HAL
open science

# KONTRAST : création d'un glossaire contrastif à partir d'un corpus de normes internationales

Martin Lafréchoux, Brigitte Juanals, Jean-Luc Minel

▶ **To cite this version:**

Martin Lafréchoux, Brigitte Juanals, Jean-Luc Minel. KONTRAST : création d'un glossaire contrastif à partir d'un corpus de normes internationales. Université de Liège. JADT 2012, Jun 2012, Liège, Belgique. pp.563-575, 2012. halshs-00709145

**HAL Id: halshs-00709145**

**https://shs.hal.science/halshs-00709145**

Submitted on 18 Jun 2012

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# KONTRAST : création d'un glossaire contrastif à partir d'un corpus de normes internationales

Martin Lafréchoux[1], Brigitte Juanals[1], Jean-Luc Minel[1],

[1]MoDyCo UMR 7114– Belgique Université Paris Ouest Nanterre La Défense - CNRS - France.

**Abstract**

The work discussed in this paper was commissioned by the French National Research Agency (ANR). The NOTSEG project studies standardization and global security. It aims to identify influence factors at play during the drafting phase of national and international standards. In this paper we present KONTRAST, an ontological contrastive glossary created using a corpus of 18 standards relating to business continuity. The approach used to create KONTRAST was shaped by the semantic and interpretative issues we encountered during the construction process, most notably the difference between terminology and ontology. After a brief discussion of previous and related works, we present the unique characteristics of terminology in the context of process standardization. We then describe the knowledge representation model upon which our glossary is built using RDF, OWL and SKOS. We then present the series of techniques used to create the contrastive glossary. Finally we present a use case study illustrating how this new ontological and terminological resource can be leveraged to detect influence.

**Résumé**

Cet article présente un travail développé dans le cadre du projet ANR NOTSEG. Ce projet vise notamment à identifier des influences dans le processus d'élaboration des normes de management. Nous présentons tout d'abord les contraintes terminologiques spécifiques au domaine de la normalisation. Nous expliquons ensuite la démarche adoptée pour construire automatiquement un glossaire contrastif ontologique, dénommé KONTRAST, élaboré à partir d'un corpus de 18 normes nationales et internationales du domaine de la continuité d'activité (*business continuity*) en langue anglaise. Ce processus de construction soulève des problèmes de sémantique et d'interprétation complexes et l'article illustre les différences de nature entre une représentation ontologique et une représentation terminologique. Nous décrivons le modèle de représentation adopté qui s'appuie sur RDF, OWL et SKOS. Nous présentons un cas d'utilisation qui illustre comment ce type d'outil permet de détecter des indices d'influence.

**Mots-clés :** ontology ; terminology ; business continuity ; process standardization

## 1. Introduction

Ce travail s'inscrit dans le projet de recherche ANR "NOTSEG"[1]. Le projet NOTSEG étudie la normalisation industrielle de la sécurité et des risques, et plus particulièrement le domaine de la continuité d'activité (*business continuity*), au travers d'un corpus de textes composé d'un ensemble de vingt normes nationales et internationales, en anglais, dans le domaine de la continuité d'activité et de la gestion des risques et des crises. La continuité d'activité désigne l'ensemble des procédures et mesures mises en place par une organisation (entreprise, institution, collectivité...) pour s'assurer que ses fonctions critiques seront préservées en cas de sinistre ou de désastre (catastrophe naturelle, attentat, risque sanitaire…). Notre corpus comporte des textes nationaux et internationaux de manière à ce qu'il soit possible de confirmer ou d'infirmer, en procédant de manière comparative, l'hypothèse de jeux d'influence entre des textes normatifs (Juanals 2011). Selon une approche socio-pragmatique (Chateauraynaud, 2011) adaptée aux thèmes en construction et aux controverses dont nous traitons, le terrain et les acteurs sont privilégiés pour construire le corpus de textes. Le corpus est envisagé comme une « archive

[1] Normalisation et Sécurité Globale : la formulation du concept de sécurité globale dans la normalisation (appel ANR-CSOG 2009).

dynamique » faisant l'objet de révisions continues au fur et à mesure de la progression de la recherche. De ce fait, une méthode inductivo-déductive inspirée de la linguistique référentielle (Condamines 2005), qui amène à constituer un corpus autour d'un macro-thème, serait insuffisante. De plus, elle ne rendrait pas compte des affrontements d'acteurs. Dans notre recherche, les termes ne sont pas assez stables pour servir de repère à la construction d'un corpus ; par exemple, les distinctions entre *emergency preparedness, business continuity, service continuity, resilience,* font l'objet de controverses.

La finalité du projet est de dresser une cartographie des cadres (institutionnels, juridiques, techniques, géopolitiques) de normalisation existants et de leurs acteurs ; parallèlement, le projet doit fournir des outils d'analyse, de veille et de pilotage à l'intention d'organismes publics, d'industriels et d'entreprises utilisatrices : outils d'aide à la consultation des normes, et outils d'exploration du domaine. Dans cet article, nous présentons la création de l'un de ces outils : un glossaire contrastif ontologique, dénommé KONTRAST, du domaine de la continuité d'activité. KONTRAST est une ressource terminologique qui prend la forme d'un réseau de concepts représentée à l'aide des langages RDF, OWL et SKOS. Sa structure complexe permet de représenter ensemble les différents systèmes terminologiques concurrents (glossaires, guides terminologiques, etc.) de notre corpus, afin de pouvoir les explorer et étudier leurs éventuelles dépendances.

La section 2 présente des travaux connexes articulant ressources terminologiques et outils de gestion de la connaissance en fonction des problématiques particulières de différents domaines (industrie, finance, recherche médicale). La section 3 présente le contexte de notre travail et justifie nos choix de modélisation, puis détaille la chaîne de traitements utilisée pour réaliser KONTRAST. La section 4 présente un cas d'utilisation. Enfin, la section 5 conclut en discutant les résultats obtenus et les perspectives de recherche.

## 2. Travaux connexes

La terminologie, "pratique théorisée" (Roche, 2005), a connu un développement rapide au cours des dernières décennies, notamment pour répondre aux besoins des industries manufacturières (Condamines, 2005). Cette forte opérationnalisation de la terminologie a conduit à la naissance d'une relation "symbiotique" entre terminologie et ingénierie des connaissances (Condamines, 2005). Les approches mêlant ontologie formelle et ressources terminologiques qui en résultent sont très diverses, et dépendent fortement de leur contexte de réalisation.

Pour les industriels travaillant sur plusieurs sites ou avec de nombreux sous-traitants, par exemple dans les secteurs de l'automobile ou de l'aviation, l'emploi d'un vocabulaire sans équivoque, partagé et contrôlé est absolument nécessaire. Récemment, (Omrane et al., 2011) ont présenté un prototype de système de gestion de règles métier (Business Rules Management System, ou BRMS) fondé sur une ontologie OWL et des concepts exprimés selon le formalisme SKOS. Développé pour Audi dans le cadre du projet européen ONTORULE, le BRMS s'appuie sur une "ontologie lexicalisée" pour formaliser des règles métier écrites en langage naturelle par des experts métier. Ce système doit permettre d'accélérer le partage d'informations entre des départements qui utilisent parfois des termes différents pour désigner les mêmes procédures.

(Roche, 2007 ; Damas et Tricot, 2010) ont proposé le modèle plus général de l'ontoterminologie, une "terminologie dont le système notionnel est une ontologie formelle". Dans ce cadre, le terminologue doit travailler avec les utilisateurs et les experts du domaine pour établir une

conceptualisation du domaine susceptible de faire consensus. Cette représentation est ensuite transposée sous forme d'ontologie, puis des termes de spécialité fournis et/ou validés par des experts sont associés à chaque concept.

Dans un cadre très proche de celui de NOTSEG, (Gresser, 2010) a présenté une ontologie des risques financiers, domaine qui recoupe en partie la continuité d'activité. Son ontologie est basée sur une représentation conceptuelle du domaine qu'il a élaboré en se fondant sur sa propre expertise. La composante terminologique de son travail est limitée à deux "listes de vocabulaire" accompagnant l'ontologie, sous forme de glossaires alphabétiques.

Dans le domaine médical, la multiplication des terminologies de spécialité est parfois problématique, notamment pour les tâches d'indexation. La solution la plus simple consiste à recourir au vocabulaire UMLS[2]. Cependant, pour (Roumier et al., 2011), vouloir recourir à tout prix à un vocabulaires contrôlé unifié est "naïf", et pourrait "nuire à la santé des patients". Ils proposent un système permettant de préserver les caractéristiques propres des différentes terminologies utilisées dans un système de recherche d'informations médicales. Leur système utilise une "terminologie d'interface" multilingue et interdisciplinaire, couplée à des terminologies spécialisées et généralement monolingues, destinées aux utilisateurs finaux. Dans une optique proche, (Grosjean et al., 2011) ont présenté EHTOP (European Health Terminology/Ontology Portal), un service offrant un accès unifié à trente-deux terminologies médicales, destinés à des utilisateurs humains ou à des agents logiciels. EHTOP intègre les différentes terminologies en les transposant dans un méta-modèle générique, puis en effectuant des "harmonisations sémantiques inter et intra-sémantiques".

Comme l'illustrent ces exemples, dans les domaines où l'ingénierie des connaissances a connu le plus de succès, les conflits et incohérences sont résolus grâce à la présence d'une autorité centrale (la direction d'un groupe industriel), à une expertise reconnue (expert mandaté par une organisation) ou à un consensus autour d'un vocabulaire unifié préexistant (par exemple UMLS dans le domaine médical). Il n'existe rien de comparable dans le domaine de la normalisation dans lequel se situe le projet NOTSEG : l'ISO ne dispose pas de l'autorité nécessaire pour imposer un vocabulaire commun, et chaque instance de normalisation demeure libre de définir comme elle l'entend les termes employés dans ses normes.

## 3. L'approche du projet NOTSEG

Notre approche doit être adaptée aux problématiques spécifiques du domaine (la normalisation des processus) et permettre de répondre aux exigences particulières du projet NOTSEG. Le volet veille du projet nécessite notamment d'étudier ensemble les vocabulaires issus de chaque norme, afin de pouvoir les comparer et les explorer. Nous avons donc conçu notre ontologie de manière à ce qu'elle puisse représenter simultanément plusieurs systèmes notionnels parallèles. Si notre approche s'appuie sur les propriétés opérationnelles des ontologies en RDF/OWL, elle n'a pas recours à la démarche terminologique telle que décrite par (Roche, 2007). C'est la raison pour laquelle nous ne reprenons pas le terme d'ontoterminologie, auquel nous préférons celui de glossaire contrastif ontologique.

[2] Unified Medical Language System - http://www.nlm.nih.gov/research/umls/

### *3.1. Le contexte de la normalisation des normes de management*

La normalisation des processus des normes de management se trouve dans une situation très différente de celle des produits industriels. Dans l'industrie, les terminologies désignent par des termes des réalités physiques mesurables. A l'inverse, l'objet de la normalisation des processus des normes de management relève de réalité abstraite que les experts du domaine appellent des concepts (le risque, la résilience, etc.). Il n'existe pas de référent tangible auquel se reporter en cas de désaccord ou d'incompréhension sur ce que désigne ce concept. Comme le souligne (Roche, 2005) il est impossible de "décrire l'objet tel qu'il nous apparaît, sans spéculation d'aucune sorte". En l'absence de référentiel physique objectivement quantifiable ou d'une autorité capable de faire consensus, les enjeux terminologiques de la normalisation des processus se trouvent subordonnés à l'introspection des experts qui prennent part au processus de normalisation, ainsi qu'à des influences externes d'ordre stratégique et technopolitiques (Minel et Juanals, 2010). Le résultat est une multiplication des référentiels, qui conduit à la cohabitation de différents systèmes terminologiques concurrents. Cette situation se traduit par des emplois et des définitions divergentes d'un même terme, selon qu'on se réfère à une norme ou à une autre. Les différents vocabulaires peuvent présenter des correspondances entre eux, mais, en l'absence d'une autorité d'arbitrage , il n'y a aucune raison pour que ces vocabulaires soient homogènes.

### *3.2. Les sources langagières du glossaire*

#### *3.2.1. La terminologie dans les normes*

Les terminologies et glossaires sont extrêmement fréquents dans le champ de la normalisation des processus. Chaque norme dispose de son propre vocabulaire, employé dans un objectif de "stipulation" (Depecker, 2005) : au début de chaque norme se situe une section "Termes et définitions" (T&D) qui présente, généralement sous forme d'un glossaire alphabétique, les termes employés dans le texte de la norme. Le choix de ces termes et de la définition qui leur est donnée sont de première importance : cette opération apparemment abstraite conditionne la mise en oeuvre concrète de la norme.

Les termes définis peuvent être spécifiques au domaine, ou au contraire des termes courants. Dans ce second cas, la définition vise à préciser l'acception dans laquelle le terme sera employé dans le cadre de la norme.

Exemple [3]

`Activité (BP – Z – 400) : ensemble de processus qui concourent à la réalisation d'objectifs bien définis`

Les définitions données dans les T&D d'une norme s'appliquent à cette norme, et cette norme seulement. Il est donc fréquent que différentes normes offrent des définitions différentes d'un même terme[4] :

Exemple :

`Crisis (ISO/IEC Guide 81:2010) : incident(s), human-caused or natural, that requires urgent attention and action to protect life, property, or environment`

`Crisis (AS/NZS 5050:2010) : Situation that is beyond the capacity of normal`

[3] La référence de la norme source est indiquée entre parenthèse.
[4] Par ailleurs, comme le montre l'exemple, des notes viennent parfois compléter ou modifier une définition.

management structures and processes to deal with effectively. – NOTE: A crisis may require significant diversion of management time, attention and resources away from normal, routine operations to respond to the situation.

En pratique, il serait cependant fastidieux et peu économique de rédiger systématiquement une nouvelle définition pour chaque terme utilisé dans une norme. Il arrive donc fréquemment que la section T&D d'une norme cite ou reprenne explicitement les définitions d'une norme préexistante :

- Lorsque tous les termes définis dans une autre norme sont employés, la section T&D précise simplement "Les définitions de la norme X s'appliquent".
- Dans d'autres cas, notamment lorsqu'une norme ne reprend qu'une partie des termes d'une autre norme, la définition d'un terme est reproduite partiellement ou en intégralité.
- On trouve également des définitions explicitement 'adaptées' d'une autre définition.
- Au sein des définitions elles-mêmes, il est parfois fait référence à d'autres définitions d'une norme, notamment au sein des guides terminologiques.

Ces réseaux de citations, qui ne sont pas toujours explicités, forment un système complexe d'emprunts et de références, tant au sein de chaque norme (réseau interne) qu'entre les normes du corpus (réseau externe).

*3.2.2. Le glossaire source*

L'un des livrables du projet NOTSEG est un glossaire structuré synthétique et représenté en XML (Malik 2011) élaboré manuellement à partir du corpus des 18 normes, où sont consolidés les différents glossaires et ressources terminologiques du corpus. C'est à partir de ce glossaire que KONTRAST a été construit.

Le glossaire rassemble :

- Les termes et définitions des normes du corpus
- Le contenu des guides terminologiques
- Les relations entre les termes décrites au paragraphe précédent

Pour chaque terme, le glossaire indique les différentes définitions provenant de l'ensemble des normes du corpus, ainsi que les termes alternatifs éventuels. Les relations d'emprunts et d'héritage sont également représentées lorsqu'elles étaient explicites et ont pu être relevées lors de la constitution du glossaire. Au total, ce glossaire contient 726 définitions de 291 termes issus des 18 normes du corpus. La DTD utilisée s'inspire de travaux sur les ressources terminologiques de spécialité en XML, notamment (Joseph, 2010). Ce glossaire ne contient pas de liens hypertextes.

***3.3. Choix techniques et méthodologiques***

Dans le contexte dans lequel NOTSEG s'insère, le travail d'harmonisation terminologique fait partie du coeur de métier des instances de normalisation. C'est à elles qu'il appartient de désigner le mot 'juste' et de le définir, au cours d'un processus de rédaction qui fait intervenir différents experts du domaine. Pour chaque nouvelle norme, les instances de normalisation produisent une terminologie qui se veut complète et cohérente, et aspire à faire consensus. Ce travail, bien qu'effectué par des non-terminologues, nous paraît relever d'une démarche comparable à celle décrite par (Roche, 2007). Les termes définis en ouverture de la norme sont par exemple

organisés dans des diagrammes de type PDCA (Plan-Do-Check-Act) qui ne sont pas sans rappeler les réseaux conceptuels des terminologues.

Nous formulons l'hypothèse que ces sections "Termes et définitions" sont la manifestation du système notionnel utilisé par les experts qui les ont rédigées. La réalisation de KONTRAST vise à représenter le plus fidèlement possible les différentes conceptualisations exprimées par les définitions données aux termes de spécialité. Nous n'avons pas cherché à choisir entre les différentes variations des définitions données à un même terme, ces variations constituant l'objet de notre étude sur les influences.

Les concepts représentés dans KONTRAST sont donc issus des termes relevés dans les glossaires qui accompagnent les normes (sections "Termes & définitions"). Dans un glossaire destiné à être consulté par des utilisateurs humains, le fait qu'il existe une confusion entre terme et concept ne pose pas de problème particulier. Cependant, la nature abstraite de la normalisation des processus rend la ligne de démarcation entre le terme et le concept nettement plus floue que dans un cadre industriel. Paradoxalement, l'absence de référent physique fait des termes de spécialité le seul élément invariant et observable du domaine.

Dans KONTRAST, à un seul terme de spécialité correspondent plusieurs concepts, qui prennent leur sens particulier en contexte : à la fois dans leur contexte d'emploi (la norme) et par rapport aux autres termes définis avec eux. Du point de vue formel, nous avons choisi de représenter les termes de spécialité comme des **méta-concepts sans valeur sémantique**, tandis que les concepts correspondent à **l'emploi d'un terme dans le contexte d'une norme** (relation entre une forme graphique, un contexte d'emploi, et une définition). Chaque concept appartient donc à deux groupes : les termes définis par la même norme que lui, et les termes employant la même forme graphique. C'est sa position dans ces deux réseaux qui constitue sa spécificité et fait de lui un concept distinct.

Enfin, KONTRAST contient des informations relatives aux instances de normalisation ayant rédigé ou publié les normes du corpus. L'un des objectifs est en effet d'aider à l'analyse des jeux d'influence entre les instances de normalisation au niveau organisationnel. Cette composante organisationnelle est reliée à la partie terminologique par le biais des individus[5] représentant les normes.

### *3.4. Modélisation*

#### *3.4.1. Modélisation des informations organisationnelles*

Dans un premier temps, nous avons modélisé dans KONTRAST les relations entre instances de normalisation nationales et internationales, ainsi que les différents comités et groupes travaillant à l'élaboration des normes de notre corpus. La retranscription a été effectuée de manière relativement simple et directe. En effet, les structures hiérarchisées telles que celles des instances de normalisations se prêtent bien à une représentation sous forme de graphe de relations. Cette partie de KONTRAST utilise le vocabulaire spécialisé OWL dcterms[6], correspondant aux métadonnées du Dublin Core. Lorsque dcterms n'était pas suffisamment précis, par exemple pour définir le rôle exact joué par différentes instances de normalisation dans l'élaboration d'une norme, nous avons créé des sous-relations explicitant le lien que nous voulions représenter. On

[5] Au sens où on l'entend en Ingénierie des connaissances
[6] http://dublincore.org/documents/dcmi-terms/

peut ainsi bénéficier de l'interopérabilité de Dublin Core sans rien sacrifier en termes de précision. Certaines propriétés spécifiques à notre ontologie, précisant par exemple le statut (/hasStatus/) et la portée (/hasReach/) de la norme, sont définies en parallèle (cf. figure 1).

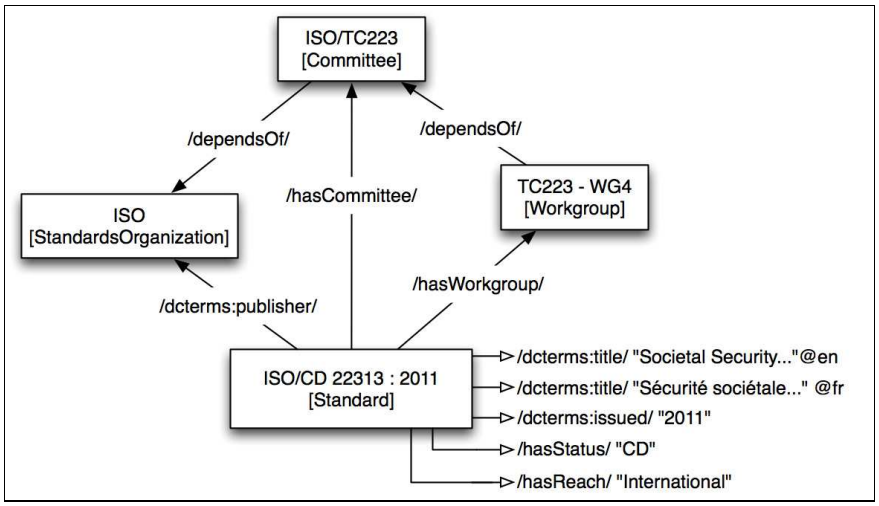


*Figure 1 : Propriétés et relations d'une norme [Standard]*

*3.4.2. Modélisation des informations terminologiques*

Les sections termes et définitions correspondent à la définition d'un thésaurus donnée par la norme ISO 25964–1:2011 : "vocabulaire contrôlé et structuré dans lequel les concepts sont représentés par des termes, organisés de façon à ce que des relations entre les concepts soient explicitées, et dont les termes préférentiels sont accompagnés par des entrées vers leurs synonymes ou quasi-synonymes."[7] Nous avons donc choisi d'utiliser le vocabulaire SKOS[8] pour définir le schéma de KONTRAST. Le glossaire de chaque norme du corpus a donc été transposé dans KONTRAST sous forme d'un 'ConceptScheme' SKOS.

***3.5. Chaîne de traitement***

KONTRAST a été peuplée automatiquement, à partir du contenu du glossaire structuré en XML présenté à la section 3.2.2. Les termes et relations décrites dans le glossaire ont été transformés en utilisant des templates XSLT pour les transformer en RDF/OWL. Lors de la transformation XSLT, les entrées du glossaire sont distribuées entre :

- la forme graphique du terme ("entree_glossaire"), de type [skos:Collection], qui regroupera tous les concepts utilisant le même terme de spécialité ;
- les emplois dans ce terme dans les normes, de type [skos:Concept] ;
- les différentes propriétés du concept, rendues par diverses propriétés SKOS ;

[7] La Partie 2 de la norme ISO 25964–1:2011 portera sur l'interopérabilité entre thésaurus et d'autres ressources terminologiques, notamment les ontologies. A l'heure actuelle, cette seconde partie en est toujours au stade de Comittee Draft.
[8] http://www.w3.org/2004/02/skos/

- la ou les normes auxquelles le concept appartient, rendues par un individu [skos :ConceptScheme] ;
- et les relations entre individus, rendues par des propriétés /skos:related/ ou /skos:Match/.

Chaque individu dans KONTRAST (cf. figure 3) reçoit un identifiant unique, stable et référençable calculé en appliquant une procédure unique de normalisation et de concaténation des informations du glossaire. Par exemple, l'identifiant unique d'un individu [skos:Concept] est formé en concaténant ses deux caractéristiques distinctives : le terme de spécialité correspondant au concept et l'identifiant de la norme où il est employé. Pour éviter toute ambiguïté, on ajoute le type d'individu à la fin de l'identifiant ('-Concept').

Exemple :

Les informations ci dessous présentes dans le glossaire structuré :

```
<feat type="entree_glossaire">communication and consultation</feat>
<UT type="norme">
<feat type="name">NF/ISO 31000 :2009</feat>
</UT>
```

sont utilisées pour calculer l'identifiant :

```
rdf:about="http://www.notseg.fr/business-
continuity.owl#communication_and_consultation-NF_ISO_31000-2009-Concept"
```

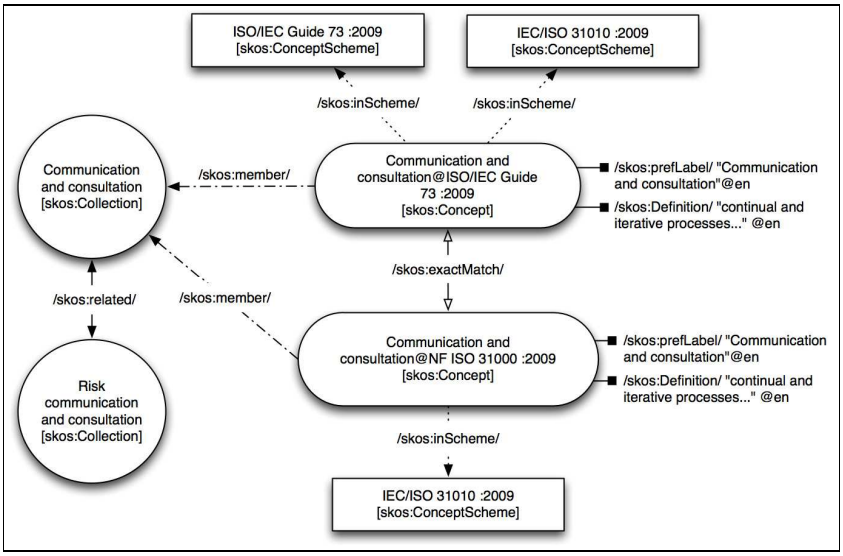


*Figure 2 : Organisation dans KONTRAST des différentes composantes extraites du glossaire*

Les différentes références à la même entité au sein du glossaire (par ex. une norme) doivent donc conduire au calcul d'un même identifiant unique dans l'ontologie. Cependant, en dépit du soin apporté à la constitution du glossaire, quelques variations ont échappé à la vigilance du rédacteur. Si un lecteur humain ignore ces différences insignifiantes, par contre la procédure de calcul produit deux identifiants différents. Ainsi, on rencontre par exemple les graphies suivantes au sein de notre corpus pour désigner l'ISO Guide 73:2009 :

```
ISO/IEC Guide 73
ISO Guide 73:2009
```

Guide 73: 2009
ISOGuide73:2009
ISO/IEC Guide 73: 2009
ISO GUIDE 73

Les variations de ce type rendent difficile la production automatique d'un identifiant unique pour chacune de ces formes graphiques. Nous avons donc développé des algorithmes implémentés sous forme d'expressions régulières à l'aide du logiciel TextMate pour résoudre ces variantes.

- Ajout de relations non-explicites

Lors de la transposition du glossaire, des relations sont automatiquement extraites afin de lier les concepts ayant des définitions proches ou identiques, mais qui n'étaient pas explicitement référencées comme telles. Ces relations implicites n'ont pas été relevées manuellement lors de la constitution du glossaire. La proximité de ces définitions est évaluée uniquement en fonction de leur contexte, sans prise en compte de leur contenu sémantique. Pour extraire les relations, nous effectuons une série de tests qui comparent les définitions correspondant aux mêmes termes, en allant du test le plus précis (identité parfaite - /skos:closeMatch/) au plus lâche (parenté - /skos:relatedMatch/). Lorsqu'un test réussit, la relation identifiée est transcrite sous forme de triplet RDF/OWL. Cette approche permet de maximiser la précision des résultats.

- **/skos:exactMatch/ :** Une relation /skos:exactMatch/ correspond à une identité complète entre deux définitions. Pour extraire ces relations, nous avons calculé la distance de Levenshtein entre le texte de deux définitions.
- **/skos:closeMatch/ :** Souvent, une définition est citée en totalité, mais avec de petites variations graphiques, les conventions typographiques différant d'un organisme de normalisation à l'autre. Ces relations de quasi-identité sont bien exprimées par **/skos:closeMatch/ :** Les variations formelles, par exemple le mode de numérotation des paragraphes, font augmenter considérablement la distance entre deux définitions par ailleurs identiques. Levenshtein ne constituait donc pas une approche satisfaisante. Pour le test 2, on normalise le texte des définitions (tokenization, nettoyage de la ponctuation, ~~passage en minuscule~~uniformisation de la casse, stemming) avant de les comparer.
- **/skos:relatedMatch/ :** Deux définitions peuvent avoir une première partie identique, et n'être différenciées que par les notes qui suivent cette première partie. Dans KONTRAST, ces relations sont rendues par /skos:relatedMatch/. Le test 3 commence par analyser la structure de la définition pour en distinguer la partie principale et les notes, puis il applique le même traitement qu'au test précédent.
- **Résultats :** Dans un contexte tel que la normalisation, la fiabilité des informations extraites automatiquement était critique. Les traitements choisis privilégient donc la précision sur le rappel. La taille relativement restreinte de notre corpus de travail a permis d'évaluer les résultats obtenus par rapport aux relations entrées manuellement. Avec des règles strictes, nos tests permettent d'identifier 389 relations avec une précision de 100% :

| | **/skos:exactMatch/** | **skos:closeMatch/** | **/skos:relatedMatch/** |
|---|---|---|---|
| **Rappel** | 1.0 | 0.38 | 0.21 |
| **Précision** | 1.0 | 1.0 | 1.0 |

Le faible rappel des tests 2 et 3 s'explique essentiellement par la latitude laissée aux personnes chargées d'identifier manuellement les relations, qui pouvaient estimer proches deux définitions trop différentes l'une de l'autre pour nos tests.

- Intégration dans KONTRAST

Toutes les données extraites dans les étapes précédentes, représentées en RDF, ont été importées dans l'éditeur Protégé, ce qui permet de tester l'intégrité des données et d'interroger le SPARQL Endpoint.

## 4. Cas d'utilisation

KONTRAST permet de modéliser ensemble toutes les informations terminologiques relevées dans les documents de notre corpus, ainsi que les informations organisationnelles sur les instances de normalisation elles-mêmes. Totalisant plus d'un millier d'individus (normes, concepts, institutions), il constitue ainsi un outil puissant pour l'analyse du domaine de la continuité d'activité.

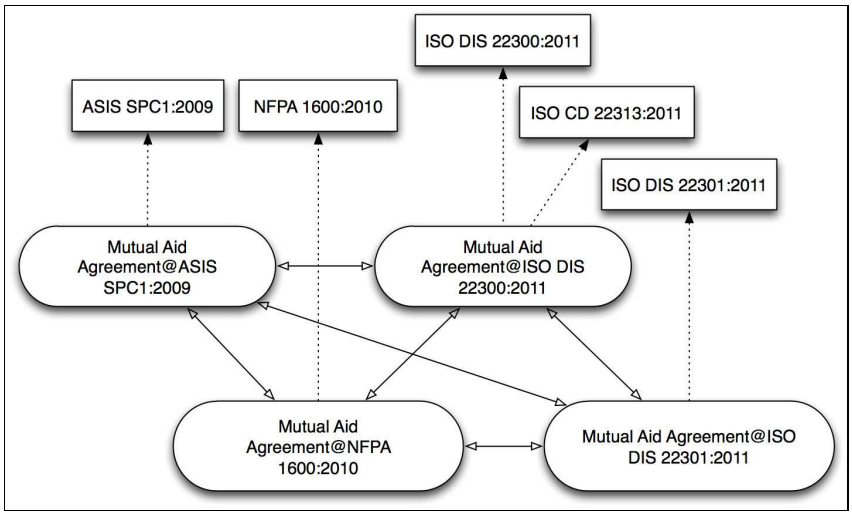


*Figure 3 : Résultat de la requête pour les définitions de « Mutual Aid Agreement »*

L'hypothèse formulée dans le cadre du projet NOTSEG (Juanals 2011) est que la possibilité de citer ou de reprendre une définition issue d'une norme ou d'un référentiel préexistant dessine des réseaux de citations et d'influence. Remonter le fil de ces héritages doit permettre d'identifier des familles de normes, et donc d'élucider les choix terminologiques ayant présidé à la rédaction de chaque norme de notre corpus. L'exploration des relations représentées dans KONTRAST répond à ce besoin en permettant de mettre en évidence les chemins entre différents individus de celle-ci. Pour illustrer cet usage, formulons l'hypothèse qu'un utilisateur cherche à mettre en évidence des indices d'influence d'une norme nationale sur une norme internationale. On pourra par exemple rechercher des paires de concepts ayant des définitions très similaires, l'une provenant d'une norme nationale et l'autre d'une norme internationale, en formulant la condition que la définition issue d'une norme nationale est plus ancienne que celle provenant de la norme internationale.

Sur la figure 3 ci-dessus, qui représente le résultat de la requête sous forme graphique, on voit par exemple que les différentes définitions de "Mutual Aid Agreement" sont liées entre elles par

des relations de similarité (/skos:relatedMatch/ ou /skos:closeMatch/). La définition la plus ancienne est celle de la norme ASIS, publiée en 2009. Elle a ensuite été reprise dans la norme NFPA en 2010, avant de servir de base aux définitions des normes ISO de la série 22300, qui sont en cours de rédaction en 2011. Un tel réseau de relations constitue un indice d'influence qui pourra être soumis à l'attention d'experts pour être confirmé ou infirmé.

## 5. Conclusion

Dans cet article, nous avons d'abord présenté plusieurs techniques combinant ontologies OWL et ressources terminologiques dans différents contextes (industrie, finance, recherche médicale) afin d'illustrer l'importance d'une instance ayant autorité pour exercer un arbitrage. Nous avons ensuite présenté le contexte de notre propre projet, dans lequel une telle instance n'existe pas et la solution que nous avons conçue pour répondre à ces problèmes spécifiques. Nous avons détaillé nos choix de modélisation, puis la chaîne de traitements utilisée pour réaliser KONTRAST. Enfin, nous avons présente un cas d'utilisation qui illustre les potentialités offertes par ce mode de représentation.

La faible taille de notre corpus explique la taille modeste de la ressource terminologique produite (651 concepts uniques représentant 291 termes, liés par plusieurs milliers de relations). Néanmoins, cette taille doit être rapportée à celle d'autres ressources terminologiques du domaine : l'ISO Guide 73:2009, qui fait référence dans le domaine, ne comporte par exemple que 51 définitions.

Par ailleurs**,** la valeur ajoutée de KONTRAST, par rapport à un glossaire, réside dans les possibilités offertes par l'exploration transversale du contenu, qu'elles soient visuelles ou logicielles (requêtes SPARQL complexes). Par rapport à une base de données relationnelle classique, les progrès sont plus marginaux. Il s'agit essentiellement des possibilités d'interconnexion avec le *linked data* global du web sémantique. C'est dans ce but que nous avons cherché à nous plier aux contraintes de vocabulaires existants. Le site DBPedia définit déjà plusieurs normes internationales comparables à celles de notre corpus, mais pas encore celles que nous avons étudiées. Lorsque ce sera le cas, nous prévoyons de lier KONTRAST à ces données en utilisant des assertions /owl:sameAs/.

Nous envisageons plusieurs directions de recherche afin de répondre aux besoins des différents acteurs du projet NOTSEG.

• Experts : Enrichir KONTRAST grâce à des relations extraites semi-automatiquement des textes du corpus. Ces relations seraient validées par des experts du domaine avant d'être intégrées à l'ontologie. La structure particulière de notre ontologie permettra de 'typer' ces relations, en intégrant le contexte institutionnel de leur rédaction.

• Utilisateurs : Analyser les besoins des utilisateurs et bâtir une plateforme d'aide à la navigation textuelle dans les normes basée sur l'ontologie, par exemple en utilisant le projet Annotation Ontology (AO)[9] de l'université de Harvard. AO est un vocabulaire permettant d'intégrer une ontologie avec un corpus de documents, sous forme d'annotations. Il s'agit de progresser en termes d'interface / ergonomie pour se mettre à portée des experts du domaine, qui doivent gérer seuls les difficultés d'interprétation de textes contradictoires.

• Institutions : Analyser les jeux d'influence qui président à la rédaction des normes

[9] http://code.google.com/p/annotation-ontology/

internationales. Cet objectif majeur du projet NOTSEG implique une analyse socio-pragmatique (Chateauraunaud 2010) basée sur l'observation du travail des experts et l'étude directe du fonctionnement des instances de normalisation d'une part, et d'autre part une analyse textométrique des textes de notre corpus. Cette analyse quantitative des textes de normes, en cours de réalisation, nous permettra d'y suivre la circulation d'expressions, de collocations et de fragments de texte, afin de mettre en évidence les réseaux de citation et d'influence, conscients ou non, qui traversent la normalisation du domaine.

## Références


Chateauraynaud F. (2010). Lost in Arlington, carnet de recherche « Socio-informatique et argumentation », *Hypotheses.org*. En ligne: http://socioargu.hypotheses.org/1533

Condamines A. (2005). Linguistique de corpus et terminologie. Langages, 157(1), L. Depecker éd., pp. 36-47

Damas L., & Tricot C. (2010). L'ontoterminologie pour la recherche d'information sémantique, *Actes de la conférence TOTh 2010*, C. Roche éd., Annecy, Institut Porphyre, pp. 101-116.

Depecker L. (2005). Contribution de la terminologie à la linguistique. *Langages*, 157(1), L. Depecker éd., pp. 6-13

Gresser J.-Y. (2010). Ontologies des risques financiers — Continuité d'activité, gestion de crise, protection des infrastructures critiques financières, *Actes de la conférence TOTh 2010*, C. Roche éd., Annecy, Institut Porphyre, pp. 155-176.

Grosjean J., Merabti T., Griffon N., Dahamna B. & Darmoni S. (2011). Multiterminology cross-lingual model to create the European Health Terminology/Ontology Portal, *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, Kyo Kageura & Pierre Zweigenbaum éd., Paris, INALCO, pp. 119-122.

Joseph J. (2010). Le projet NucSTML Structuration d'un dictionnaire de spécialité en vue de sa publication sur internet Bénéfices du langage XML, *Actes de la conférence TOTh 2009*, C. Roche éd., Annecy, Institut Porphyre, pp. 181-196

Juanals B. (2011). Les acteurs et les textes de la normalisation internationale de la sécurité sociétale. Construction d'une approche communicationnelle, article soumis à *Revue Sciences de la Société*.

Juanals B. & Minel J.-L. (2011). Writing and Monitoring in International Standardization, Theoretical Choices and Methodological Tools, *Proceedings of IMET2011*, Florida, USA.

Malik M., (2011). Construction d'un glossaire contrastif, *Livrable ANR-NOTSEG*.

Omrane N., Nazarenko A. & Rosina P. (2011). Lexicalized ontology for a business rules management platform: An automotive use case, *Proceedings of the 5th International Symposium on Rules (RuleML, Industry focused session)*, Fort Lauderdale, Florida.

Roche C. (2005). Terminologie et ontologie. *Langages*, 157(1), L. Depecker éd., pp. 48-62

Roche C. (2007).Le terme et le concept : fondements d'une ontoterminologie, *Actes de la conférence TOTh 2007*, C. Roche éd., Annecy, Institut Porphyre, pp. 1-22

Roumier J., Vander Stichele R.,& Romary L. (2011). Approach to the Creation of a Multilingual, Medical Interface Terminology, *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, Monique Slodzian & al. éd., Paris, INALCO , pp. 13-15