » identifié comme un nom de personne, comment savoir de quel individu il s'agit parmi des milliers d'homonymes ? Répondre à cette question implique de pouvoir associer l'entité à un référent extralinguistique [Stern 10]. Ce dernier pourrait être une base de données classique, ou plus fréquemment une base de connaissance (*knowledge base*). Une fois cette résolution effectuée, il devient alors possible de relier entre eux, non pas des documents comme c'est le cas dans le web d'aujourd'hui, mais les *données* citées dans ces documents. La reconnaissance d'entités nommées devient sous cet angle un élément fondamental du web sémantique, ou du *Linked Open Data*, notions que

2. Notamment l'extraction d'événements, quand ceux-ci impliquent une "action" effectuée à une certaine date, dans un lieu et avec des participants [Serrano 12]. Certains, comme [Piskorski 11], résument même l'extraction d'événements à l'identification d'entités nommées et des relations qu'elles entretiennent entre elles.

nous présenterons au deuxième chapitre.

Questions de recherche

Dans ce mémoire, notre champ d'études sera restreint à ces textes bien précis que sont les dépêches d'agences. Du point de vue de l'extraction d'informations, ces genres journalistiques possèdent plusieurs propriétés intéressantes. Bien qu'il s'agisse d'un texte non structuré, une dépêche obéit à des règles formelles plus rigoureuses que la plupart des autres types d'articles. Destinée à informer plus qu'à plaire, elle respecte scrupuleusement (du moins en théorie) le principe de la *pyramide inversée* (voir chapitre 3). Les informations essentielles et leurs éléments constitutifs sont censés figurer dès le premier paragraphe, que l'on nomme en jargon le *lead* de la dépêche. Ces particularités, tout comme l'importance des dépêches d'agences dans le cycle de l'information, en ont fait un terrain d'expérimentation privilégié en extraction d'entités nommées.

Compte tenu de ces éléments, nos recherches prendront deux directions. D'une part, nous tenterons de déterminer dans quelle mesure la règle de la pyramide inversée peut faciliter la reconnaissance d'entités nommées. Sachant que ces entités peuvent répondre à au moins trois des cinq questions traditionnelles du journalisme (qui ? où ? et quand ?, à défaut de pourquoi ? et comment ?), leur taux de présence devrait être plus élevé dans le *lead* des dépêches, où sont censées apparaître les réponses à ces questions essentielles. Nous nous intéresserons par ailleurs à la nature et à la quantité des entités contenues dans ces *leads*, ainsi que dans les titres et les textes qui les accompagnent. L'hypothèse que nous émettons peut être formalisée de la manière qui suit :

Hypothèse 1 Le *lead* d'une dépêche d'agence doit répondre aux questions principales que soulève l'information. Il devrait donc contenir un taux d'entités nommées plus important que le corps du texte. De plus, ces entités devraient être plus représentatives que les autres du thème central de la dépêche.

Une deuxième phase consistera à comparer les résultats de nos analyses avec une base de données de noms propres, en l'occurrence Cumuleo.be. Cette dernière recense à l'heure actuelle quelque 15 000 noms de mandataires belges, les fonctions

qu'ils ont exercées depuis 2004 et le nom des organismes privés ou publics par lesquels ils ont transité durant cette période. Selon notre hypothèse, cette base de données pourrait contenir une part non négligeable des entités nommées contenues dans des dépêches belges d'informations générales.

Hypothèse 2 Des dépêches d'actualité générale centrée sur un petit pays comme la Belgique devraient contenir bon nombre d'entités nommées inconnues des grandes bases de connaissance internationales. Nous présumons qu'une part plus ou moins importante de ces entités figure dans Cumuleo, a fortiori si le thème de la dépêche est politique. Cette base de données pourrait dès lors jouer un rôle non négligeable dans la résolution d'entités nommées au sein de textes journalistiques belges.

Plan de l'exposé

Ce mémoire est structuré en deux parties, l'une purement théorique, l'autre plus applicative. Chacune est divisée en deux chapitres. Le chapitre 1 reviendra plus en détail sur la notion d'entités nommées, sur l'historique de ce concept et sur le problème de sa définition précise. Nous passerons en revue les principales méthodes et outils qui permettent, selon l'état de l'art, de reconnaître et de catégoriser automatiquement ces entités. Nous aborderons également leurs grandes typologies et introduirons le concept de désambiguïsation.

Ce dernier sera développé plus avant au chapitre 2, consacré aux bases de connaissance qui permettent de l'effectuer. Après avoir rappelé les fondements technologiques et conceptuels sur lesquels reposent ces KB, nous en présenterons quelques-unes et clarifierons l'idée de « résolution » des entités nommées.

La deuxième partie sera consacrée à l'étude quantitative d'un corpus de dépêches d'agence belges francophones, annoté à la main. Le chapitre 3 détaillera notre méthodologie, notamment la manière dont ce corpus a été constitué et traité. Dans le chapitre 4, enfin, nous tenterons sur la base de ce matériau de confirmer ou d'infirmer les questions de recherche énoncées plus haut.

Nos résultats, ainsi que les apports et limites de notre méthode, feront l'objet de la conclusion.

Partie 1 : Fondements théoriques

« What's in a name ?
That which we call a rose
By any other name
would smell as sweet. »

(William Shakespeare, Romeo and Juliet, *1597)*

Chapitre 1

Les entités nommées

1.1. Introduction 6
1.2. Historique 7
1.3. Définition 11
1.4. Typologies 13
1.5. Méthodes de reconnaissance 14
1.6. Évaluation des systèmes 18
1.7. Quelques systèmes de NERC 22
1.8. NERC : une tâche résolue ? 24
1.9. La désambiguïsation sémantique 26

1.1. Introduction

La reconnaissance et la classification d'entités nommées (désormais NERC, selon l'acronyme anglophone de *Named Entity Recognition and Classification*) peut être définie comme l'activité qui consiste à retrouver et à catégoriser dans des documents les noms de personnes, de lieux et d'organisations, mais aussi parfois les mentions de dates, d'heures, etc.

Le « etc. » a toute son importance : faute de commun dénominateur entre ces unités lexicales ou ces syntagmes, leur définition se résume souvent à une liste énumérative ouverte [Ehrmann 08, p. 73].

Ce flou trouve son origine dans l'histoire même des entités nommées, concept issu non pas du champ de la linguistique, mais de celui de l'extraction d'informations. Les informaticiens qui ont forgé le terme étaient moins soucieux de rigueur conceptuelle que d'efficacité opérationnelle. Il s'agissait avant tout pour eux de remplir des tâches clairement définies et d'intérêt stratégique, voire militaire.

À mesure que le champ des entités nommées s'est affranchi des textes de type journalistique, sa relation étroite avec le concept de noms propres s'est distendue. Dans le domaine de la bio-informatique, les types d'entités recherchées s'apparentent plutôt à des noms communs désignant des protéines ou des substances pharmaceutiques [Settles 04]. En outre, les documents dans lesquels sont recherchées les entités nommées ne se limitent plus dorénavant à des textes écrits, mais peuvent englober des productions audiovisuelles [Hatmi 14].

Dans ce chapitre, nous commencerons par retracer l'historique du concept au travers de celui des campagnes d'évaluation qui ont permis de le définir. Nous verrons également certaines des typologies mises au point jusqu'ici, ainsi que les méthodes destinées à évaluer les performances d'un système de reconnaissance d'entités nommées. Quelques-uns d'entre eux seront brièvement présentés. Enfin, après avoir abordé les domaines de recherche toujours ouverts en NERC, nous introduirons le concept de « désambiguïsation » et de « résolution d'entité », qui feront l'objet d'une étude plus détaillée au chapitre 2.

1.2. Historique

Selon [Friburger 02, p. 23], le premier travail informatique relatif à la détection de noms propres semble être celui de [Borkowski 66]. Ce dernier proposait alors un système d'identification des noms de personnes et de leurs titres dans les journaux en utilisant des listes de « marqueurs » (preuves internes). Mais les noms propres ne constituent qu'un sous-ensemble de ce que l'on appelle « entités nommées ». Pour cerner ce concept flou, issu du domaine de l'extraction d'informations, il est nécessaire de se replonger dans les campagnes d'évaluation qui l'ont vu naître.

1.2.1. Les conférences MUC

Le terme *Named Entities* fut forgé aux États-Unis à l'occasion de la 6e *Message Understanding Conference* (MUC), qui s'est tenue en novembre 1995 à Columbia, dans le Maryland [1]. En dépit de leur nom, ce sont moins les conférences elles-mêmes que leurs campagnes d'évaluation préparatoires qui ont marqué l'histoire de l'extraction d'informations [Grishman 96]. Financées par l'ARPA-DARPA (*Defense Advanced Research Projects Agency*), une agence départementale de la Défense états-unienne [2], les MUC (1987-1998) tiraient leur nom du projet originel, qui visait à promouvoir la recherche autour de la compréhension automatique de messages militaires textuels. Dès MUC-3, en 1991, le champ de recherche s'élargit aux textes journalistiques, en l'occurrence afin de détecter automatiquement les mentions d'actes terroristes en Amérique latine [Chinchor 93]. C'est également au cours de MUC-3 que fut mis en place un système d'évaluation semi-automatisé basé sur la précision et le rappel [Chinchor 91] (voir Section 1.6.2).

Dans chaque cas, les participants se voyaient remettre un corpus d'entraînement, à charge pour eux de développer un système capable d'en extraire automatiquement une série d'informations précises. Les systèmes étaient ensuite évalués sur un corpus de test. Enfin, les résultats étaient présentés et débattus lors de la conférence finale.

C'est ainsi que lors de MUC-6 [3] et 7, l'une des tâches (*Named Entity task*) consista à identifier dans des textes en anglais sept catégories d'informations, regroupées en trois classes créées pour l'occasion :

— ENAMEX (*Named Entities Expressions*) : les noms de personnes, de lieux et d'organisations ;
— NUMEX (*Number Expressions*) : les expressions monétaires et les pourcentages ;
— TIMEX (*Temporal Expressions*) : les dates et les heures.

On remarquera comme [Nouvel 15, p. 26] le « flottement historique » à l'œuvre, dès l'origine, dans la définition du terme « entités nommées » :

« Désignant à l'origine les seuls noms propres de personnes, de lieux et d'organisations rassemblés au sein de la catégorie ENAMEX de MUC, l'appellation "named

1. http ://www.cs.nyu.edu/cs/faculty/grishman/muc6.html
2. Rebaptisée DARPA en 1996, l'ARPA est également à l'origine du réseau d'ordinateurs ARPANET (1972), ancêtre d'internet [Volle 06, pp. 333-335].
3. http ://cs.nyu.edu/faculty/grishman/NEtask20.book_1.html

entity" s'étend rapidement aux autres grandes classes des expressions numériques (NUMEX) et temporelles (TIMEX), cette dérivation ayant lieu au sein même des participants aux premières conférences. »

Listing 1.1 – Exemple de balisage SGML extrait de MUC-6

```
Mr.<ENAMEX TYPE="PERSON">Dooner</ENAMEX> met with <ENAMEX TYPE="PERSON">Martin
    Puris</ENAMEX>, president and chief executive officer of <ENAMEX TYPE="
    ORGANIZATION">Ammirati & Puris</ENAMEX>, about <ENAMEX TYPE="ORGANIZATION">
    McCann</ENAMEX>'s acquiring the agency with billings of <NUMEX TYPE="MONEY
    ">$400 million</NUMEX>, but nothing has materialized.
```

1.2.2. ACE

Organisée de 1999 à 2008 par le National Institute of Standards and Technology (NIST), la campagne ACE (*Automatic Content Extraction*) [4] visait à développer des technologies pour déduire automatiquement, à partir de données humaines linguistiques, les entités mentionnées, les relations entre ces entités et les événements dans lesquels ces entités participent. En plus de textes, les sources de données comprenaient des documents audio et des images océrisées, tandis que l'arabe et le chinois s'ajoutaient à l'anglais [Doddington 04].

ACE se distinguait également de MUC par la nature des entités recherchées. Celles-ci ne se limitaient plus à des noms propres, mais englobaient également des descriptions (« papa », « le boucher »...) et les pronoms qui s'y rapportent, ce qui ajoutait au processus de reconnaissance d'entités toute la complexité de la résolution des coréférences [5]. Par ailleurs, ACE distinguait clairement la reconnaissance d'entités (*Entity Detection and Tracking*) et celle des expressions temporelles ou chiffrées.

Aux trois grandes classes d'entités nommées définies lors des conférences MUC (personnes, lieux, organisations), ACE a ajouté les armes, les véhicules, les *Facilities* (bâtiments et autres structures créées par l'homme) et les *Geo-Political Entities*

4. https ://www.ldc.upenn.edu/collaborations/past-projects/ace

5. La coréférence désigne en linguistique le lien entre différentes expressions dans un texte qui se rapportent à la même entité. Elle ne doit pas être confondue avec l'anaphore, élément du discours qui fait référence à un autre élément qui précède. Dans les phrases « Comment va Odette ? - Ma mère ? je l'ai vue hier, elle va bien », les pronoms « elle » et « l' » sont des anaphores de leur antécédent « mère ». « Odette », « elle », « l' » et « mère » sont, eux, coréférents [Kübler 15, p. 118].

(GPEs).

Ces deux dernières classes visaient à simplifier l'annotation humaine en levant certaines ambiguïtés. *Facility* permet ainsi de traiter le cas de lieux nommés par métonymie sous un nom d'organisation ou de personne (ONU pour désigner le siège des Nation-Unies à New York, Brugmann pour qualifier l'hôpital bruxellois du même nom...). Quant à GPE, elle englobe toute mention géographique qui peut désigner à la fois un lieu et une communauté politique ou sociale - comme dans la phrase « Bagdad a réagi à l'ultimatum de Washington », où Bagdad et Washington représentent respectivement les gouvernements irakien et américain.

Enfin, les sept types d'entités ont été divisés en sous-types, ce qui permettait une catégorisation plus fine. Dans sa dernière version [ACE 08], le guide d'annotation d'ACE incluait ainsi pas moins de 45 sous-types :

1. ***Person :*** *Group, Indeterminate, Individual ;*
2. ***Location :*** *Address, Boundary, Celestial, Land-Region-Natural, Region-General, Region-International, Water-Body ;*
3. ***Organization :*** *Commercial, Educational, Entertainment, Government, Media, Medical-Science, Non-Governmental, Religious, Sports ;*
4. ***Facility :*** *Airport, Building-Grounds, Path, Plant, Subarea-Facility ;*
5. ***Geo-Political :*** *Continent, County-or-District, GPE-Cluster, Nation, Population-Center, Special, State-or-Province ;*
6. ***Vehicle :*** *Air, Land, Subarea-Vehicle, Underspecified, Water ;*
7. ***Weapon :*** *Biological, Blunt, Chemical, Exploding, Nuclear, Projectile, Sharp, Shooting, Underspecified.*

1.2.3. Autres conférences

Bien que très largement dominée par l'anglais, la recherche en matière d'entités nommées connut très vite des déclinaisons dans d'autres langues. Dès 1996, dans le sillage de MUC-6 et 7, les conférences MET et MET 2 (*Multilingual Entity Task*) se penchèrent sur des corpus en espagnol, chinois et japonais [Merchant 96]. L'on pourrait également mentionner la campagne IREX (japonais), HAREM (portugais),

EVALITA (italien), GermEval (allemand) ou encore ConLL 2002 et 2003 pour l'espagnol et le néerlandais, puis l'anglais et l'allemand. Un récapitulatif des principales campagnes d'évaluation peut être trouvé dans [Nadeau 07b] et [Nouvel 15].

Pour le français, langue qui nous intéresse plus particulièrement, les campagnes de référence sont ESTER 1 (2003-2005) et ESTER 2 (2007-2009), qui visaient à évaluer des systèmes de transcriptions d'émissions radiophoniques [6] [Galliano 09]. Elles furent suivies en 2011 par le projet ETAPE [Galibert 14], axé sur « l'évaluation des performances des technologies vocales appliquées à l'analyse des flux télévisés en langue française » [7].

1.3. Définition

Nous avons vu que le concept d'entités nommées désignait à l'origine certaines catégories de noms propres (les noms de personnes, de lieux et d'organisations), mais aussi des expressions temporelles et numériques. Par la suite sont venus s'ajouter des noms communs (ceux d'armes par exemple), des descriptions, des pronoms... Cette constatation faite, quelle unité peut-on trouver dans un ensemble d'objets lexicaux si hétéroclite ? [Nadeau 07b] livre une tentative de définition en empruntant au logicien Saul Kripke le concept de *désignateur rigide* [Kripke 80, p. 48], à savoir « quelque chose qui désigne le même objet dans tous les mondes possibles ».

C'est ainsi que, selon Kripke, le nom propre « Richard Nixon » peut être considéré comme un désignateur rigide, mais pas « le président des Etats-Unis », puisque le référent de cette expression change périodiquement.

Nadeau et Sekine sont toutefois contraints d'admettre que cette définition s'applique imparfaitement aux dates [8]. « It is arguable that the NE définition is loosened in such cases for practical reasons », concluent-ils.

La linguiste Maud Ehrmann a sans doute mené le travail de définition le plus étoffé. Elle reprend en partie la notion de désignateur rigide (qu'elle rebaptise désignateur stable), mais en le liant aux notions de corpus et de modèle applicatif :

6. http ://www.afcp-parole.org/camp_eval_systemes_transcription/
7. http ://www.afcp-parole.org/etape.html
8. « 2001 » peut être considéré comme un désignateur rigide de « la 2001e année du calendrier grégorien », mais pas l'expression « en juin dernier ».

> « Étant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus. » [Ehrmann 08, p. 168].

Examinons de plus près cette proposition de définition. Elle implique donc :

- Un modèle applicatif : C'est-à-dire une conceptualisation d'une partie du réel à visée opérationnelle, par exemple un système automatique de traitement d'informations relatives à un domaine précis.
- Un corpus : l'ensemble des données langagières auxquelles est appliqué le traitement, par exemple des articles de journaux.
- Des expressions linguistiques : noms propres, descriptions définies, expressions temporelles…
- Des entités uniques du modèle : « Une entité nommée est une expression linguistique monoréférentielle », précise Ehrmann. Elle renvoie à un référent unique dans le cadre d'un modèle. Autrement dit, « le président de la République » peut se référer à une entité unique (François Hollande), mais ne devient entité nommée que si un modèle applicatif requiert son identification et annotation.
- Autonomes : Elles peuvent, par leurs seules ressources, évoquer un référent. L'expression « le président » peut se référer à une entité unique, mais son identification nécessitera des connaissances extralinguistiques ou des éléments de contexte, contrairement à la description définie « le président de la République française en 2005 » [Nouvel 15, p. 46].

Selon cette définition, rien n'est donc entité nommée par nature, et rien n'est entité nommée s'il n'existe aucun besoin de l'identifier et de l'annoter. Pures constructions du TAL, les EN sont apparues au croisement de la linguistique et de l'informatique ; elles réclament ce double contexte pour exister. Ce sont aussi des constructions empiriques. Comme le souligne déjà [De Wilde 15, p. 30], on peut appliquer à la reconnaissance d'EN le principe de *fitness for use* (adéquation aux usages) issu du domaine de la qualité des données : « if an entity is useful for our needs, then it can be argued that it is valid, notwithstanding epistemological considerations. »

1.4. Typologies

De la définition qui précède, l'on peut déduire que, si rien n'est entité nommée par nature, tout peut le devenir pour peu qu'un modèle applicatif le nécessite. Dans ces conditions, tenter d'établir une typologie reviendrait à dresser l'inventaire stérile et jamais définitif des objets linguistiques qui ont pu être considérés comme entités nommées au cours des campagnes d'évaluation menées depuis vingt ans.

Pourrait-on toutefois imaginer une typologie applicable à tout corpus « généraliste », par exemple de type journalistique? C'était toute l'ambition des travaux de [Sekine 02], qui ont abouti à une vaste classification de 150, puis 200 types d'entités nommées [Sekine 04b].

Au vu des campagnes MUC-7, IREX et ACE, les auteurs en ont conclu qu'une typologie de sept ou huit entités s'avère insuffisante pour traiter un contexte général – sans même parler, donc, d'un contexte spécialisé tel que la génomique.

Pour aboutir à leur « hiérarchie étendue », ils ont d'abord conçu trois hiérarchies initiales, puis les ont fusionnées. La première était basée sur l'annotation d'entités nommées dans un corpus de journaux, la deuxième sur des classifications déjà utilisées dans des tâches précédentes et la troisième sur des thésaurus tels que WordNet [9] [Miller 98] et le Roget [Roget 11]. Le tout a enfin été affiné après l'annotation d'un nouveau corpus et le développement d'annotateurs automatiques. La seule relation thésaurale retenue entre les entités est de type hyperonomique [10].

Afin de limiter les ambiguïtés lors de l'annotation automatique, les auteurs ont emprunté à ACE le concept de GPE (*Geographical and Political Entity*) et introduit celui de GOE (*Geographical and Organizational Entity*), qui rassemble les lieux dont le nom désigne également une organisation - par exemple un musée ou un aéroport.

En dépit de ces précautions, note l'auteur principal dans [Sekine 04a], une hiérarchie d'entités si étendue présente à la fois des avantages et des inconvénients. Là où des notions telles que GOE ou GPE facilitent l'annotation, la multiplication de catégories fines et sémantiquement proches la complique. Par exemple, comment classer un typhon léger quand les classes « natural phenomena » et « natural disaster » coexistent?

9. https ://wordnet.princeton.edu/
10. A (concept spécifique) EST UN B (concept générique). cf. [Hudon 13, p. 155].

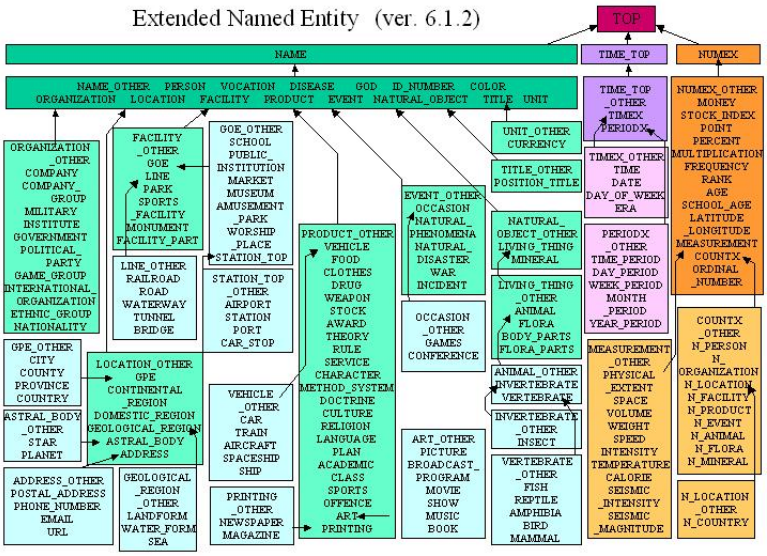


FIGURE 1.4.1. – Hiérarchie étendue d'entités nommées, tirée de [Sekine 10]

Autre problème, que nous détaillerons à la section suivante : plus le nombre de classes augmente, plus il devient difficile d'en répertorier un nombre suffisant d'instances dans un corpus annoté manuellement. Or certaines méthodes de reconnaissance des entités nommées réclament de grandes quantités d'exemples annotés.

1.5. Méthodes de reconnaissance

A l'image de la plupart des applications TAL [Poibeau 11, p. 24], les systèmes de reconnaissance d'entités nommées peuvent se diviser globalement en trois grandes catégories : les systèmes à base de règles, les systèmes fondés sur l'apprentissage automatique et les méthodes hybrides [Poibeau 01].

1.5.1. Les systèmes à base de règles

Ces méthodes dites aussi « linguistiques » ou « symboliques » impliquent de confectionner à la main des patrons d'extraction, à la manière d'une expression régulière [11].

11. Les expressions régulières (souvent abrégées « regex » en anglais) constituent des notations en langage formel permettant de désigner des classes de chaînes de caractères [Jurafsky 14, p. 18]. L'expression régulière `\d+` correspond ainsi à n'importe quelle suite ininterrompue d'un ou plusieurs

Ceux-ci peuvent être fondés sur des preuves internes (l'abréviation inc dans Microsoft inc en tant que nom d'entreprise) ou externes (les termes « Monsieur » ou « Madame » devant des noms de personnes) [Friburger 02, p. 16]. Certains de ces ensembles de règles sont couplés à des lexiques, notamment des dictionnaires de noms propres.

L'on peut citer comme exemple de ces méthodes le système CasEN [Maurel 11] pour la reconnaissance d'entités nommées en français, qui fait appel à une cascade de transducteurs dans le logiciel CasSys de la plateforme Unitex[12]. Comme en témoigne la Figure 1.5.1, définir le patron d'un seul sous-type d'entité (ici les organisations politiques) peut réclamer une quantité considérable de travail et d'expertise linguistique.

Dominants jusqu'au début des années 90, les systèmes à base de règles ont peu à peu cédé du terrain face aux méthodes probabilistes (section 1.5.2), sans pour autant devenir obsolètes. Plus ardus à concevoir et à maintenir, « ils prédominent pour les langues ou les typologies pour lesquelles il n'existe pas de corpus de données annoté de taille suffisante », soulignent [Nouvel 15, p. 91]. Généralement très précis, ils permettent en outre un contrôle plus fin sur les règles de détection. Ils ont toutefois l'inconvénient d'offrir un rappel plus faible. Chaque type de détection devant être prévu à l'avance et codifié, ils s'adaptent mal à un contexte nouveau.

1.5.2. Les systèmes à apprentissage automatique

Issues des travaux en IA, les méthodes par apprentissage automatique (*Machine Learning*) reposent sur un principe aussi simple que séduisant : plutôt que d'écrire soi-même un programme qui effectue une action quelconque, par exemple annoter un texte, il est plus avantageux d'en écrire un qui apprend à le faire tout seul à partir d'exemples [Tellier 10, p. 95]. Autrement dit, un programme qui *s'améliore* à mesure qu'il reçoit de nouvelles données. Pour reprendre la définition formelle de l'un des

chiffres, tandis que `(?<=<b>)\w+(?=</b>)` identifie n'importe quels mots situés entre deux balises HTML [Goyvaerts 09, p. 84]. Issues des travaux de Stephen C. Kleene [Kleene 51], les expressions régulières entretiennent un rapport étroit avec les transducteurs finis utilisés en TAL, puisque tout langage définissable par l'un de ces automates peut l'être par une expression régulière, et vice-versa [Hopcroft 01, p. 90]. Le site http ://hackingoff.com/compilers/regular-expression-to-nfa-dfa permet de vérifier visuellement cette équivalence.

12. http ://www-igm.univ-mlv.fr/~unitex/

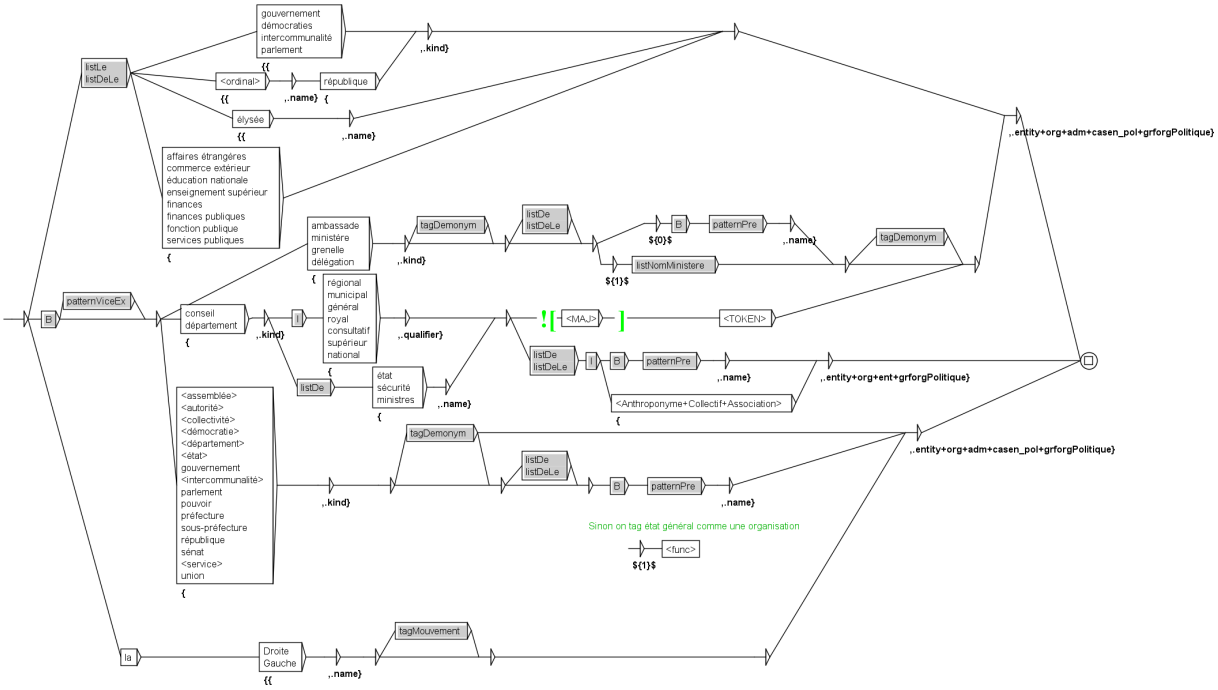


FIGURE 1.5.1. – Exemple de transducteur CasEN destiné à reconnaître les organisations politiques. Les boîtes grisées indiquent un lien vers un autre graphe ou vers un dictionnaire [Maurel 11]

pionniers de ces systèmes :

> « A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E. » [Mitchell 97, p. 2].

Ces méthodes, parfois appelées « probabilistes » ou « fondées sur les données », peuvent elles-mêmes se subdiviser en trois grandes catégories [Pustejovsky 13, p. 141] :

— **Les méthodes non supervisées**, qui visent à découvrir une structure dans un jeu de données non annoté. La classification automatique [Sebastiani 02] et le regroupement (*clustering*) en constituent deux applications courantes. [Etzioni 05] ou encore [Nadeau 06] fournissent des exemples d'applications à la classification d'entités nommées ;

— **Les méthodes supervisées**, dans lesquelles il s'agit de reproduire automatiquement un schéma d'annotation appris d'un corpus annoté manuellement. Les systèmes qui y ont recours peuvent faire appel à une large gamme de modèles d'apprentissage automatique, tels que les champs conditionnels aléatoires (CRF), les modèles de Markov cachés (HMM), les machines à vecteurs de support (SVM), les classifieurs d'entropie maximale (MaxEnt), etc. ;

— **Les méthodes semi-supervisées**, qui combinent les deux précédentes en ingérant à la fois des données annotées et des données brutes. Pour un exemple concret, voir [Nadeau 07a].

En matière de reconnaissance d'entités nommées, les systèmes à apprentissage automatique supervisé se sont principalement développés grâce à la publication de vastes corpus annotés, issus notamment des campagnes d'évaluation déjà citées. Pour les langues peu dotées, en revanche, la confection de tels corpus *ex nihilo* peut s'avérer rédhibitoire. Bien que l'annotation réclame moins de compétences que la confection de règles linguistiques, obtenir un F-score (voir Section 1.6) satisfaisant impose des volumes de données considérables. [Poibeau 01] cite plusieurs exemples de systèmes qui nécessitaient plus d'un million de mots annotés afin obtenir un taux de rappel/précision avoisinant 0.90. Par ailleurs, un système entraîné sur un corpus précis offrira de piètres performances sur un autre. [Ciaramita 05], par exemple, a constaté une chute de performances considérable dans un système entraîné sur des

dépêches *Reuters* (ConLL 2003) appliqué à un corpus d'articles du *Wall Street Journal,* le F1-score étant passé de 0.908 à 0.643.

1.5.3. Les systèmes mixtes

Enfin, rien n'interdit qu'un ensemble de règles apprises automatiquement soit ensuite affiné par un linguiste, ou l'inverse. [Béchet 11], par exemple, témoigne de la complémentarité de deux démarches en associant un système à base de règles, conçu sur des dépêches de l'Agence France-Presse, avec un système probabiliste entraîné sur des transcriptions de l'oral issues du corpus ESTER. Selon leur conclusion, cette adaptation leur a permis de disposer de deux systèmes, l'un privilégiant la précision, l'autre le rappel, « sans nécessiter aucune annotation supplémentaire, illustrant la complémentarité des méthodes numériques et symboliques pour la résolution de tâches linguistiques ».

Ce type de métissage semble constituer la voie la plus prometteuse. Loin de se cantonner à la reconnaissance d'entités nommées, le croisement des méthodes à base de règles et des systèmes probabilistes est d'ailleurs devenu une tendance centrale dans l'ensemble du TAL [Tellier 09, Watrin 06].

1.6. Évaluation des systèmes

1.6.1. Principes de base

Comme son nom l'indique, la tâche de NERC est en réalité composée deux sous-tâches. A partir d'un texte donné, il s'agit :

1. de *reconnaître les entités*, autrement dit d'identifier le début et la fin des segments de texte qui les composent ;
2. de *les classer* dans des catégories prédéfinies (personne, lieu, date, etc.).

Au terme du processus, un système de reconnaissance renverra le texte annoté dans un format balisé, comme nous l'avons vu dans Listing 1.1, ou encore au format IOB [Tjong Kim Sang 00, Ramshaw 95]. Dans cette notation, les éléments identifiés par un B (*beginning*) marquent le début d'un segment reconnu, le I (*internal*) les éléments à l'intérieur du segment et le O (*outside*) les éléments étrangers.

Prenons la phrase :

```
John Kennedy a été assassiné à Dallas le 22 novembre 1963.
```

L'une de ses interprétations au format IOB pourrait être :

John	B–PER
Kennedy	I–PER
a	O
été	O
assassiné	O
à	O
Dallas	B–LOC
le	O
22	B–DATE
novembre	I–DATE
1963	I–DATE
.	O

On le devine, le logiciel commencerait par segmenter le texte en *tokens* (unités lexicales) avant d'attribuer à chacun une étiquette morphosyntaxique - limitée ici à son appartenance ou non à l'une des catégories d'entités nommées recherchées, ainsi qu'à sa position initiale ou non dans le segment identifié.

Si cette reconnaissance s'effectuait dans le cadre d'une campagne d'évaluation, le résultat pourrait être aussitôt transmis à un logiciel « scoreur » [Douthat 98], qui se chargerait d'évaluer le système.

1.6.2. Métriques d'évaluation

Dès MUC-3, nous l'avons dit, le besoin se fit sentir d'évaluer les systèmes en lice selon des mesures objectives. Le choix se porta sur des indicateurs classiques en extraction d'informations, à commencer par le rappel et la précision. Ces deux notions sont fondées sur celle de *pertinence* d'un résultat par rapport à un besoin d'information, et donc sur celles de *bruit* et de *silence*. Lors d'une recherche d'informations, les résultats renvoyés ou non par un système d'information peuvent en effet être divisés en quatre catégories (Tableau 1.1).

Résultats	Pertinents	Non pertinents
Extraits	TP : *true positives*	FP : *false positives*
Non extraits	FN : *false negatives*	TN : *true negatives*

TABLE 1.1. – Tableau de contingence des résultats d'une recherche documentaire, repris de [Boydens 15, p. 155]

Les résultats de la catégorie TP sont ceux qui répondent aux attentes. Ceux de la catégorie FP (extraits, mais non pertinents) relèvent du *bruit*, tandis que ceux de la catégorie FN (pertinents, mais non extraits) du *silence*. La catégorie TN (non pertinents et non extraits) est quant à elle rarement mesurable et présente un intérêt limité.

Compte tenu de cette matrice, le rappel désigne le taux d'entités pertinentes effectivement identifiées, ou taux d'exhaustivité. Il peut être calculé selon la formule :

$$R = \frac{TP}{TP + FN} \tag{1.6.1}$$

La précision (taux d'entités pertinentes parmi les entités identifiées, ou taux de pertinence) équivaut quant à elle à :

$$P = \frac{TP}{TP + FP} \tag{1.6.2}$$

Ces deux indicateurs sont souvent agrégés en un seul, la « F-measure » [Van Rijsbergen 79], qui correspond à leur moyenne harmonique pondérée :

$$F_\beta = (1 + \beta^2) * \frac{2 * R * P}{(\beta^2 * P) + R} \tag{1.6.3}$$

β désigne ici un réel positif. Sa valeur indique l'importance relative que l'on souhaite accorder au rappel par rapport à la précision (si $\beta = 2$, le rappel aura deux fois plus de poids que la précision ; si $\beta = 0.5$, deux fois moins). Quand l'importance des deux est identique (β=1), on parle alors de « F1-score », calculable selon la formule simplifiée suivante :

$$F_1 = \frac{2 * R * P}{P + R} \tag{1.6.4}$$

La manière dont cette formule standard peut être appliquée varie toutefois d'une

campagne à l'autre, ce qui brouille les tentatives de comparaisons entre elles. Dans MUC-6, par exemple, la reconnaissance d'une entité (le segment de texte) et sa catégorisation (le type d'entité nommée auquel il appartient) faisaient l'objet d'évaluations distinctes, regroupées in fine.

Dans ConLL 2002 et 2003, en revanche, une reconnaissance n'était considérée comme correcte que si elle attribuait la bonne catégorie au bon segment de texte. Plus simple et plus claire, cette méthode se révèle parfois trop sévère. Comme le souligne [Konkol 12], « if the original entity is « The United States » and « United States » is marked by the system, then « The United States » is considered as FN and « United States » as FP. The result is, that the system is penalized in two ways for almost good answer ».

Plus fondamentalement, [Makhoul 99] ont démontré que la mesure F telle qu'elle était utilisée dans les évaluations de MUC [Chinchor 95] tend à minimiser le poids de certaines erreurs par rapport à d'autres. Le problème n'est pas lié aux notions de précision et de rappel, mais à leur fusion dans une moyenne unique. S'inspirant du *Word Error Rate* utilisé en reconnaissance de la parole [McCowan 04], les auteurs ont proposé une métrique nommée *Slot Error Rate* (SER), qui est égale à la somme des trois types d'erreurs prises en compte dans MUC – substitutions (S), omissions (D) et insertions (I) - divisée par le nombre d'entités dans la référence (R) :

$$SER = \frac{S + D + I}{R} \tag{1.6.5}$$

Contrairement à la mesure F, le SER mesure un taux d'erreurs, et non de performance. Il devrait donc idéalement se rapprocher de zéro.

Enfin, signalons pour mémoire l'existence de métriques plus fines encore, comme ETER [Jannet 14], qui vise à mieux évaluer la reconnaissance et la classification d'entités structurées selon une hiérarchie complexe.

1.6.3. L'accord inter-annotateurs

Évaluer un système de NERC implique de comparer les résultats du système avec ceux d'un ou plusieurs annotateurs humains, qui constituent le corpus de référence (*Gold Standard Corpus*). Mais comment estimer la qualité de l'annotation humaine, puisque comme le souligne [Garside 97, p. 2], l'annotation constitue une *interprétation* du texte ? Aucun mécanisme strictement objectif ne permet ainsi de décider

quel label appliquer à tel phénomène linguistique. Pour pallier ce problème, la solution retenue consiste à comparer plusieurs annotations humaines. Si les résultats sont suffisamment identiques, et donc s'il existe un consensus suffisant entre humains, alors seulement est-il envisageable qu'une machine parvienne à reproduire ces annotations. Plusieurs méthodes ont vu le jour afin de mesurer l'accord inter-annotateurs, notamment la famille des *kappas*. Ces méthodes reposent sur la comparaison entre l'accord observé entre annotateurs et l'accord qui aurait pu se produire par le seul fait du hasard. Le plus connu des coefficients kappa est celui de Cohen [Cohen 60], destiné à mesurer l'accord entre deux annotateurs seulement [13]. Il se calcule selon la formule suivante :

$$K = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{1.6.6}$$

Où `Pr(a)` désigne l'accord observé entre annotateurs et `Pr(e)` l'accord attendu entre annotateurs si chacun avait choisi un label au hasard. Tous deux peuvent être calculés à l'aide d'une matrice de confusion. Le résultat peut aller de 0 (aucun accord) à 1 (accord parfait). On estime généralement qu'un kappa de 0.80 est suffisant pour tirer des conclusions fiables [Carletta 96].

Il est également possible de calculer un accord *intra*-annotateur (de l'annotateur avec lui même), ce qui fournit un indice sur la cohérence et la reproductibilité du travail effectué [Gut 04, Krippendorff 04, p. 215].

1.7. Quelques systèmes de NERC

Plusieurs outils librement disponibles permettent d'atteindre une précision et un rappel honorables, du moins sur des textes généraux, en langues dominantes et pour des entités classiques. Certains systèmes peuvent toutefois être entraînés sur d'autres corpus et d'autres types d'entités. Nous nous limiterons à en citer cinq parmi les plus connus.

— **Stanford Named Entity Recognizer (NER)** [14], qui implémente en Java des modèles de reconnaissance basés sur des champs conditionnels aléatoires (CRF)

13. Tandis que l'accord entre trois annotateurs ou plus peut être mesuré à l'aide du kappa de Fleiss [Fleiss 71].
14. http ://nlp.stanford.edu/software/CRF-NER.shtml

linéaires [Finkel 97]. Ces modèles ont été entrainés sur les corpus de dépêches de ConLL 2003, MUC-6 et 7 et une partie d'ACE 2002. En anglais, ils peuvent détecter selon les cas de trois (Location, Person, Organization) à sept types d'entités (Location, Person, Organization, Money, Percent, Date, Time). D'autres modèles sont disponibles pour l'espagnol, l'allemand et le chinois. Inclus dans la suite d'outils Stanford CoreNLP [15], Stanford NER peut également être employé seul ou avec les systèmes Apache Tika [16], UIMA [17] ou encore avec la bibliothèque Python NLTK [18] [Bird 09].

— **Illinois Named Entity Tagger** [19] [Ratinov 09], un annotateur capable d'identifier en anglais quatre classes d'entités (person, organization, location et miscellaneous), mais dont les versions récentes peuvent aller jusqu'à 18 classes (celles utilisées dans le corpus annoté OntoNotes [20] [Weischedel 13]). Il emploie notamment des lexiques extraits de Wikipedia. Décrit par ses concepteur comme robuste, il peut atteindre un F1-score de 90.8 sur les corpus CoNLL 2003. Il peut être utilisé seul ou en tant que composant de l'Illinois NLP Curator [21].

— **Apache OpenNLP** [22], une suite TAL basée sur des algorithmes d'apprentissage automatique (maxEnt et Perceptron). La reconnaissance d'entités nommés constitue l'une de ses fonctionnalités parmi d'autres (tokenization, segmentation en phrases, POS-tagging, résolution de coréférence, etc). Son module « Name Finder » dispose de modèles pré-entraînés sur divers corpus en anglais, néerlandais et espagnol [23]. Un modèle francophone entraîné sur le corpus ESTER est disponible sur la plateforme Github [24] [Azpeitia 14].

— **LingPipe** [25], une suite de bibliothèques Java dédiées au TAL, développée par la société américaine Alias-I. LingPipe est capable de détecter des entités classiques de type personnes, lieux et organisations en anglais, mais dispose aussi

15. http://stanfordnlp.github.io/CoreNLP/
16. https://tika.apache.org/
17. https://uima.apache.org/
18. http://www.nltk.org/
19. https://cogcomp.cs.illinois.edu/page/software_view/NETagger
20. https://catalog.ldc.upenn.edu/LDC2013T19
21. http://cogcomp.cs.illinois.edu/page/software_view/Curator
22. https://opennlp.apache.org/
23. http://opennlp.sourceforge.net/models-1.5/
24. https://github.com/opener-project/nerc-fr
25. http://alias-i.com/lingpipe/

de modèles plus spécialisés, notamment pour la détection de noms de gènes. Il peut par ailleurs être entraîné sur d'autres langues et d'autres types d'entités. Une licence gratuite perpétuelle est disponible pour les chercheurs.

— **ANNIE** [26] (acronyme de *A Nearly-New IE system*), qui fait partie intégrante de la plateforme logicielle GATE (*General Architecture for Text Engineering*) [Cunningham 14], très utilisée en analyse de corpus. Ce système repose notamment sur des algorithmes à états finis, sur des lexiques et sur un langage nommé JAPE (*Java Annotation Patterns Engine*), qui permet d'établir des règles d'annotation à l'aide d'expressions régulières. ANNIE peut reconnaître divers types d'entités nommées (personnes, lieux, organisations, dates, adresses, etc.), essentiellement en anglais. Il peut aussi être adapté à d'autres langues [Maynard 03].

1.8. NERC : une tâche résolue ?

Vingt ans après MUC-6, la tâche de reconnaissance et de classification des entités nommées semble avoir atteint sa maturité. Dès le milieu des années 2000, les taux de succès à 95 % obtenus par certains systèmes, équivalent à ceux d'un humain [27], laissaient penser à certains auteurs que la tâche était pratiquement résolue [Cunningham 05]. Toutefois, nuance [Marrero 12], cet optimisme ne vaut que pour des cas précis :

> « *NER has been considered a solved problem when the techniques achieved a minimum performance with a handful of NE types, document genre and usually in the journalistic domain. We do not know how well current techniques perform with other types of NE and different kinds of documents.* »

Si l'on y ajoute le problème du multilinguisme, nous avons là un parfait résumé des trois principales difficultés pour lesquelles la NERC reste une tâche ouverte.

Langues Tant les systèmes à apprentissage supervisé que ceux à base de règles sont étroitement dépendants de la langue pour laquelle ils ont été conçus. Par ailleurs, comme l'a démontré [Palmer 97], la difficulté de la tâche peut varier du

26. https ://gate.ac.uk/sale/tao/splitch6.html#chap :annie

27. Lors de MUC-7, le meilleur système atteignait une mesure F globale de 93.39, contre 97.6 et 96.95 pour les annotateurs humains [Marsh 98].

simple au double selon les langues considérées [28]. L'un des axes de recherche en vogue consiste à mettre au point des systèmes indépendants de la langue, ou tout au moins multilingues. Parmi les pionniers, [Cucerzan 99] ont testé sur cinq langues [29] un algorithme semi-supervisé dont les F-scores se révélaient encourageants, quoique très variables. Parmi les tentatives plus récentes, citons [Al-Rfou 15], dont l'annotateur multilingue basé sur Wikipedia et Freebase fonctionne dans 40 langues, sans recours à des corpus annotés ou à des règles écrites à la main.

Domaines La NERC s'est longtemps focalisée sur des textes journalistiques, puis sur la littérature biomédicale [30]. Voilà encore dix ans, [Nadeau 07a, p. 12] notait que l'impact du genre (journalistique, scientifique, informel...) et du domaine (jardinage, sport, économie...) était un champ d'études plutôt négligé dans la littérature sur les EN. Tout au plus pouvait-il citer les travaux de [Maynard 01], qui portaient sur des emails et des textes scientifiques ou religieux, ainsi que ceux de [Minkov 05], également sur des emails. Dans les deux cas, notait Nadeau, ces expériences démontraient que si n'importe quel domaine pouvait raisonnablement être couvert, appliquer un système à un domaine ou à un genre différent demeurait un défi sérieux. Avec l'essor de la communication mobile et des réseaux sociaux, de nouveaux domaines de recherche sont apparus autour de la reconnaissance d'EN dans des SMS [Ek 11] ou des tweets [Liu 13]. Ces deux types de textes ont la particularité d'être rédigés dans une langue très informelle [31] et d'être entourés d'un contexte très étriqué.

Types Nous l'avons vu à la section 1.4, les sept ou huit types d'entités classiques se sont révélés rapidement insuffisants pour couvrir certains domaines du TAL, notamment celui des systèmes de questions-réponses. Or certaines entités peuvent présenter des difficultés spécifiques. [Guerraz 15], par exemple, montre que les EN de type « film » sont plus délicate à segmenter et à classer que des noms de

28. En l'occurrence le chinois, l'anglais, le français, le japonais, le portugais et l'espagnol.
29. Roumain, anglais, grec, turc et hindi.
30. En partie grâce à la diffusion du corpus Genia [Kim 03], dont la version 3.0 est constituée de 2000 résumés annotés extraits de la base de données bibliographique MEDLINE (*Medical Literature Analysis and Retrieval System Online*).
31. A tel point que certaines méthodes de normalisation consistent à les traduire auparavant en anglais standard, comme s'il s'agissait d'une langue étrangère [Kaufmann 10].

lieux ou de personnes. Il en va de même pour tous les types d'entités complexe, comme les titres de livres [Downey 07]. Outre la multiplication des types d'entités (pour rappel, jusqu'à 200 dans la typologie de Sekine), on assiste également à leur hiérarchisation de plus en plus fine. Comme le souligne [Desmet 13], « *such subtype classification would be especially valuable for applications that involve question answering, information retrieval or the automatic construction of ontologies* ». A noter que ces derniers auteurs se sont également penchés sur la résolution de métonymies, par exemple lorsqu'un nom de pays est utilisé pour désigner une équipe sportive. Ce champ de recherche, qui présente un intérêt fondamental pour de nombreuses autres tâches du TAL [Markert 02], constitue lui aussi l'un des défis ouverts de la NERC.

1.9. La désambiguïsation sémantique

Revenons sur la phrase `John Kennedy a été assassiné à Dallas le 22 novembre 1963`, examinée au point 1.6. Nous avons vu que le système de reconnaissance a annoté le segment « John Kennedy » comme un nom de personne. Pour y parvenir, il lui a fallu au préalable effectuer une désambiguïsation minimale du mot « Kennedy », puisque ce nom propre peut désigner aussi bien un patronyme, un toponyme [32] qu'un nom d'entreprise.

Comment résoudre ces ambiguïtés ? Un système à base de règles, par exemple, pourrait s'appuyer sur le contexte gauche du mot. A l'aide de dictionnaires de prénoms et de toponymes, il lui serait possible d'écarter l'idée qu'un mot inconnu précédé de « John » soit un nom de lieu. Le contexte droit, à savoir le groupe verbal « a été assassiné », permettrait quant à lui de trancher entre un nom de personne et un nom d'entreprise [33]. Une fois la même désambiguïsation appliquée au mot « Dallas », et le segment « 22 novembre 1963 » reconnu comme une date en trois parties, l'opération

32. C'est d'ailleurs le nom d'un district de la ville de Bogota, en Colombie : https ://fr.wikipedia.org/wiki/Kennedy_(Bogota)

33. Dans les faits, comme le souligne [Friburger 02, p. 80], l'expérience montre que beaucoup de verbes a priori réservés aux humains peuvent être employés métonymiquement avec un nom d'organisation. Exemple : « Volkswagen reconnaît que huit millions de véhicules sont concernés par le scandale en Europe » (france24.com, 06/10/2015, http ://www.france24.com/fr/20151006-volkwagen-scandale-nombre-voiture-europe-logiciel-fraude-anti-polluants-automobile-allemagn., consulté le 18 juin 2016).

de NERC est formellement terminée.

Cependant, rien de tout cela ne saurait répondre à la simple question : de quel John Kennedy parle-t-on ? Ou pour le dire de manière plus formelle, vers quel élément du réel, quel *référent* extralinguistique, pointe le signe « John Kennedy » ?

Car identifier cet élément réclame des connaissances extérieures à la phrase. Si la réponse nous paraît ici évidente, c'est parce que l'assassinat du 35e président des Etats-Unis fait partie de cette somme de connaissances contextuelles nommée « culture générale ». A la rigueur, « assassiné » et « Dallas » nous suffiraient pour produire l'image mentale appropriée, et sans doute l'un des deux termes serait-il interdit dans l'un de ces jeux de société qui visent à faire deviner un concept à l'aide de quelques mots-clés. Mais qu'en serait-il pour la phrase « John Doe a été assassiné à Acapulco le 23 mai 2004 », dont la structure est pourtant rigoureusement identique ?

1.9.1. Une ambiguïté à deux niveaux

L'opération qui consiste à relier une entité à un référent extralinguistique peut s'appeler « désambiguïsation d'entité » (de l'anglais *Entity desambiguation*) ou encore « résolution » d'entité [Stern 13, p. 73]. Nous privilégierons ce dernier terme, qui limite les confusions avec la désambiguïsation lexicale (*word sense desambiguation*) [34].

Contrairement à la phase de NERC, la résolution d'entité ne saurait se satisfaire d'un simple dictionnaire de noms propres, puisqu'il existe rarement une relation biunivoque entre la mention d'une entité et l'entrée d'un lexique [35]. Comme le rappelle [Blume 05], « *Entity disambiguation inherently involves resolving many-to-many relationships* ». D'un côté, l'entité réelle connue comme « 35e président des Etats-Unis » peut être désignée de multiples manières : « John Kennedy », mais aussi « John Fitzgerald Kennedy », « JFK », « M. Kennedy », « Jack Kennedy », etc. De l'autre, chacune de ces étiquettes peut s'appliquer à de multiples entités. Si l'on en croit le site howmanyofme.com, les Etats-Unis à eux seuls compteraient à ce jour quelque 3400 « John

34. Laquelle consiste à déterminer le sens d'un mot dans une phrase quand il peut en avoir plusieurs, comme dans le cas des homographes catégoriels. Par exemple, « nuit » dans « fumer nuit » est-il un nom ou un verbe ?

35. De même que cette relation n'est pas garantie entre un élément d'une base de données et le réel empirique correspondant. Voir [Boydens 11, pp. 19-21].

Kennedy »[36]. Même en se limitant aux personnes qui disposent de leur propre page Wikipedia, une telle dénomination pourrait correspondre à une soixantaine d'individus plus ou moins célèbres[37]. Cette profusion d'homonymie n'est évidemment pas exceptionnelle. Aux Etats-Unis toujours, 100 millions de personnes peuvent se partager 90 000 noms différents, selon des statistiques de l'US Census Bureau citées par [Artiles 07].

1.9.2. Reconnaissance vs résolution d'entités

Bien qu'il s'agisse en théorie de deux tâches distinctes, la NERC et la résolution d'entité entretiennent des liens évidents. On peut considérer que la seconde va au-delà de la première [Dai 12] et l'englobe. Comme nous le verrons au chapitre suivant, déterminer qu'une entité telle que « John Kennedy » correspond au sujet de la page Wikipedia https ://fr.wikipedia.org/wiki/John_Fitzgerald_Kennedy permet de lui attribuer par la suite une série de caractéristiques, parmi lesquelles celle de « personne ». Dans la pratique, la NERC constitue le plus souvent l'étape préalable à sa résolution. Une fois un segment de texte identifié comme entité, puis sa catégorie définie, il est plus simple de déterminer de quelle entité en particulier il s'agit (voir par exemple [Stern 12]). Mais le raisonnement inverse est également possible. Dans le sillage de [Guo 13] et [Yamada 15a], [Yamada 15b] décrit ainsi un système dans lequel la résolution d'entités s'effectue avant leur reconnaissance, dans le but d'améliorer cette dernière. Il s'agit toutefois d'un contexte précis (l'extraction d'entités dans des tweets) pour lequel les méthodes classiques de NERC montrent leurs limites.

Nous n'aborderons pas ici les méthodes de résolution d'entités, qui feront l'objet de la Section 2.4.

36. http ://howmanyofme.com/people/John_Kennedy/, consulté le 18 juin 2016.
37. https ://en.wikipedia.org/wiki/John_Kennedy_(disambiguation)

Chapitre 2

Web de données et annotation sémantique

2.1. Du web... 30
2.2. ...au web sémantique... 31
2.3. ...puis au Linked Open Data 39
2.4. Résolution d'entités 42
2.5. Limites de Wikipedia 49

Nous avons vu au chapitre précédent les limites d'une simple reconnaissance et classification des entités nommées. Si attribuer à un segment de texte des labels tels que PER, ORG ou LOC constitue une fonctionnalité déterminante en TAL, cet étiquetage basique ne permet guère d'améliorer notre gestion de la connaissance informatisée. Quel intérêt documentaire peut avoir la phrase `<LOC>Paris</> se situe au cœur d'un vaste bassin sédimentaire` si Paris peut désigner aussi bien la capitale française que 22 villes américaines homonymes [1] ? Là où un être humain pourrait éventuellement s'aider du contexte, un programme informatique ne dispose d'au-

1. https ://fr.wikipedia.org/wiki/Liste_des_villes_s%27appelant_Paris

cun moyen pour résoudre l'ambiguïté. Que l'on songe à la quantité d'information non structurée qui échappe ainsi à l'analyse. A l'échelle d'une entreprise, un ordre de grandeur souvent cité veut que 80 % des informations qui circulent soient non structurées (contre 20 % seulement dans des bases de données) [Dupoirier 09]. Mais à l'échelle du web, ce répertoire ouvert du savoir humain, ce sont des milliards de pages qui resteront inaccessibles aux machines. A moins, peut-être, que chaque élément significatif de leur texte puisse être enrichi de métadonnées qui précisent son sens, ainsi que les relations qu'il entretient avec les autres segments de texte, le tout d'une manière interprétable par une machine... En apparence farfelue, cette idée constitue pourtant le rêve d'ingénieurs des plus sérieux, ceux-là même qui ont défini le web. Et si ce rêve est aujourd'hui loin d'être devenu une réalité, les travaux accomplis pour tenter de le concrétiser ont abouti à des réalisations concrètes, dont quelques-unes seront utilisées dans ce travail. Afin de mieux les comprendre, nous nous proposons dans ce chapitre de les replacer dans leur contexte, dont les racines remontent à la création même du web. Une fois présentées les notions de « web sémantique » et de « Linked Open Data », nous verrons de quelle manière elles peuvent nous permettre de répondre à la question laissée en suspens au chapitre 1 : comment identifier l'objet précis du réel auquel fait référence une entité nommée ? [2]

2.1. Du web...

Le web a connu d'indéniables améliorations techniques depuis son invention par Tim Berners-Lee et Robert Cailliau au tournant des années 90 [Berners-Lee 89, Berners-Lee 90]. En 1995, l'apparition des feuilles de style en cascade (CSS) [3] a permis d'enfin découpler le contenu d'une page de sa présentation. Lancés la même année, les langages JavaScript (côté navigateur) ou PHP (côté serveur) sont venus apporter du dynamisme à des pages auparavant statiques. En 1998, la première recommandation du XML (*Extensible Markup Language*) [4], plus simple et mieux adaptée au web que

2. La structure de ce chapitre s'inspire de [De Wilde 15, p. 51-87], qui nous a été d'une précieuse aide méthodologique et nous a permis de clarifier nos idées.
3. Devenues une recommandation du W3C en 1996 : https ://www.w3.org/TR/REC-CSS1/
4. https ://www.w3.org/TR/1998/REC-xml-19980210

son ancêtre SGML (*Standard Generalized Markup Language*) [5], fournit à la Toile un format qui permet l'échange de données structurées. Ces innovations conduiront près d'une décennie plus tard [6] aux technologies Ajax (*Asynchronous JavaScript and XML*), permettant entre autres de mettre à jour une page sans la recharger. C'est également au milieu des années 2000 que, sous le nom de « web 2.0 » [O'Reilly 05], s'amorça le virage vers plus de simplicité et d'interactivité, qui conduira à l'hégémonie des réseaux sociaux. Sans oublier bien sûr les différentes moutures du langage HTML, qui en est à sa 5e version [7].

Toutefois, un internaute plongé en 1996 dans un sommeil de vingt ans ne serait pas trop dépaysé à son réveil. En dépit de ces innovations techniques, le web de 2016 reste fondamentalement une collection de *documents*, de pages web localisées par un URL (*Uniform Resource Locator*) [8] et reliées entre elles par des liens hypertextes. Cette structure, qui nous paraît désormais aller de soi, est parfaitement adaptée à une consultation humaine. En revanche, elle ne se prête guère à un traitement par des machines. Pour reprendre un exemple de [van Hooland 14, p. 32], un moteur de recherche retrouverait facilement une page consacrée à Pablo Picasso. Une requête du type « œuvres peintes par Picasso » s'avérerait déjà plus compliquée, car elle nécessiterait de tester plusieurs formulations possibles. Mais obtenir la liste des « peintures réalisées par des artistes qui ont rencontré Picasso » serait tout bonnement impossible : le moteur de recherche peut isoler des mots, mais il est incapable de démêler leur sens, leur sémantique.

2.2. ...au web sémantique...

L'idée d'un web compréhensible par des machines n'a pourtant rien d'inédit. On peut estimer qu'elle remonte à l'origine même du projet. En 1994, année de la fon-

5. http ://www.iso.org/iso/fr/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=16387

6. Le terme Ajax est apparu en 2005 dans un article de Jesse James Garrett : http ://adaptivepath.org/ideas/ajax-new-approach-web-applications/ Les XMLhttpRequests, éléments essentiels de cette technologie, sont devenues une recommandation du W3C en 2006 : https ://www.w3.org/TR/2006/WD-XMLHttpRequest-20060405/.

7. https ://www.w3.org/TR/html5/

8. https ://url.spec.whatwg.org/

dation du World Wide Web Consortium (W3C) [9], Tim Berners-Lee rappela lors d'une conférence que « réduire le web à un espace documentaire avec des liens entre les documents, c'est ne prendre en considération qu'un seul plan de la problématique » [Gandon 12]. On retrouve cette idée d'incomplétude en 1999 dans son livre *Weaving The Web* :

> « I have a dream for the Web... and it has two parts. In the first part, the Web becomes a much more powerful means for collaboration between people. (...) In the second part of the dream, collaborations extend to computers. Machines become capable of analyzing all the data on the Web—the content, links, and transactions between people and computers. A "Semantic Web," which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy, and our daily lives will be handled by machines talking to machines, leaving humans to provide the inspiration and intuition. » [Berners-Lee 99, p. 157]

Berners-Lee développa sa pensée deux ans plus tard dans un article phare du *Scientific American*. Il précise alors que, pour qu'un tel web fonctionne, les ordinateurs devraient avoir accès à des collections structurées d'informations et à des ensembles de règles d'inférence qu'ils peuvent utiliser pour effectuer des raisonnements automatisés [Berners-Lee 01]. En ce qui concerne la structure, précisait-il, deux importantes technologies étaient déjà en place : le langage XML et le RDF (*Ressource Description Framework*), dont la première recommandation était parue en 1999 [10]. Restait à produire les règles d'inférence, à savoir des ontologies (voir 2.2.4) reliées entre elles.

Cet ensemble de technologies, enrichies par la suite de nouvelles normes, forme ce que l'on appelle la pile de standards du web sémantique (*Semantic Layer Cake* ou *Semantic Web Stack*), représentée à la Figure 2.2.1.

Décrire l'ensemble des concepts évoqués dans ce graphique dépasserait le cadre de ce travail. Toutefois, il est indispensable pour la suite de l'exposé de présenter brièvement certaines notions essentielles.

9. L'organisme international chargé d'établir les standards du web, appelés « recommandations » : https ://www.w3.org/
10. https ://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

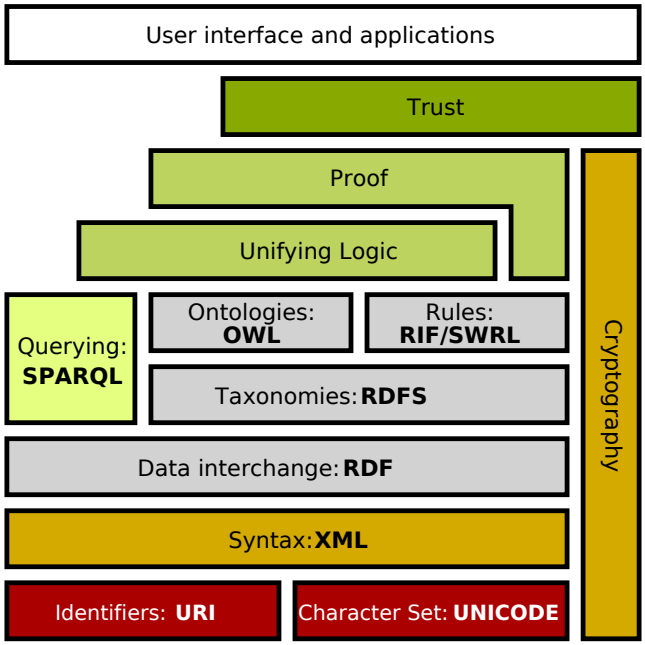


FIGURE 2.2.1. – La pile de standards du web sémantique, reprise de Wikipedia (user :Marobi1)

2.2.1. Les URIs

Un URI (*Uniform Resource Identifier*) [11] est défini comme « une séquence compacte (N.D.A. : sans espace) de caractères qui identifie une ressource abstraite ou physique » [12]. Par ressource abstraite ou physique, il faut entendre tout concept du monde réel, aussi bien « la beauté » qu'un personnage historique (« Napoléon ») ou encore un document présent sur le web. Dans ce dernier cas, on parlera plutôt d'URL (*Uniform Resource Locator*), laquelle n'est qu'un sous-ensemble particulier de la notion d'URI permettant à la fois d'identifier et de localiser une ressource. Il est important de préciser qu'un URI n'est pas forcément stocké sur le web. Le cas échéant, on précisera donc qu'il s'agit d'un URI HTTP (*Hypertext Transfer Protocol*).

Distinguer un URI d'un URL n'est pas toujours évident. Supposons qu'une entreprise compte une employée nommée Alice, dont la page web est localisée à l'adresse

11. On parle aussi d'IRI (*Internationalized Resource Identifier*) pour désigner un URI autorisant des caractères Unicode (http ://unicode-table.com/fr/#control-character), là où un URI classique est limité aux caractères ASCII. Voir https ://tools.ietf.org/html/rfc3987.
12. https ://tools.ietf.org/html/rfc3986

www.example.com/people/alice Il s'agit bien là d'un URL, et donc par extension d'un URI ; mais est-ce l'URI de l'être humain nommé Alice ? Certainement pas, puisque cela reviendrait à confondre une personne et une page web qui lui est consacrée [13].

Afin de lever ce genre d'ambiguïté, le W3C préconise deux méthodes. D'une part les *Hash URIs*, dans lesquels le nom de la ressource est précédé d'un octothorpe (../about#alice). Cette méthode permet de distinguer une ressource d'un document web consacré à cette ressource. Elle a toutefois le désavantage, notamment, de produire des URIs difficiles à stocker. L'autre solution consiste à utiliser la réponse HTTP 303 SEE OTHER, qui permet de rediriger une requête. De cette manière, l'URL ../id/alice peut renvoyer à la fois vers une page web HTML qui décrit une ressource et vers l'URI qui l'identifie.

Dès lors que les URIs sont souvent longs et fastidieux à écrire, une pratique courante consiste à les rédiger sous une forme abrégée nommée "CURIEs" (*compact URIs*). Dans celles-ci, l'espace de noms est synthétisé par une abréviation suivie de deux points (:). C'est ainsi que les URIs http ://dbpedia.org/resource/Rudy_Demotte et http ://dbpedia.org/page/Elio_Di_Rupo, par exemple, seront le plus souvent retranscrits sous une forme du type `dbr:Rudy_Demotte` et `dbr:Elio_Di_Rupo`, en précisant préalablement que "dbr" signifie http ://dbpedia.org/resource/. Pour plus de détails sur la syntaxe des CURIEs, voir [Segaran 09, p. 76].

2.2.2. Le modèle RDF

Le *Resource Description Framework*, déjà brièvement évoqué, permet de représenter une information complexe de manière simple, à savoir un ensemble de triplets sujet-prédicat-objet reliés entre eux. « Les Aventures de Tom Sawyer (sujet) ont été écrites (prédicat) par Mark Twain (objet) » ; « Mark Twain (cette fois sujet) est (prédicat) un écrivain (objet) », etc.

Cette manière de rédiger des triplets est parfaitement compréhensible par un humain, qui sait ce qu'est l'action d'écrire et qui devinera, s'il ignorait, que Tom Sawyer est le protagoniste d'une œuvre littéraire. Il en va tout autrement pour une machine, incapable de saisir pareils concepts et d'effectuer de tels raisonnements. C'est pourquoi la représentation du RDF passe par des URIs : en associant chaque partie des

13. Exemple repris de [Sauermann 08].

triplets à une ressource qui définit le concept évoqué, les deux mêmes assertions deviendraient cette fois :

```
<http://dbpedia.org/resource/The_Adventures_of_Tom_Sawyer>
        <http://dbpedia.org/ontology/author>
            <http://dbpedia.org/resource/Mark_Twain>

<http://dbpedia.org/resource/Mark_Twain>
        <https://www.w3.org/1999/02/22-rdf-syntax-ns#type>
            <http://dbpedia.org/ontology/Writer>
```

Contrairement au XML, qui possède son propre format de fichier, le RDF est un modèle abstrait. Sa représentation conceptuelle passe par un graphe orienté et étiqueté (figure 2.2.2) et sa représentation informatique par plusieurs syntaxes de sérialisation. Le W3C recommande le format RDF/XML [14], ce qui explique la présence de XML dans la pile des standards du web sémantique. Dans la pratique, cette syntaxe rigide et verbeuse est souvent délaissée au profit des notations N-Triples [15] (celle que nous avons utilisée dans l'exemple ci-dessus), N3 [16] ou Turtle [17].

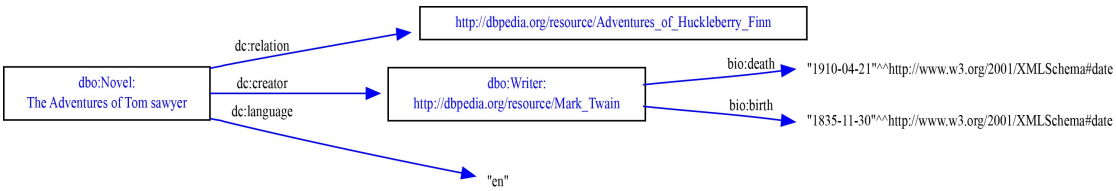


FIGURE 2.2.2. – Représentation visuelle d'un graphe RDF

2.2.3. SPARQL

Le modèle RDF dispose de son propre langage de requêtes baptisé SPARQL [18], acronyme récursif de *SPARQL Protocol and RDF Query Language*. Sa syntaxe inspirée du SQL tient compte des spécificités du RDF, où il s'agit d'interroger un graphe et non des tables reliées entre elles. Rédiger une requête SPARQL revient ainsi à créer

14. https://www.w3.org/TR/rdf-syntax-grammar/
15. https://www.w3.org/TR/n-triples/
16. https://www.w3.org/TeamSubmission/n3/
17. https://www.w3.org/TR/turtle/
18. https://www.w3.org/TR/sparql11-overview/

un graphe RDF, dans une syntaxe proche de Turtle, graphe dans lequel les éléments recherchés sont remplacés par des noms de variables arbitraires précédés d'un point d'interrogation. Cette requête doit être adressée à un « SPARQL Endpoint », une base de données de triplets RDF (*triple store*) accessible sur le web.

Dans l'exemple 2.1, inspiré de [DuCharme 13, p. 222], la requête tentera de retrouver les films où les acteurs Al Pacino et Robert De Niro ont joué ensemble, ainsi que, s'il est connu, le nom de la ville natale de leurs réalisateurs. A l'intérieur de la condition `where` se trouvent deux triplets RDF dont l'un des éléments (le sujet) est remplacé par la variable `?pacinoFilm`. La requête parcourra le graphe afin de retrouver un motif qui répond aux deux conditions : une ressource reliée par la propriété `dbo:starring` (définie dans http ://dbpedia.org/ontology/, donc) aux ressources `dbr:Al_Pacino` et `dbr:Robert_De_Niro` (Figure 2.2.3). Elle cherchera ensuite le réalisateur de ces films, l'URI de leur ville d'origine et, enfin, le nom « humain » de celles-ci. Si la ville d'origine du réalisateur n'est pas indiquée, la clause `optional` autorise à renvoyer tout de même le nom des réalisateurs en question. Si le nom de la ville existe en plusieurs langues, la clause `filter` impose de ne garder que la version française. Les réponses renvoyées seront dans ce cas-ci l'URI du film et éventuellement le nom de la ville d'origine du réalisateur, puisque la clause `select` n'exige que celles-là. Cet exemple peut être testé sur le SPARQL Endpoint suivant : http ://dbpedia.org/sparql.

Listing 2.1 – Exemple de requete SPARQL

```
PREFIX dbr: <http://dbpedia.org/resource/>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbp: <http://dbpedia.org/property/>
PREFIX dbr: <http://dbpedia.org/resource/>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbp: <http://dbpedia.org/property/>

SELECT ?pacinoFilm, ?directorCityName
WHERE
{ ?pacinoFilm dbo:starring dbr:Al_Pacino .
  ?pacinoFilm dbo:starring dbr:Robert_De_Niro .
  ?pacinoFilm dbp:director ?director .
  OPTIONAL {?director dbo:birthPlace ?directorCity .}
  OPTIONAL {?directorCity rdfs:label ?directorCityName}
  FILTER (LANG(?directorCityName ) = "fr" )}
```

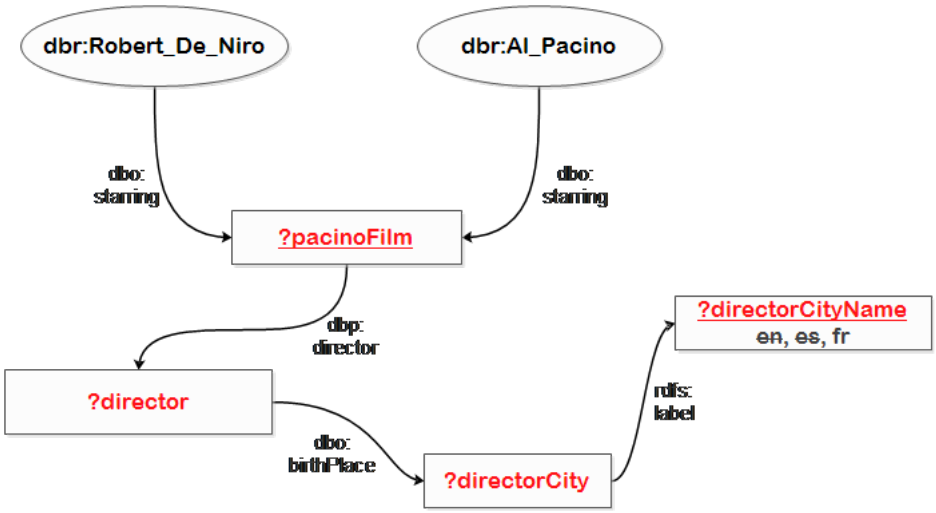


FIGURE 2.2.3. – Représentation du graphe requête

2.2.4. Les langages d'ontologies RDFS et OWL

L'ontologie désigne traditionnellement la « partie de la philosophie qui a pour objet l'étude des propriétés les plus générales de l'être, telles que l'existence, la possibilité, la durée, le devenir »[19]. A partir des années 90, le terme fut emprunté par l'informatique, notamment dans le domaine de l'IA, pour décrire un système de gestion de la connaissance que [Gruber 95] définit comme « une spécification explicite d'une conceptualisation »[20]. De manière plus terre-à-terre, on pourrait dire comme [Maedche 12, p. 11] que :

> « *an ontology refers to an engineering artifact, constituted by a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary.* »

Un exemple concret aidera à mieux cerner ce concept. Nous avons vu que le modèle RDF permet de formaliser des propositions du type « Mark Twain (sujet) a écrit (prédicat) *Les Aventures de Tom Sawyer* (objet)». En lisant cette courte phrase, un être

19. *Trésor de la langue française informatisé*, http ://atilf.atilf.fr/
20. « *An ontology is an explicit specification of a conceptualization.* »

humain qui ne connaît ni Mark Twain ni Tom Sawyer sera pourtant capable d'en déduire quantité d'informations. A commencer par le fait que Mark Twain désigne sûrement un être humain, et non un animal, car seuls les hommes savent écrire. Que ses yeux sont donc probablement noirs, bruns, bleus, gris ou verts, mais pas orange vif. Qu'il est probablement de sexe masculin, comme l'immense majorité des porteurs du prénom Mark. Que *Les Aventures de Tom Sawyer* est le titre d'une production littéraire, et non d'une peinture, etc.

Décrire le réel afin qu'une machine soit capable à son tour d'effectuer ce type d'inférence, voilà le principal objet des ontologies informatiques. L'enjeu est de taille, puisque l'inférence produit automatiquement un savoir nouveau (i.e. : de nouveaux triplets) par déductions successives à partir des éléments connus.

Les standards du web sémantique offrent deux outils pour confectionner de telles ontologies. Le premier, plus basique et plus ancien, est le RDF Schema (RDFS), créé en 1998 et devenu une recommandation en 2004 [21]. A l'aide de la syntaxe RDF, RDFS permet de spécifier des « ontologies légères » [Gandon 12] limitées à quinze primitives de type « est une sous-classe (ou une sous-propriété) de », « ne s'applique qu'à », « voir aussi »...

[van Hooland 16] nous fournit un exemple tout fait pour la propriété « a écrit », que l'on pourrait formuler en RDFS de la manière suivante :

Listing 2.2 – Exemple d'utilisation de RDFS

```
@prefix ex: <http://example.org/ontology#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.

ex:hasWritten rdf:type rdf:Property;
              rdfs:domain ex:Person;
              rdfs:range ex:LiteraryWork;
              rdfs:subPropertyOf ex:hasCreated.
```

On voit que quatre triplets RDF suffisent pour préciser que `ex:hasWritten` est une propriété, qu'elle ne s'applique qu'à des ressources de la classe `ex:Person`, qu'elle ne concerne que des œuvres littéraires et, enfin, qu'il s'agit d'une sous-propriété de `ex:hasCreated` : cette dernière l'englobe, ou plus précisément la *subsume*, ce qui

21. https://www.w3.org/TR/2004/REC-rdf-schema-20040210/

permet à la propriété `ex:hasWritten` d'*hériter* des caractéristiques de `ex:hasCreated`, tout en lui apportant des nuances.

Une fois cette propriété spécifiée en RDFS, un algorithme d'inférence (« raisonneur ») confronté au triplet `:Mark_Twain ex:hasWritten :The_Adventures_of_Tom_Sawyer` sera capable d'effectuer une bonne partie des déductions de notre lecteur humain, et donc de créer seuls de nouveaux triplets – par exemple, `:Mark_Twain a ex:Person`[22].

Pourrait-il en déduire lui aussi que les yeux de Mark Twain ne sont certainement pas jaune fluorescent? Oui, à condition que l'éventail de couleurs possibles des yeux humains soit stipulé dans la classe `ex:Person`. Or définir une classe et des propriétés dépasse les possibilités de RDFS, qui se borne à établir des relations hiérarchiques. Ce type d'ingénierie nécessiterait d'utiliser le langage d'ontologie OWL (*Ontology Web Langage*), bien plus puissant et expressif. C'est notamment lui qui, au travers de la propriété `owl:sameAs`, permet d'exprimer la notion d'*équivalence* entre concepts :

Listing 2.3 – Exemple d'équivalence entre classes exprimée en OWL

```
<owl:Class rdf:ID="FormationPolitique">
  <owl:sameAs rdf:resource="http://dbpedia.org/ontology/PoliticalParty"/>
</owl:Class>
```

Nous reviendrons dans la section suivante sur cet élément essentiel de l'architecture du web sémantique. Pour le reste, la description de ces schémas lourds sortirait du cadre de ce travail. Nous invitions le lecteur intéressé à consulter les recommandations OWL et OWL 2 du W3C [Dean 04, Hitzler 12].

2.3. ...puis au Linked Open Data

Nous avons vu au point 2.2 que le web sémantique décrit par Tim Berners-Lee en 2001 impliquait de créer de vastes ontologies, à charge pour les internautes de créer des ponts entre elles. Mais une ontologie universelle gérée par ses propres utilisateurs est-elle concevable? N'était-ce pas ressusciter l'utopie des grandes classifications universelles à la Dewey et Otlet que les bibliothécaires, comme le rappelle

22. Le prédicat `a` est une reformulation communément admise de la propriété `rdf:type`.

[Bates 02], avaient abandonnée depuis le début du XXe siècle ? De nombreuses critiques plus ou moins pertinentes vinrent doucher cette vision peut-être trop logico-mathématique de la complexité du réel. Citons par exemple celles du journaliste américain Clay Shirky :

> « *The list of factors making ontology a bad fit is, also, an almost perfect description of the Web - largest corpus, most naive users, no global authority, and so on. The more you push in the direction of scale, spread, fluidity, flexibility, the harder it becomes to handle the expense of starting a cataloguing system and the hassle of maintaining it, to say nothing of the amount of force you have to get to exert over users to get them to drop their own world view in favor of yours.* » [Shirky 05]

De fait, cet idéal d'un web uniformément « sémantisé », où des machines dialogueraient avec d'autres machines, peina à se concrétiser. C'est pourquoi, en 2006, Berners-Lee reformula ses ambitions de manière plus pragmatique, en promouvant la notion de *Linked Data* [Berners-Lee 06]. Contrairement au Web documentaire, où des pages HTML sont connectées entre elles par des liens hypertextes, le « web de données » relie des objets du monde, décrits au format RDF à l'aide de liens typés. En conséquence, « *the result, which we will refer to as the Web of Data, may more accurately be described as a web of things in the world, described by data on the Web* ». [Bizer 09].

Concrètement, le *Linked Data* décrit par Berners-Lee consiste en une série de bonnes pratiques destinées à favoriser la publication des données sur le Web. Ces bonnes pratiques se fondent sur quatre principes :

1. Utiliser des URIs pour nommer les choses ;
2. Utiliser des URIs HTTP afin que les gens puissent consulter ces choses ;
3. Quand quelqu'un consulte un URI, lui fournir des informations utiles en utilisant les standards (RDF, SPARQL) ;
4. Inclure des liens vers d'autres URIs de sorte qu'il puisse découvrir plus de choses.

Cette notion de données liées s'enrichit très vite par celle de données liées *sous licence ouverte* (*Linked Open Data*), dont la qualité peut être estimée sur une échelle allant d'une à cinq étoiles (Figure 2.3.1).

FIGURE 2.3.1. – Les cinq étoiles du web de données liées ouvertes, selon le W3C

En janvier 2007, le W3C Semantic Web Education and Outreach Group lança le *Linking Open Data Project*[23], dont l'objectif était de développer un web de données en identifiant les sets de données libres existants, les convertir au format RDF et les publier sur le web [Bizer 09].

En août 2014, le *Mannheim Linked Data Catalog*[24] identifiait 1 091 jeux de données liées qui forment le *Linked Open data Cloud*[25]. L'extrait publié à la Figure 2.3.2 montre la place centrale qu'occupent les grandes bases de connaissance telles que DBpedia et ses déclinaisons linguistiques, que nous présenterons plus en détail au point 2.4.2, mais aussi l'importance relative de données ouvertes issues d'entreprises privées ou semi-privées, comme le grand quotidien *New York Times*[26] ou la radio-télévision publique britannique BBC[27].

23. http ://esw.w3.org/topic/SweoIG/TaskForces/ CommunityProjects/LinkingOpenData
24. http ://linkeddatacatalog.dws.informatik.uni-mannheim.de/about
25. http ://lod-cloud.net/
26. http ://data.nytimes.com/, inaccessible en juillet 2016.
27. https ://datahub.io/fr/dataset?q=bbc

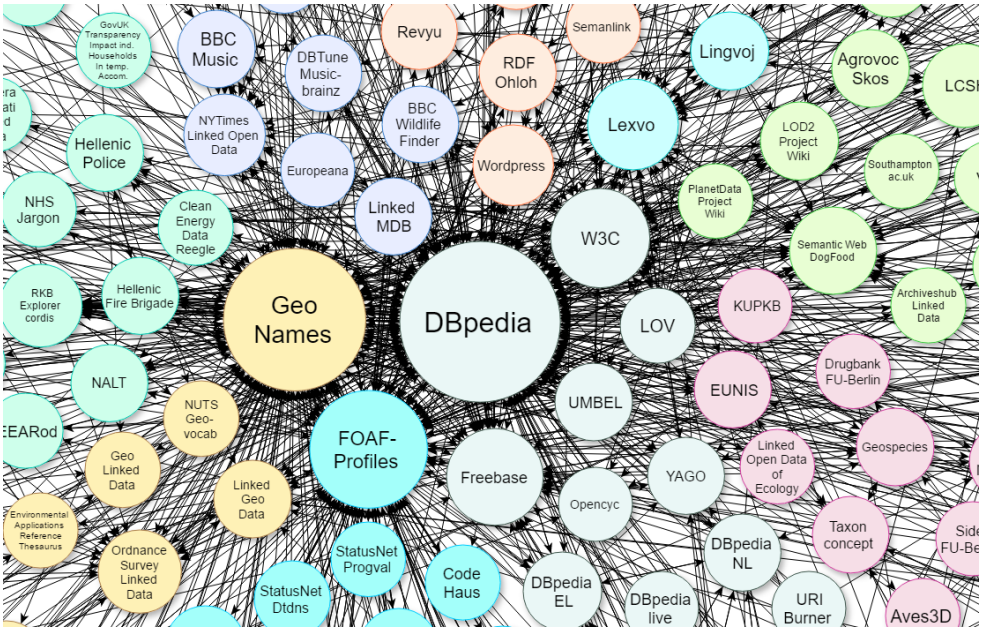


FIGURE 2.3.2. – *Linked Open Data Cloud* (détail), extrait de [Cyganiak 14]

2.4. Résolution d'entités

2.4.1. « Entity Linking »

La notion de « liage d'entités » [Stern 13, p. 18], plus connue sous le terme anglais *entity linking* (EL), peut être définie comme l'opération qui consiste à trouver la correspondance entre la mention textuelle d'une entité et une entrée dans une base de connaissances, entrée considérée comme la forme canonique pour cette entité [28] [Rao 13, p. 96]. Selon les domaines de recherche et les auteurs, l'EL est parfois désigné sous le nom d'*objet identification*, *data de-duplication*, *entity resolution*, *entity disambiguation*, *name matching*, *term identification*, *normalization*, etc. [Dai 12]. Dans tous les cas, ce liage implique que l'entité en question soit auparavant désambiguïsée (ou « résolue », selon la terminologie employée au chapitre 1).

La Figure 2.4.1 représente le schéma général d'un système d'EL : à partir d'un texte en langage naturel, le système interroge Wikipedia afin d'identifier les éventuelles

28. « *We define entity linking as matching a textual entity mention, possibly identified by a named entity recognizer, to a KB entry, such as a Wikipedia page that is a canonical entry for that entity.* »

entités présentes dans l'un et dans l'autre.

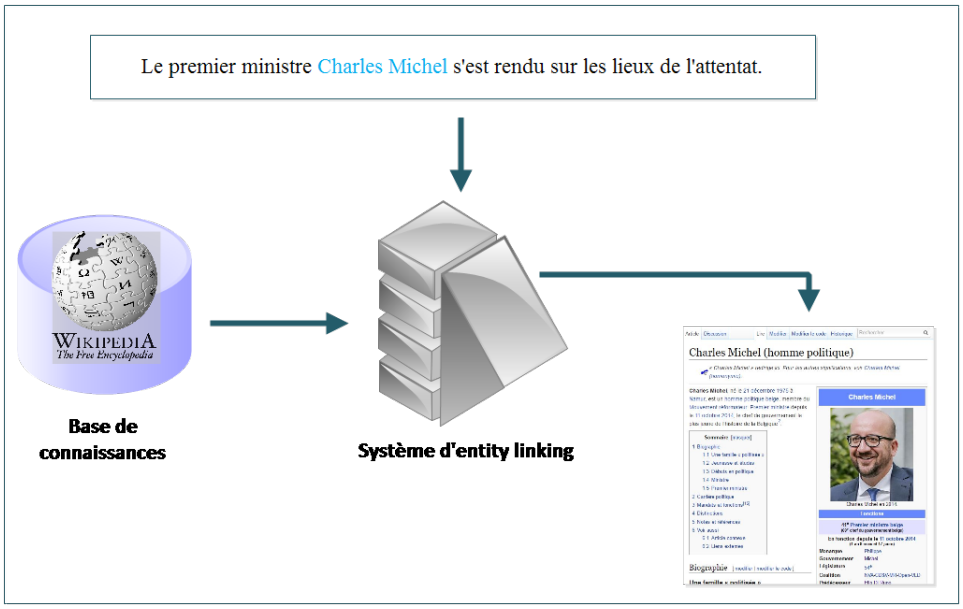


FIGURE 2.4.1. – Schéma général d'un système d'*Entity linking*

Cet exemple très simple montre déjà quelque-unes des difficultés auxquelles doit faire face un tel système. A commencer par l'empan de l'entité reconnue. Où commence-t-elle et où finit-elle ? S'il est évident pour un locuteur humain que « Charles Michel » constitue une unité de sens, un algorithme qui se contenterait de comparer chaque mot du texte au contenu de la base de connaissance pourrait renvoyer, à tort, la fiche Wikipedia de Charles [29] puis celle de Michel [30].

Une fois cet empan correctement identifié, le système doit en outre pouvoir désambiguïser l'entité. Dans le cas du nom « Charles Michel », la version francophone de Wikipedia recense au moins six personnalités qui portent exactement ce nom, dont quatre possèdent leur propre page [31]. Pour renvoyer la référence correcte, le système d'EL devra se montrer capable de tenir compte du contexte, par exemple de la présence dans la phrase du terme « Premier ministre ». Enfin, le système devrait idéalement pouvoir interpréter les différentes formes et graphies sous laquelle se présentent les entités, notamment les abréviations (« M. Michel »), les surnoms

29. https ://fr.wikipedia.org/wiki/Charles
30. https ://fr.wikipedia.org/wiki/Michel
31. https ://fr.wikipedia.org/wiki/Charles_Michel_(homonymie)

et sobriquets (comme « Big Loulou » dans le cas du père du Premier ministre belge actuel), voire les descriptions définies (« le bourgmestre empêché de Wavre »). Nous renvoyons à [Shen 15] pour un état de l'art à jour de la question.

Dans le point qui suit, nous présenterons successivement les principales bases de connaissance utilisées par les systèmes d'EL et quelques-uns de ces systèmes.

2.4.2. Bases de connaissance

Issu des recherches en intelligence artificielle (AI), le terme « base de connaissance » (*knowledge base* ou *knowledge base management system*) désigne un système de gestion de l'information qui contient à la fois des données et un système de référence permettant d'interpréter ces données, ou d'en tirer des inférences. Comme le résume [Brodie 86] :

> « *knowledge bases have to have an associated theory for interpreting the information they contain, whereas databases are expected to meet performance standards for response time, robustness, reliability, security, etc. (...) To summarize then, a knowledge base contains knowledge about something represented in a suitable notation, whereas a database stores (large amounts of) shared data.* »

Si l'on s'en tient à cette définition, Wikipedia ne peut à proprement parler être considérée comme une base de connaissance. Il s'agit plutôt d'une base de données de contenu semi-structuré. Bien qu'utilisé en *entity linking* [Hachey 13], son contenu est avant tout destiné à être lu (ou édité) par des humains, non par des machines. En revanche, la vaste encyclopédie collaborative[32] sert de socle à plusieurs bases de connaissance stricto sensu, dont les plus connues sont DBpedia, Freebase, Yago ou encore Wikidata. Précisons, selon les mots de [Hengchen 15], que « ces bases de connaissance sont majoritairement anglophones ou, lorsqu'elles sont multilingues, sont beaucoup plus fournies en anglais ».

32. Wikipedia propose plus de 30 millions d'articles dans plus de 280 langues. Début juillet 2016, la version anglaise se prévalait de 5 185 654 articles (https ://en.wikipedia.org/wiki/Wikipedia :Statistics), contre 1 769 691 pour la version francophone (https ://fr.wikipedia.org/wiki/Wikip%C3%A9dia :Statistiques).

DBpedia

Lancée en 2007[33], DBpedia est le fruit d'un travail collaboratif qui vise à extraire de Wikipedia des informations structurées afin de les mettre à disposition librement sur le Web, les relier à d'autres bases de connaissance et permettre de les interroger à l'aide de requêtes informatiques [Lehmann 14]. Selon ses statistiques les plus récentes,

> « the full DBpedia data set features 38 million labels and abstracts in 125 different languages, 25.2 million links to images and 29.8 million links to external web pages ; 80.9 million links to Wikipedia categories, and 41.2 million links to YAGO categories. DBpedia is connected with other Linked Datasets by around 50 million RDF links ». [34]

Cette masse de liens vers des ressources RDF extérieures, mais aussi les liens entrants, explique pourquoi DBpedia est souvent qualifiée de noyau du Web de données [Auer 07].

Concrètement, DBpedia utilise des scripts informatiques afin d'extraire le résumé des pages Wikipedia, les images ou encore les autres langues dans lesquelles la page est disponible. Mais sa source première reste le contenu des éventuelles « infoboxes », ces encadrés qui affichent les principales informations factuelles sur l'entité décrite dans une page.

Ces informations sont ensuite associées manuellement à une ontologie. Celle-ci contient à l'heure actuelle une hiérarchie de 685 classes décrites par 2 795 propriétés[35][36].

Pour une description plus détaillée de l'ensemble du processus, voir [Jentzsch 09].

YAGO

Proche de DBpedia, YAGO (*Yet Another Great Ontology*) est également basée sur Wikipedia, mais aussi sur l'ontologie lexicale anglophone WordNet[37], ainsi que sur

33. A l'initiative conjointe de chercheurs et développeurs de la Freie Universität Berlin, de l'Universität Leipzig et de l'entreprise OpenLink Software.
34. http ://wiki.dbpedia.org/about
35. http ://wiki.dbpedia.org/services-resources/ontology
36. http ://mappings.dbpedia.org/server/ontology/classes/
37. https ://wordnet.princeton.edu/

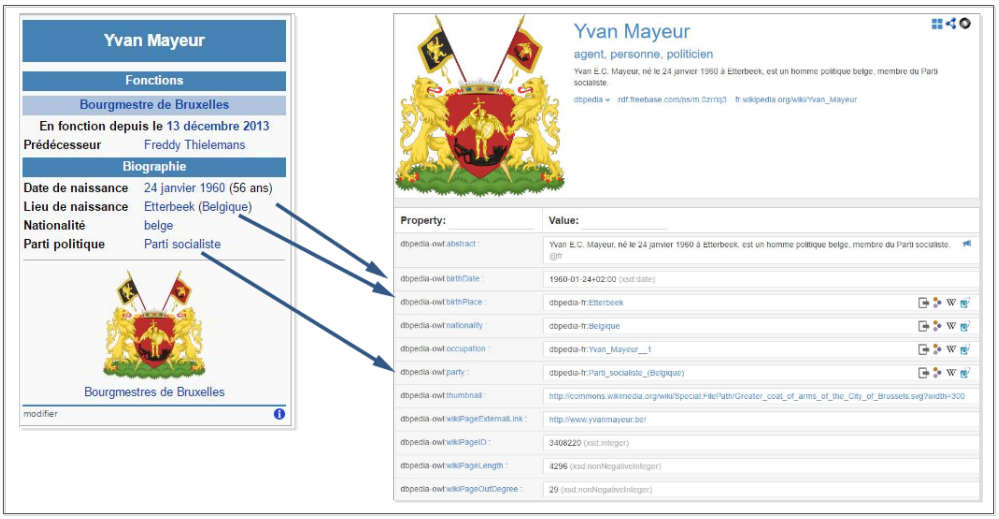


FIGURE 2.4.2. – Infoboxe de Wikipedia et la page correspondante de DBpedia version française

la base de données géographique GeoNames [38]. Issue de l'Institut Max-Planck d'informatique de Sarrebruck ([Kasneci 06]), cette base de connaissance contient plus de 120 millions de prédicats concernant plus de 10 millions d'entités (personnes, organisations, lieux...) [39]. A la différence de DBpedia, où chaque déclinaison locale dispose de son autonomie, les entités de YAGO [40] combinent dix versions linguistiques de Wikipedia, ce qui explique son plus grand nombre de prédicats [Färber 16]. Par ailleurs, ces entités ne sont pas mappées à une ontologie de manière manuelle, mais automatiquement reliées aux catégories de WordNet et de Wikipedia. C'est ainsi que YAGO parvient à compter plus de 350 000 classes. Pour un aperçu plus complet des différences entre YAGO et DBpedia, notamment en matière de qualité des données, voir [Hoffart 13].

Wikidata

Selon Wikipedia, combien la Ville de Bruxelles compte-t-elle d'habitants? Pour l'utilisateur francophone de l'encyclopédie, l'infoboxe de la page mentionne 168 230

38. http ://www.geonames.org/
39. http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/
40. Du moins depuis la version YAGO3 [Mahdisoltani 14].

habitants au 1er janvier 2015 [41]. Un utilisateur anglophone aura droit curieusement au chiffre plus à jour de 178 552 habitants (1er janvier 2016) [42], mais pas à leur répartition par sexe ou à leur pyramide des âges. Un locuteur espagnol, lui, devra se contenter du chiffre de 2013, sans autre précision, à savoir 166 497 habitants [43].

Ces différences entre les différentes versions de Wikipedia semblent inévitables dans la mesure où chaque infoboxe est éditée manuellement, et cela dans plus de 280 langues. Le projet Wikidata, lancé en 2012 par le chapitre allemand de Wikipedia et soutenu par la Wikimedia Foundation, vise notamment à résoudre ce genre d'incohérence.

> « *For all the different language versions of Wikipedia, Wikidata should be a data repository, where the basic facts are centralized, shared and reused by the all the infoboxes on the same topic, across all the language versions.* » [Yu 11, p. 568].

Cette ambition spécifique explique quelques-unes des différences entre DBpedia et Wikidata. Cette dernière est éditable par la communauté d'utilisateurs elle-même et possède des mécanismes qui permettent de gérer les informations contradictoires. Ainsi, plutôt qu'une « population de Bruxelles », on y trouvera la « population de Bruxelles indiquée par la Ville en 2011 »[Vrandečić 14]. Mais c'est sa manière de gérer le multilinguisme qui, le plus ostensiblement, distingue Wikidata des autres bases de connaissance. Pour garder l'exemple de Bruxelles, on sait que la ville peut, selon les langues, s'appeler Brussels, Brüksel, Brusela, Brysselin kaupunki... Quelle appellation unique choisir ? Wikidata a résolu cette difficulté en remplaçant chaque concept par un identifiant numérique. Bruxelles-Ville devient ainsi Q239 [44], tandis que la notion de « ville de Belgique » s'exprimera par Q493522 ou celle de « capitale » par Q5119. Un exemple de requête SPARQL dans Wikidata, fourni en annexe, montre à quel point l'opération semble plus abstraite que dans DBpedia...

Déjà riche de quelque 18 671 750 concepts [45], Wikidata pourrait connaître une subite croissance à deux chiffres après la future intégration de Freebase (lire ci-dessous).

41. https ://fr.wikipedia.org/wiki/Ville_de_Bruxelles
42. `https://en.wikipedia.org/wiki/City_of_Brussels`
43. https ://es.wikipedia.org/wiki/Bruselas
44. https ://www.wikidata.org/wiki/Q239
45. Chiffre affiché sur la page d'accueil de Wikidata en juillet 2016 : https ://www.wikidata.org/wiki/Wikidata :Main_Page.

Freebase

Avec Freebase, lancée en 2007, l'ambition de l'entreprise Metaweb Technologies n'était rien de moins que de créer un répertoire public structuré des connaissances mondiales [Bollacker 08, Bollacker 07]. Inspirée du principe de Wikipedia, cette base de connaissance offrait la possibilité aux utilisateurs d'ajouter et d'éditer des informations. Freebase était en outre alimentée d'extraits de l'encyclopédie libre, mais également de sources aussi diverses que l'agrégateur de célébrités NNDB [46], le Fashion Model Directory [47] ou encore l'encyclopédie musicale MusicBrainz [48]. Près de deux milliards de triplets ont ainsi été stockés dans un modèle de graphe propriétaire, interrogeable à l'aide d'un langage spécifique nommé MQL (*Metaweb Query Language*). Chaque concept dispose d'un identifiant unique appelé « mid » (pour *Machine ID*), d'un ou plusieurs types et de propriétés. Par exemple, la ville de Bruxelles (mid : /m/02rnbv) possède plusieurs types tels que /location/citytown ou /travel/travel_destination, ainsi que des propriétés comme « has a Population ».

Intéressé par cette base de connaissance riche de millions d'entités, le moteur de recherche Google racheta Metaweb en 2010. C'est ainsi que Freebase servit de fondement au Knowledge Graph de Google, cette base de connaissance propre au moteur de recherche qui offre à ses utilisateurs, en plus des résultats de leurs requêtes, des suggestions issues du Web sémantique [49].

Après le lancement de Wikidata, qui obéit à des principes similaires, Google annonça en décembre 2014 la fusion de Freebase dans la base de connaissance de la Wikimedia Fundation. Annoncée pour le 30 juin 2015, la mise hors service de son API devrait finalement avoir lieu le 31 août 2016 [50]. Les « dumps » de la base de données resteront toutefois disponibles [51]. Pour plus de détails sur la « grande migration » que représente l'intégration de Freebase dans Wikidata, nous renvoyons à [Pellissier Tanon 16].

46. http ://www.nndb.com/
47. http ://www.fashionmodeldirectory.com/
48. https ://musicbrainz.org/
49. https ://googleblog.blogspot.be/2012/05/introducing-knowledge-graph-things-not.html
50. https ://developers.google.com/freebase/
51. https ://plus.google.com/109936836907132434202/posts/bu3z2wVqcQc

2.5. Limites de Wikipedia

Résoudre une entité à l'aide d'une base de connaissance suppose évidemment que cette dernière possède une entrée correspondante. A défaut, le système renverra la mention `NIL` [McNamee 09]. Se pose alors la question de la couverture des KB fondées sur Wikipedia, puisque l'encyclopédie ne pourra jamais répertorier la totalité des entités nommées possibles. Pour l'anglais, langue la mieux dotée, [Zhou 10] ont montré sur la base de 1000 articles *Yahoo! News* annotés en `PER`, `ORG`, `LOC` que près d'un quart des entités n'étaient pas référencées dans Wikipedia. Et encore, remarque [Sil 12], s'agissait-il de textes journalistiques, pourtant plus susceptibles de mentionner des entités connues. Lors de la campagne TAC 2009 [McNamee 09], sur un mélange d'articles de journaux et de blogs, 57 % des 3 904 entités annotées n'avaient pas d'équivalent dans Wikipedia [Hachey 13].

Le tableau 2.1 fournit un aperçu du nombre d'instances des types PER, ORG, LOC et WORK (livres, films, albums musicaux, *singles*, logiciels, émissions télévisées, etc.) répertorié par quelques-unes des versions linguistiques de DBpedia en 2014 [52]. Le nombre d'entités total est constitué de l'addition de 28 versions linguistiques, doublons compris [53].

	en	it	es	fr	TOTAL 28
Personnes	763 643	145 060	65 337	62 942	1 471 254
Lieux	572 728	141 101	132 961	80 602	818 539
Organisations	192 832	4142	11 710	17 513	266 551
Ouvrages	333 269	51 918	36 484	39 195	462 124

TABLE 2.1. – Répartition de quelques types d'entités nommées dans DBpedia

Afin d'accroître le taux de résolution, l'une des solutions peut consister à utiliser plusieurs bases de connaissance plutôt qu'une. C'est l'objet des recherches de [Pereira 14], dont les résultats finaux ne sont pas encore publiés. L'auteure constate que l'Internet Movie Database (IMDB) [54] contient plus de 8 millions d'entités relatives au cinéma, tandis que Music Brainz [55] en recense 30 millions liées à la musique.

52. http ://wiki.dbpedia.org/services-resources/datasets/data-set-38/data-set-statistics
53. http ://wiki.dbpedia.org/services-resources/datasets/cross-language-overlap-statistics
54. http ://www.imdb.com/
55. https ://musicbrainz.org/

En dépit de leur spécialisation, ces deux bases de données renferment donc sept fois plus d'entités que la totalité de Wikipedia (5 millions de pages en anglais).

Toutefois, la résolution d'entités nécessite que la base de connaissance utilisée dispose de liens entre ses entités. Les algorithmes de désambiguïsation fondent en effet leurs calculs sur les similarités de contextes et sur les liens sémantiques entre entités. Or beaucoup de bases de connaissance spécialisées ne disposent pas de ces liens. Pour pallier ce problème, [Li 16] propose une méthode de fouille de textes destinée à récolter des indices qui simulent des liens dans une KB de documents peu structurés. Avant cela, [Sil 12] s'est penché sur la désambiguïsation d'entités à l'aide de n'importe quelle base de données, du moins si elle est en forme normale de Boyce-Codd [Elmasri 11, p. 529]. Le système mis au point, Open-DB NED, a obtenu des résultats concluants dans les domaines sportif et cinématographique.

Ces travaux et quelques autres ([Jin 14, Pantel 11, Shen 14]...) laissent penser que la désambiguïsation d'entités à l'aide de bases de données, ou de réseaux de documents peu structurés, constituera encore dans les années à venir une piste de recherches prometteuse.

2.5.1. Les services d'*entity linking*

Plusieurs services web *freemium* permettent de rechercher les entités nommées dans un texte, tout comme ceux présentés au point 1.7, mais aussi de les désambiguïser avec des bases de connaissance. Citons par exemple Dandelion (ex-DataTXT) [56], DBpedia Spotlight [57], TextRazor [58], AlchemyAPI [59] ou encore OpenCalais [60].

Chacun de ces systèmes propose des APIs REST [Fielding 00] auxquelles des textes courts peuvent être envoyés par requêtes HTTP. Le résultat de l'annotation est renvoyé aux formats JSON ou XML [61].

56. https ://dandelion.eu/, offre gratuite limitée à 1000 requêtes par jour.
57. http ://dbpedia-spotlight.github.io/demo/, libre et gratuit, mais souvent évalué comme moins efficace que les outils payants.
58. www.textrazor.com/, offre gratuite limitée à 500 requêtes par jour.
59. http ://www.alchemyapi.com/, 1000 requêtes gratuites par jour. Il est capable de reconnaître et de catégoriser les entités nommées en français, mais pas (encore) de les désambiguïser dans cette langue.
60. http ://www.opencalais.com/opencalais-api/, gratuit, il résout les entités avec des URIs propriétaires.
61. Pour les utilisateurs d'Open Refine, l'extension NER, que nous évoquerons au chapitre 3, permet

Lors d'une récente évaluation qui portait également sur des dépêches d'agence en français [Rizza 15b], quelques-uns de ces services ont obtenu des résultats satisfaisants. Le vainqueur de ce test (TextRazor) a atteint un F1-Score de 0.94 en pure NERC, proche de celui d'un humain donc, et un taux de résolutions correctes de 81 % avec des pages Wikipedia. Il s'agissait toutefois d'un corpus réduit (20 dépêches) centré sur l'actualité internationale. Il pourrait être instructif de reproduire cette étude sur le corpus annoté de dépêches belges que nous présenterons au chapitre 2. Cela sortirait toutefois du cadre restreint de ce mémoire.

Il faut préciser, pour conclure, que ces services travaillent de manière totalement opaque (*black boxes*). L'utilisateur n'a que peu de prise sur leurs paramètres, et doit se contenter des réponses renvoyées.

d'utiliser quatre d'entre eux dans un environnement graphique.

Partie 2 : Étude de cas

Chapitre 3

Analyse du corpus : présentation et aspects méthodologiques

3.1. Présentation du corpus 54
3.2. Récupération et traitement du corpus 60
3.3. Traitement des résultats 70

Notre objectif dans cette seconde partie du mémoire sera d'examiner les entités nommées de type `PER`, `ORG` et `LOC` contenues dans un échantillon représentatif de dépêches d'agence belges francophones, en l'espèce celles de l'agence Belga.

Le choix de nous cantonner à ce seul type de production journalistique s'explique par trois motifs. D'une part, puisqu'il s'agit de se livrer notamment à une analyse statistique, un corpus hétérogène aurait risqué de brouiller les conclusions. Sur un échantillon de taille modeste, comme ce sera le cas du nôtre (voir 3.2.1), la ligne rédactionnelle spécifique d'un seul quotidien sur trois ou quatre risquerait de parasiter les résultats. D'autre part, les dépêches d'agence possèdent, en théorie, une forme plus régulière que la plupart des autres types d'articles. Tant la longueur du titre que celle du texte obéissent à des règles uniformes. Mais c'est surtout leur structure qui

nous a conduit à privilégier ce matériau. Là où un « papier » de quotidien peut appartenir à une large gamme de genres [1], chacun doté de ses propres conventions stylistiques, une dépêche suivra le plus souvent [2] la trame suivante : il s'agit d'un article strictement *factuel*, rédigé selon la méthode de la *pyramide inversée* et introduit par un *lead* qui résume l'essentiel de l'information. Ces éléments seront détaillés dans la section qui suit.

Enfin, les dépêches d'agence constituent le matériau journalistique le plus répandu. Tous les grands médias sont abonnés à au moins une agence et réutilisent plus ou moins abondamment leur production. Pour les quotidiens gratuits, au modèle commercial fondé uniquement sur la publicité, elles constituent même l'essentiel de l'information fournie dans leurs pages [Thomas 10]. Avec le développement des sites d'information gratuits, les dépêches se sont également multipliées sur le web. Source d'information rapide, neutre, le plus souvent exacte, peu coûteuse, elles constituent le principal point commun entre des médias aussi différents que lesoir.be, sudinfo.be ou rtbf.be. Aux Etats-Unis, entre 2001 et 2006, la proportion de dépêches d'agences reprises par les grands médias en ligne est ainsi passée de 34 % à 50 %, et même de 68 % à 85 % dans le cas des portails et agrégateurs, comme Yahoo! ou Google [Paterson 07] [3].

3.1. Présentation du corpus

3.1.1. Belga

Il n'existe en Belgique qu'une seule agence de presse généraliste, Belga, qui couvre l'actualité de l'ensemble du pays, ainsi que celle des institutions européennes installées à Bruxelles. Fondée le 20 août 1920 par Pierre-Marie Olivier et Maurice Travailleur, l'agence [4] est détenue depuis 1948 par un actionnariat majoritairement com-

1. Compte-rendu, portrait, reportage, enquête, interview, chronique judiciaire, éditorial, etc. De plus, ces grands genres peuvent obéir à des règles encore plus spécifiques selon la rubrique qui les héberge (sport, politique, économie...). Voir [solidaires 12].

2. Certaines agences internationales telles que l'AFP (France) peuvent parfois produire des dépêches moins factuelles, ce qui n'est pas le cas de l'agence belge Belga.

3. L'étude de Chris Paterson se limitait toutefois à la couverture de l'actualité internationale.

4. Sa dénomination officielle est toujours « Agence Belga-Agence Télégraphique Belge de Presse, S.A.-Agentschap Belga-Belgisch Perstelegraaf Agentschap N.V. » : http ://kbopub.econo-

posé des groupes de presse écrite belges. Dernière rédaction réellement bilingue du pays, elle est divisée depuis 1970 en deux *desks*, l'un francophone, l'autre néerlandophone [Shrivastava 07, p. 193], chacun traitant l'actualité de sa communauté linguistique. Les dépêches susceptibles d'intéresser l'ensemble du pays sont traduites, mais chaque journaliste ne rédige que dans sa propre langue. Belga emploie aujourd'hui 100 journalistes, francophones et néerlandophones. Elle fait également appel à des « correspondants » indépendants pour couvrir l'actualité judiciaire et régionale.

Sa clientèle d'abonnés est constituée à 67.5 % de médias, tandis le tiers restant se partage entre pouvoirs publics (16 %) et entreprises (16.5 %) [5]

En 2015, Belga a produit 145 620 dépêches en français et 133 201 en néerlandais [6]. Depuis 2010, celles-ci se divisent en quatre grandes catégories :

— **Alerte :** dépêche de maximum 150 signes qui annonce une information importante. Exemple : « ALERT : Le Royaume-Uni a décidé d'envahir l'Irak de manière prématurée en 2003 (rapport) ». Il ne s'agit pas d'un synonyme de *breaking news,* puisqu'une directive interne que nous avons pu consulter précise : « Une ALERT sur une information d'un autre média ne peut être donnée qu'en cas de « breaking news » ».

— **Ultra-short :** petit texte d'environ un paragraphe destiné à développer rapidement l'alerte, idéalement dans les dix minutes qui suivent.

— **Short (ou brief) :** les dépêches les plus courantes, en théorie limitées à 250 mots. Une version adaptée de ces shorts, illustrée par des photos d'actualité et *ready to publish*, peut être envoyée aux rédactions web sous le nom de *brief*s. Ce sont ces derniers qui constituent l'essentiel des fils infos sur le web belge. Par défaut, tous les *shorts* se voient attribuer le mot-clé `brief` [7]. « Il revient à chaque journaliste qui valide une dépêche de le supprimer s'il estime que l'information n'a pas sa place dans le fil online », précise un document interne de l'agence.

— **Texte :** toute autre dépêche excédant 250 mots, le plus souvent des développe-

mie.fgov.be/kbopub/toonondernemingps.html?ondernemingsnummer=403481693

5. « A propos de Belga », site web officiel de l'agence. URL : http ://www.belga.be/fr/a-propos-de-belga/, consulté le 25 juillet 2016.

6. Informations de sources internes.

7. La présence de ce mot-clé implique que même si aucun journaliste n'édite le *short* pour l'adapter au web, il sera tout de même envoyé automatiquement aux clients online une demi-heure après sa validation.

ments d'information, des synthèses ou des analyses. En théorie, un texte long doit toujours être précédé d'un *short*.

Structure d'une dépêche

Une dépêche est constituée de quatre éléments essentiels, clairement distingués dans le système de production de dépêches de Belga (Figure 3.1.2).

— **Le titre.** La quintessence de l'information qui fait l'objet de la dépêche, résumée en environ une ligne (maximum 90 signes dans Belga). Les agences n'utilisent que des titres factuels. Dans Belga, il s'agit de l'une des différences entre un *short* et un *brief*, le titre de ce dernier étant souvent adapté aux besoins spécifiques du web.

— **Le *lead.*** Il s'agit du premier paragraphe, ce que l'on pourrait appeler en français l'« attaque » ou l'« accroche ». Mais alors qu'il existe une grande variété d'attaques journalistiques [8], le *lead* d'une dépêche d'agence obéit à des règles uniformes et universelles. Son objectif est moins d'accrocher le lecteur que de lui fournir immédiatement l'essentiel de l'information [9]. Il devrait donc logiquement mettre en avant la même information que le titre. Dans leurs manuels internes, certaines grandes agences telles que l'AFP en définissent le contenu de manière stricte : « Le lead est composé d'une phrase unique, écrite avec des mots simples, précis (pensez à la traduction), bien choisis. Il ne doit contenir qu'une seule idée, clairement exprimée, en 30 mots maximum. (...) » ([France-Presse 04, p. 15], cité par [Grevisse 08, p. 103]). Il n'existe pas de manuel de style équivalent au sein de l'agence Belga, l'apprentissage des nouveaux journalistes se faisant plutôt par socialisation [10]. Du point de vue qui est le nôtre, le *lead* d'une dépêche présente un intérêt particulier. S'il répond réellement aux questions

8. Dans son *Writing Solutions* [Fensch 01], Thomas Fensch en recense 25 types sans pour autant épuiser le sujet.

9. Un article de quotidien peut évidemment employer la même méthode. Les anglo-saxons utilisent dans ce contexte l'expression *summary lead*, par opposition à toute la gamme de *feature leads* employés dans des articles moins étroitement factuels [Scanlan 00, pp. 115-151]. On retrouve parfois l'expression *5-W lead*, en référence aux cinq questions fondamentales auxquelles devrait répondre tout article factuel *(Who, What, When, Where, Why)*.

10. Au sens sociologique du terme, à savoir « le processus d'acquisition (...) des 'manières de faire, de penser, de sentir' propres aux groupes, à la société où une personne est appelée à vivre » [Rocher 68].

qui? et où?, nous devrions en effet y trouver un taux élevé d'entités nommées de type PER, ORG et LOC. On peut également présumer que les entités en question seront les plus constitutives de l'événement sur lequel porte l'information. Très souvent, ces deux éléments sont accessibles directement dans les fils infos diffusés au format RSS (*Really Simple Syndication*). Leur traitement automatisé se révélerait donc plus simple que celui d'un texte entier.

— **Le corps du texte.** Une fois les éléments essentiels de l'information fournis dans le *lead*, le reste de la dépêche apporte des précisions par ordre décroissant d'importance. Les faits les plus anecdotiques figureront donc à la fin. Cette structure, baptisée *pyramide inversée,* est utilisée par toutes les agences de presse. Plusieurs auteurs, comme [Flesch 49], y voient un héritage de la guerre de Sécession (1861-1865), à savoir d'une époque où les transmissions télégraphiques pouvaient s'interrompre à tout moment. De nos jours, son avantage principal est d'ordre pratique : lorsqu'un journal ne dispose pas de la place nécessaire pour publier une dépêche entière, un secrétaire de rédaction peut la réduire facilement en retranchant des paragraphes à partir de la fin. Sur le web, où l'espace est moins compté, cette structure semble adaptée aux usages « zappeurs » des internautes pressés [redactiweb 10].

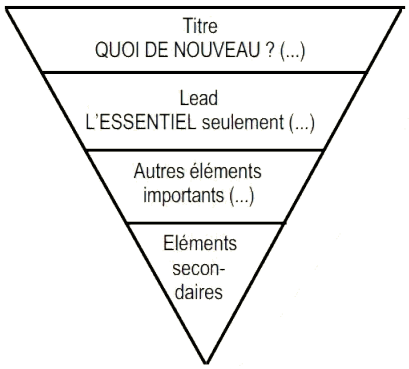


FIGURE 3.1.1. – La pyramide inversée selon [Ross 94], repris de [Labasse 12]

— **Les métadonnées.** Elles remplissent un rôle essentiel dans une dépêche d'agence, tant pour des questions d'archivage que pour assurer le suivi de la production - chez Belga comme chez ses clients, le traitement des dépêches s'effectue de manière largement automatisée. C'est ainsi qu'une dépêche se verra attribuer

par exemple les catégories INT (actualité intérieure), EXT (international), SPF (sport francophone), puis la sous-catégorie GEN, ECO, POL... La ville où se déroule l'information est également indiquée. Les rédacteurs sont ensuite invités à ajouter un maximum de mots-clés dit « libres », mais en réalité puisés dans une liste contenant une centaine de termes (justice, santé, logement, chimie cinéma, décès, etc). Enfin, la dépêche mentionne les initiales du journaliste qui l'a rédigée [11], ainsi que celles du *deskeur* qui l'a relue ou traduite.

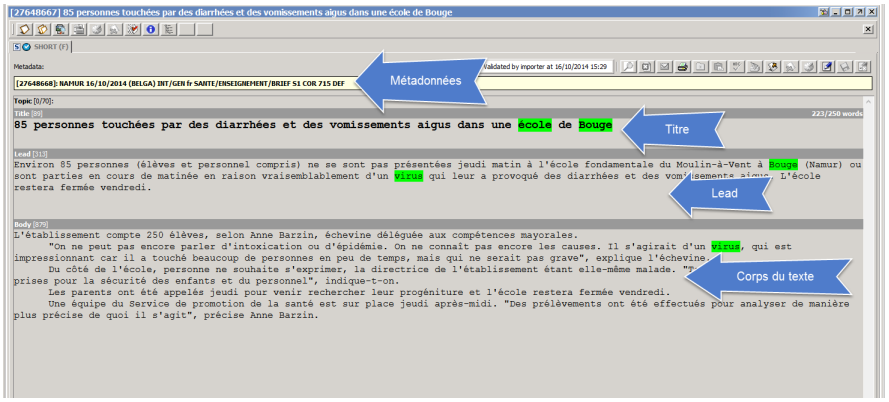


FIGURE 3.1.2. – Aperçu d'une dépêche dans le système interne *Belga360*

3.1.2. 7sur7.be

L'idéal, nous en sommes conscient, aurait été d'extraire notre corpus directement du fil de dépêches de l'agence. Mais les spécificités de son système informatique, ainsi que la nécessité pour notre travail d'obtenir un échantillon aléatoire réparti sur au moins une année complète, nous ont poussé à emprunter une voie détournée. On sait d'expérience que, parmi les sites web d'information, certains utilisent les dépêches Belga plus que d'autres [12]. C'est notamment le cas de *7sur7.be*, propriété du groupe de presse flamand De Persgroep, qui édite également le quotidien populaire *Het Laatste Nieuws*. Le pendant néerlandophone de *7sur7* est d'ailleurs hln.be. L'un et l'autre « s'articulent autour de quatre grands axes rédactionnels : l'actualité belge,

11. Ou le numéro d'identification du correspondant, s'il s'agit d'un texte envoyé par un journaliste extérieur.
12. Les sites en question sont parfois qualifiés péjorativement de « canons à dépêches ».

l'information internationale, le sport et le showbiz »[13]. Depuis 2008, les deux sites disposent d'une équipe de rédacteurs à Sydney, en Australie, ce qui leur permet d'assurer à moindre coût, grâce au décalage horaire, un service 24h sur 24 et sept jours sur sept[14] [15]. En matière d'actualité belge, le modèle économique de *7sur7* repose presque entièrement sur la diffusion la plus rapide possible de dépêches Belga. Sa production propre est très limitée. Le site dispose d'une page où sont regroupées ses archives depuis 2008[16], classées par jour, mois et année, mais aussi par catégories (*Belgique, Monde, Sport, Insolite,* etc.)

A priori, la section « Belgique » de 7sur7 devrait constituer un échantillon représentatif du fil INT/GEN francophone de Belga. Pour nous en assurer, nous avons examiné à la date du 1er juin 2016 la production de *briefs* Belga et celle de 7sur7. Sur 45 articles publiés par ce dernier en catégorie « Belgique »[17], 43 sont sourcés « Belga » (93.3 %) et trois ont été rédigés en interne sur la base d'informations puisées dans les JT (RTBF, RTL et VTM). Le même jour, Belga a publié 156 *briefs*, dont 48 peuvent classés en catégorie « actualité intérieure générale » (hors international, économie et sport donc). Certains textes de 7sur7 sourcés Belga n'apparaissent pas dans cette liste, signe qu'ils ont été repris dans les dépêches classiques (*short* ou *texte*).

Bien que les deux ne se recoupent pas parfaitement[18], cet aperçu nous conforte dans l'idée que les publications de 7sur7 sont représentatives de celles de Belga, à tout le moins des *briefs*.

Signalons que les titres de 7sur7 modifient parfois ceux des dépêches originelles,

13. « 7sur7.be », Journal des sites web, URL : http ://www.journaldessitesduweb.be/site/dossiers/les-sites-d-information/29-7sur7-be.html, consulté le 28 juin 2016.

14. « HLN.BE dispose aussi d'une rédaction à Sydney », 7sur7.be, URL : http ://www.7sur7.be/7s7/fr/1502/Belgique/article/detail/246507/2008/04/17/HLN-BE-dispose-aussi-d-une-redaction-a-Sydney.dhtml, consulté le 28 juin 2016.

15. Belga s'est d'ailleurs inspirée de ce modèle. Depuis septembre 2014, elle dispose de quatre journalistes belges (deux francophones, deux Flamands) dans un bureau loué à Sydney. Auparavant, deux journalistes assuraient des gardes de nuit dans les locaux bruxellois de l'agence.

16. http ://www.7sur7.be/7s7/fr/1481/archives/integration/nmc/frameset/archive/archiveYear.dhtml?archiveYear=2016

17. http ://www.7sur7.be/7s7/fr/1481/archives/integration/nmc/frameset/archive/archiveDay.dhtml?archiveDay=20160601

18. Certains *briefs* concernant un même événement, comme la grève à la SNCB et les perturbations aux TEC, ont été regroupés par 7sur7 en un article unique. Les *briefs* qui n'ont pas du tout été repris sont des informations d'importance mineure, du type « L'artisan de la rénovation de la vieille Havane fait Chevalier de l'Ordre de Leopold » ou « Des élèves de Champion et de Virton lauréats de la 8e édition de "Journalistes en herbe" ».

tout en respectant l'essentiel de leur structure [19].

3.2. Récupération et traitement du corpus

Une fois la source du corpus identifiée, l'objectif sera d'extraire un échantillon d'articles suffisamment étoffé et représentatif. Nous avons pour cela suivi un flux de travail résumé à la Figure 3.2.1, et dont les étapes sont décrites dans les points qui suivent.

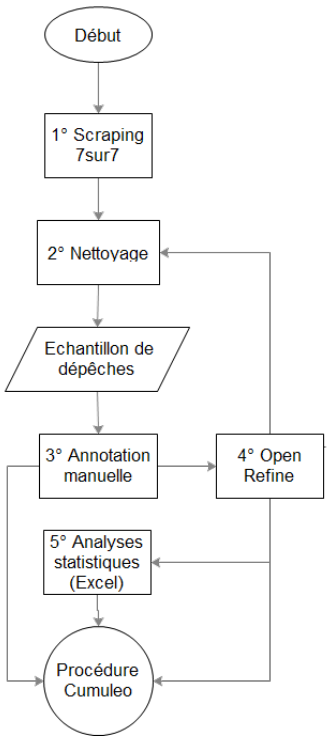


FIGURE 3.2.1. – Schéma du processus

19. Par exemple, la dépêche « Grève à la SNCB : la direction prête à suspendre la circulaire litigieuse à condition que la grève s'arrête » est devenue « Grève à la SNCB : la circulaire litigieuse sera suspendue si la grève s'arrête ».

3.2.1. Scraping

L'extraction de données du web, ou *web scraping*, est le processus qui consiste à convertir automatiquement des ressources présentes sur la Toile en un format structuré [Kushmerick 09]. Son ancêtre direct est le *screen scraping*, méthode de rétro-ingénierie qui permet de récupérer, grâce à leur affichage sur écran, des informations renfermées dans des systèmes informatiques obsolètes [Naumann 09]. Bien que mal considérée, notamment en raison des questions de légalité qu'elle soulève, cette méthode est souvent la seule solution qui permette de récupérer des données sur un site web qui ne dispose pas d'API [Laiacano 12]. Pour reprendre une formule expressive du site School of Data : « *An API is a key ; a scraper is a crowbar* » [Veltman 13].

La plupart des langages de script - mais aussi certains langages compilés comme Java - disposent d'une ou plusieurs bibliothèques spécialisées dans le *web scraping* et le *parsing* du HTML. Parmi les plus connues, citons BeautifulSoup en Python [20], Nokogiri en Ruby [21] ou encore cURL en PHP [22]. En parallèle, plusieurs dizaines d'éditeurs se partagent le marché des logiciels commerciaux génériques destinés aux utilisateurs plus ou moins néophytes.

Par souci de reproductibilité, nous avons utilisé lors de ce travail un outil gratuit, open source et basé sur un langage de programmation très répandu dans le monde des humanités numériques, à savoir Python (version 2.7) et le module BeautifulSoup. Notre script, fourni en annexes, effectue les opérations suivantes.

1 Récupération des liens

A partir de la page d'accueil des archives de 7sur7.be [23], le script extrait les liens de tous les articles de la catégorie « Belgique » publiés au cours d'une période donnée. En l'espèce, notre corpus sera composé d'articles parus de janvier 2013 à décembre 2015 inclus, ce qui constitue 28 741 URLs.

20. https ://www.crummy.com/software/BeautifulSoup/
21. http ://www.nokogiri.org/
22. http ://php.net/manual/fr/book.curl.php
23. http ://www.7sur7.be/7s7/fr/1481/archief/integration/nmc/frameset/archive/archiveYear.dhtml?archiveYear=2016

2 Échantillonnage

Il est évidemment impossible pour une personne isolée d'annoter près de 30 000 dépêches. C'est pourquoi notre corpus de travail sera composé d'un échantillon des articles identifiés lors de l'étape précédente. Afin de déterminer la taille de cet échantillon, il nous faut commencer par définir une marge d'erreur maximale et un niveau de confiance. En sciences sociales, pour des variables qualitatives (comme c'est le cas des entités nommées dans un texte), une marge d'erreur de 5 % à un niveau de confiance de 95 % est considérée comme une approximation acceptable ([Krejcie 70], cité par [Barlett 01]).

Une fois ces valeurs adoptées, la taille de l'échantillon peut être calculée à l'aide de la formule suivante [Cochran 77] :

$$TE = \frac{Z^2 * p * (1-p)}{m^2} = \frac{(1.96)^2 * 0.5 * 0.5}{(0.05)^2} = 384 \tag{3.2.1}$$

où :

TE = la taille de l'échantillon ;

Z = la valeur Z (1.96 pour un niveau de confiance à 95 %) ;

p = la proportion qui sera testée, sachant que 0.5 (50 %) donnera la marge d'erreur la plus élevée ;

m = la marge d'erreur maximale acceptable (0.05 = ±5 %).

Autrement dit, toute proportion observée dans un échantillon aléatoire de 384 dépêches devrait nous permettre de tirer des conclusions relatives à cette même proportion au sein de la population totale (en l'occurrence, les articles de 7sur7 classés en catégorie Belgique) dans un intervalle de confiance d'au maximum 10 % (la marge d'erreur fois deux) et avec un risque d'erreur alpha de 5 %. Mais cette marge d'erreur maximale, qui vaut pour des proportions proches de 50 %, est considérablement plus réduite à mesure que les taux se rapprochent de 0 % ou de 100 %. Si l'analyse démontre, par exemple, que les mots de l'échantillon sont composés à 5 % d'entités nommées du type « organisation », nous pourrions en déduire, avec une chance sur vingt de se tromper, que l'ensemble des dépêches est constitué de 2,8 % à 7,8 % d'entités « organisation » (5 % ± une marge d'erreur de 2.2 points de pour cent).

3 Extraction des textes

Le script Python parcourt ensuite les 384 liens, sélectionnés précédemment de manière aléatoire, afin d'en extraire les éléments structurés qui constitueront notre corpus : texte des articles, mais aussi titre, chapeau, date de publication, source (généralement « Belga »). Nous avons également extrait les métadonnées inscrites dans le code source des pages, qui se révéleront utiles pour classer les articles en catégories (société, politique, etc.).

ID	URL	INFO	META	TITLE	LEAD	TEXT
318	www...	Par : rédaction 8/01/15 -23h15 Source : Belga	justice, bruxelles	Manifestation nationale : peines de travail de 300 heures	©belga Trois hommes qui avaient lancé des pierres et autres projectiles sur les policiers le ...	Deux des trois hommes avaient reconnu leur participation aux bagarres ...

TABLE 3.1. – L'une des dépêches du corpus au format CSV (lignes tronquées)

3.2.2. Nettoyage

L'extraction ou la recherche d'informations constituent un sempiternel ajustement entre rappel et précision, les deux étant en général inversement proportionnels - lorsque l'un augmente, l'autre diminue et vice-versa [Salaün 10, p. 145]. En favorisant la précision, notre script récupérerait exactement les informations nécessaires, mais avec le risque de ne renvoyer aucun résultat lorsque les éléments HTML d'un article particulier diffèrent légèrement de la norme. A l'inverse, en favorisant le rappel, nous nous assurerons de récupérer l'ensemble des informations requises, mais assorties d'éléments excédentaires (comme des balises superflues) ou agglutinées dans des blocs de textes mal structurés (comme la source et la date récupérées dans un même champ).

La seconde solution, soit le taux de rappel maximal, constitue dans notre cas un moindre mal, nettoyer le corpus étant moins fastidieux que de récolter manuellement des éléments omis lors du *scraping*. Nous avons effectué ce nettoyage dans le

logiciel Open Refine, que nous décrirons plus en détail à la Section 3.3. Nous ne mentionnerons ici, en guise d'exemple, que la formule GREL destinée à ôter des textes les crédits photo et les encadrés « lire aussi » en fin d'articles, particulièrement pernicieux dans la mesure où ils contiennent souvent eux-mêmes des entités nommées susceptibles de parasiter une annotation :

Listing 3.1 – Exemple de formule GREL

```
value.replace(/^©.+\. ?/, '').replace(/Lire aussi :.+/, '')
```

D'autres formules de ce type sont disponibles dans la documentation d'Open Refine, ainsi que dans l'ouvrage de référence [Verborgh 13].

Une fois toutes les opérations effectuées, le logiciel offre la possibilité d'exporter le résultat au format .txt, en encadrant de balises chaque élément constitutif de la dépêche (titre, chapeau, texte...). Ces métadonnées permettront, après annotation, d'identifier dans quelle partie du texte se situaient les entités nommées.

3.2.3. Annotation

Pour [Garside 97, p. 2], l'annotation de corpus peut être définie comme la pratique consistant à ajouter des informations linguistiques interprétatives à un corpus électronique de documents langagiers textuels ou oraux. Annotation peut aussi désigner le résultat final de ce processus [24]. Nous avons vu au chapitre 1 que l'annotation manuelle d'entités nommées a accompagné toute l'histoire de la NERC, puisque les campagnes d'évaluation qui ont défini le concept reposaient sur la comparaison entre annotations humaines et automatiques. En dépit de la somme d'expérience accumulée, annoter un corpus reste une tâche très sujette à interprétations.

Dans notre cas, l'objectif premier était d'identifier dans le corpus de dépêches toutes les mentions de personnes, de lieux et d'organisations. En apparence limpide, l'opération produit rapidement quantité de questions et d'hésitations. Nous pourrions classer les problèmes rencontrés en trois grandes catégories, en grande partie déjà décrites dans [Ehrmann 08] et [Fort 09] :

24. *« It can be defined as the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data. 'Annotation' can also refer to the end-product of this proces. »*

— **Sens absolu et sens en contexte.** Dans la phrase « Environ 85 personnes ne se sont pas présentées jeudi matin à l'**école fondamentale du Moulin-à-Vent** à Bouge », le segment en gras possède toutes les caractéristiques d'un nom de lieu, et l'on serait tenté de l'annoter aussitôt comme tel. Mais un paragraphe plus loin, lorsqu'il est question de « la direction de l'école du Moulin-à-Vent », la même entité apparaît cette fois clairement comme une organisation. Le dilemme eût été plus vif encore si nous avions défini une classe d'entités « bâtiments », puisque le même terme, selon le contexte, aurait pu appartenir à trois types distincts. L'incertitude s'accroît lorsque les types d'entités sont structurés en hiérarchie. Cela peut sembler une bonne idée de distinguer, lors de l'annotation, les « lieu-région » des « lieu-province » et « lieu-commune ». Mais dans la pratique, des entités telles que « Bruxelles » peuvent adopter successivement les trois sens dans une même dépêche, en plus de celui d'organisation (comme dans le titre « Bruxelles fait la chasse aux logements vides », où le mot désigne par métonymie le gouvernement bruxellois). Les descriptions définies et les abréviations constituent une variante du même problème. Des termes comme « Cour constitutionnelle » (organisation) sont la plupart du temps mentionnés une première fois sous leur dénomination longue, puis abrégés en « la Cour ». Par souci d'éviter les répétitions, ils peuvent aussi être évoqués par périphrases (« la gardienne de la constitutionnalité des lois », « l'ancienne Cour d'arbitrage »...). L'une est-elle une entité nommée et les autres pas, alors que toutes font référence au même objet du réel?

— **Frontière des entités.** Dans le segment « … a déclaré le ministre Marcourt à l'agence Belga, confirmant une information du quotidien La DH », « Belga » et « Dernière Heure » désignent clairement des organisations. Mais faut-il annoter également le mot « agence », vu que l'appellation officielle de ce média est « Belga News Agency »? Si oui, pourquoi ne pas annoter « quotidien »? Et que faire de l'article « la » qui précède « DH », alors que l'intitulé exact du journal est désormais *DH Les Sports*, sans article? Dans le même ordre d'idée, les organisations dotées de filiales régionales constituent une source importante d'hésitations. Si « parquet de Bruxelles » semble constituer une organisation à part entière, distincte de « parquet » tout seul, le même raisonnement peut-il s'appliquer à « les pompiers carolos » ou à « les TEC liégeois »? En apparence

vétilleuses, ces questions révéleront toute leur importance lorsqu'il s'agira de retrouver automatiquement ces entités dans des bases de connaissances. Elles rejoignent finalement celles que se pose tout taxonomiste, notamment lors du choix d'un type de vocabulaire pré- ou post-coordonné [Hedden 10]. Par exemple, si « parquet de Liège » est considéré comme une entité ORG à part entière, mais qui n'existe dans aucune base de connaissance, comment évaluer la reconnaissance automatique si celle-ci renvoie un URI correct pour « parquet (Belgique) » et un autre pour « Liège (arrondissement) » ?

— **Entités entremêlées.** L'expression « Hôtel de police de la zone Namur » fait référence à un lieu, mais il contient également un nom d'organisation (« police de la zone Namur ») et un autre nom de lieu (« Namur »). Faut-il annoter les deux enchâssées ou uniquement la première ? Les entités imbriquées constituent une source d'incertitude analogue. Si « les Régions wallonne et flamande » désigne manifestement deux organisations (ou deux lieux…), comment les annoter ? « Région**s** wallonne » paraîtrait contenir une faute d'orthographe, tandis que « flamande » seul est un simple adjectif. Dans certains cas, que nous n'avons pas rencontrés, l'ellipse de l'une des deux entités peut même être totale, comme dans « M. et Mme Chirac » [Ehrmann 08, p. 60].

Il n'existe pas de réponse unique aux questions posées ci-dessus. Chaque campagne d'évaluation a défini ses propres règles d'annotations, et elles sont souvent divergentes entre elles. Dans le cas des entités enchâssées, par exemple, MUC préconisait d'annoter uniquement la plus grande. L'entreprise « Boston Chicken Corp » serait ainsi étiquetée uniquement comme ORG, malgré la présence d'un nom de ville dans sa dénomination. Selon le guide d'annotation ESTER, au contraire, il aurait fallu annoter également « Boston » comme LOC, mais uniquement car il s'agit d'un type d'entité distinct de ORG [25]. Lorsque des entités enchâssées sont de même type, comme « Palma » (LOC) et « Majorque » (LOC) dans « Palma de Majorque » (LOC), seule la principale doit être annotée, selon le guide ESTER.

Nous avons utilisé pour notre part une version adaptée et simplifiée de ce dernier. L'une des principales modifications porte sur le contexte pris en compte et sur le traitement des entités enchâssées. Contrairement aux consignes d'ESTER, nous avons annoté toutes les entités contenues dans une autre, quel que soit leur type. En

25. Exemple emprunté à [Ehrmann 08].

revanche, nous n'avons tenu compte que des entités dont la forme de surface étaient suffisante pour les identifier, sans recours à d'autre contexte que les mots présents à droite ou à gauche et le cadre général belge de ce corpus. En cas de doute, nous nous sommes posé les questions heuristiques suivantes : « Cette entité est-elle susceptible de posséder une fiche Wikipedia ? Et si oui, cette fiche ferait-elle référence à une entité nommée ou à un concept général ? » C'est ainsi que le segment « les pompiers ont déclaré… » n'a pas été annoté, car pompier ne désigne pas à nos yeux une organisation unique localisable, contrairement à « pompiers de Liège ». Le même raisonnement vaut pour « syndicat socialiste » ou « la Région » sans plus de précision. En revanche, « gouvernement » a été annoté lorsqu'il désigne le fédéral, le contexte général du corpus permettant de l'identifier. De même, les noms propres abrégés (« M. Kir », « Furlan »…) ont été considérés comme désambiguïsables, et donc annotés.

Sensible aux recommandations de [Fort 09], nous avons également pris soin de choisir un outil adéquat et d'évaluer rapidement l'annotation.

Outils. Il existe une large gamme de logiciels d'annotation gratuits, sinon *open source*. Parmi ceux que nous avons testés, citons GATE Developer [26], Glozz [27], Brat [28] ou encore FuuTag [29], spécifiquement dédié à l'annotation selon la hiérarchie d'entités nommées de Satoshi Sekine [30].

Notre choix s'est finalement porté sur UAM Corpus Tool [31] [O'Donnell 08]. Par certains aspects limité [32], il compense ses lacunes par son aisance d'utilisation. Comme le souligne [Fort 12, p. 229], il a notamment l'avantage de permettre l'édition du schéma d'annotation en cours de projet, ce qui constitue un atout indéniable lorsque la typologie des entités nommées se construit au fur et à mesure de l'annotation. Les résultats peuvent être exportés en TSV (*tab-separated values*) ou sous la forme d'un fichier XML dans lequel apparaît la totalité du texte, les éléments annotés figurant entre balises. Cette dernière solution était la mieux adaptée à notre corpus,

26. https ://gate.ac.uk/family/developer.html
27. http ://glozz.free.fr/
28. http ://brat.nlplab.org/
29. http ://nlp.cs.nyu.edu/ene/
30. Pour une liste d'outils plus complète, voir les annexes de [Fort 12] et de [Pustejovsky 13], ou encore le Google Doc https ://docs.google.com/spreadsheets/d/127aLQ_kXEcqnQUGmo6O1-rpYsFH-2g-YMHmBi5z87OY/edit?usp=sharing
31. http ://www.corpustool.com/
32. Il ne permet pas, par exemple, d'indiquer une relation entre les entités, ce qui aurait proscrit son usage si nous avions voulu annoter les coréférences d'entités.

puisque, nous l'avons dit, certains éléments étaient déjà balisés avant l'annotation (Figure 3.2.2).

```
<?xml version='1.0' encoding='utf-8'?>
<document>
    <element>
        <ID>1</ID>
        <Url>
        http://www.7sur7.be/7s7/fr/1502/Belgique/article/detail/2091752/2014/10/16/85-personnes-touc
        hees-par-un-virus-dans-une-ecole-a-Bouge.dhtml</Url>
        <MotsCles>belgique, bouge, sante et forme, province de namur, namur</MotsCles>
        <Titre>85 personnes touchées par un virus dans une école à <segment id='1' features=
        'entities;lieu;commune;section' state='active'>Bouge</segment></Titre>
        <Date>16/10/14</Date>
        <Source>Source: Belga</Source>
        <Chapeau>Environ 85 personnes (élèves et personnel compris) ne se sont pas présentées
        jeudi matin à l'<segment id='2' features='entities;lieu;poi' state='active'>école
        fondamentale du Moulin-à-Vent</segment> à <segment id='3' features=
        'entities;lieu;commune;section' state='active'>Bouge</segment> (<segment id='4' features=
        'entities;lieu;commune' state='active'>Namur</segment>) ou sont parties en cours de
        matinée en raison vraisemblablement d'un virus qui leur a provoqué des diarrhées et des
        vomissements aigus. L'école restera fermée vendredi.
        </Chapeau>
        <Texte>L'établissement compte 250 élèves, selon <segment id='5' features=
        'entities;personne;politicien' state='active'>Anne Barzin</segment>, <segment id='6'
        features='fonction;politique' state='active'>échevine déléguée aux compétences mayorales
        </segment>. On ne peut pas encore parler d'intoxication ou d'épidémie. On ne connaît pas
        encore les causes. "Il s'agirait d'un virus, qui est impressionnant car il a touché
        beaucoup de personnes en peu de temps, mais qui ne serait pas grave", explique l'<segment
        id='49' features='fonction' state='active'>échevine</segment>. Du côté de l'école,
        personne ne souhaite s'exprimer, la <segment id='7' features='fonction' state='active'>
        directrice de l'établissement</segment> étant elle-même malade. "Toutes les mesures ont
        été prises pour la sécurité des enfants et du personnel&quot;, indique-t-on. Les parents
        ont été appelés jeudi pour venir rechercher leur progéniture et l'école restera fermée
        vendredi. Une équipe du <segment id='8' features='entities;incertain' state='active'>
        Service de promotion de la santé</segment> est sur place jeudi après-midi. &quot;Des
        prélèvements ont été effectués pour analyser de manière plus précise de quoi il s'agit",
        précise <segment id='48' features='entities;personne;politicien' state='active'>Anne Barzin
        </segment>.
        </Texte>
    </element>
```

FIGURE 3.2.2. – Exemple de dépêche annotée au format XML

Évaluation. Plutôt que de solliciter un second avis, nous avons préféré calculer rapidement l'accord intra-annotateur (c.f. 1.6.3). Quelques jours après les premières annotations selon la hiérarchie définitive, nous sommes revenus sur les vingt premières dépêches et les avons annotées de nouveau. Le *kappa* de Cohen obtenu est de 0.871 (IC95% [0.8172 ; 0.9254]) [33], ce qui nous a convaincu que les consignes d'annotation et la définition des entités étaient claires dans notre esprit. Dans ce cas précis, puisque notre objectif est avant tout statistique, la stabilité de l'annotation nous a paru plus importante que sa reproductibilité par un annotateur extérieur.

33. Calculé à l'aide l'outil VassarStat : http ://vassarstats.net/kappa.html

3.2.4. Hiérarchie d'entités

Il nous est rapidement apparu qu'une annotation limitée à trois grands types d'entités ne nous offrirait pas la granularité nécessaire pour en tirer des conclusions pertinentes. De plus, la résolution d'entités dans une base de connaissance nécessite souvent, pour traiter les homonymes, de préciser la classe ontologique dans laquelle l'entité doit être cherchée. Ces classes peuvent être générales, comme `dbo:Person`, voire `owl:Thing`. Elles peuvent également être aussi précise que `yago:SocialistParty(francophone`

L'avantage des secondes est de faciliter le *matching* automatique : alors qu'il pourrait exister plusieurs « Elio Di Rupo » répertoriés dans DBpedia ou Wikipedia comme `Person`, il est moins probable que plus d'un soit membre du PS belge. Sans viser ce niveau de granularité, nous avons mis à profit les possibilités qu'offre UAM Corpus Tool pour confectionner, en cours même d'annotation, une hiérarchie en deux niveaux. Dès qu'un type d'entité nous paraissait suffisamment spécifique et récurrent, nous lui avons créé une sous-classe. La définition de ces dernières obéissait moins à une logique taxonomique qu'à nos besoins pratiques. Par exemple, nous avons jugé utile d'introduire la sous-classe `PER/avocat`, à première vue inconséquente (`PER/agent_public`, situé au même niveau, est nettement moins spécifique). Il ressortait toutefois de l'annotation que les avocats sont souvent mentionnés dans les dépêches par leur patronyme seul, précédé du titre « Me » (comme dans « Me Vergès »). Cette particularité nous a paru suffisante pour isoler ces entités, afin de pouvoir les traiter plus aisément par la suite. La classe de 2e niveau `LOC/Commune/section` répond à des besoins similaires. Les dépêches portant sur un fait-divers localisent souvent les événements relatés dans un lieu-dit ou une ancienne commune d'avant la fusion de 1977, dont le nom est suivi entre parenthèses de celui de l'entité administrative officielle. Il pouvait être intéressant de distinguer les deux, afin d'expliquer plus facilement un éventuel problème de reconnaissance de ces lieux (les simples sections de communes sont moins susceptibles de posséder leur propre page Wikipedia). Contrairement aux noms de régions comme « Wallonie » ou « Région bruxelloise », annotés selon leur sens (`LOC` ou `ORG`), les noms de communes ont tous été considérés comme des `LOC`[34].

34. Wikipedia ne distingue pas les deux dans le cas des communes. Que le nom évoque le lieu ou le collège communal, l'URI renvoyé en cas de désambiguïsation réussie serait le même.

Notre méthode taxonomique, finalement proche de l'une des trois utilisées par [Sekine 02] (voir chapitre 1), s'est moins souciée de cohérence que de pragmatisme. Elle fait écho à cette réflexion de [Lambe 14, p. 153] : « *In a knowledge management context, pragmatism will always trump tidiness, and the connection with organisational effectiveness must never be lost* ».

La hiérarchie finale est présentée dans la Table 3.2.

3.3. Traitement des résultats

La liste des entités nommées présentes dans le corpus étant récoltée, il nous faudra déterminer si ces noms possèdent une entrée dans une base de connaissance classique, telle que DBpedia ou Wikidata (voir Section 2.4.2). Afin d'automatiser cette recherche, nous avons mis au point une méthode fondée sur le logiciel Open Refine et le langage SPARQL (Section 2.2.3), vérifiée ensuite à l'aide d'un script Python. Cette méthode fera l'objet du point qui suit.

3.3.1. Open Refine

Open Refine (ex-Google Refine, ex-Freebase Gridworks [35]) s'est frayé en quelques années une place en vue dans la boîte à outils des humanités numériques [van Hooland 14, Southwick 15]. A mi-chemin entre un tableur et un système de gestion de base de données, cette application Java multi-plateformes (Windows, Linux, Mac) et *open source* offre une série de fonctionnalités efficaces pour identifier les problèmes de qualité de données (*data profiling*), puis les résoudre (*data cleaning*). Ses avantages sur un tableur tel que Microsoft Excel reposent sur plusieurs spécificités : logique basée sur les colonnes et les enregistrements (*records*) plutôt que sur les lignes, tri des données par facettes, utilisation avancée des expressions régulières, algorithmes de

35. Conçu en 2010 au sein de la société Metaweb Technologies, son ancêtre Freebase Gridworks avait comme but premier de faciliter l'insertion de nouvelles informations dans Freebase (voir 2.4.2). Après le rachat de Metaweb par Google, en 2010, Freebase Gridworks prit le nom de Google Refine et fut maintenu durant deux ans, avant d'être abandonné en 2012 à sa communauté d'utilisateurs. L'outil, renommé Open Refine, continue d'être perfectionné, à un rythme toutefois plus lent. Ses évolutions peuvent être suivies et commentées librement sur la plateforme Github, qui rassemble également sa documentation technique : https ://github.com/OpenRefine/OpenRefine. La version utilisée dans ce travail est Open Refine 2.6 RC2.

TABLE 3.2. – Hiérarchie d'entités issue de l'annotation

NIVEAU1	NIVEAU2	DESCRIPTION
PER		
	POLITIQUE	*Élu ou mandataire présenté comme tel.*
	EXPERT	*ex. : professeur d'université consulté à titre documentaire.*
	AGENT PUBLIC	*Fonctionnaire, policier, porte-parole du parquet…*
	CÉLÉBRITÉ	*Personnalité connue (Ban Ki Moon, le roi Philippe...)*
	LOBBYISTE	*Porte-parole d'un groupe d'intérêts (syndicat, patronat...)*
	REPRÉSENTANT	*Porte-parole d'entreprise ou d'autres ORG que* `publique` *et* `corporation`.
	JUSTICIABLE	*Auteur ou victime de faits-divers*
	ANONYME	*Témoin sans autre rôle défini*
	AVOCAT	*Souvent précédés du titre « Me ».*
	AUTRE	*Anonyme ou catégorie peu représentée (ex : journalistes).*
LOC		
	POI/LIEU	*« rue Jean Jaurès », « intra-muros montois »…*
	COMMUNE	*Municipalité belge, au sens géographique ou politique.*
	COMMUNE/SECTION	*Ancienne commune ou lieu-dit*
	RÉGION/GEO	*Hauts-Pays, Hesbaye, Wallonie au sens géographique...*
	PROVINCE	*Limbourg, Hainaut...*
	PAYS	*Belgique, France, Mali...*
	RÉGION/MONDE	*Lieu administratif (ville, province…) étranger*
	NATUREL	*La Meuse, la Côte, Saturne…*
ORG		
	ENTREPRISE	*Carrefour, BMW…*
	SPORTIVE	*Standard de Liège*
	CORPORATION	*FGTB, FEB, groupes d'intérêt…*
	POI/ORG	*Organisation et lieu à la fois (musée, école, hôpital, aéroport...)*
	PUBLIQUE	*SNCB, parquet de Namur, Région Bruxelles-Capitale au sens politique...*
	PARTI	*PS, Action fouronnaise, Mouvement réformateur...*
	MÉDIA	*Le Soir, la DH, Belga…*
	ENSEIGNEMENT	*Université, école...*
	INTERNATIONALE	*ONU, US Army, SNCF…*
	AUTRE	*Groupe musical, ASBL indéterminée, Sharia4Belgium…*

clustering, enregistrement systématique des opérations effectuées, possibilité d'annuler ces dernières à tout moment, de les sauvegarder ou de les reproduire sur un autre fichier, exportation et partage de projet...

A l'image des formules d'Excel, Open Refine possède son propre mini-langage de script, GREL (*General Refine Expression Language*), dont la syntaxe proche du JavaScript est relativement simple.

L'extension RDF

Le code source ouvert d'Open Refine et de ses prédécesseurs [36] a permis à sa communauté d'utilisateurs de l'enrichir à l'aide de modules complémentaires [37]. Citons ainsi l'extension NER [van Hooland 13], qui facilite la jonction entre Open Refine et plusieurs services web de reconnaissance d'entités nommées [38]. Celle qui nous intéresse ici est l'extension RDF de DERI (*Digital Enterprise Research Institute*), un institut de recherche irlandais aujourd'hui englobé dans l'*Insight Centre for Data Analytics* [39].

Une première version de cette extension permettait d'exporter des données tabulaires au format RDF [Cyganiak 10]. Une autre est venue lui ajouter la réconciliation [40] avec le *Linked Open Data Cloud*. Cette dernière s'effectue à l'aide de requête SPARQL, basées selon les cas sur des expressions régulières ou sur une recherche *full text* (voir section 3.3.1).

Cette extension, hélas, n'est plus mise à jour. L'utiliser dans des versions récentes d'Open Refine entraîne régulièrement des soucis techniques. Son ergonomie, notamment celle de sa recherche manuelle, est discutable. Par ailleurs, sa simplicité d'emploi va de pair avec un certain manque de précision [Rajabi 14]. Son fonction-

36. Conçu en 2010 au sein de la société Metaweb Technologies, son ancêtre Freebase Gridworks avait comme but premier de faciliter l'insertion de nouvelles informations dans Freebase (voir 2.4.2). Après le rachat de Metaweb par Google, en 2010, Freebase Gridworks prit le nom de Google Refine et fut maintenu durant deux ans, avant d'être abandonné en 2012 à sa communauté d'utilisateurs. L'outil, renommé Open Refine, continue d'être perfectionné, à un rythme toutefois plus lent. Ses évolutions peuvent être suivies et commentées librement sur la plateforme Github, qui rassemble également sa documentation technique : https ://github.com/OpenRefine/OpenRefine.
37. https ://github.com/OpenRefine/OpenRefine/wiki/Extensions
38. https ://github.com/RubenVerborgh/Refine-NER-Extension
39. https ://www.insight-centre.org/
40. Dans l'environnement d'Open Refine, la fonction dite de « réconciliation » (à l'origine avec Freebase seule) correspond à ce que nous avons défini jusqu'ici sous le nom d'*entity linking*.

nement peut toutefois être reproduit de manière plus fine à l'aide la méthode décrite dans le point qui suit.

Réconciliation semi-automatique dans Open Refine

Dans le langage SPARQL, identifier une ressource DBpedia correspondant, par exemple, à une ville ou une commune belge pourrait ressembler à ceci :

Listing 3.2 – Requête SPARQL "communes belges"

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbp: <http://dbpedia.org/property/>
PREFIX dbr: <http://dbpedia.org/resource/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT DISTINCT ?uri
WHERE {
?uri dbo:country dbr:Belgium ;
rdfs:label ?label
FILTER EXISTS { ?uri dbp:mayor ?m }
FILTER langMatches(lang(?label), "FR" )
FILTER regex(str(?label), "\\bBruxelles\\b", "i") .}
LIMIT 5
```

Autrement dit : retrouver un maximum de cinq URIs qui possèdent comme propriété `label` une valeur littérale taguée « français » et contenant le mot « Bruxelles » (en majuscules ou en minuscules), dont la propriété `country` est « Belgium » et qui sont dotés d'une propriété `mayor`.

Comme le laisse deviner le mot `regex` dans le dernier `FILTER`, cette requête utilise une expression régulière afin de trouver les occurrences du mot « Bruxelles ». Cette manière de procéder, standard dans SPARQL, a le défaut de solliciter fortement le serveur. Une fois le graphe `?uri dbo:country dbr:Belgium` identifié, et les résultats sans propriété `mayor` exclus, il reste en effet à tester l'expression régulière sur chaque candidat [41]. Pour réduire leur temps d'exécution, certaines bases de données RDF utilisent des moteurs d'indexation *full text*. La manière de les utiliser, déjà évoquée au sein d'un groupe de travail du W3C [42], n'a toutefois pas été stan-

41. http://stackoverflow.com/questions/13572915/how-to-make-my-sparql-query-with-regex-faster
42. https://www.w3.org/2009/sparql/wiki/Feature:FullText

dardisée dans la dernière recommandation SPARQL [Harris 13]. OpenLink Virtuoso, le serveur de DBpedia, autorise ainsi une recherche textuelle à l'aide de l'opérateur `bif:contains`[43], tandis que l'extension LARQ d'Apache Jena [Jena 15] possède une propriété `pf:textMatch`.

C'est pourquoi l'extension RDF propose plusieurs options au moment de créer un service de réconciliation basé sur un SPARQL Endpoint (Figure 3.3.1). « Generic SPARQL (poor performance) » correspond à une recherche par expressions régulières, tandis que les autres utilisent les syntaxes *full text search* (FTS) propres à trois systèmes.

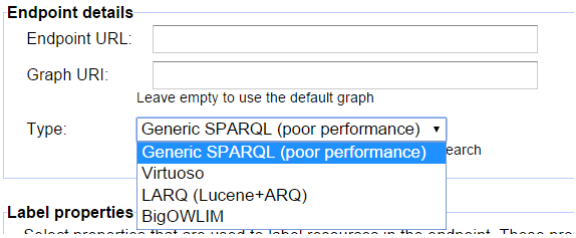


FIGURE 3.3.1. – Choix du type de service dans Open Refine

Dans Virtuoso, voici à quoi ressemble la requête SPARQL sur le terme « Bruxelles », mais cette fois envoyée par l'extension RDF[44] :

Listing 3.3 – Requête SPARQL Virtuoso avec l'extension RDF

```
# Classe choisie : <http://dbpedia.org/ontology/Location>
SELECT DISTINCT ?entity ?label ?score1
WHERE {
?entity ?p ?label.
?label <bif:contains> "'+Bruxelles'"
OPTION(score ?score1).
FILTER (?p=<http://www.w3.org/2000/01/rdf-schema#label>).
?entity a ?type.
FILTER (?type IN (<http://dbpedia.org/ontology/Location>)).
FILTER isIRI(?entity). }
ORDER BY desc(?score1)
```

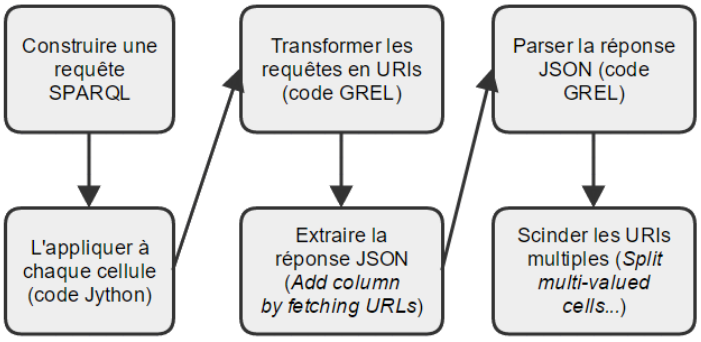
43. Le préfixe `bif` (*built-in functions*) est qualifié implicitement dans Virtuoso. Il n'est associé à aucun URI dans l'espace de noms, ce qui transgresse la norme SPARQL 1.1.

44. Requête récupérée en examinant les activités réseau d'Open Refine. D'autres exemples sont disponibles sur le site web de l'extension RDF : http://refine.deri.ie/searchReconDocs

```
LIMIT 6
```

On notera la présence d'une option `score`, qui permet d'évaluer la proximité des résultats (ou plutôt de leurs n-grammes) avec le terme recherché [45]. Cette fonctionnalité, propre à la recherche *full text,* détermine le classement que dresse l'extension RDF d'Open Refine entre les candidats à la réconciliation. La méthode que nous proposons, décrite plus en détail dans [Rizza 15a], consiste à reproduire ce type de requête SPARQL et à traiter le résultat en six opérations, résumées dans la Figure 3.3.2.

FIGURE 3.3.2. – Synthèse de la procédure dans Open Refine



Cette procédure peut être adaptée à tout SPARQL Endpoint qui autorise des requêtes HTTP GET, notamment ceux de DBpedia et de Wikidata [46].

3.3.2. Vérification dans Wikipedia FR et NL

Le complétude de DBpedia ou de Wikidata n'étant pas assurée, nous avons vérifié si les entités nommées non reconnues immédiatement possédaient une entrée dans la version francophone et néerlandophone [47] de Wikipedia. Nous nous sommes

45. http ://docs.openlinksw.com/virtuoso/creatingtxtidxs.html#hitscores
46. https ://query.wikidata.org/. Dans le cas de Wikidata, il peut être plus simple d'essayer auparavant une recherche par labels à l'aide de son API. Le code GREL suivant permet d'automatiser la procédure dans Open Refine : `https://www.wikidata.org/w/api.php?action=wbsearchentities&search=" + value.escape('url') + "&language=en&type=item&limit=6&format=json`
47. Certains noms de personnes, notamment des personnalités politiques flamandes, n'existent que dans cette version linguistique.

appuyé notamment sur un script Python, fourni en annexes, ainsi que sur le programme Java *Reconcile-CSV*[48]. Ce dernier permet d'effectuer un *fuzzy matching* (concordance floue) entre une colonne d'un projet Open Refine et celle d'un fichier tabulaire.

48. http ://okfnlabs.org/reconcile-csv/

Chapitre 4

Analyse des résultats et *matching* avec Cumuleo

4.1. Corpus général 78
4.2. Entités nommées 80
4.3. Cumuleo 87
4.4. Réconciliation avec le corpus 90

Ce dernier chapitre portera sur l'analyse statistique des résultats récoltés selon les méthodes exposées au chapitre 3. Nous tenterons de valider les deux hypothèses de travail exposées en introduction, qui sont pour rappel :

— Le *lead* d'une dépêche d'agence doit répondre aux questions principales que soulève l'information. Il devrait donc contenir un taux d'entités nommées plus important que le corps du texte. De plus, ces entités devraient être plus représentatives que les autres du thème central de la dépêche.

— Des dépêches d'actualité générale centrée sur un petit pays comme la Belgique devraient contenir bon nombre d'entités nommées inconnues des grandes bases de connaissance internationales. Nous présumons qu'une part plus ou moins

importante de ces entités figure dans Cumuleo, a fortiori si le thème de la dépêche est politique. Cette base de données pourrait dès lors jouer un rôle non négligeable dans la résolution d'entités nommées au sein de textes journalistiques belges.

À notre connaissance, aucune étude ne s'est encore penchée spécifiquement sur la première question, et certainement pas dans le cas de la Belgique. Bien que les travaux de [Stern 10] consacrés à l'AFP fournissent un bref aperçu chiffré des entités nommées présentes dans les dépêches de l'agence française, ceux-ci ne sont mentionnés qu'en passant. La taille réduite de son corpus (100 dépêches) et les inévitables ambiguïtés dans la définition des classes d'entités interdisent toute comparaison avec notre travail.

[Matei 16], publié après que nous avons défini nos propres questions de recherche, s'intéresse précisément à la distribution d'entités nommées de type `PER`, `ORG`, `LOC` dans des articles de presse. Mais outre que son corpus est en anglais, les distributions recherchées étaient centrées sur la position des entités nommées au sein d'un texte, problématique relativement éloignée de la nôtre.

La réconciliation d'entités nommées avec une base de données a quant à elle fait l'objet de plusieurs recherches, déjà évoquée à la Section 2.5.

4.1. Corpus général

Notre corpus contient 384 articles de 7sur7.be classés en section « Belgique », dont 383 comptent au moins une entité nommée [1]. Chaque article possède une source, un titre, un *lead* [2], un corps de texte et des métadonnées cachées dans le code source de la page.

Dans 356 articles (92.7 %), la source mentionnée est explicitement *Belga*. Mais 19 articles dont la source indiquée est « par : rédaction » semblent pour la plupart des textes de l'agence. Nous pouvons donc estimer qu'environ 97.5 % sont issus de dépêches, le solde étant constitué d'articles signés par un journaliste interne ou dont la source est un autre média.

1. L'exception est une courte dépêche météorologique.
2. Sur 7sur7.be, le *lead* des dépêches Belga est représenté en HTML comme un « chapeau » - un paragraphe introductif mis en gras.

La longueur moyenne des éléments textuels [3], exprimée en nombre de caractères avec espace, figure dans le Tableau 4.1.

	Car.
TITRE	53
LEAD	308
BODY	1090
TOTAL TEXTE	1398

TABLE 4.1. – Longueur moyenne des articles du corpus en caractères, espaces compris

Les métadonnées (*tags*) associées à l'article se sont révélées arbitraires et incohérentes. Lors de l'annotation, nous avons donc subjectivement attribué à chaque article l'une des dix catégories retranscrites dans le Tableau 4.2. Cette catégorisation, en partie subjective, s'inspire de celle pratiquée au journal *Le Soir*. Sans les décrire toutes, « Faits-Divers » correspond ainsi aux dépêches rédigées sur la base d'un suivi quotidien des activités de la police ou des services de secours. « Judiciaire » à celles qui proviennent manifestement d'un suivi des cours et tribunaux, qu'ils soient pénaux, civils ou administratifs. « Social » reprend des textes qui portent généralement sur un conflit dans lequel interviennent des syndicats. Quant à la catégorie « Politique », elle englobe toute initiative ou polémique dont la principale caractéristique est d'avoir un ou des personnalités politiques comme principaux acteurs, à l'exclusion des articles déjà repris dans une autre catégorie [4].

Additionnées, les catégories « Faits-Divers » et « Judiciaire » constituent plus de la moitié du total. Cette prédominance de sujets populaires, que nous soupçonnions, ne nous était pas apparue si vive lors de la comparaison sur une journée de la production de Belga et de 7sur7 (Section 3.1.2). Il nous faudra en tenir compte lors de l'analyse des catégories d'entités nommées.

Nous n'avons pas jugé utile de scinder le corpus en phrases ou en paragraphes.

3. Calculée dans Open Refine à l'aide de la fonction GREL :`value.length()`.

4. Les dépêches relatives à l'ex-député Bernard Wesphael, accusé d'avoir tué sa compagne, sont ainsi classées en « judiciaire ».

	#	%
FAITS-DIVERS	126	32,8
JUDICIAIRE	75	19,5
SOCIÉTÉ	67	17,4
POLITIQUE	59	15,4
SOCIAL	36	9,4
RÉGION	8	2,1
INTERNATIONAL	5	1,3
MÉTÉO	4	1,0
ÉCONOMIE	3	0,8
DIVERS	1	0,3

TABLE 4.2. – Répartition des articles par thématiques

4.2. Entités nommées

Après nettoyage des incohérences et erreurs d'annotation manifestes, notre fichier de données contient 4734 occurrences entités nommées et 1823 entités uniques. Leur répartition dans les trois grandes catégories visées est indiquée dans le Tableau 4.3. On remarquera la sous-représentation des noms de personnes par rapport à ceux de lieux et d'organisations. Rappelons toutefois que notre classification comporte une part d'arbitraire, puisque certaines organisations, telles que les gares, les aéroports ou les musées, ont été classées comme `Lieu/poi`. Toute comparaison dans des travaux similaires sera donc à prendre avec prudence.

	#	%
LOC	1859	39.2
ORG	1706	36.0
PER	1181	24.9

TABLE 4.3. – Répartition entre les trois grandes classes d'entités

La répartition par sous-classes (Figure 4.2.1), quoique elle aussi subjective, offrira un point de comparaison moins sujet à interprétation. Outre la prépondérance des noms de communes, on notera celles des organisations publiques et des personnalité politiques. Le taux de `PER/Justiciable` paraît modéré en regard de la proportion de sujets judiciaires ou policiers évoquée à la Figure 4.2. Quant à `la catégorie`

`Lieu/Poi`, rappelons qu'elle englobe les noms de rues, ainsi que certaines entités qui pourraient être considérées, sous un autre regard, comme des organisations.

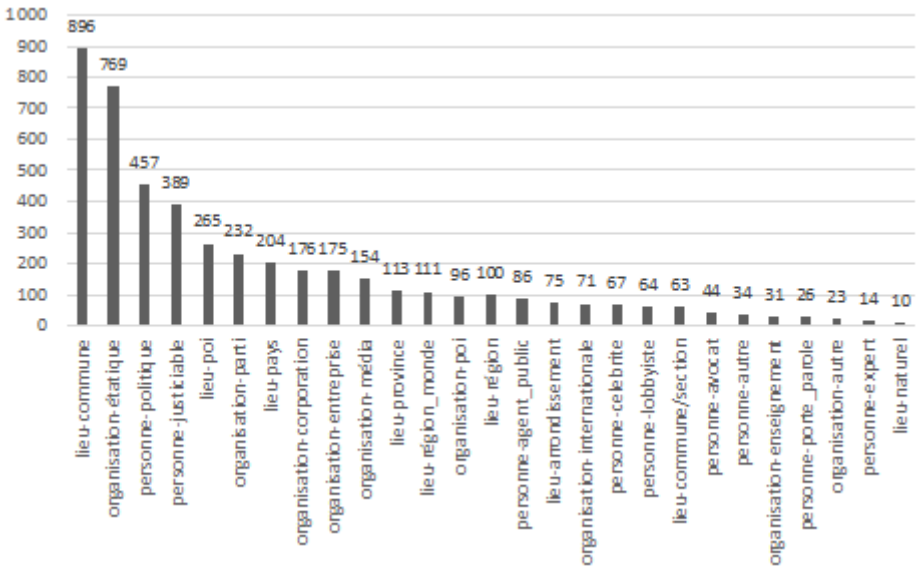


FIGURE 4.2.1. – Distribution des sous-classes d'entités

4.2.1. Répartition par éléments de texte

Aspects quantitatifs

Alors que la longueur des titres épouse une distribution normale (en forme de « cloche »), celles de chapeaux et de textes sont nettement asymétriques. Le rapport *lead*/texte va de 7 % (*lead* 14 fois plus court que le texte donc) à 218 % (deux fois *plus long*) [5]. Ce dernier cas de figure n'est pas exceptionnel. Plus d'une dépêche sur cinq (21.8 %) dans notre échantillon possède un *lead* dont la longueur est supérieure ou égale à celle du texte. C'est le plus souvent le cas de dépêches courtes dont le premier paragraphe offre l'essentiel du contenu. Outre le nombre d'entités nommées par éléments structurels de la dépêche, nous avons donc calculé également leur taux moyen par 1000 caractères.

Les résultats (Figure 4.2.2) confirment l'intuition selon laquelle le *lead* constitue une source privilégiée d'entités nommées. Si leur nombre absolu est deux fois infé-

5. Un article dépourvu de chapeau mis à part.

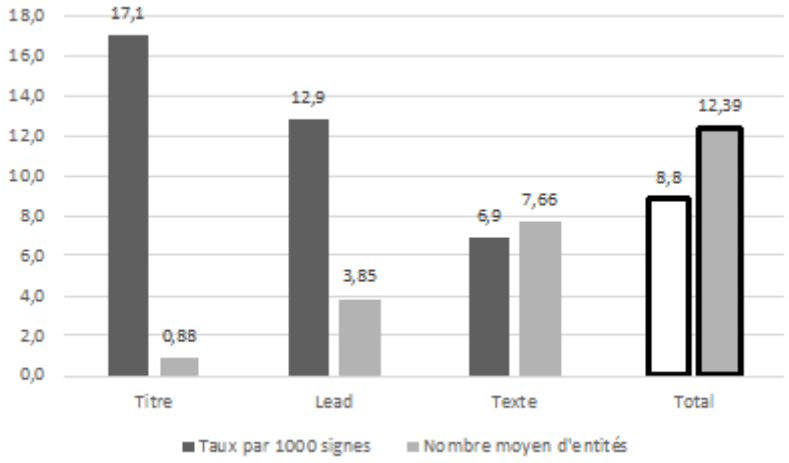


FIGURE 4.2.2. – Nombre et taux moyens d'entités nommées par élément de la dépêche

rieur à celui du texte (ce dernier étant, rappelons-le, en moyenne trois fois et demie plus long), leur taux par nombre de signes est de l'ordre du double.

Si l'on observe la répartition relative par grandes catégories d'entités (Figure 4.2.3), on constate une surreprésentation de noms de lieux dans les titres, au détriment des noms d'organisations. La remarque semble valoir également pour les chapeaux/-*leads*, quoique de manière moins marquée. Dans le cas de ces derniers, on remarquera surtout le faible taux relatif de noms de personnes comparé au corps du texte ($p<0.05$), compensé par un léger excédent de lieux et/ou d'organisations.

Les entités nommées contenues dans l'association « titre + chapeau » sont-elles, à elles seules, représentatives de l'ensemble de la dépêche? Pour tenter de le savoir, nous avons dédoublonné notre liste d'entités nommées, en ne conservant qu'un exemplaire unique de chacune par élément de la dépêche (titre/*lead*/texte). Nous avons ensuite comparé, pour chaque dépêche, les entités nommées contenues dans le corps de texte qui étaient déjà présentes dans le titre/chapeau, et vice-versa [6]. Nos

6. La méthode utilisée, semi-automatique, repose sur une « comparaison floue » (*Fuzzy Matching*). Il a ainsi été possible de faire correspondre des noms dont la graphie change entre le titre/*lead* et le corps de texte (par exemple « ministre Antoine » qui devient ensuite « André Antoine », ou « Bruxelles-Nord » et « gare du Nord »).

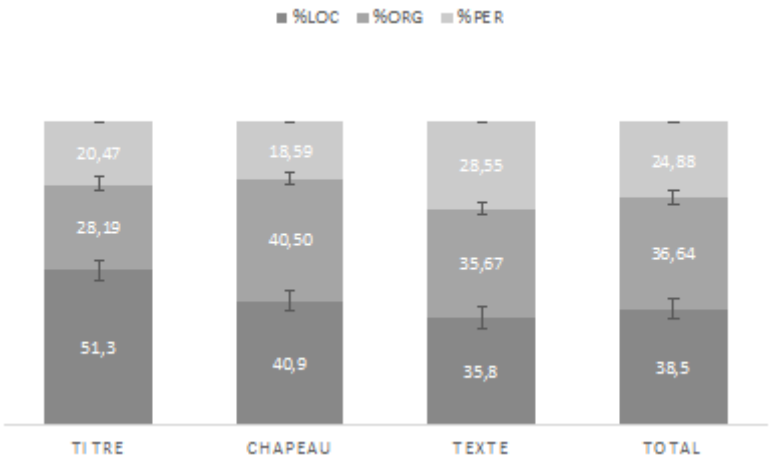


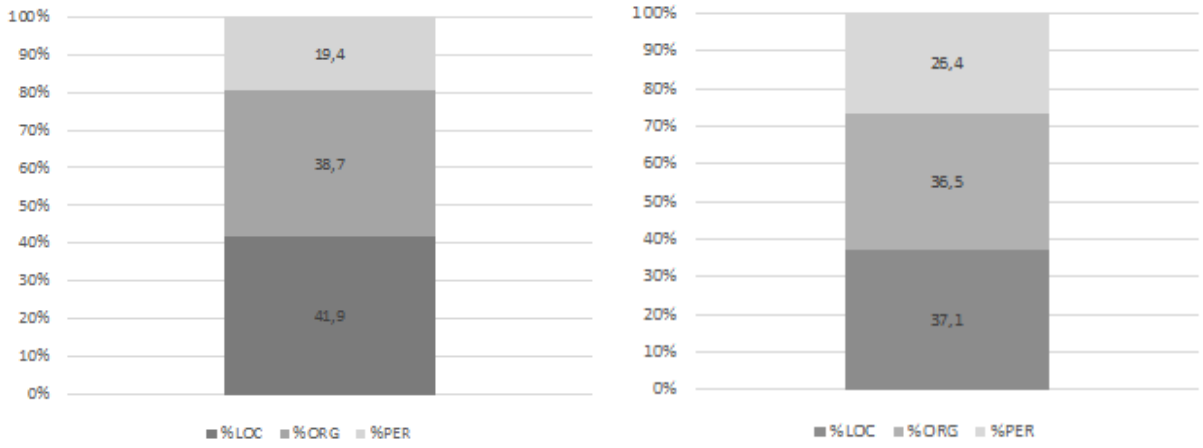
FIGURE 4.2.3. – Pourcentages par catégories d'entités

résultats montrent que 758 entités sur les 1712 (44.3 %) contenues dans les titres/-*leads* n'apparaissent plus par la suite dans le corps du texte. Ce taux monte à 47.2 % si on se limite aux entités contenues dans les *leads* seuls, à l'exclusion des titres. La répartition de ces entités en catégories ne montre pas de différences significatives entre celles contenues uniquement dans les titres/*leads* et celles mentionnées uniquement dans le corps de texte (Figure 4.2.4). La constatation vaut pour les sous-catégories, que nous n'afficherons pas ici.

Nous n'avons trouvé aucune corrélation évidente entre la longueur d'une dépêche et le nombre d'entités nommées présentes uniquement dans le titre et le *lead* ($R^2 = 0.0724$). Cela contredit l'intuition suivant laquelle plus la dépêche est courte, plus le *lead* est susceptible de contenir toute l'information.

Aspects qualitatifs

Nous avons déjà vu que dans un *lead* et un titre, le taux d'entités nommées par nombre de signes était environ deux fois plus élevé que celui du corps de texte. Mais qu'en est-il de la plus grande proximité avec la thématique centrale du texte? Répondre à cette question est indispensable pour valider notre première hypothèse. Cela impliquerait, idéalement, de définir avec précision ladite thématique, ce qui implique une large part de subjectivité. Nous n'entrerons pas ici dans le domaine



(a) Uniquement titres et *leads*

(b) Uniquement corps de texte

FIGURE 4.2.4. – Répartition des entités non redondantes

du *topic modelling*[7], qui nécessiterait un mémoire à lui seul, ou à tout le moins un chapitre entier. Le besoin s'en fait d'autant moins sentir que les thématiques abordées dans notre corpus de dépêches ont fait l'objet de deux annotations humaines. La première est due aux journalistes de *7sur7* eux-mêmes, qui ont pourvu chaque dépêche de mots-clés[8]. Nous avons déjà vu toutefois que ces mots-clés sont largement incohérents. On peut soupçonner qu'ils répondent davantage à une logique de SEO (*Search Engine Optimization*) que de classement taxonomique.

L'autre annotation humaine, tout aussi subjective, plus générale, mais aussi sans doute plus cohérente, est l'une des dix rubriques dans lesquelles nous avons rangé chaque dépêche lors de l'annotation (4.2).

Après avoir suivi plusieurs fausses pistes, nous ne sommes pas parvenus à trouver une méthode qui permette de comparer les deux de manière rigoureuse. L'examen des entités nommées titre/chapeau nous laisse pourtant penser plus que jamais que notre première hypothèse pourrait se révéler juste : on sent que le thème de certaines dépêches pourrait être « deviné » à l'aide de ces seuls mot-clés (Tableau 4.4). Mais les

7. Méthode informatique qui vise à découvrir automatiquement les thématiques « cachées » d'un texte, en examinant les mots qui le composent [Song 14].

8. Entre 1 et 11 mots-clés, avec une moyenne de 4.43 et une médiane de 4. Nous en avons compté quelque 365 différents.

méthodes que nous sommes pour l'heure capable de mettre en œuvre se révèlent insuffisantes pour confirmer cette intuition.

ID_Article	EN Titre/*Lead*
1	Bouge
2	cdH
2	PS
3	église Sainte-Croix
3	Ixelles
4	KB Lux
...	...
13	Vervoort
13	Région wallonne
13	PS
13	Sénat
13	Demotte
13	Magnette
13	Fédération Wallonie-Bruxelles
14	Plan Belgique

TABLE 4.4. – Exemples d'entités nommées uniques contenues dans les titres/*leads*

4.2.2. Entités reconnaissables

Notre seconde hypothèse portait sur l'apport possible d'une base de données constituée de noms de mandataires belges (Section 4.3) afin de désambiguïser les entités nommées de notre corpus. Pour la vérifier, il nous faut cette fois déterminer lesquelles de nos entités figurent dans l'une des grandes bases de connaissance classiques. La méthode que nous avons employée a déjà été définie à la section 3.3.

Résultats

Les méthodes que nous avons utilisées nous ont permis d'identifier 3372 entités (71.2 %). Nous ne sommes toutefois pas certain de leur exhaustivité. En termes d'entités uniques, le taux est de 1031 reconnaissances sur 1823 (56.6 %). Cela ne signifie pas que plus d'une entité sur deux dans notre corpus sera reconnaissable dans tous les cas par un logiciel performant. Mais, en tout cas, elles pourraient l'être, dès lors

qu'elles possèdent une fiche Wikipedia ou à tout le moins une ressource dans DBpedia ou Wikidata.

De notre point de vue, toutefois, les entités *non réconciliées* présentent un intérêt supérieur. Leur distribution en sous-types (par entités uniques) est présentée à la Figure 4.2.5.

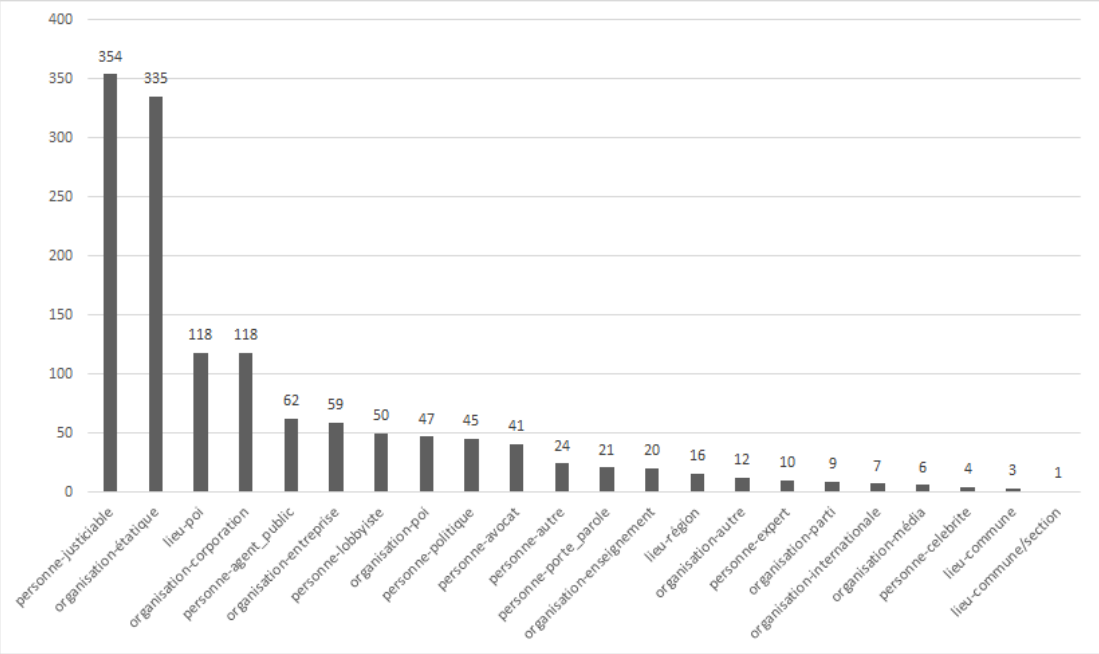


FIGURE 4.2.5. – Distribution des entités non reconnues, classées par sous-types

Ce graphique appelle quelques commentaires. Si la première place occupée par les sous-types `personne-justiciable` était attendue (il s'agit le plus souvent d'anonymes, dont le nom est souvent même réduit à des initiales), la forte représentation d'organisations étatiques peut surprendre. Elle s'explique à la fois par la nature du corpus et par notre méthode d'annotation. Les dépêches reprises par 7sur7 étant très axées sur les thèmes judiciaires et les faits-divers (voir 4.2), elles mentionnent donc très souvent des entités telles que « parquet de Bruxelles », « police de la zone Botha », « cour d'assises de Liège », « tribunal correctionnel de Namur », « pompiers de Huy », etc. Or, dans la plupart des cas, ces entités ne possèdent pas d'entrée spécifique dans Wikipedia. S'il existe bien des fiches consacrées à leur thématique [9], notre méthode

9. Tribunal correctionnel pourrait ainsi correspondre à la fiche « tribunal de Première instance (Belgique) » : https ://fr.wikipedia.org/wiki/Tribunal_de_premi%C3%A8re_instance_(Belgique)

d'annotation précoordonnée imposait de retrouver une entrée exacte dans les bases de connaissance interrogées. Le même raisonnement vaut pour les « CPAS », très souvent cités dans les dépêches. Mais reconnaître le terme générique « Centre public d'action sociale » ne suffisait pas selon le point de vue que nous avons adopté.

La catégorie `lieu-poi` a quant à elle souffert de notre choix d'inclure les noms de rues. Hormis dans le cas d'artères célèbres, il ne s'agit manifestement pas d'entités reconnaissables par un système basé sur Wikipedia [10].

La forte représentation du sous-type`organisation-corporation` nous a également étonné, puisqu'il s'agit dans la plupart des cas de syndicats connus. Ce faible taux de résolution peut s'expliquer par le recours fréquent dans la presse à des sigles et acronymes, ainsi que par la granularité très fine des entités mentionnées. Car s'il existe bien une fiche Wikipedia consacrée à la « Fédération générale du travail de Belgique (FGTB) », ses six centrales (SETCa, CGSP...) sont présentées dans le corps de l'article et ne possèdent pas de page propre. Pourtant, ce sont souvent ces dernières qui font l'actualité sociale sur le terrain, et donc dans la presse.

Les autres sous-types ne seront pas détaillés ici. Nous relèverons toutefois qu'en mettant de côté les justiciables, les `personne-autre` et les célébrités [11], le taux de reconnaissance des entités `PER` est de 53.5 % (193 reconnues sur 361 entités distinctes). Ce constat confirme l'intuition selon laquelle les grandes bases de connaissance internationales montrent rapidement leurs limites pour désambiguïser des noms dans des articles de presse centrés sur l'actualité nationale. Des bases de données plus spécialisées pourraient-elles permettre de reconnaître les termes en question ? A la section suivante, nous nous intéresserons à l'une d'elles, Cumuleo, focalisée sur les personnalités politiques et les mandataires belges.

4.3. Cumuleo

Depuis 2005, la Cour des comptes publie au Moniteur belge les déclarations de mandats, fonctions et professions exercés l'année précédente par tout mandataire

10. Du moins si la recherche se borne aux titres des pages, car ces noms de rue sont souvent cités dans le corps des articles Wikipedia.

11. Il s'agit en réalité d'une catégorie fourre-tout destinée à ranger les personnalités étrangères qui ont fait l'actualité en Belgique, sans que cela n'implique une grande notoriété.

public. Il s'agit pour ces derniers d'une obligation légale [12]. Ces déclarations, ainsi que le nom des personnes en défaut, sont réunies dans un document PDF d'un millier de pages, mis en ligne chaque année à la mi-août [13]. En pleine torpeur estivale, ces publications sont devenues un «marronnier» pour la presse, friande de classements des « plus grands cumulards ». Le travail des journalistes est considérablement facilité depuis qu'un webdesigner indépendant, Christophe Van Gheluwe, a entrepris de transformer ces déclarations en une base de données consultable librement sur le web et mise à jour chaque année [14]. Cumuleo [15], du nom du site qui lui est dédié, ne se contente pas de présenter les documents de la Cour des comptes dans un format unifié, structuré et accessible. Avec l'aide d'une traductrice rémunérée, le concepteur du site s'est également livré à un important travail de traduction des noms de fonctions, puisque chaque mandataire les fournit dans une seule des trois langues nationales (français, néerlandais, allemand). Des informations additionnelles, souvent puisées dans la presse, ont été ajoutées afin d'étoffer le profil de chaque mandataire. Notamment son rôle linguistique, l'éventuel parti politique dont il se réclame, ou encore les mandats qu'il exerce de notoriété publique et qu'il aurait omis de signaler... Ces données additionnelles permettent à Cumuleo de publier régulièrement des statistiques et des analyses détaillées.

Nous avons pu obtenir un « *dump* » SQL de Cumuleo, à savoir une photographie de la base de données telle qu'elle se présentait le 9 décembre 2015. Une version simplifiée et dénormalisée de son schéma est reproduite à la Figure 4.3.1, ce qui donne une idée des champs disponibles. A titre de comparaison, la Figure 4.3.2 montre un extrait des champs disponibles dans les documents originaux de la Cour des comptes (édition 2015, mandats 2014).

12. Inscrite dès 1995 dans une loi ordinaire et une loi spéciale, qui en énonçaient les grands principes, cette obligation n'est devenue effective qu'après les lois du 26 juin 2004 qui précisent ses modalités [Mendola 05].

13. L'édition 2015 est accessible à l'adresse http ://www.ejustice.just.fgov.be/mopdf/2015/08/14_1.pdf

14. La prochaine mise à jour est prévue le 12 août 2016, soit quelques jours après la remise de ce travail.

15. www.cumuleo.be

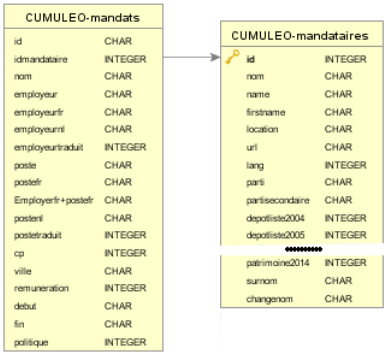


FIGURE 4.3.1. – Schéma simplifié de Cumuleo

Aelgoet Jean-Michel					
	Froidchapelle (Commune de)	Conseiller communal	Rémunéré		
	AIESH	Administrateur	Rémunéré		
	AIESH	Membre du comité de gestion	Non rémunéré	30/10/2014	
	AIMAC	Administrateur	Non rémunéré		
	INTERSUD	Administrateur	Rémunéré		
	Foyer culturel communal	Administrateur	Non rémunéré		
Aelterman Guy					
	Vlaamse Regering, Onderwijs, Jeugd, Gelijke Kansen en Brussel	Kabinetschef	Bezoldigd		24/07/2014
	Qanu	Bestuurder	Bezoldigd		

FIGURE 4.3.2. – Extrait de la déclaration de mandats 2015 telle que publiée en PDF au Moniteur belge

4.3.1. Aperçu général

Mandataires

La base de données contient 15 406 noms de mandataires, chacun associé à un ID unique. Il s'agit de toute personne qui a déclaré en Belgique au moins un mandat public entre 2004 et 2014. Tous ne sont donc pas des « politiques » au sens strict. En dépit des efforts fournis par le concepteur de la base de données, 3436 noms (22 %) n'ont pu être associés à une formation politique (Figure 4.3.3), le plus probablement parce qu'ils ne se sont jamais réclamé d'aucune [16].

16. Si l'on s'en tient au compte brut du nombre de mandats différents, l'une des mandataires les plus dotées est Marie-Hélène Ska, secrétaire générale du syndicat CSC. Il ne s'agit pourtant pas d'une personnalité politique ni d'une élue. La plupart de ses fonctions déclarées dérivent presque naturellement de la principale. Il en va souvent de même pour les bourgmestres ou députés provinciaux, désignés ès qualités administrateurs non rémunérés d'une kyrielle d'ASBL.

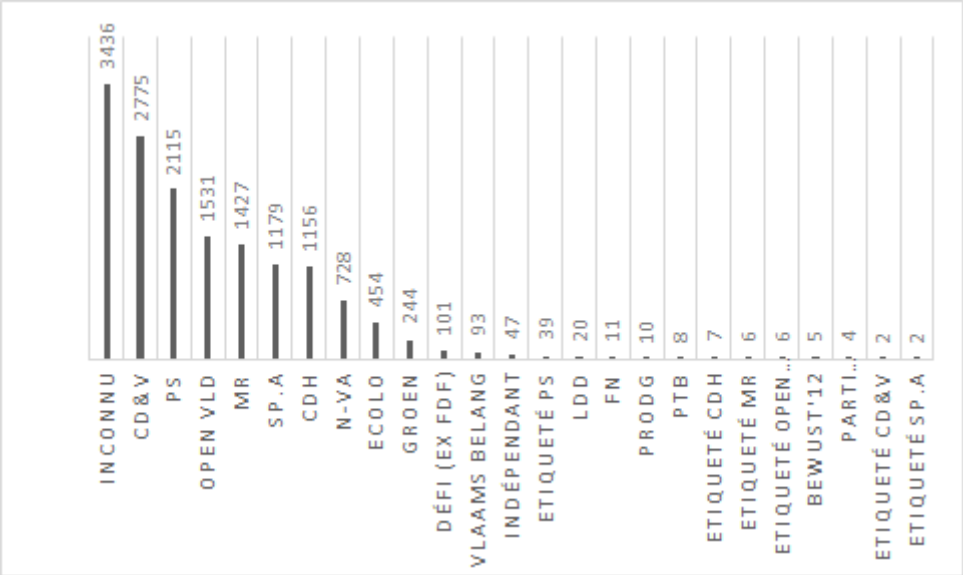


FIGURE 4.3.3. – Répartition des mandataires par formation politique

4.4. Réconciliation avec le corpus

En dépit de nos nombreuses tentatives, nous n'avons pas trouvé de moyen satisfaisant pour réconcilier une telle quantité de noms avec une grande base de connaissance. Les méthodes déjà testées sur notre corpus de dépêches n'ont permis de résoudre que 1578 noms, soit un peu plus de 10 %. L'une des difficultés du « matching » avec DBpedia ou Wikipedia est la gestion de l'homonymie. Celle-ci est beaucoup plus difficile à traiter de manière automatisée lorsque les seuls signes distinctifs sont une nationalité, un lieu de résidence [17] et une activité parfois accessoire [18].

En outre, c'est surtout l'inverse qui nous occupe : vérifier dans quelle mesure ces noms de personnes, d'organisations ou de lieux peuvent être appariés avec les entités non reconnues de notre corpus.

17. Rarement mentionné dans le « résumé » d'une page, contrairement au lieu de naissance.
18. Même si une personne possède sa propre fiche sur Wikipedia, cette dernière n'indique pas forcément sa qualité de conseiller communal ou d'administrateur d'une entité publique.

Noms de mandataires

La seule méthode à notre disposition pour comparer les deux listes consiste à effectuer un « fuzzy matching » entre elles [19]. Dans ce cas-ci également, la gestion de l'homonymie pose problème. Mais s'agissant de quelques centaines de noms, elle peut être traitée manuellement. La principale difficulté réside cette fois dans les discordances de graphie entre le nom « public » d'une personne (utilisé dans la presse) et son nom officiel (utilisé dans les documents de la Cour des comptes). On connaît le problème classique des noms d'épouse [20]. Celui des diminutifs est particulièrement prégnant pour les personnalités flamandes. De nombreuses personnes connues publiquement sous le prénom de « Tine » ou « Huub » reprennent dans Cumuleo leur prénom de baptême [21]. La base de données possède certes une colonne `surnom`, qui répertorie différentes appellations d'un même mandataire, mais seules les personnalités renommées (et dont le surnom est connu) peuvent en faire l'objet. Ce n'est pas le cas, par exemple, pour le secrétaire fédéral de la FGTB Jean-François Tamellini [22], mentionné dans une dépêche sous le prénom de François [23].

Hors célébrités et justiciables, la méthode nous a toutefois permis de réconcilier 17.03 % (31 sur 182) des noms de personnes qui ne disposent pas d'une fiche Wikipedia. Leur distribution par sous-classes est présentée à la Figure 4.4.1. Sans surprise, les personnalités politiques sont largement en tête, suivies des agents publics et des acteurs sociaux.

Noms de lieux et d'organisations

Notre version de Cumuleo contient un total de 150 711 mandats, chacun composé de l'association entre un employeur (par exemple « Ville de Mons ») et d'une

19. Opération effectuée en créant un service de réconciliation dans Open Refine, à l'aide du programme *Reconcile-CSV* déjà évoqué au chapitre 5.

20. Comme dans le cas de feu Anne-Marie Lizin, dont le nom officiel est « Anne-Marie Vanderspeeten » : https ://www.cumuleo.be/mandataire/11354-anne-marie-vanderspeeten.php

21. C'est le cas par exemple du bourgmestre de Fourons, Huub Broers, dont même la fiche Wikipedia en néerlandais ne mentionne pas qu'il se prénomme en réalité Hubert : https ://www.cumuleo.be/mandataire/8411-hubert-broers.php

22. https ://www.cumuleo.be/mandataire/15088-jean-francois-tamellini.php

23. Mais il s'agit peut être simplement d'une erreur du journaliste. Même ceux d'une agence en commettent. Le député socialiste Philippe Blanchart est ainsi rebaptisé Blanchert à deux reprises dans une dépêche.

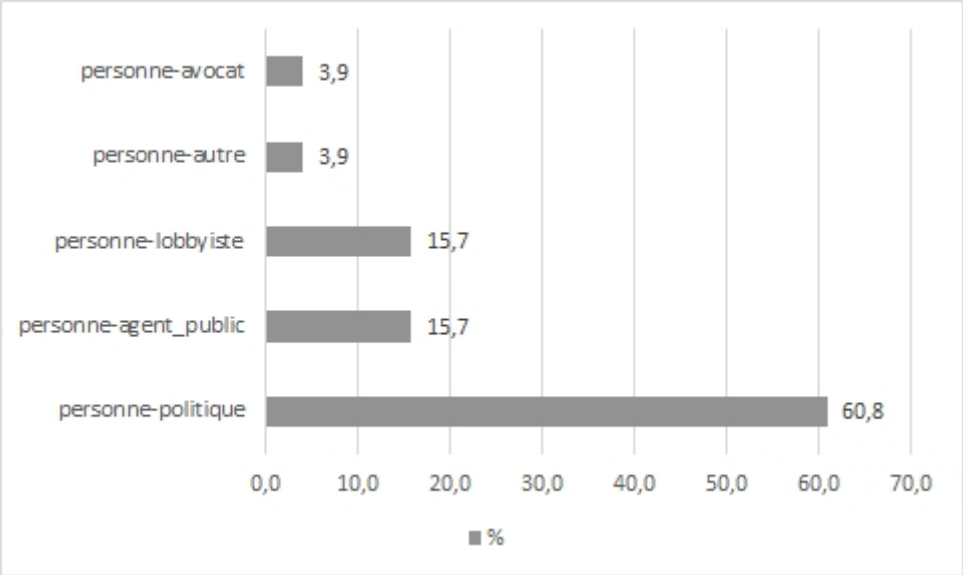


FIGURE 4.4.1. – Distribution des PER reconnus par Cumuleo seul

fonction (par exemple « échevin » ou « conseiller communal »). Mais le nombre de lieux ou d'organisations dans lesquelles ils s'exercent est difficile à estimer. Contrairement aux noms de personnes, qui sont normalisés, la graphie de « l'employeur » et du poste occupé varient considérablement. Les mandataires sont libres de l'écrire sous la forme qu'ils souhaitent. Au moins trois colonnes des déclarations de mandats ont donc été remplies par des milliers de personnes différentes, en trois langues, avec toute la richesse, la liberté et la diversité qu'offre le langage naturel. Nous avons tenté de normaliser au maximum la colonne `EmployeurFR`, à l'aide des algorithmes de *clustering* que propose Open Refine. L'utilisation successive des algorithmes *Key collision fingerprint* et *Ngram-fingerprint* nous a permis de réduire leur nombre à 61 070 entités *a priori* distinctes (Tableau 4.5).

	VALEURS « UNIQUES »
EMPLOYEURFR	61 070
POSTEFR	8947
POSTEFR+' '+EMPLOYEURFR	115 413

TABLE 4.5. – Nombre de fonctions distinctes après normalisation sommaire

Ces deux méthodes permettent de réconcilier rapidement les légères différences

de graphies, notamment le mélange de majuscules et de minuscules. On peut leur faire en général une confiance quasi aveugle. Il en va tout autrement pour l'algorithme *Nearest neighbor Levenshtein*, très utile pour fusionner des appellations proches, notamment les formes masculines et féminines (« président» et «présidente »). Son utilisation réclame cependant beaucoup de prudence, un temps de calcul très supérieur aux deux premiers et, dans tous les cas, une vérification humaine [24]. La charge de travail qu'impliquerait une normalisation complète de la colonne est par conséquent rédhibitoire [25]. Nous nous sommes donc limité à utiliser un fichier peu normalisé.

En dépit de cette imperfection, la réconciliation avec les entités ORG non résolues s'est déroulée de manière satisfaisante, puisque 171 sur 355 ont pu être retrouvées. Ce taux de 48.17 % est près de trois fois supérieur à celui des noms de personnes.

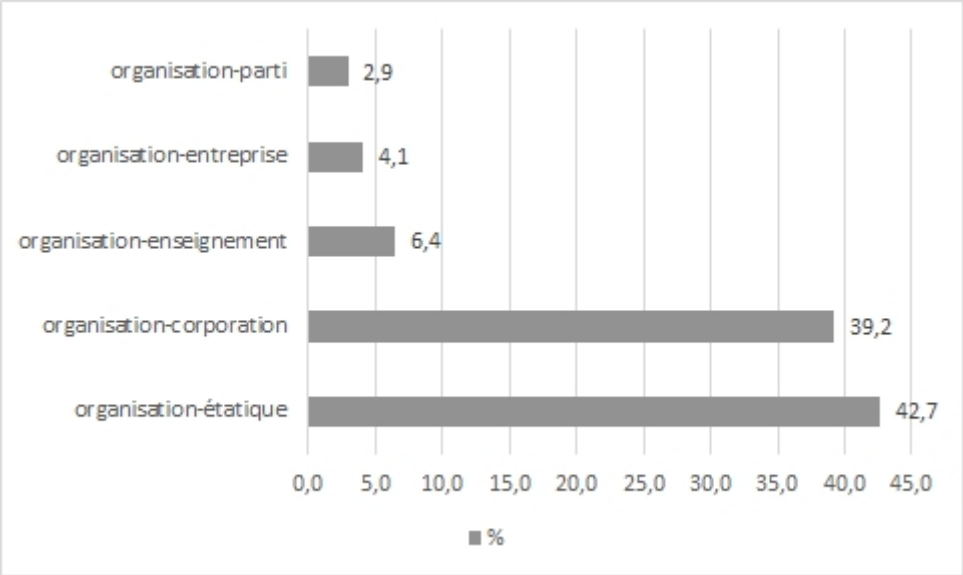


TABLE 4.6. – Distribution des ORG reconnus par Cumuleo seul

La tentative de réconciliation avec les noms de lieux non résolus s'est quant à elle révélée décevante (3 réussites sur 119). Malgré la présence de certaines organisations, rappelons que cette catégorie est en grande partie constituée de noms de rues.

24. Son cousin *Nearest neighbor PPM* peut quant à lui faire d'un «vice-président» un «président», voire d'un « fonctionnaire » un « actionnaire »...

25. A fortiori celle des noms de fonctions, qui entrainerait des explosions combinatoires. Nous avions pourtant annoté ces fonctions dans le corpus, sans en faire usage.

Conclusion

Rappel des objectifs

Les deux hypothèses qui ont guidé nos travaux puisaient leur source dans de pures intuitions, issues de notre expérience journalistique. Pour avoir rédigé quantité de "chapeaux" et de leads, nous sentions que ces paragraphes introductifs devaient posséder quelque caractéristique intéressante au regard des sujets qui nous occupaient, à savoir les entités nommées et le web sémantique. Or nulle production journalistique ne possède de lead plus rigoureux qu'une dépêche d'agence.

C'est également de nos travaux passés qu'est née cette idée, ou plutôt cet embryon d'idée, suivant laquelle bon nombre de personnes, d'organisations ou de lieux cités dans des articles de presse, et pas uniquement dans des sujets politiques, pouvaient se retrouver dans la base de données Cumuleo.be. Confirmer ou infirmer ces hypothèses impliquait d'examiner un corpus de dépêches belges d'actualité générale. Seules celles de l'agence Belga répondaient à ces deux critères.

Avant d'entamer leur analyse, nous devions poser des fondements théoriques solides. C'était l'objet de la première partie de ce mémoire. Nous avons ainsi parcouru au moins un quart de siècle de recherches, qui vont de la naissance du web aux réseaux sociaux, des tâtonnements de MUC 6, quand fut forgé le concept d'entité nommée, aux recherches actuelles sur leur identification dans des tweets.

Reste que les entités nommées sont un concept pragmatique, né de la recherche

opérationnelle et guidé avant tout par des préoccupations utilitaires. A tel point qu'en 2008, alors que le terme avait douze ans d'âge, la linguiste française Maud Ehrmann éprouva le besoin de consacrer une thèse à sa définition. Aborder ces concepts flous par le seul biais de la théorie ne permet guère d'en saisir toutes les nuances. C'est en les brassant à pleine pâte, lors d'une annotation manuelle de corpus, que leur nature enchevêtrée et contradictoire s'expose au grand jour.

Comprendre implique aussi souvent de compter, scinder, regrouper, hiérarchiser, classer. Nous nous y sommes attelé dans la seconde partie de ce travail.

Bilan des travaux effectués

Compter et classer les entités nommées dans cent dépêches eût été trop peu ; dans mille, excessif. Il nous a fallu trouver un juste milieu, finalement estimé à un peu moins de 400 articles. Encore fallait-il obtenir ces textes, en nous assurant de leur représentativité au regard des centaines de milliers d'autres que nous n'examinerions jamais. C'est ainsi que nous avons développé quelques outils et méthodes pour récupérer et échantillonner un corpus. Il nous a fallu également choisir un « fournisseur » de dépêches, puisqu'accéder directement au fil de l'agence Belga, et surtout à son historique, présentait de lourdes difficultés. Nous avons ainsi estimé que le site web 7sur7.be, dont la production est constituée à plus de 90 % de dépêches, pouvait constituer un succédané du fil Belga.

Le corpus réuni, nous devions l'annoter ; et pour cela définir ce qui méritait de l'être ou non, ce qui pouvait éclairer l'analyse et ce qui au contraire l'embrouillerait. Nous avons ainsi mis au point, pour l'occasion, une hiérarchie d'entités nommées à deux niveaux, axée sur les grandes catégories classiques que sont les noms de personnes, de lieux et d'organisations.

Les résultats bruts de l'annotation ne se prêtaient guère aux analyses que nous souhaitions mener. Du moins pas sans un solide nettoyage au cours duquel nous avons largement sollicité le logiciel Open Refine. C'est grâce à lui également, avec l'aide d'une méthode propre, que nous avons vérifié si les entités extraites du corpus existaient dans les grandes bases de connaissance issues de Wikipédia, ou à défaut dans l'encyclopédie elle-même.

Apports, limites et perspectives

Le dernier chapitre de ce mémoire répond à quelques questions, mais en soulève plus encore. Nous avons vu ainsi que le taux d'entités nommées semble significativement plus important dans le lead d'une dépêche que dans son corps de texte. Toutefois, nous étions mal outillé conceptuellement pour tirer sur ce fil et pousser les conclusions un pas plus loin : les entités nommées contenues dans le lead sont-elles finalement plus représentatives que les autres du thème central de la dépêche ? Nous n'avons pas pu répondre à cette question, qui impliquerait un important investissement dans l'étude de domaines tels que le topic modelling. Si notre première intuition de départ est sortie renforcée des travaux d'annotation, elle ne repose toujours sur aucune démonstration réfutable et reproductible.

La comparaison entre les entités du corpus et celles contenues dans Cumuleo a livré des résultats intéressants. Nous avons vu ainsi que la base de données de mandats et de mandataires pouvait aider, dans une certaine mesure, à désambiguïser les noms inconnus de Wikipedia. Mais alors que nous attendions surtout des résultats en matière de noms de personnes, ce sont les organisations qui ont fourni le plus large contingent de réconciliations réussies. Ce sont pourtant, au sein de Cumuleo, les catégories dont la graphie est parmi les moins normalisées. Peut-être cette diversité est-elle justement un avantage : en multipliant les dénominations pour une même organisation, on multiplie également ses chances de correspondre à la forme de surface utilisée dans les articles de presse. D'autant que, là où DBpedia ou Wikipedia emploient des dénominations officielles, Cumuleo répertorie des noms plus usuels. S'il ne s'agit ici encore que d'une intuition, elle mériterait à notre sens d'être vérifiée sur un corpus plus vaste. Les noms de fonctions, que nous avons renoncé à examiner, mériteraient d'être traités dans la foulée.

De tels travaux réclameraient une méthode plus fine et, peut-être, un ensemble de textes plus hétérogène. Car ce travail a également instillé en nous un doute sur la représentativité de notre corpus, dans lequel l'actualité judiciaire et policière occupe une place très (trop ?) importante. Les résultats auraient-ils été identiques avec un corpus de dépêches régionales ?

Il a également mis en lumière les failles de notre méthode d'annotation, dont certains choix se sont révélés soit inutiles (l'annotation des noms de rues), soit dis-

cutables (annoter les entités de manière précoordonnée, ne pas tenir compte de celles dont la forme de surface ne permet pas une identification en dehors de tout contexte). La méthode que nous avons adoptée pour réconcilier les entités nommées avec des bases de connaissance présente aussi des défauts, qui nous ont imposé bien des adaptations manuelles. Nous avons par ailleurs sous-estimé la difficulté que représente une recherche automatisée dans Wikipedia.

Ces imperfections méthodologiques sont toutefois riches d'enseignements. Commettre quelques faux pas sur un corpus limité, et dans un projet qui mettait en œuvre une seule base de données, nous a permis d'entrevoir une méthodologie plus rigoureuse. Une fois développée, celle-ci pourrait permettre d'aborder des bases de données plus nombreuses, plus diversifiées et plus vastes.

Annexes

Annexe A

Scripts Python

Listing A.1 – Extraction et échantillonnage du corpus 7sur7

```
#!/usr/bin/env python2
# -*- coding: utf-8 -*-

import re
import unicodecsv as csv #Le module csv de la standard
   library pose problèmes avec des caractères accentués
from bs4 import BeautifulSoup
from urllib2 import urlopen
from datetime import datetime,timedelta
from random import sample


# URL de départ
base_url = \
```

```
    "http://www.7sur7.be/7s7/fr/1481/archives/
       integration/nmc/frameset/archive/archiveDay.dhtml
       ?archiveDay="

################################################

def create_url_days():
    """
    crée une série de dates, puis d'url, au format 7sur7
        Yeardaymonth
    """
    date_depart = raw_input("Introduisez la date de
       départ au format 31-12-2000: ")
    date_fin = raw_input("Introduisez la date de fin au
       format 31-12-2000: ")
    start = datetime.strptime(date_depart,"%d-%m-%Y")
    end = datetime.strptime(date_fin,"%d-%m-%Y")
    step = timedelta(days=1)
    dates = []
    while start <= end:
        dates.append(start.date().strftime('%Y%m%d'))
        start += step

    url_dates = []
    for jour in dates:
        url_dates.append(base_url + jour)
    return url_dates

################################################

def create_bs4_object(url):
    """
    crée un objet BeautifulSoup à partir d'un URL
```

```
    """
    html = urlopen(url).read()
    soup = BeautifulSoup(html,"lxml")
    return soup

################################################

def get_url_articles(url_days):
    """
    extrait les url des articles Belgique à partir d'une
        page calendrier 7sur7
    """
    soup = create_bs4_object(url_days)
    links = []
    for liens in soup.find_all(href=re.compile(".*\/(
       belgique|bruxelles)\/article\/.*",re.IGNORECASE))
       :
        links.append(liens.get("href"))
    print links
    return links

################################################

def get_text_articles(links):
    """
    Extrait les informations d'un article à partir de
       son URL et les place dans un dictionnaire
    """
    Results = dict(URL='Not Found',infos='Not Found',
       lead='Not Found',text='Not Found', metadata='Not
       Found')
    soup = create_bs4_object(links)
```

```
    try:
        Results['URL'] = links
        Results['infos'] = soup.find("span",class_="
           author").getText()
        Results['lead'] = soup.find("p",class_="intro").
           getText()
        Results['text'] = soup.find("p",class_="intro").
           nextSibling.nextSibling.getText()
        Results['metadata'] = soup.find("meta", {"
           property":"article:tag"})['content']

        return Results

    except:
        # Renvoie le dictionnaire initial
        return Results

#####################################################

def export_dict_list_to_csv(data,filename):
    """
    Exporte une liste de dictionnaires en CSV
    """
    with open(filename,'wb') as f:
        # Les dictionnaires de la liste doivent avoir
           les mêmes clés
        headers = sorted([k for k,v in data[0].items()])
        csv_data = [headers]

        for d in data:
            csv_data.append([d[h] for h in headers])

        writer = csv.writer(f)
```

```
        writer.writerows(csv_data)

########################################
#  Test  ###############################
########################################

if __name__ == '__main__':
    """
    Récupère les informations de chaque article
    publié au cours d'une période données
    """
    liste_liens = []
    for elements in create_url_days():
        for liens in get_url_articles(elements):
            liste_liens.append(liens)

    # sauvegarde la liste de liens dans un fichier
    with open("liens.csv",mode="wt") as myfile:
        myfile.write('\n'.join(str(liens) for liens in
           liste_liens))

    # Echantillonne 384 lignes de liens Belgique et
       récupère les infos
    sample = sample(liste_liens,384)
    sample_text = []
    for liens in sample:
        sample_text.append(get_text_articles(liens))
    print sample_text

    # sauvegarde l'échantillon dans un fichier CSV
    export_dict_list_to_csv(sample_text,"sample_brut.csv
       ")
```

Listing A.2 – Recherche automatique dans Wikipedia

```
#!/usr/bin/env python2
# -*- coding: utf-8 -*-

"""
Ce script (horriblement rédigé, mais fonctionnel)
   récupère une liste de noms dans un fichier texte et
   vérifie d'abord s'il existent dans #Wikipedia.fr,
   puis dans Wikipedia.nl. Le résultat sera nettoyé dans
    Open Refine.
"""

import codecs
import wikipedia as wiki  # voir https://pypi.python.org
   /pypi/wikipedia/
import unicodecsv as csv

def append_csv(value, liste):
    liste.append(
        str(count) + "|||" + "Find" + "|||" + wiki.page(
           value, auto_suggest=False).title
        + "|||" + wiki.page(value,auto_suggest=False).
           url + \
        "|||" + wiki.page(value,auto_suggest=False).
           summary)

with codecs.open('C:/Users/Ettore2/Desktop/
   mandataires_cumuleo.tsv', 'r', encoding='utf-8') as f
   :
    liste = f.read().splitlines()

lines = []
```

```
count=1

for names in liste:
    try:
        append_csv(names, lines)
    except:
        try:
            wiki.set_lang("nl")
            append_csv(names, lines)
        except:
            lines.append( str(count) + "|||" + "notFind"
                + "|||" + names)
    print count, names
    count+=1

with open('C:/Users/Ettore2/Desktop/resultats.csv', 'wb'
   ) as file:
    writer = csv.writer(file, delimiter='\t', quoting=
       csv.QUOTE_ALL, lineterminator='\n')
    for val in lines:
        writer.writerow([val])
```

Annexe B

Opérations dans Open Refine

Listing B.1 – Code Jython : créer une requête SPARQL pour chaque élément d'une colonne

```
if value is not None:
    return """
SELECT DISTINCT ?entity ?country ?score1
WHERE {?entity ?p ?label.
OPTIONAL {?entity <http://dbpedia.org/ontology/country> ?country}
FILTER exists{?entity <http://dbpedia.org/property/mayor> ?m}
?label <bif:contains> "'+%s'"
OPTION(score ?score1).
FILTER (lang(?label) = "fr" || lang(?label) = "nl")
FILTER (?p=<http://www.w3.org/2000/01/rdf-schema#label>).
FILTER isIRI(?entity).}
ORDER BY desc(?score1)
LIMIT 6
""" %value
```

Listing B.2 – Code GREL : transformer une requête SPARQL en requête HTTP

```
if (isNotNull(value),
"http://dbpedia.org/sparql?defaultgraphuri=http%3A%2F%2Fdbpedia.org&query="
+
value.escape('url')
+
"&format=application%2Fsparqlresults%2Bjson", null)
```

Listing B.3 – Code GREL : extraire les URIs

```
forEach(value.parseJson().results.bindings, e, e.entity.value)
    .uniques()
    .join('||')
    .unescape('url')
```

Annexe C

Exemple de requête SPARQL

Listing C.1 – Retrouver des hommes politiques belges dans Wikidata

```
politician (Q82955)
has occupation (P106)
country of citizenship (P27)
Belgium (Q31)
Position Held (P39)
Wikipedia (?site)

#selectionner des gens qui ont comme activité politicien et
#comme nationalité "Belgique". Retrouver leur "position held" et
#leur page Wikipedia
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wikibase: <http://wikiba.se/ontology#>
PREFIX p: <http://www.wikidata.org/prop/>
PREFIX v: <http://www.wikidata.org/prop/statement/>
PREFIX q: <http://www.wikidata.org/prop/qualifier/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX schema: <http://schema.org/>
```

```
SELECT DISTINCT ?p ?pLabel ?w ?posLabel ?site WHERE {
    ?p wdt:P106 wd:Q82955 .
    ?p wdt:P27 wd:Q31 .
    ?p wdt:P39 ?pos .
          OPTIONAL { ?site schema:about ?p . ?site schema:inLanguage "fr" .}
          SERVICE wikibase:label {
     bd:serviceParam wikibase:language "fr" .
   }
}
```

Bibliographie

[ACE 08] ACE. *English Annotation Guidelines for Entities, Version 6.6.* Linguistic Data Consortium, 2008. www

[Al-Rfou 15] Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi & Steven Skiena. *Polyglot-NER : Massive multilingual named entity recognition.* In Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada. SIAM, 2015.

[Artiles 07] Javier Artiles, Julio Gonzalo & Satoshi Sekine. *The SemEval-2007 WePS Evaluation.* Proceedings of SemEval, 2007.

[Auer 07] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak & Zachary Ives. Dbpedia : A nucleus for a web of open data. Springer, 2007.

[Azpeitia 14] Andoni Azpeitia, Montse Cuadros, Seán Gaines & German Rigau. *NERC-fr : Supervised Named Entity Recognition for French.* In Text, Speech and Dialogue, pages 158–165. Springer, 2014.

[Barlett 01] James E Barlett, Joe W Kotrlik & Chadwick C Higgins. *Organizational research : Determining appropriate sample size in survey research.* Information technology, learning, and performance journal, vol. 19, no. 1, page 43, 2001.

[Bates 02] Marcia Bates. *After the Dot-Bomb : Getting Web Information Retrieval Right This Time*. First Monday, vol. 7, no. 7, 2002. www

[Béchet 11] Frédéric Béchet, Benoît Sagot & Rosa Stern. *Coopération de méthodes statistiques et symboliques pour l'adaptation non-supervisée d'un système d'étiquetage en entités nommées*. In TALN'2011-Traitement Automatique des Langues Naturelles, 2011.

[Bergman 01] Michael K Bergman. *The deep web : surfacing hidden value*. Journal of electronic publishing, vol. 7, no. 1, 2001.

[Berners-Lee 89] Tim Berners-Lee. *Information management : A proposal*. 1989. www

[Berners-Lee 90] Tim Berners-Lee & Robert Cailliau. *WorldWideWeb : Proposal for a HyperText project*. 1990. www

[Berners-Lee 99] Tim Berners-Lee, Mark Fischetti & Michael L Foreword By-Dertouzos. Weaving the web : The original design and ultimate destiny of the world wide web by its inventor. HarperInformation, 1999.

[Berners-Lee 01] Tim Berners-Lee, James Hendler, Ora Lassila*et al. The semantic web*. Scientific american, vol. 284, no. 5, pages 28–37, 2001.

[Berners-Lee 06] Tim Berners-Lee. *Design issues : Linked data*, 2006. https://www.w3.org/DesignIssues/LinkedData.html

[Bird 09] Steven Bird, Ewan Klein & Edward Loper. Natural language processing with python. " O'Reilly Media, Inc.", 2009. www

[Bizer 09] Christian Bizer, T Heath & T Berners-Lee. *Linked data-the story so far*. International journal on Semantic Web and Information Systems, vol. 5, no. 3, pages 1–22, 2009. www

[Blume 05] Matthias Blume. *Automatic entity disambiguation : Benefits to ner, relation extraction, link analysis, and inference*. In International Conference on Intelligence Analysis, 2005.

[Bollacker 07] Kurt Bollacker, Patrick Tufts, Tomi Pierce & Robert Cook. *A platform for scalable, collaborative, structured information in-*

tegration. In Intl. Workshop on Information Integration on the Web (IIWeb'07), 2007.

[Bollacker 08] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge & Jamie Taylor. *Freebase : a collaboratively created graph database for structuring human knowledge*. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pages 1247–1250. ACM, 2008.

[Borkowski 66] Casimir George Borkowski. A system for automatic recognition of names of persons in newspaper texts. IBM Watson Research Center, 1966.

[Bosch 16] Antal Bosch, Toine Bogers & Maurice Kunder. *Estimating search engine index size variability : a 9-year longitudinal study*. Scientometrics, vol. 107, no. 2, pages 839–856, 2016.

[Boydens 11] Isabelle Boydens, Arnaud Hulstaert & D Van Dromme. *Gestion intégrée des anomalies*. Série des" deliverables" de la Section Recherche-Smals, vol. 2011, 2011.

[Boydens 15] Isabelle Boydens. *Documentologie*. 2015.

[Brodie 86] Michael L Brodie & John Mylopoulos. *Knowledge bases vs databases*. In On Knowledge Base Management Systems, pages 83–86. Springer, 1986.

[Carletta 96] Jean Carletta. *Assessing agreement on classification tasks : the kappa statistic*. Computational linguistics, vol. 22, no. 2, pages 249–254, 1996.

[Chinchor 91] Nancy Chinchor. *MUC-3 Evaluation Metrics*. In Proceedings of the 3rd Conference on Message Understanding, MUC3 '91, pages 17–24, Stroudsburg, PA, USA, 1991. Association for Computational Linguistics. www

[Chinchor 93] Nancy Chinchor, David D Lewis & Lynette Hirschman. *Evaluating message understanding systems : an analysis of the third message understanding conference (MUC-3)*. Computational linguistics, vol. 19, no. 3, pages 409–449, 1993.

[Chinchor 95] Nancy Chinchor. *Four scorers and seven years ago : the scoring method for MUC-6*. In Proceedings of the 6th conference on Message understanding, pages 33–38. Association for Computational Linguistics, 1995.

[Ciaramita 05] Massimiliano Ciaramita & Yasemin Altun. *Named-entity recognition in novel domains with external lexical knowledge*. In Proceedings of the NIPS Workshop on Advances in Structured Learning for Text and Speech Processing, 2005.

[Cochran 77] William G Cochran. Sampling techniques, 3rd ed. John Wiley & Sons, 1977.

[Cohen 60] Jacob Cohen. *A coefficient of agreement for nominal scales*. Educational and Psychological Measurement, 1960.

[Cucerzan 99] Silviu Cucerzan & David Yarowsky. *Language independent named entity recognition combining morphological and contextual evidence*. In Proceedings of the 1999 Joint SIGDAT Conference on EMNLP and VLC, pages 90–99, 1999.

[Cunningham 05] Hamish Cunningham. *Information extraction, automatic*. Encyclopedia of language and linguistics„ pages 665–677, 2005.

[Cunningham 14] H Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Ian Roberts*et al. Developing Language Processing Components with GATE Version 8*. https ://gate.ac.uk/sale/tao/split.html, vol. 8, no. 1, 2014. WWW

[Cyganiak 10] Richard Cyganiak, Fadi Maali & Vassilios Peristeras. *Self-service linked government data with dcat and gridworks*. In Proceedings of the 6th International Conference on Semantic Systems, page 37. ACM, 2010.

[Cyganiak 14] Richard Cyganiak & Anja Jentzsch. *Linking open data cloud diagram*. LOD Community (http ://lod-cloud. net/), vol. 12, 2014. WWW

[Dai 12] Hong-Jie Dai, Chi-Yang Wu, R Tsai, W Hsu*et al. From entity recognition to entity linking : a survey of advanced entity linking*

techniques. In The 26th Annual Conference of the Japanese Society for Artificial Intelligence, pages 1–10, 2012.

[De Wilde 15] Max De Wilde. *From Information Extraction to Knowledge Discovery : Semantic Enrichment of Multilingual Content with Linked Open Data.* PhD thesis, Université Libre de Bruxelles, 2015.

[Dean 04] Mike Dean, Guus Schreiber, Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L McGuinness, Peter F Patel-Schneider & L Andrea Stein. *OWL web ontology language reference.* W3C Recommendation February, vol. 10, 2004. www

[Desmet 13] Bart Desmet & V??ronique Hoste. *Fine-grained Dutch named entity recognition.* Language Resources and Evaluation, vol. 48, no. 2, pages 1–37, 2013.

[Doddington 04] George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel & Ralph M Weischedel. *The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation.* In LREC, 2004.

[Douthat 98] Aaron Douthat. *The message understanding conference scoring software user's manual.* In Proceedings of the 7th Message Understanding Conference (MUC-7), 1998. www

[Downey 07] Doug Downey, Matthew Broadhead & Oren Etzioni. *Locating Complex Named Entities in Web Text.* In IJCAI, volume 7, pages 2733–2739, 2007.

[DuCharme 13] Bob DuCharme. Learning sparql. O'Reilly Media, Inc., 2013.

[Dupoirier 09] Gérard Dupoirier. Valorisation de l'information non-structurée. Ed. Techniques Ingénieur, 2009.

[Ehrmann 08] Maud Ehrmann. *Les entités nommées, de la linguistique au TAL.* PhD thesis, Paris 7, 2008.

[Ek 11] Tobias Ek, Camilla Kirkegaard, Håkan Jonsson & Pierre Nugues. *Named entity recognition for short text messages.* Procedia-Social and Behavioral Sciences, vol. 27, pages 178–187, 2011.

[Elmasri 11] Ramez Elmasri & Shamkant B. Navathe. *Fundamentals of database systems, 6th édition*, 2011.

[Etzioni 05] Oren Etzioni, Michael Cafarella, Doug Downey, Ana Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld & Alexander Yates. *Unsupervised named-entity extraction from the Web : An experimental study.* Artificial Intelligence, vol. 165, no. 1, pages 91–134, 2005.

[Färber 16] Michael Färber, Basil Ell, Carsten Menne, Achim Rettinger & Frederic Bartscherer. *Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO.* Semantic Web Journal, 2016. www

[Fensch 01] Thomas Fensch. Writing solutions : Beginnings, middles & endings. Xlibris Corporation, 2001.

[Fielding 00] Roy Thomas Fielding. *Architectural styles and the design of network-based software architectures.* PhD thesis, University of California, Irvine, 2000.

[Finkel 97] Jenny Rose Finkel, Trond Grenager & Christopher Manning. *Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling*, 1997.

[Fleiss 71] Joseph L Fleiss. *Measuring nominal scale agreement among many raters.* Psychological bulletin, vol. 76, no. 5, page 378, 1971.

[Flesch 49] Rudolf Franz Flesch*et al. Art of readable writing.* 1949.

[Fort 09] Karën Fort, Maud Ehrmann & Adeline Nazarenko. *Vers une méthodologie d'annotation des entités nommées en corpus ?* In Traitement Automatique des Langues Naturelles 2009, 2009.

[Fort 12] Karën Fort. *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus.* PhD thesis, Université Paris 13 - Sorbonne, 2012.

[France-Presse 04] Agence France-Presse. *Le manuel de l'agencier*, 2004.

[Friburger 02] Nathalie Friburger. *Reconnaissance automatique des noms propres : application à la classification automatique de textes journalistiques.* PhD thesis, Tours, 2002. www

[Galibert 14] Olivier Galibert, Jeremy Leixa, Gilles Adda, Khalid Choukri & Guillaume Gravier. *The ETAPE speech processing evaluation.* In LREC, pages 3995–3999. Citeseer, 2014.

[Galliano 09] Sylvain Galliano, Guillaume Gravier & Laura Chaubard. *The ester 2 evaluation campaign for the rich transcription of French radio broadcasts.* In Interspeech, volume 9, pages 2583–2586, 2009.

[Gandon 12] Fabien Gandon, Olivier Corby & Catherine Faron-Zucker. Le web sémantique : Comment lier les données et les schémas sur le web ? Dunod, 2012.

[Garside 97] Roger Garside, Geoffrey N Leech & Tony McEnery. Corpus annotation : linguistic information from computer text corpora. Taylor & Francis, 1997.

[Goyvaerts 09] Jan Goyvaerts & Steven Levithan. Regular expressions cookbook. Oreilly, 2009.

[Grevisse 08] Benoît Grevisse. Écritures journalistiques : Stratégies rédactionnelles, multimédia et journalisme narratif. De Boeck Supérieur, 2008.

[Grishman 96] Ralph Grishman & Beth Sundheim. *Message Understanding Conference-6 : A Brief History.* COLING, vol. 1, pages 466–471, 1996. www

[Gruber 95] Thomas R Gruber. *Toward principles for the design of ontologies used for knowledge sharing ?* International journal of human-computer studies, vol. 43, no. 5, pages 907–928, 1995.

[Guerraz 15] Olivier Collin Aleksandra Guerraz. *Classification d'entités nommées de type «film».* 22e Traitement Automatique des Langues Naturelles, Caen, 2015. www

[Guo 13] Stephen Guo, Ming-Wei Chang & Emre Kiciman. *To Link or Not to Link? A Study on End-to-End Tweet Entity Linking.* In HLT-NAACL, pages 1020–1030, 2013.

[Gut 04] Ulrike Gut & Petra Saskia Bayerl. *Measuring the reliability of manual annotations of speech corpora.* In Speech Prosody 2004, International Conference, 2004.

[Hachey 13] Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal & James R. Curran. *Evaluating entity linking with wikipedia.* Artificial intelligence, vol. 194, pages 130–150, 2013. www

[Harris 13] Steve Harris, Andy Seaborne & Eric Prud'hommeaux. *SPARQL 1.1 query language.* W3C Recommendation, vol. 21, 2013.

[Hatmi 14] Mohamed Hatmi. *Reconnaissance des entités nommées dans des documents multimodaux.* PhD thesis, Université de Nantes, 2014.

[Hedden 10] Heather Hedden. The accidental taxonomist. Information Today, Incorporated, 2010.

[Hengchen 15] Simon Hengchen, Seth van Hooland, Ruben Verborgh & Max De Wilde. *L'extraction d'entités nommées : une opportunité pour le secteur culturel?* I2D–Information, données & documents, vol. 52, no. 2, pages 70–79, 2015.

[Hitzler 12] Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F Patel-Schneider & Sebastian Rudolph. *OWL 2 web ontology language primer. W3C recommendation, 11 December 2012*, 2012. https://www.w3.org/TR/owl2-overview/

[Hoffart 13] Johannes Hoffart, Fabian M Suchanek, Klaus Berberich & Gerhard Weikum. *YAGO2 : A spatially and temporally enhanced knowledge base from Wikipedia.* Artificial Intelligence, vol. 194, pages 28–61, 2013.

[Hopcroft 01] John E Hopcroft, Rajeev Motwani & Jeffrey D Ullman. *Introduction to automata theory, languages, and computation, 2nd edi.* 2001.

[Hudon 13] Michèle Hudon. Analyse et représentation documentaires. Presses de l'Université du Québec, 2013.

[Jannet 14] Mohamed Ben Jannet, Martine Adda-Decker, Olivier Galibert, Juliette Kahn & Sophie Rosset. *Eter : a new metric for the evaluation of hierarchical named entity recognition.* In Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 3987–3994, 2014.

[Jena 15] Apache Jena. *A free and open source Java framework for building Semantic Web and Linked Data applications.* 2015. www

[Jentzsch 09] Anja Jentzsch. *DBpedia–Extracting structured data from Wikipedia.* Presentation at Semantic Web In Bibliotheken (SWIB 2009), Cologne, Germany, 2009.

[Jin 14] Yuzhe Jin, Emre Kıcıman, Kuansan Wang & Ricky Loynd. *Entity linking at the tail : sparse signals, unknown entities, and phrase models.* In Proceedings of the 7th ACM international conference on Web search and data mining, pages 453–462. ACM, 2014.

[Jurafsky 14] Dan Jurafsky & James H Martin. Speech and language processing. Pearson, 2014.

[Kasneci 06] Gjergji Kasneci, Fabian Suchanek & Gerhard Weikum. *Yago-a core of semantic knowledge.* 2006.

[Kaufmann 10] Max Kaufmann & Jugal Kalita. *Syntactic normalization of twitter messages.* In International conference on natural language processing, Kharagpur, India, 2010.

[Kim 03] J-D Kim, Tomoko Ohta, Yuka Tateisi & Jun'ichi Tsujii. *GENIA corpus—a semantically annotated corpus for bio-textmining.* Bioinformatics, vol. 19, no. suppl 1, pages i180–i182, 2003.

[Kleene 51] Stephen Cole Kleene. *Representation of events in nerve nets and finite automata.* Rapport technique, DTIC Document, 1951.

[Konkol 12] Michal Konkol. *Named entity recognition : technical report no. DCSE/TR-2012-04.* 2012.

[Krejcie 70] Robert V Krejcie & Daryle W Morgan. *Determining sample size for research activities.* Educ psychol meas, 1970.

[Kripke 80] Saul A Kripke. Naming and necessity. Harvard University Press, 1980.

[Krippendorff 04] Klaus Krippendorff. Content analysis : An introduction to its methodology, 2d edition. Sage, 2004.

[Kübler 15] Sandra Kübler & Heike Zinsmeister. Corpus linguistics and linguistically annotated corpora. Bloomsbury Publishing, 2015.

[Kushmerick 09] Nicholas Kushmerick. *Languages for Web Data Extraction.* In Encyclopedia of Database Systems, pages 1595–1600. Springer, 2009.

[Labasse 12] Bertrand Labasse. *Structures narratives et congruence cognitive : cas du summary lead et de la pyramide inversée.* Canadian Journal for Studies in Discourse and Writing/Rédactologie, vol. 24, no. 1, 2012.

[Laiacano 12] Adam Laiacano. *(Re)Organizing the Web's Data.* In Ethan Q. McCallum, editeur, Bad Data Handbook, chapitre 5, pages 69–82. O'Reilly Media, Inc., 2012.

[Lambe 14] Patrick Lambe. Organising knowledge : taxonomies, knowledge and organisational effectiveness. Elsevier, 2014.

[Lehmann 14] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer & Christian Bizer. *DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia.* Semantic Web, vol. 1, pages 1–5, 2014.

[Li 16] Yang Li, Shulong Tan, Huan Sun, Jiawei Han, Dan Roth & Xifeng Yan. *Entity Disambiguation with Linkless Knowledge Bases.* In Proceedings of the 25th International Conference on World Wide Web, pages 1261–1270. International World Wide Web Conferences Steering Committee, 2016.

[Liu 13] Xiaohua Liu, Furu Wei, Shaodian Zhang & Ming Zhou. *Named entity recognition for tweets.* ACM Transactions on Intelligent Systems and Technology (TIST), vol. 4, no. 1, page 3, 2013.

[Maedche 12] Alexander Maedche. Ontology learning for the semantic web. Springer Science & Business Media, 2012.

[Mahdisoltani 14] Farzaneh Mahdisoltani, Joanna Biega & Fabian Suchanek. *Yago3 : A knowledge base from multilingual wikipedias.* In 7th Biennial Conference on Innovative Data Systems Research. CIDR Conference, 2014.

[Makhoul 99] John Makhoul, Francis Kubala, Richard Schwartz, Ralph Weischedel*et al. Performance measures for information extraction.* In Proceedings of DARPA broadcast news workshop, pages 249–252, 1999.

[Markert 02] Katja Markert & Malvina Nissim. *Towards a Corpus Annotated for Metonymies : the Case of Location Names.* In LREC. Citeseer, 2002.

[Marrero 12] Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, Juan Miguel Gómez-Berb\'is, Juan Miguel Gómez-Berbís, Juan Miguel Gómez-Berb\'is, Juan Miguel Gómez-Berbís, Juan Miguel Gómez-Berb\'is & Juan Miguel Gómez-Berbís. *Named Entity Recognition : Fallacies, challenges and opportunities.* Computer Standards & Interfaces, vol. 35, no. 5, pages 482–489, September 2012. www

[Marsh 98] Elaine Marsh & Dennis Perzanowski. *MUC-7 evaluation of IE technology : Overview of results.* In Proceedings of the seventh message understanding conference (MUC-7), volume 20, 1998.

[Matei 16] Liviu Sebastian Matei & Stefan Trausan Matu. *Named entities distribution in Newspaper articles.* In The International Scientific Conference eLearning and Software for Education, volume 1, page 231. " Carol I" National Defence University, 2016.

[Maurel 11] Denis Maurel, Nathalie Friburger, Jean-Yves Antoine, Iris Eshkol & Damien Nouvel. *Cascades de transducteurs autour de la reconnaissance des entités nommées*. Traitement automatique des langues, vol. 52, no. 1, pages 69–96, 2011.

[Maynard 01] Diana Maynard, Valentin Tablan, Cristian Ursu, Hamish Cunningham & Yorick Wilks. *Named entity recognition from diverse text types*. In Recent Advances in Natural Language Processing 2001 Conference, pages 257–274, 2001.

[Maynard 03] Diana Maynard & Hamish Cunningham. *Multilingual adaptations of ANNIE, a reusable information extraction tool*. In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2, pages 219–222. Association for Computational Linguistics, 2003.

[McCowan 04] Iain A McCowan, Darren Moore, John Dines, Daniel Gatica-Perez, Mike Flynn, Pierre Wellner & Hervé Bourlard. *On the use of information retrieval measures for speech recognition evaluation*. Rapport technique, IDIAP, 2004.

[McNamee 09] Paul McNamee & Hoa Trang Dang. *Overview of the TAC 2009 knowledge base population track*. In Text Analysis Conference (TAC), volume 17, pages 111–113, 2009.

[Mendola 05] Luigi Mendola. *Déclaration de mandats et de patrimoine - Le point de la situation*, 2005. http://www.uvcw.be/articles/3,17,2,0,873.htm

[Merchant 96] Roberta Merchant, Mary Ellen Okurowski & Nancy Chinchor. *The multilingual entity task (MET) overview*. In Proceedings of a workshop on held at Vienna, Virginia : May 6-8, 1996, pages 445–447. Association for Computational Linguistics, 1996.

[Miller 98] George Miller & Christiane Fellbaum. *Wordnet : An electronic lexical database*, 1998.

[Minkov 05] Einat Minkov, Richard C Wang & William W Cohen. *Extracting personal names from email : applying named entity recognition*

to informal text. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages 443–450. Association for Computational Linguistics, 2005.

[Mitchell 97] Tom M Mitchell*et al. Machine learning. WCB*, 1997.

[Moens 06] Marie-Francine Moens. Information extraction : algorithms and prospects in a retrieval context, volume 21. Springer Science & Business Media, 2006.

[Nadeau 06] David Nadeau, Peter Turney & Stan Matwin. *Unsupervised named-entity recognition : Generating gazetteers and resolving ambiguity.* 2006.

[Nadeau 07a] David Nadeau. *Semi-Supervised Named Entity Recognition.* PhD thesis, University of Ottawa, 2007.

[Nadeau 07b] David Nadeau & Satoshi Sekine. *A survey of named entity recognition and classification.* Lingvisticae Investigationes, vol. 30, no. 1991, page 3, 2007. www

[Naumann 09] Harald Naumann. Screen scraper, pages 2511–2512. Springer US, Boston, MA, 2009. www

[Nouvel 15] Damien Nouvel, Maud Erhmann & Sophie Rosset. Les entités nommées pour le traitement automatique des langues. Iste Editions, 2015.

[O'Donnell 08] Mick O'Donnell. *The UAM CorpusTool : Software for corpus annotation and exploration.* In Proceedings of the XXVI Congreso de AESLA, Almeria, Spain, pages 3–5, 2008.

[O'Reilly 05] Tim O'Reilly. *What is Web 2.0.* 2005. www

[Palmer 97] David D Palmer & David S Day. *A statistical profile of the named entity task.* In Proceedings of the fifth conference on Applied natural language processing, pages 190–193. Association for Computational Linguistics, 1997.

[Pantel 11] Patrick Pantel & Ariel Fuxman. *Jigs and lures : Associating web queries with structured entities.* In Proceedings of the 49th An-

nual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1, pages 83–92. Association for Computational Linguistics, 2011.

[Paterson 07] Chris Paterson. *International news on the internet : Why more is less.* Ethical Space : The International Journal of Communication Ethics, vol. 4, no. 1/2, pages 57–66, 2007.

[Pellissier Tanon 16] Thomas Pellissier Tanon, Denny Vrandečić, Sebastian Schaffert, Thomas Steiner & Lydia Pintscher. *From Freebase to Wikidata : The Great Migration.* In Proceedings of the 25th International Conference on World Wide Web, pages 1419–1428. International World Wide Web Conferences Steering Committee, 2016.

[Pereira 14] Bianca Pereira. *Entity linking with multiple knowledge bases : An ontology modularization approach.* In International Semantic Web Conference, pages 513–520. Springer, 2014.

[Piskorski 11] Jakub Piskorski, Hristo Tanev, Martin Atkinson, Eric Van Der Goot & Vanni Zavarella. *Online news event extraction for global crisis surveillance.* In Transactions on computational collective intelligence V, pages 182–212. Springer, 2011.

[Poibeau 01] T Poibeau. *« Deconstructing Harry» : une evaluation des systemes de reperage d'entites nommees.* Revue de l'electricite et de l'electronique, no. 7, pages 65–73, 2001.

[Poibeau 11] Thierry Poibeau. Traitement automatique du contenu textuel. Lavoisier, 2011.

[Pustejovsky 13] J Pustejovsky & a Stubbs. Natural language annotation for machine learning. 2013.

[Rajabi 14] Enayat Rajabi, Miguel-Angel Sicilia & Salvador Sanchez-Alonso. *An empirical study on the evaluation of interlinking tools on the Web of Data.* Journal of Information Science, vol. 40, no. 5, pages 637–648, 2014.

[Ramshaw 95] Lance A Ramshaw & Mitchell P Marcus. *Text Chunking using Transformation-Based Learning*. In eprint arXiv : cmp-lg/9505040, volume 1, page 5040, 1995.

[Rao 13] Delip Rao, Paul McNamee & Mark Dredze. *Entity linking : Finding extracted entities in a knowledge base*. In Multi-source, multilingual information extraction and summarization, pages 93–115. Springer, 2013.

[Ratinov 09] L. Ratinov & D. Roth. *Design Challenges and Misconceptions in Named Entity Recognition*. In CoNLL, 6 2009. www

[redactiweb 10] redactiweb. *La pyramide inversée : la règle d'or de l'écriture web*, 2010. http://www.redactiweb.com/blog/la-pyramide-inversee-la-regle-d%E2%80%99or-de-l%E2%80%99ecriture-web

[Rizza 15a] Ettore Rizza. *Du prêt-à-porter au sur-mesure : une méthode de réconciliation manuelle dans Open Refine*, 2015. http://student.ulb.ac.be/~erizza/docs/assignmentFinal_Modelisation.pdf

[Rizza 15b] Ettore Rizza. *Évaluation de services Web pour la reconnaissance et la résolution d'entités nommées dans des dépêches d'agence en français*, 2015. http://student.ulb.ac.be/~erizza/docs/Entites-nommees-ASI.pdf

[Rocher 68] Guy Rocher. Introduction à la sociologie générale. HMH, Ltée, Québec), 1968.

[Roget 11] Peter Mark Roget. Roget's thesaurus of english words and phrases... TY Crowell Company, 1911.

[Ross 94] Line Ross. L'écriture de presse : l'art d'informer. Montréal : G. Morin, 1994.

[Salaün 10] Jean-Michel Salaün & Clément Arsenault. Introduction aux sciences de l'information. La Découverte, 2010.

[Sauermann 08] Leo Sauermann & Richard Cyganiak. *Cool URIs for the Semantic Web*, 2008. http://www.w3.org/TR/cooluris/

[Scanlan 00] Christopher Scanlan. Reporting and writing : Basics for the 21st century. Harcourt College Publishers, 2000.

[Sebastiani 02] Fabrizio Sebastiani. *Machine learning in automated text categorization*. ACM computing surveys (CSUR), vol. 34, no. 1, pages 1–47, 2002.

[Segaran 09] Toby Segaran, Colin Evans & Jamie Taylor. Programming the semantic web. O'Reilly Media, Inc., 2009.

[Sekine 02] Satoshi Sekine, Kiyoshi Sudo & Chikashi Nobata. *Extended Named Entity Hierarchy*. In LREC, pages 1818–1824, 2002. www

[Sekine 04a] Satoshi Sekine. *Named entity : History and future*. Project notes, New York University, 2004. www

[Sekine 04b] Satoshi Sekine & Chikashi Nobata. *Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy*. In LREC, pages 1977–1980, 2004.

[Sekine 10] Satoshi Sekine. *Sekine's extended named entity hierarchy*. 2010. www

[Serrano 12] Laurie Serrano, Thierry Charnois, Stephan Brunessaux, Bruno Grilheres & Maroua Bouzid. *Combinaison d'approches pour l'extraction automatique d'événements*. In 19e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2012), pages 423–430, 2012.

[Settles 04] Burr Settles. *Biomedical named entity recognition using conditional random fields and rich feature sets*. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, pages 104–107. Association for Computational Linguistics, 2004.

[Shen 14] Wei Shen, Jiawei Han & Jianyong Wang. *A probabilistic model for linking named entities in web text with heterogeneous information networks*. In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, pages 1199–1210. ACM, 2014.

[Shen 15] Wei Shen, Jianyong Wang & Jiawei Han. *Entity linking with a knowledge base : Issues, techniques, and solutions.* Knowledge and Data Engineering, IEEE Transactions on, vol. 27, no. 2, pages 443–460, 2015.

[Shirky 05] Clay Shirky. *Ontology is overrated : Categories, links, and tags.* 2005. www

[Shrivastava 07] K.M. Shrivastava. News agencies from pigeon to internet. Sterling Publishers Pvt. Ltd, 2007.

[Sil 12] Avirup Sil, Ernest Cronin, Penghai Nie, Yinfei Yang, Ana-Maria Popescu & Alexander Yates. *Linking named entities to any database.* In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 116–127. Association for Computational Linguistics, 2012.

[solidaires 12] Reporters solidaires & Christine Cognat. Les rubriques du journalisme : décrypter, organiser et traiter l'actualité. Presses universitaires de Grenoble, 2012.

[Song 14] Min Song & Ying Ding. *Topic modeling : Measuring scholarly impact using a topical lens.* In Measuring Scholarly Impact, pages 235–257. Springer, 2014.

[Southwick 15] Silvia B Southwick. *A Guide for Transforming Digital Collections Metadata into Linked Data Using Open Source Technologies.* Journal of Library Metadata, vol. 15, no. 1, pages 1–35, 2015.

[Stern 10] Rosa Stern & Benoît Sagot. *Détection et résolution d'entités nommées dans des dépêches d'agence.* In Traitement Automatique des Langues Naturelles : TALN 2010, 2010.

[Stern 12] Rosa Stern, Benoît Sagot & Frédéric Béchet. *A joint named entity recognition and entity linking system.* In Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data, pages 52–60. Association for Computational Linguistics, 2012.

[Stern 13] Rosa Stern. *Identification automatique d'entités pour l'enrichissement de contenus textuels*. PhD thesis, Université Paris-Diderot-Paris VII, 2013.

[Tellier 09] Isabelle Tellier. *Apprentissage automatique pour le TAL : Préface*. Traitement Automatique des Langues, vol. 50, no. 3, pages 7–21, 2009.

[Tellier 10] Isabelle Tellier. *Introduction au TALN et à l'ingénierie linguistique*. Polycopié de cours : Université de Lille 3, 2010.

[Thomas 10] L Thomas. *Le journal gratuit Métro et sa stratégie commerciale*. Publications Études & Analyses, 2010. www

[Tjong Kim Sang 00] Erik F Tjong Kim Sang & Sabine Buchholz. *Introduction to the CoNLL-2000 shared task : Chunking*. In Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7, pages 127–132. Association for Computational Linguistics, 2000.

[van Hooland 13] Seth van Hooland, Max De Wilde, Ruben Verborgh, Thomas Steiner & Rik Van de Walle. *Exploring entity recognition and disambiguation for cultural heritage collections*. Literary and linguistic computing, 2013.

[van Hooland 14] Seth van Hooland & Ruben Verborgh. Linked data for libraries, archives and museums : How to clean, link and publish your metadata. London, Facet Publishing, 2014.

[van Hooland 16] Seth van Hooland, Florence Gillet, Simon Hengchen & Michael E. Sinatra. Introduction aux humanités numériques : méthodes et pratiques : sciences humaines et sociales. De Boeck supérieur, 2016.

[Van Rijsbergen 79] CJ Van Rijsbergen. *Information Retrieval. Dept. of Computer Science, University of Glasgow*. URL : citeseer. ist. psu. edu/-vanrijsbergen79information. html, 1979.

[Veltman 13] Noah Veltman. *Scraping the web | School of Data - Evidence is Power*, 2013. http://schoolofdata.org/2013/11/25/scraping-the-web/

[Verborgh 13] Ruben Verborgh & Max De Wilde. Using open refine. Packt Publishing Ltd, 2013.

[Volle 06] Michel Volle. De l'informatique : savoir vivre avec l'automate. Economica Paris, 2006.

[Vrandečić 14] Denny Vrandečić & Markus Krötzsch. *Wikidata : a free collaborative knowledgebase*. Communications of the ACM, vol. 57, no. 10, pages 78–85, 2014. www

[Watrin 06] Patrick Watrin. *Une approche hybride de l'extraction d'information : sous-langages et lexique-grammaire*. PhD thesis, UCL, 2006.

[Weischedel 13] Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini*et al. Ontonotes release 5.0 LDC2013T19*. Linguistic Data Consortium, Philadelphia, PA, 2013.

[Yamada 15a] Ikuya Yamada, Hideaki Takeda & Yoshiyasu Takefuji. *An end-to-end entity linking approach for tweets*. In CEUR-WS, 2015.

[Yamada 15b] Ikuya Yamada, Hideaki Takeda & Yoshiyasu Takefuji. *Enhancing Named Entity Recognition in Twitter Messages Using Entity Linking*. ACL-IJCNLP 2015, page 136, 2015.

[Yu 11] Liyang Yu. A developer's guide to the semantic web. Springer Science & Business Media, 2011.

[Zhou 10] Yiping Zhou, Lan Nie, Omid Rouhani-Kalleh, Flavian Vasile & Scott Gaffney. *Resolving surface forms to wikipedia topics*. In Proceedings of the 23rd International Conference on Computational Linguistics, pages 1335–1343. Association for Computational Linguistics, 2010.