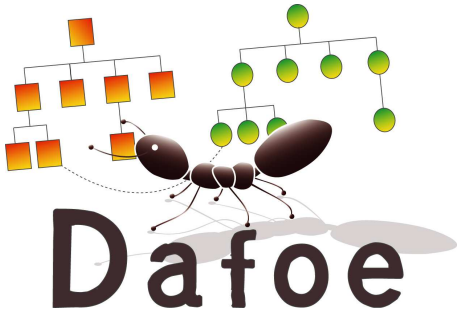
# Projet DaFOE4App DIFFERENTIAL AND FORMAL ONTOLOGIES EDITOR FOR APPLICATIONS. Cahier des charges scientifique et technique de la plateforme DaFOE. Chapitre Modèle de données

Sylvie Szulman, Nathalie Aussenac-Gilles, Adeline Nazarenko, Jean Charlet




**Projet DaFOE4App**

DIFFERENTIAL AND FORMAL ONTOLOGIES EDITOR FOR APPLICATIONS

**Dossier A.1 / Document A.1.1**

# Cahier des charges scientifique et technique de la plateforme DaFOE

*Chapitre Modèle de données*

**19 mai 2009**

Responsable du lot : Sylvie Szulman
Coordination de la rédaction : Sylvie Szulman
Auteurs : N. Aussenac-Gilles (IRIT) – A. Nazarenko, S. Szulman (LIPN) – J. Charlet (INSERM)
Date de diffusion : mai 2008
Date de MAJ : 19 mai 2009

Contrat ANR, programme Technologies Logiciel : 06 TLOG 010
Coordinateur du projet : J. Charlet (INSERM - UMR_S 872, Equipe 20)

| DaFOE4App | Dossier A.1 | Lot SPA_1 |
|---|---|---|

**Tâche concernée**

| | |
|---|---|
| **Titre :** | Cahier des charges scientifique et technique de la plateforme DaFOE |
| **Calendrier :** | |
| **Responsable :**<br>**Partenaires participants :** | LIPN<br>ENST, INSERM, IRIT, LIPN, LISI, MONDECA, SUPELEC, UTC |

**Document**

| | |
|---|---|
| **Titre du document :** | A.1.1 - Cahier des charges scientifique et technique de la plateforme DaFOE– Chapitre Modèle de données |
| **Coordination de la rédaction :** | Sylvie Szulman |
| **Auteurs :** | Sylvie Szulman |

**Table des révisions**

| Révision | Date | Auteur(s) | Description |
|---|---|---|---|
| 0.0 | 15/03/2008 | S. Szulman (LIPN) | Création du document |
| 0.1 | 15/05/2008 | S. Szulman (LIPN) - J. Charlet (INSERM) | Modification du document |
| 0.2 | 11/02/2009 | S. Szulman (LIPN) - J. Charlet (INSERM) | Modification du document |
| 0.9 | 03/04/2009 | S. Szulman (LIPN) - J. Charlet (INSERM) | Modification du document |
| 0.10 | 19/05/2009 | A. Nazarenko (LIPN) | a) Lissage des sections 1, 2, 3 et 4. b) Uniformisation terminologique (on ne parle plus de "candidat terme" ; le terme de "termino-conceptuel" remplace celui de "termino-ontologique"). c) Ajout d'un champ supplémentaire dans chacune des tables des types de relations terminologiques et des types de relation termino-conceptuelles pour établir des correspondance par défaut entre les types de relations des deux couches. |

# Table des matières

**1 Modèle de données de la plateforme DaFOE** **9**

1 Structure générale . . . 9
2 Couche corpus ou couche textuelle . . . 10
3 Couche terminologique . . . 10
3.1 Les termes . . . 10
3.2 Les relations de la couche terminologique . . . 11
3.3 Les clusters . . . 13
4 Couche termino-conceptuelle . . . 13
4.1 Les concepts terminologiques (ou termino-concepts) . . . 13
4.2 Les relations termino-conceptuelles . . . 14
4.3 Le multilinguisme . . . 15
5 Couche ontologique . . . 16
6 Remarques transversales . . . 16
6.1 Descriptif d'un objet . . . 16
6.2 Statistiques . . . 16
6.3 Divers . . . 17

# Table des figures

1.1 Les couches du modèle de données. . . . . 9
1.2 Table correspondant à la couche corpus du MdD. . . . . 10
1.3 Table correspondant à la couche terminologique du MdD. . . . . 10
1.4 Table décrivant une méthode de la couche terminologique du MdD. . . . . 12
1.5 Table décrivant un type de relations au niveau de la couche terminologique du MdD. . . . . 12
1.6 Table décrivant une relation lexicale au niveau de la couche terminologique du MdD. . . . . 13
1.7 Table décrivant une relation de composition syntaxique au niveau de la couche terminologique du MdD. . . . . 13
1.8 Table décrivant un élément de la liste des types de relations termino-conceptuelles de la couche 2 du MdD. . . . . 14
1.9 Table décrivant une relation termino-conceptuelle. . . . . 15

**Résumé**

# Chapitre 1

# Modèle de données de la plateforme DaFOE

Le modèle de données de DaFOE est conçu selon une structure générique :
– constituée d'un noyau ;
– offrant des possibilités d'extension.

## 1 Structure générale

Le modèle de données (MdD) est un modèle en 4 couches (textuelle, terminologique, termino-conceptuelle, formelle). Chaque couche possède un élément indiquant pour chaque objet l'état de validation.

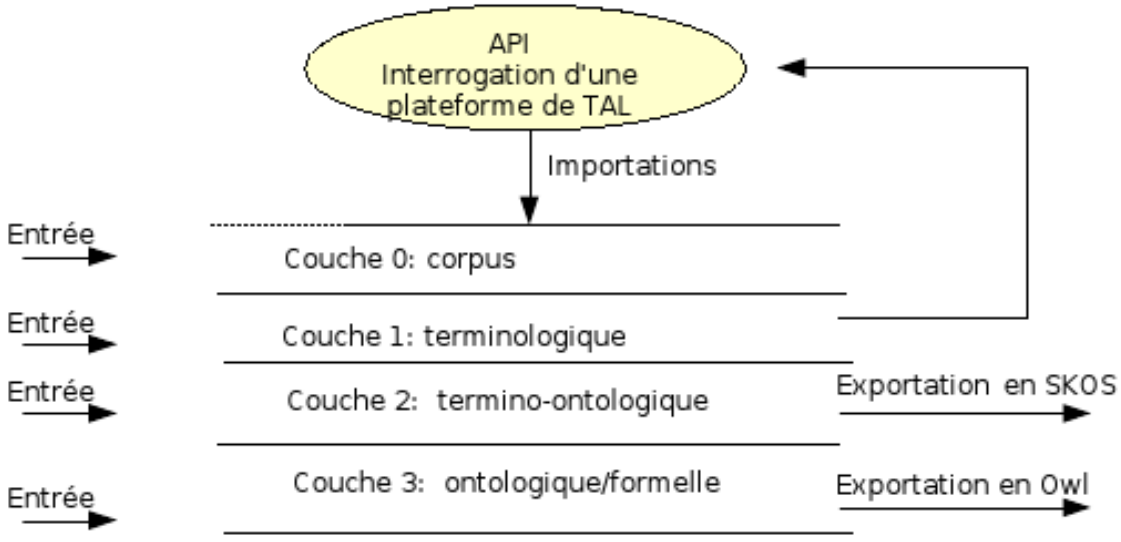



**FIG. 1.1** – Les couches du modèle de données.

Les éléments linguistiques nécessaires sont :
– les termes,

– les relations terminologiques,
– les classes de mots ou de termes (clusters),

ainsi que les propriétés linguistiques qui leur sont associées.

## 2 Couche corpus ou couche textuelle

Un corpus peut contenir plusieurs documents qui sont définis dans un ou plusieurs fichiers. Le texte brut est découpé en phrases qui sont repérées par des identifiants uniques (idPhrase). Si le corpus d'entrée est déjà segmenté, c'est ce découpage qui est repris et un lien est fait entre les identifiants de phrase du MdD (figure 1.2) et ceux du corpus annoté produit.

| idCorpus | idDocument | idPhrase | phrase |
|---|---|---|---|

**FIG. 1.2** – Table correspondant à la couche corpus du MdD.

## 3 Couche terminologique

Le principal « objet » à modéliser à ce niveau est le terme. On veut pouvoir modéliser les candidats termes issus des outils de TAL et les termes qui ont été retenus parmi les candidats termes. Les termes et les candidats termes sont tous représentés comme des termes, c'est le statut qui détermine s'ils sont validés ou non (auquel cas ce ne sont encore que des candidats termes).

### 3.1 Les termes

Un terme est décrit par les éléments suivants (voir table 1.3) :
– son statut qui peut être (neutre, détruit ou éliminé, rejeté, validé, en attente),
– ses propriétés morpho-syntaxiques (comme le genre, le nombre),
– sa forme canonique (éventuellement une variante choisie arbitrairement à défaut d'autres critères),
– ses variantes,
– ses occurrences (identifiants des phrases dans le corpus dans lesquelles le terme apparaît),
– ses critères de saillance enregistrés dans un vecteur,
– sa langue (par défaut, c'est la langue du corpus où se trouve le terme).

| idTerme | statut | liste Prop. morpho-syn | liste variantes | Ens. IdPhrases | liste critères de saillance | lang. |
|---|---|---|---|---|---|---|

**FIG. 1.3** – Table correspondant à la couche terminologique du MdD.

Les listes de propriétés morpho-syntaxiques et les listes de critères sont des listes attribut-valeur. En particulier, il doit y avoir un attribut stat_entNom indiquant si le candidat terme peut être **une entité nommée**. L'attribut stat_entNom doit avoir trois valeurs (oui/non/neutre ou EN/T/indifférencié). Un terme est par défaut un "terme".

Les listes de variantes sont des listes de chaînes de caractères.

Par défaut, les critères de saillance sont les suivants : fréquence, répartition (par ex. le poids tf.idf ou le facteur idf dans le tf.idf), nombre de variantes, productivité en tête et productivité en expansion,

poids typographique, poids de position textuelle, autre poids (pour avoir un critère prévu supplémentaire). Calculer des poids typographique ou de position permet de faire du tri facilement et d'affecter un poids de saillance global sur le terme en agglomérant la saillance de ses différentes occurrences. Cela suppose qu'on ait précalculé des poids en fonction des caractéristiques des occurrences du terme dans le corpus.

## 3.2 Les relations de la couche terminologique

### 3.2.1 Introduction

Remarques terminologiques :

- Une relation lexicale est une relation (de sens, sémantique) entre termes. Ce sont ces relations lexicales que l'on désigne couramment comme des "relations terminologiques", même si les termes peuvent être reliés par d'autres types de relations.
- Une relation de composition syntaxique est également une relation entre deux termes mais considérée du point de vue syntaxique, l'un étant un composant de l'autre : « sac » est la tête de « sac de couchage » et « couchage » est le modifieur. A noter qu'une relation entre deux temes peut être à la fois lexicale et de composition syntaxique
- Une relation syntaxique est une relation qu'un terme entretient avec un autre mot ou terme dans la phrase.
- Une relation conceptuelle est une relation entre concepts.
- Un patron (motif) est une expression régulière contenant des mots, des catégories grammaticales, des caractères spéciaux. Il peut y avoir des contraintes syntaxiques, typographiques, de distance, etc.
- Une relation sémantique est une relation unissant les termes entre eux à l'intérieur d'un thésaurus. On distingue en général les relations d'équivalence, de hiérarchie et d'association (Chaumier 1988). En linguistique, on parle aussi de relation sémantique entre mots d'une phrase, en faisant ou non intervenir les rôles associés au verbe (relations prédicatives). Dans ce document, le terme « relation sémantique » est entendu au premier sens. À noter que ces relations sémantiques sont un sous-type des relations lexicales décrites *supra*.

Comment trouver les relations lexicales (ou relations terminologiques) ?

- à la main
- par des règles d'extraction à partir de textes / patrons, éventuellement en s'appuyant sur des relations entre concepts d'une ontologie de référence
- en exploitant les relations de composition syntaxique
- à partir d'indices statistiques (termes co-occurrents, termes d'un même cluster, etc.)

Parmi les relations lexicales, les relations suivantes sont privilégiées pour élaborer des relations conceptuelles :

- relations de synonymie
- relations d'hyperonymie
- relation partie de
- relation d'antonymie
- relations propres au domaine

Pour chacune de ces relations (synonymie, hyperonymie, partie de, antonymie), il existe une base de patrons génériques, pouvant être adaptés au corpus pour commencer une modélisation. Pour les relations propres au domaine, seules l'expérience du modélisateur et l'application peuvent permettre de les élaborer.

Certaines techniques statistiques peuvent éventuellement aider à les repérer.

Le MdD de la partie relation de la couche terminologique est composé de 4 éléments :

– Un ensemble de méthodes permettant de trouver les relations,
– Un ensemble de types de relations terminologiques,
– Un ensemble de relations lexicales ou terminologiques,
– Un ensemble de relations syntaxiques.

### 3.2.2 Méthodes permettant de trouver des relations

Une méthode peut être structurelle (interne au terme) ou contextuelle (dépendante du contexte). Le contexte peut être différent selon qu'on utilise un analyseur syntaxique, un concordancier ou un outil de clustering etc.

Une méthode est décrite pour chaque « projet » ou « corpus » donné (cf. tableau 1.4). La syntaxe utilisée pour chaque outil d'extraction relationnelle est dépendante de l'outil. Dans le fichier de configuration, on indiquera l'outil utilisé pour appliquer la méthode. La Méthode dans la table est une chaîne de caractères renseignée par l'outil d'extraction de relations si l'utilisateur de la plateforme en possède un ou à la main sinon.

| idMethode | Méthode |
|---|---|

**FIG. 1.4** – Table décrivant une méthode de la couche terminologique du MdD.

Une méthode est associée à un ou plusieurs types de relation.

### 3.2.3 Types de relations lexicales (ou terminologiques)

A un type de relation terminologique sont associés :

– un identifiant,
– un nom,
– l'ensemble des méthodes permettant de le retrouver,
– un identifiant de type de relation termino-conceptuelle qui lui associé par défaut (ce champ permet de faire le lien entre les types de relations terminologiques et les types de relations termino-conceptuelles (couche termino-conceptuelle) de manière à faciliter la construction de la couche termino-conceptuelle à partir de la couche terminologiqiue.

Un type de relation est décrit par le tableau 1.5 où :

| idTypeRel | nomTypeRel | Ens.IdMethode | idTypeRelTC |
|---|---|---|---|

**FIG. 1.5** – Table décrivant un type de relations au niveau de la couche terminologique du MdD.

### 3.2.4 Relations lexicales (ou terminologiques)

Une relation est décrite par le tableau 1.6 où

– idTerme1 et idTerme2 sont les identifiants des termes entre lesquels la relation identifiée par IdRel, de type IdTypeRel,

| idRel | idTerme1 | idTerme2 | idTypeRel | origineRel | statut |
|---|---|---|---|---|---|

**FIG. 1.6** – Table décrivant une relation lexicale au niveau de la couche terminologique du MdD.

| idTerme/mot_recteur | IdRel | IdTerme/mot_régi | idTerme/recteur_regi | statut |
|---|---|---|---|---|

**FIG. 1.7** – Table décrivant une relation de composition syntaxique au niveau de la couche terminologique du MdD.

- origineRel est une liste de couples (IDPhrase, IDMethode),
- statut prend la valeur validé ou invalidé.

La référence à une terminologie est considérée comme une méthode comme une autre. Cela permet d'avoir la même relation trouvée dans une terminologie et de manière contextuelle. L'IDPhrase n'est renseigné que pour les méthodes contextuelles.

### 3.2.5 Relations syntaxiques

Une relation syntaxique est soit interne au terme (tête ou modifieur) ou externe au terme en relation avec des mots (termes) voisins liés par des relations syntaxiques.

Une relation syntaxique est décrite par un triplet (cf. tableau 1.7) où

- IdTerme/mot est un terme ou un mot du corpus,
- idTerme/recteur_regi est le terme composé ou VIDE si le terme composé n'existe pas,
- statut prend la valeur validé ou invalidé.

Cette partie du modèle devrait permettre la mise en place de l'analyse distributionnelle à condition d'avoir fait passer le corpus dans un analyseur syntaxique. Les tables décrites dans les annexes permettent de visualiser des résultats équivalents à Upery, un outil d'analyse distributionnelle créé par D. Bourigault qui fonctionne sur l'outil d'analyse syntaxique Syntex.

## 3.3 Les clusters

Un cluster est un ensemble de mots ou de termes regroupés par proximité sémantique. Un cluster peut être considéré comme une hypothèse de concept. Un plugin doit être proposé par l'équipe Supelec pour aider la construction de l'ontologie en utilisant des clusters.

# 4 Couche termino-conceptuelle

Le niveau termino-conceptuelle permet de représenter les termes désambiguïsés, ou termino-concepts, et les relations désambiguïsées.

Comme au niveau terminologique, ces ressources (concepts terminologiques ou relations termino-conceptuelles) peuvent avoir différents statuts de validation.

## 4.1 Les concepts terminologiques (ou termino-concepts)

On représente un sens de terme par un concept terminologique. Les différents synonymes sont représentés par le même concept terminologique... En cas d'ambiguïté au niveau terminologique, tout ou partie

des occurrences du niveau terminologique sont réparties entre les différents concepts terminologiques associés à un même terme. Cette couche contient tous les résultats intermédiaires et finaux résultant de la conceptualisation des termes.

À un concept terminologique (ou termino-concept, noté TC), on associe :

- le pointeur sur le terme vedette (identifiant d'un des termes associés),
- la langue,
- les termes associés (synonymes du terme vedette),
- les pointeurs sur des occurrences (liste d'idPhrases),
- du texte libre contenant la définition,
- des critères de différenciation en langage naturel : 4 champs texte précisant la différence, la similitude avec le concept père, la différence et la similitude avec les concepts frères. Ces champs pourraient pointer sur des relations entre concepts terminologiques. Un langage semi-informel pourrait être proposé. Un plugin pourrait traiter ces champs textes et remplacer les définitions en langage naturel par des définitions en utilisant un langage contrôlé,
- la ou les méthodes qui permettent de retrouver les termes associés en corpus,
- un concept formel lorsqu'il a été défini,
- un statut qui peut être validé ou invalidé.

## 4.2 Les relations termino-conceptuelles

Comme pour les termes, il existe un correspondant au niveau termino-conceptuel des relations terminologiques. Une relation termino-conceptuelle est définie entre n TC ou entre un TC et une valeur.

### 4.2.1 Types de relations termino-conceptuelles

Une liste de types de relations termino-conceptuelles est prédéfinie qui pourra être complétée par l'utilisateur de la plateforme. Cette liste est décrite dans le modèle de données par une table 1.8 où

- idTypeRelTC est l'identifiant du type de relation termino-conceptuelle considéré,
- nom est le nom de ce type,
- nom_pere est le nom du type de relation père (ce qui permet de décrire une hiérarchie simple),
- idTypeRel est l'identifiant du type de relation terminolgoique qui est associé par défaut au type de relation termino-conceptuelle (ce champ permet de faire le lien entre les couches termino-conceptuelle et terminologique et de dériver la seconde de la première). A noter que ce champ fait le pendant du champ idTypeRelTC dans la table 1.5. Cette double correspondance définie par défaut ne signifie pas qu'il y a bijection entre les deux ensembles de types de relations puisque le type par défaut peut corrigé à tout moment quand on établit la correspondance entre les relations terminologiques et les relations termino-conceptuelles (voir le champ origineRel dans la table 1.9 ci-après).

| idTypeRelTC | nom | nom_pere |
|---|---|---|

**FIG. 1.8** – Table décrivant un élément de la liste des types de relations termino-conceptuelles de la couche 2 du MdD.

### 4.2.2 Relations termino-conceptuelles

Une relation termino-conceptuelle est décrite par (voir figure 1.9) :
- un identifiant
- l'identifiant (idTypeRelTC) du type de la relation qui permet de retrouver le nom du type de la relation termino-conceptuelle dans la table 1.8,
- le n-uplet de termino-concepts qui reliés par la relation (en général et, en tout cas, dans la 1ère version de la plateforme Dafoe, il s'agira de couples de termino-concepts mais il faut se ménager la possibilité de représenter au niveau termino-conceptuel des relations n-aires) ou le couple formé d'un termino-concept et d'une valeur.
- une mention de l'origine de la relation qui assure la traçabilité des relations : origineRel identifie la relation terminologique (idRel) qui a donné naissance à cette relation termino-conceptuelle,
- un statut qui peut prendre la valeur validé ou invalidé.

| idRelTC | idTypeRelTC | liste ordonnée de TC ou (TC, valeur) | origineRel | statut |
|---|---|---|---|---|

**FIG. 1.9** – Table décrivant une relation termino-conceptuelle.

Remarques :
- La liste ordonnée de TC est un couple de TC dans la première version Dafoe mais on devrait pouvoir représenter des relations n-aire à l'avenir.
- (TC, valeur) représente un lien entre un TC et une valeur.
- La table 1.8 décrit toutes les relations entre termino-concepts y compris la relation hiérarchique.

Ces relations permettent :
- d'assurer la traçabilité entre relations conceptuelles et le corpus,
- de donner une sémantique qui permettrait des inférences (transitives, fonctionnelles, ...),
- de limiter le nombre de relations dans l'ontologie,
- d'éviter d'utiliser des relations termino-conceptuelles différentes mais ayant la même sémantique.

Ces relations termino-conceptuelles permettent de modéliser au niveau termino-conceptuel les relations terminologiques validées dans la couche 1 mais elles permettent aussi d'encoder les relations en provenance d'un thesaurus. En effet, les relations du thesaurus peuvent être encodées de la manière suivante :
- Broader_Than peut être défini comme un type de relation et donné comme le père du type de relation Narrower_Than (voir table 1.8).
- Related_Term/Related_Term peut être défini comme un type de relation termino-conceptuelle et déclaré comme le "père" de toutes sortes de types de relations, la sémantique de la relation Related_Term/Related_Term étant sous-spécifiée.
- Compounds (ou partie/tout) peut être représenté explicitement comme un type de relation.
- Les relations de domaine doivent être représentées comme autant de types particuliers de relations termino-conceptuelles.

## 4.3 Le multilinguisme

Le support du multilinguisme se fait à ce niveau : on peut choisir que les concepts terminologiques soient définis selon une approche multilingue ou selon une approche monolingue avec des relations de traductions. Les deux approches semblent nécessaires.

- Version 1 : les concepts sont multilingues. On veut pouvoir présenter la même ressource "localisée" dans différentes langues.
  - On associe au concept terminologique plusieurs termes-vedettes, un dans chaque langue et pour chacun, tout ce qui est dépendant de la langue : l'ensemble des informations qui lui sont associées (synonymes, définition en langage naturel, définition ...)
  - On lui associe des informations conceptuelles (relations conceptuelles, différenciation)
- Version 2 : les concepts sont monolingues mais ils peuvent être mis en correspondance par des relations de traduction ou d'équivalence. On veut articuler des ressources différentes, chacune étant liée à une langue et pouvant avoir sa structure propre.
  - Il peut y avoir autant de concepts terminologiques que de langues
  - Les concepts terminologiques peuvent être en relation de traduction (plus ou moins exacte)
  - On leur associe des informations conceptuelles (relations conceptuelles, différenciation)

Dans les deux versions, un concept est repéré par un identifiant. Dans la première version, à un identifiant correspond un concept multilingue. On a donc une seule ontologie qui peut être présentée dans différentes langues. Dans la deuxième version, deux solutions sont envisagées selon qu'on considère 1/ un identifiant unique pour plusieurs concepts jugés équivalents, ou 2/ un identifiant par concept terminologique, des concepts terminologiques associés à différentes langues pouvant alors être reliés explicitement par une relation d'équivalence.

La première version de la plateforme implémente la version 1 ci-dessus.

# 5 Couche ontologique

Le langage de modélisation est au moins équivalent à Owl-DL. Il contient des concepts et des propriétés, des instances de concepts et des instances de propriétés et des valeurs (DataTypes). Dans une première version, les types les plus courants (string, int, float, boolean, date (version simple)) seront définis. Un concept est défini ou primitif. Un concept formel ou l'un de ses attributs contient un lien sur le termino-concept s'il existe.

# 6 Remarques transversales

## 6.1 Descriptif d'un objet

Chaque objet (y compris les instances) a :

1. un numéro de version
2. un id universel généré automatiquement ( peut être une URI)... Cependant chaque utilisateur pourra définir sa règle concernant les identifiants d'objet. Cette règle sera conservée dans un fichier de configuration.
3. une date
4. un auteur

## 6.2 Statistiques

Un module statistiques sera constitué de manière transversale à toutes les couches.

### 6.3 Divers

Prévoir un fichier de logs pour les actions effectuées dans la plateforme qui pour le moment est mono-utilisateur.Le fichier de log doit contenir des informations utiles au greffon de mainteance d'une ontologie.

**Lien corpus ontologie** La traçabilité entre le corpus et les objets de l'ontologie (classes, propriétés, instances) est mis en oeuvre par l'intermédiaire les concepts termino-ontologiques et les relations termino-ontologiques.