# Gestion de l'incertitude dans le cadre d'une extraction des connaissances à partir de texte

Fadhela Kerdjoudj, Olivier Curé

# Gestion de l'incertitude dans le cadre d'une extraction des connaissances à partir de texte

Fadhela Kerdjoudj[*,**] Olivier Curé[*]

[*]Université Paris-Est Marne-La-Vallée, LIGM, CNRS UMR 8049, France
fadhela.kerdjoudj@univ-mlv.fr, ocure@univ-mlv.fr
[**]GEOLSemantics 12 rue Raspail, 94250, Gentilly

**Résumé.** Le domaine de l'extraction de connaissances à partir de texte nécessite des méthodes permettant de détecter et de manipuler l'incertitude. En effet, de nombreux textes contiennent des informations dont la véracité peut être remise en cause. Il convient alors de gérer de manière efficace ces informations afin de représenter les connaissances de manière explicite. Une première démarche consiste à identifier les différentes formes d'incertitudes pouvant intervenir durant un processus d'extraction. Puis, nous proposons une représentation RDF basée sur une ontologie développée destinée à modéliser l'incertitude.

## 1 Introduction

La multiplication de sources textuelles sur le Web offre un champ pour l'extraction de bases de connaissances. Dernièrement, de nombreux travaux dans ce domaine sont apparus ou se sont intensifiés, Dong et al. (2014), Niu et al. (2012). Dans le domaine de la construction automatique de bases de connaissances, il est nécessaire de faire collaborer des approches linguistiques, pour extraire certains concepts relatifs aux entités nommées, aspects temporels et spatiaux, à des méthodes issues des traitements sémantiques afin de faire ressortir la pertinence et la précision de l'information véhiculée. Pour présenter un intérêt à l'échelle du Web, les traitements linguistiques doivent être multi-sources et inter-lingues.

GEOLSemantics est une entreprise qui s'appuie sur une expérience cumulée de plusieurs années dans le monde de la linguistique et de la sémantique. Cette société propose une solution logicielle de traitement linguistique basée sur une analyse linguistique profonde. Le but est d'extraire automatiquement, d'un ensemble de textes, des connaissances structurées, localisées dans le temps et l'espace, des concepts, des relations et des événements impliquant des Entités Nommées. Pour représenter les connaissances extraites du texte, nous avons opté pour les technologies du web sémantique. Nous représentons donc nos extractions sous forme de triplets RDF et exploitons une ontologie que nous avons spécifiquement développées pour couvrir certains domaines. Cette approche permet de relier les résultats de nos extractions à des connaissances externes contenues dans des bases de références du Linked Open Data, Bizer et al. (2008), tels que Dbpedia et Geonames, et ainsi d'effectuer un certain nombre de vérifications. Les textes actuellement traités par GEOLSemantics sont issus d'articles de presse et traitent de sujets divers allant de la politique au sport en passant par des faits divers.

Gestion de l'incertitude.

Lors de l'analyse linguistique, il arrive que l'information traitée contienne des imperfections. En effet, l'information peut être bruitée, biaisée, implicite, imprécise, incohérente ou incertaine. Dans notre travail, nous accordons un intérêt particulier à l'incertitude. Il s'agit de l'une des imperfections les plus courantes et qui apporte une importante information à la connaissance acquise. Notre première contribution porte sur une catégorisation de l'incertitude lors des différentes phases d'extraction. Notre seconde contribution se situe au niveau de la représentation de l'incertitude dans le graphe RDF.

Cet article est organisé comme suit : Dans la Section 2, nous introduisons les grandes lignes du système d'extraction de GEOLSemantics. La Section 3 détaille notre catégorisation de l'incertitude. Ensuite, nous motivons et présentons notre représentation de l'incertitude dans un graphe RDF. Nous concluons et présentons nos travaux futurs en Section 5.

## 2 Acquisition de l'information et représentation des connaissances

L'acquisition de l'information comporte plusieurs étapes distinctes allant du simple découpage du texte en mots à la représentation de son contenu. Ces étapes consécutives consistent en :

- *L'analyse morpho-syntaxique :* il s'agit de la mise en évidence des structures d'agencement des catégories grammaticales (nom, verbe, adjectif, etc.), afin d'en découvrir les relations formelles ou fonctionnelles (Exemple : sujet, verbe et complément).
- *L'analyse sémantique :* il s'agit de l'étude linguistique du sens. L'objectif principal de cette analyse est de déterminer le sens de chaque mot dans la phrase.
- *L'extraction de connaissances :* permet de mettre en évidence des entités nommées et des relations relatives à un concept particulier. Dans cette étape, nous nous basons sur la notion de marqueur ou déclencheur. Il s'agit d'un terme ou une expression indiquant la présence d'un concept donné. Exemple : les termes *partir, aller, voyager* sont des déclencheurs indiquant un *déplacement*. Ces déclencheurs indiquent qu'une relation relative à un concept est présente et peut être extraite. A chaque concept de l'ontologie correspondra une liste de déclencheurs possibles, ceci permet de sélectionner les règles d'extraction de connaissances à appliquer.
- *La mise en cohérence :* permet de consolider les connaissances extraites, notamment le regroupement des entités nommées et la résolution des dates relatives.
- *L'enrichissement :* permet de compléter l'information à partir des données du Linked Open Data.

A l'issue de ces traitements, grâce à la représentation RDF, Cyganiak et al. (2013), nous disposons d'un ensemble de triplets qui permettra d'utiliser par la suite la connaissance exprimée dans le texte étudié. Le format RDF présente l'avantage d'une syntaxe simple sous forme de triplets (sujet-prédicat-objet). Il repose sur des URI (Uniform Resource Identifier) qui permettent d'identifier de manière unique chaque ressource et de faciliter la distribution et la publication des données sémantiques sur le web à travers le Linked Open Data [1]. Néanmoins, ce traitement suppose que les données fournies sont toutes fiables et sûres. Ceci n'est pas toujours garanti. Le but de notre travail est de pondérer la connaissance extraite en fonction de la fiabilité de

1. http://linkeddata.org/

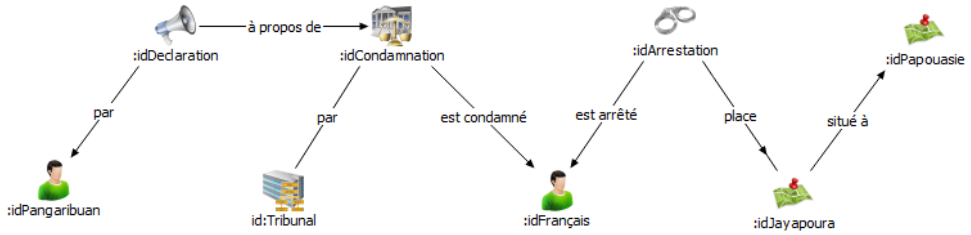


FIG. 1 – *Représentation graphique des triplets RDF après extraction.*

l'information initiale. En effet, celle-ci peut être remise en cause suivant différents critères que nous catégorisons dans la section suivante. L'exemple 2.1 permettra d'illustrer nos propos tout au long de cet article :

**Exemple 2.1** *Le figaro, Par Jeanne Fremin Du Sartel Publié le 02/09/2014 à 16 :43*
***Détenus** dans un **centre** de rétention à Jayapura, **capitale** de la Papouasie, les **Français** vont très probablement devoir **comparaître** devant un **tribunal** pour avoir triché sur la nature de leur visa, prévient leur **avocat** Aristo Pangaribuan.*

Dans cet exemple, nous remarquons la présence de plusieurs déclencheurs que nous avons mis en gras :

— détenus : permet d'identifier une arrestation.
— centre, capitale : annonceurs d'entités nommées de type Lieu.
— comparaître : permet d'identifier un jugement.
— avocat : entité nommée de type Personne
— tribunal : entité nommée de type Organisation

La figure 1 propose une représentation graphique de l'extraction textuelle de notre exemple. Celle-ci décrit la déclaration de l'avocat concernant l'arrestation en Papouasie puis la condamnation des français. Les noeuds représentent les instances de concepts de l'ontologie, alors que les arcs formalisent les liens entre ces concepts.
Cependant, la condamnation est annoncée comme étant "probable", il est donc nécessaire d'introduire une incertitude concernant cet événement. Dans la suite de cet article nous décrirons les étapes qui nous permettent d'identifier et d'inclure cette incertitude dans notre graphe RDF.

# 3 Catégorisation de l'incertitude

La fiabilité de l'information est très souvent remise en cause. En effet, l'imprécision, l'incertitude ou encore l'incomplétude sont des problèmes récurrents dans le traitement de l'information. Dans cet article, nous accordons un intérêt particulier à l'aspect incertain de l'information acquise. Notre démarche est de considérer les modalités d'acquisition et d'expression de l'information qui constituera la connaissance, ainsi que le traitement de cette dernière jusqu'à la génération d'un graphe RDF permettant de la stocker dans des bases de connaissances afin de pouvoir la réutiliser par la suite. Nous considérons que le degré de confiance accordé à une

information est influencé par différentes phases du traitement. Pour cela, nous identifions trois niveaux où peut intervenir de l'incertitude dans le traitement de l'information.

## 3.1 Incertitude pré-extraction

Dans cette partie, nous considérons la source du texte ainsi que les métadonnées qui lui sont associées. Ces métadonnées peuvent faire référence à l'auteur de l'article, l'organisation chargée de le publier, le contexte abordé... Ces métadonnées sont obtenues lors de la récupération du texte. L'information peut provenir de sources variées avec différents niveaux de fiabilité, il s'agira donc lors de cette étape de qualifier la fiabilité de la source utilisée. En effet, la cotation d'informations est une tâche qui vise à mesurer la qualité d'une information, Mombrun et al. (2010), et en particulier la confiance qu'on peut lui accorder. Cette confiance peut varier en fonction des informations qu'elles produisent, la compétence des auteurs, la certitude qu'elles expriment, la vraisemblance du contenu ou l'existence de confirmations ou d'infirmations...

Pour modéliser ces métadonnées nous nous basons sur l'ontologie PROV-O [2], un standard RDF certifié par le W3C. PROV-O est une ontologie qui permet de modéliser les informations liées à la source des données. Elle décrit les entités, les activités et les agents impliqués dans la production d'informations, ainsi que la qualité, la fiabilité et la confiance associée, Lebo et al. (2013). Dans Missier et al. (2013), les auteurs proposent un tutoriel qui permet de modéliser les métadonnées ainsi que la confiance associée à la source.

Par ailleurs, nous disposons d'une base de connaissances sur la confiance accordée aux sources. À chaque source sera associé un degré de confiance évaluant la fiabilité des informations qu'elle fournit. Ce degré de confiance dépend également de l'utilisateur. En effet, pour une même source, la confiance accordée peut varier suivant l'utilisateur.
Nous avons alors créé une base de connaissances (que nous nommerons Trust) permettant de stocker la confiance associée à une source pour chaque utilisateur. L'ontologie de cette base comporte trois classes : *Source, User et Trustworthiness*.
La classe *Source* correspond à la super-classe d'une hiérarchie de concepts comportant des classes décrivant des auteurs et éditeurs, e.g., bloggeur, journaliste, chaîne TV, journal. La classe *User* désigne l'utilisateur du programme. Enfin, la classe *Trustworthiness* décrit le degré de confiance qu'attribue l'utilisateur à la source considérée.
Le degré de confiance permet d'évaluer la fiabilité de la connaissance véhiculée. Nous avons décidé de représenter le degré de confiance par un nombre réel compris entre 0 et 1. La représentation en nombres continus (intervalle) poserait par la suite plus de problèmes, lors du raisonnement.

**Exemple 3.1** *Un utilisateur donné peut croire le journal "Le figaro" avec un degré de confiance de* $0.8$ *alors qu'un autre utilisateur décidera de croire cette source à* $0.6$.

Les informations fournies dans cet exemple nous permettent d'ajouter les triplets suivants dans notre base :
: $idTrustworthiness1 - hasSource - : idLeFigaro$;
: $idTrustworthiness1 - hasUser - : idUserX$;
: $idTrustworthiness1 - hasTrust - 0.8$;
: $idTrustworthiness2 - hasSource - : idLeFigaro$;

2. http://www.w3.org/TR/prov-o/

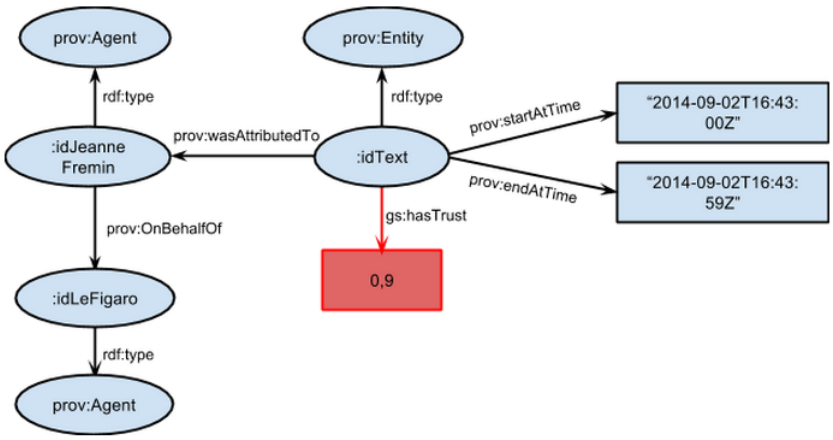


FIG. 2 – *Représentation de l'exemple 2.1 avec Prov-o.*

: $idTrustworthiness2 - hasUser -$ : $idUserY$;
: $idTrustworthiness2 - hasTrust - 0.6$;

La figure 2 décrit le graphe RDF associé à la représentation des métadonnées de l'exemple 2.1, en utilisant la représentation recommandée dans Prov. La propriété *wasAttributedTo* permet de lier le contenu du texte à l'auteur de la publication. Cet auteur est, quand à lui, lié à l'organisme de publication par la propriété *onBehalfOf*. Enfin, les proprétés *startAtTime* et *endAtTime* permettent respectivement d'indiquer la date de début et de fin de la publication. Le degré de confiance *hasTrust* est attribué après interrogation de la base de connaissances Trust. Les URI des objets des triplets dont le prédicat est *prov :wasAttributedTo* et/ou *prov :OnBehalfTo* permettent d'identifier la source. L'utilisateur avant d'évaluer la certitude des connaissances du texte, devra en premier vérifier dans Trust si le degré de confiance qu'il attribue à cette source lui convient ou pas.

## 3.2 Pendant l'extraction

Le deuxième niveau où une incertitude peut être exprimée concerne le contenu même du texte. En effet, durant l'analyse du texte, quelques imperfections de l'information peuvent être identifiées. De plus, les règles d'analyse et d'extraction de connaissances peuvent elles aussi être incertaines. Un degré d'incertitude devient alors nécessaire pour évaluer la certitude et la qualité de la connaissance extraite.

### 3.2.1 Imperfection au niveau de l'information

Le but dans cette partie est de considérer le caractère même de l'information. Celle-ci peut être objective (provenant d'une machine tel qu'un capteur ou radar) ou bien subjective (provenant d'une déclaration faite par un agent). L'information objective ne peut être remise en cause alors que l'information subjective peut être considérée de manière différente suivant

les personnes. L'information peut contenir de l'incertitude, des imprécisions ou encore être incomplètes... Dans Smets (1997), l'auteur distingue trois variantes d'imperfections

1. *L'incertitude* liée à la relation observée entre la donnée et l'univers pris en compte indiquant lorsque la donnée est possible.
   Exemple : "L'ancien président, Nicolas Sarkozy **pourrait** se présenter aux présidentielles 2017".
2. *L'imprécision*, causée par des données statistiques imprécises. Elle diffère d'une interprétation à une autre. Elle est généralement décrite grâce à des termes linguistiques décrivant le monde réel.
   Exemple : "Pierre a l'air **jeune**"
   "Jeune" étant une définition abstraite, l'âge de la personne peut varier de 18 à 25 ans par exemple.
   L'espace temporel peut lui aussi être exprimé de manière floue. Dans ce cas nous représenterons la date avec un intervalle temporel comprenant cette date.
   Exemple : " Pierre est né en **2003** à Paris".
   La date de naissance n'étant pas précise, un intervalle d'incertitude permettra alors de modéliser cette connaissance.
3. *Ignorance totale ou partielle*, il arrive qu'une information soit livrée sans que tous les détails la concernant ne soient décrits. L'information devient alors incomplète.
   Exemple : "Pierre est allé en **Allemagne**".
   Dans cet exemple, le lieu d'arrivée n'est pas bien précisé et les autres informations telles que la date ou encore le lieu de départ ne sont pas précisés. Il s'agit donc ici d'une information partielle.

Dans notre travail, nous avons mis l'accent sur la première forme d'imperfection, à savoir l'incertitude.
L'incertitude qualifie la connaissance de l'auteur sur l'information fournie. Cette dernière est soit vraie ou fausse à un moment donné, mais si l'auteur peut ne pas avoir connaissance de cet état ni de sa véracité, il exprimera alors une certaine incertitude lors de son récit. L'incertitude peut être employée pour exprimer une intention, une volonté, une supposition, une éventualité, un doute, une hésitation, une indécision, une croyance, une préférence, une émotion...
Comme pour l'extraction de connaissances, nous nous basons sur la notion de marqueurs telle que nous l'avons décrite dans la section 2. Les déclencheurs permettent de repérer l'incertitude que peut exprimer l'auteur lors de son récit. Une liste d'indicateurs d'incertitude est établie. Nous distinguons

— Les verbes d'opinion : croire, penser, douter...
— Les verbes impersonnels : il parait que, il semble que...
— Les adjectifs : douteux, incertain, possible...
— Les adverbes : peut-être, apparemment, probablement...
— Les locutions adverbiales : éventuellement, hypothétiquement..
— Les expressions : selon lui, à mon avis, il se peut...

Chaque marqueur d'incertitude est associé à un degré qui permettra de quantifier la fiabilité de l'information véhiculée.

Le marqueur peut également être accompagné d'une valeur numérique qui facilitera alors la quantification de l'incertitude. Exemple : *"La **probabilité** qu'il pleuve demain est de **60%**."*

La difficulté rencontrée à cette étape est la portée de cette incertitude. Il faut savoir si par exemple l'incertitude porte sur une propriété d'un concept donné ou bien sur le concept en entier. Le plus souvent, l'utilisation d'une structure emphatique telle que *"c'est ...qui/que..."* permet d'identifier le premier cas.

**Exemple 3.2** *Je pense que* ***c'est*** *Paul qui emmène Julie à Paris.*

Dans cet exemple, le voyage de Julie n'est pas remis en cause, mais c'est le fait que ce soit Paul qui l'emmène est incertain.

### 3.2.2 Application des règles d'extraction

Il arrive quelques fois que l'ambiguité soit très importante à tel point que le système n'arrive pas à distinguer la meilleure solution, étant donné que même un humain peut mal interprêter ces textes.

**Exemple 3.3** *Le livre de Paul.*

Dans cet exemple, nous ne pouvons distinguer entre la relation d'appartenance ou bien d'auteur. En effet, cette phrase peut exprimer le fait que Paul soit le propriétaire du livre ou bien le fait que Paul soit l'auteur du livre. En général, les systèmes d'analyse linguistique choisissent soit une règle d'extraction au hasard ou la plus probable. Dans cet exemple, la relation la plus probable est la relation d'appartenance, mais une incertitude subsiste quand même. Avec une démarche qui prend en compte l'incertitude, un poids est associé aux règles d'extraction afin d'extraire tous les triplets possibles. L'utilisateur décidera par la suite quelle connaissance il souhaite garder et ajouter à la base de connaissances.

## 3.3 Incertitude post-extraction

### 3.3.1 Mise en cohérence

Après l'extraction des connaissances brutes, il est nécessaire d'ajouter un traitement supplémentaire qui consiste à mettre en cohérence l'ensemble des connaissances extraites. L'extraction de connaissances traite le texte phrase par phrase, sans prendre en considération de lien sémantique qui peut exister entre les phrases. Le résultat de l'extraction des connaissances est un graphe RDF faisant référence aux concepts et propriétés issus de l'ontologie intégrée dans le système. L'étape de mise en cohérence correspond aux opérations de consolidation, résolution d'ambigüités et enrichissement de ce graphe. Les principaux traitements effectués lors de cette étape sont :

— Le regroupement des entités nommées : ceci consiste à regrouper les co-occurrences de chaque entité (Personne, Organisation, Lieu) citée dans le texte.
— La résolution des dates relatives : une date non absolue n'a pas d'intérêt à être stockée dans une base de connaissances car elle ne réfère à aucune date précise. Il est alors nécessaire de calculer la date relative en fonction de la date de référence. Cette dernière peut être la date d'émission de l'article, ou bien d'une date citée à l'intérieur du texte.
— Le regroupement d'événements : un événement peut être décrit avec différents déclencheurs ce qui créera plusieurs événements au lieu d'un seul et même évènement au sein de la phrase. La règle de regroupement suppose alors que si rien n'indique le contraire

-aucune contradiction dans la description de ces deux évènements ou encore qu'aucun adverbe de temps n'a été introduit- alors il s'agit d'un seul événement.

— Inférence au niveau de la phrase : de nouvelles connaissances peuvent être déduites du contexte. Ces connaissances étant exprimées implicitement il est impossible de les extraire lors de l'analyse linguistique. Dans l'exemple 2.1, le lieu où se situe le tribunal n'est pas mentionné mais nous pouvons supposer qu'il se trouve à Jayapura, avec cependant un certain degré de réticence.

Chacun de ces traitements repose sur des règles spécifiques. Le risque d'incertitude est évalué à l'application des règles. Pour le regroupement des entités nommées, nous nous basons sur des algorithmes d'Entity Matching, Köpcke et Rahm (2010). Ces méthodes de regroupement reposent principalement sur des algorithmes de comparaison sémantique de deux entités, ou de calcul de mesures de similarités, qui permettent d'évaluer le degré de ressemblance de deux entités, telles que : *Jaccard, Levenshtein, TF-IDF, etc.* Le calcul du niveau de ressemblance permet alors de quantifier le niveau d'incertitude associé à ce regroupement.
Pour ce qui est de la résolution des dates relatives, nous avons proposé une représentation sous forme d'intervalle de toutes les dates citées dans le texte. Ainsi une date qui réfère à la semaine dernière, par exemple, aura pour date de début le lundi de la semaine passée et pour date de fin le dimanche de cette même semaine. Cependant, aucun degré d'incertitude supplémentaire n'est associé à cette extraction.
Enfin, en ce qui concerne le regroupement des évènements et l'inférence intra-phrase, le degré dépendra également de la règle appliquée. Par exemple, l'inférence d'une entité nommée telle que la date ou le lieu aura un degré de possibilité plus élevé que l'ajout d'autres propriétés telles que la cause, ou les agents intervenants dans la description d'un évènement.

### 3.3.2 Enrichissement et vérification avec les bases de références

La dernière étape de notre système est la vérification et/ou l'enrichissement des connaissances extraites. La vérification est réalisée afin de s'assurer de la cohérence et de la véracité des connaissances extraites avant de les ajouter à la base de connaissances. Ceci consiste à interroger les jeux de données du Linked Open Data (LOD) tels que DBpedia ou Geonames. Cependant, il arrive que, dans des articles de presse, les entités nommées telles que les personnes ou les organisations soient orthographiées de manières différentes. Grâce aux pages d'homonymie disponibles sur le web noil devient possible de résoudre ce problème.
Enfin, pour ce qui concerne l'enrichissement, il s'agit de pouvoir améliorer la pertinence de la connaissance extraite en complétant l'information à partir des datasets du LOD. Ceci est une démarche pour gérer l'information incomplète. Cependant, les données du LOD peuvent elles aussi être incertaines. En effet, la qualité de ces données est parfois remise en cause. Quelques travaux ont été menés dans ce cadre afin d'évaluer la qualité de ces données : Bizer et Cyganiak (2009); Hartig (2008); Flemming (2010). Zaveri et al. (2013) considère que la qualité peut varier suivant six dimensions : le contexte abordé, la confiance accordée, le caractère intrinsèque de l'information, l'accessibilité des données, la pérennité et la représentation. Malheureusement, aucun dataset n'offre un score d'une précision de 100%. Il est donc nécessaire d'évaluer la qualité des données avant de compléter la connaissance extraite. Ce n'est pas toujours le cas. Par exemple, DBpedia comporte une grande quantité d'erreurs qui remettent en cause son utilisation Knuth et al. (2012).

# 4 Représentation de l'incertitude

Dans cette section, nous présentons le formalisme adopté pour la représentation de l'incertitude dans les graphes issus de nos extractions textuelles. Comme tout formalisme de représentation des connaissances, celui-ci comporte un aspect syntaxique et un aspect sémantique. Au niveau syntaxique, nous nous intéressons essentiellement à la manière d'intégrer les valeurs d'incertitude dans nos graphes. Nous dédions une sous-section à la notion de réification qui dans un premier temps semblait être une solution évidente puis introduisons notre approche.

## 4.1 Aspect sémantique

L'interprétation de nos graphes RDF se base sur la sémantique standard des graphes RDF qui sont associés aux ontologies OWL telle qu'elles ont été évoquées précédemment, e.g., l'ontologie de l'extraction, de l'incertitude et PROV-O. La sémantique qui est attribuée à la notion d'incertitude n'est pas développée dans cet article. Nous détaillerons cet aspect dans un prochain article lorsque nous approfondirons les aspects requêtage et raisonnement sur nos graphes. Nous pouvons simplement mentionner que notre représentation est compatible avec la sémantique standard des logiques possibiliste et probabiliste.

## 4.2 Ascpect syntaxique

### 4.2.1 Réification

La réification est une recommandation RDF du W3C, Semantics et al. (2004). Elle permet de décrire des informations concernant les triplets, telles que les métadonnées par exemple. Le principe de la réification est de diviser le triplet en quatre sous-triplets ayant la même ressource en sujet. Cette ressource est désignée par un rdf :nodeID. Le premier triplet permet d'identifier le sujet, le deuxième le prédicat, le troisième l'objet et enfin le quatrième permet d'indiquer que le nodeID décrit un triplet (Statement). Ce nodeID pourra par la suite être utilisé pour ajouter des triplets supplémentaires qui permettront de décrire les informations à ajouter au triplet initial. La figure 4 illustre l'exemple 4.1, et permet de décrire l'incertitude exprimée dans la phrase.

**Exemple 4.1** *John est probablement marié avec Mary.*

A première vue cela semble intéressant car l'incertitude peut être considérée comme une information supplémentaire à ajouter à un triplet. Cependant, vu que ce triplet est divisé, il devient difficile de le relier aux autres triplets dans le graphe. Ceci pose un sérieux problème lors de l'interrogation car la syntaxe habituelle (s,p,o) n'est pas respectée. De plus, le fait de diviser le triplet en quatre augmente la taille du graphe RDF, et rend son parcours plus long, ce qui ralentit le temps de réponses aux requêtes. Dans (Hartig et Thompson, 2014), les auteurs proposent une alternative à la réification telle que l'a définie le W3C, en créant une nouvelle syntaxe qui admet de prendre en sujet un triplet et non pas une ressource. A cette syntaxe RDF*, ils définissent un langage alternatif à SPARQL, SPARQL* qui interroge des graphes RDF*. Cependant, à notre connaissance, aucun raisonneur ne permet d'inférer de nouvelles connaissances à partir des triplets RDF*.

Gestion de l'incertitude.

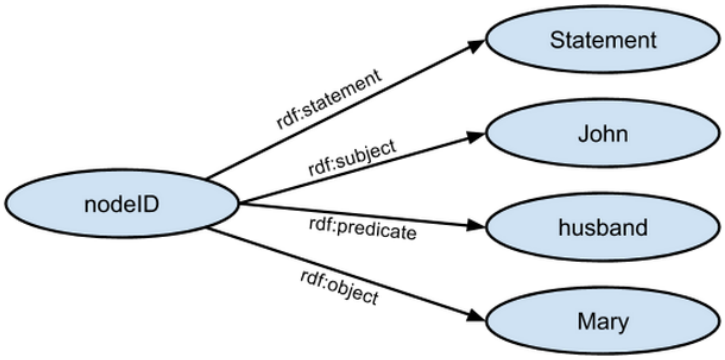


FIG. 3 – *représentation de l'incertitude avec reification.*

### 4.2.2 Notre approche

Notre approche consiste à considérer l'incertitude comme une connaissance à part entière et non pas comme une simple métadonnée à ajouter au triplet. Pour cela, nous avons décidé de créer dans l'ontologie une classe, nommée *Uncertainty*, pour modéliser cette incertitude. Elle servira à décrire ce qui est incertain dans le texte. Cette classe est décrite par trois propriétés :

- weight : une propriété littérale pour quantifier l'incertitude identifiée.
- isUncertain : propriété object qui aura pour co-domain le top-concept, cela veut dire que tout concept de l'ontologie pourra être visé par une incertitude.
- hasUncertainProp : une propriété object qui servira d'intermédiaire entre le domaine initial de la propriété et la propriété en question

La figure 4 permet de modéliser l'incertitude exprimée dans l'exemple 2.1. L'auteur de l'article indique que la condamnation des accusés est probable donc incertaine. Pour cela nous avons créé une instance de Uncertainty qui nous permet de qualifier l'information incertaine (l'ensemble de la condamnation) ainsi que de la quantifier grâce au poids associé à cette incertitude grâce au terme marqueur "probable". D'un autre coté, nous avons déduit que le tribunal pourrait peut être se situer en Papouasie, nous l'avons donc ajouté avec un degré d'incertitude adéquat.

L'ontologie que nous avons développé est indépendante de tout domaine d'application. Dès lors, elle peut être ajoutée à toute autre ontologie voulant prendre en compte l'incertitude.

À noter également que le degré de confiance associée (cf 3.1) à la source sera répercuté sur le reste des connaissances extraites du texte.

# 5 Conclusion et perspectives

Dans cet article nous nous sommes intéressés au traitement de l'information incertaine dans le cadre d'une extraction de connaissances à partir de texte. Le traitement repose sur les technologies du web sémantique pour permettre de faire le lien avec les données du Linked Open Data.

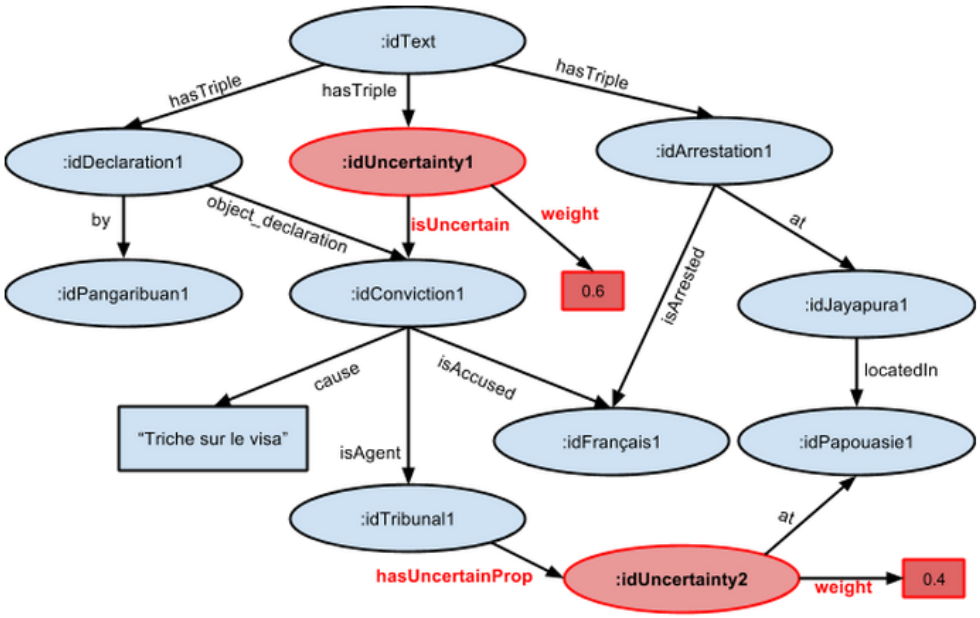


FIG. 4 – *Représentation RDF de l'incertitude.*

Notre démarche consiste à identifier les différentes situations où une incertitude remettant en cause la validité de l'information peut subsister. Nous proposons une ontologie pour modéliser l'information incertaine et la représenter au format RDF.
Nous travaillons actuellement sur développement d'un ensemble de patterns pouvant faciliter l'interrogation du graphe RDF prenant en compte notre représentation de l'incertitude. Nous prévoyons par la suite de développer un raisonneur basé sur le formalisme des logiques possibilistes afin de permettre l'inférence sur les données incertaines.

## Références

Bizer, C. et R. Cyganiak (2009). Quality-driven information filtering using the wiqa policy framework. *Web Semantics : Science, Services and Agents on the World Wide Web 7*(1), 1–10.

Bizer, C., T. Heath, K. Idehen, et T. Berners-Lee (2008). Linked data on the web (ldow2008). In *Proceedings of the 17th international conference on World Wide Web*, pp. 1265–1266. ACM.

Cyganiak, R., D. Wood, et M. Lanthaler (2013). Rdf 1.1 concepts and abstract syntax. *World Wide Web Consortium, Working Draft WD-rdf11-concepts-20130723*.

Dong, X., E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, et W. Zhang (2014). Knowledge vault : a web-scale approach to probabilistic knowledge

fusion. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pp. 601–610.

Flemming, A. (2010). Quality characteristics of linked data publishing datasources. *Master's thesis, Humboldt-Universität of Berlin*.

Hartig, O. (2008). Trustworthiness of data on the web. In *Proceedings of the STI Berlin & CSW PhD Workshop*. Citeseer.

Hartig, O. et B. Thompson (2014). Foundations of an alternative approach to reification in rdf. *arXiv preprint arXiv :1406.3399*.

Knuth, M., J. Hercher, et H. Sack (2012). Collaboratively patching linked data. *arXiv preprint arXiv :1204.2715*.

Köpcke, H. et E. Rahm (2010). Frameworks for entity matching : A comparison. *Data & Knowledge Engineering 69*(2), 197–210.

Lebo, T., S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, et J. Zhao (2013). Prov-o : The prov ontology. *W3C Recommendation, 30th April*.

Missier, P., K. Belhajjame, et J. Cheney (2013). The w3c prov family of specifications for modelling provenance metadata. In *Proceedings of the 16th International Conference on Extending Database Technology*, pp. 773–776. ACM.

Mombrun, Y., A. Pauchet, B. Grilhères, S. Canu, et al. (2010). Collecte, analyse et évaluation d'informations en sources ouvertes. In *Atelier COTA des 21es Journées francophones d'Ingénierie des Connaissances*.

Niu, F., C. Zhang, C. Re, et J. W. Shavlik (2012). Deepdive : Web-scale knowledge-base construction using statistical learning and inference. In *Proceedings of the Second International Workshop on Searching and Integrating New Web Data Sources, Istanbul, Turkey, August 31, 2012*, pp. 25–28.

Semantics, R., P. Hayes, W. W. W. Consortium, et al. (2004). W3c recommendation. *Reification, Containers, Collections and rdf : value*.

Smets, P. (1997). Imperfect information : Imprecision and uncertainty. In *Uncertainty Management in Information Systems*, pp. 225–254. Springer.

Zaveri, A., A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, et P. Hitzler (2013). Quality assessment methodologies for linked open data. *Submitted to Semantic Web Journal*.

# Summary

The knowledge representation area needs some methods that allow to detect and handle uncertainty. Indeed, a lot of text hold information whose the veracity can be called into question. These information should be managed efficiently in order to represent the knowledge in an explicit way. As first step, we have identified the different forms of uncertainty during a knowledge extraction process, then we have introduce an RDF representation for these kind of knowledge based on an ontologie that we developped for this issue.