Laboratoire d'Informatique

pour la Mécanique

et les Sciences de l'Ingénieur

NOTES et DOCUMENTS LIMSI N^{O} : 2011-04
Septembre 2011

Entités nommées structurées : guide d'annotation Quaero

Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum

limsi

Centre National de la Recherche Scientifique

LIMSI-CNRS - BP 133, F-91403 ORSAY Cedex (France)

Entités nommées structurées : guide d'annotation Quaero

Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum

Quaero, T3.2, presse écrite et orale
Version : 1.25
Date : 2011/09/08

Résumé

Ce document présente les principes et spécifications des entités nommées structurées définies dans le projet Quaero. Dans le projet, ces spécifications ont servi à annoter des corpus de presse audio et de presse ancienne, pour un total de trois millions de mots.

Le travail de définition et de structuration qui a abouti à ce guide, ainsi que les mesures d'accord inter-annotateurs, sont décrits dans Grouin et al. [2011b]. Une réflexion concernant les métriques pour l'évaluation de systèmes de détection d'entités nommées structurées, en lien avec des tâches finalisées, est proposée dans Grouin et al. [2011a]. Les problématiques liées à l'évaluation sur des transcriptions automatiques de parole sont abordées dans Galibert et al. [2011]. Enfin, un récapitulatif de l'ensemble du travail est présenté dans Rosset et al. [2011].

Mots-clés : Entités nommées, Types d'entités, Composants d'entités, entités nommées structurées, Projet Quaero, Annotation de corpus.

Abstract

This document presents the principles and specifications of the structured named entities defined in the Quaero project. In the project, these specifications were used to annotate broadcast news and old press corpora, for a total of three million words. % The definition and structuring work which led to these guidelines, along with inter-annotator agreement measures, were described in Grouin et al. [2011b]. A discussion about metrics for evaluating structured named entity detection systems, in relation to finalized tasks, is proposed in Grouin et al. [2011a]. The issues linked to the evaluation of automatic speech transcripts are addressed in Galibert et al. [2011]. Finally, a general overview of the overall work is presented in Rosset et al. [2011].

Keywords : Named Entities, Entity Types, Entity Components, Structured Named Entities, Quaero Project, Corpus Annotation.

Table des matières

Chapitre 1

Description générale

1.1 Définitions

1.1.1 Entité nommée étendue

- Rappel : les entités nommées incluent traditionnellement trois grandes classes : les noms, les quantités, les dates et durées.
- Nous nous plaçons dans le contexte d'extraction d'informations (entités, relations) servant à constituer une base de connaissances. Les entités nommées étendues forment le pivot de la base de connaissances.
- Dans une optique de transition, nous partons des entités habituelles en extraction d'entités nommées (à partir de la presse) et étendons leur inventaire de deux façons :

  1. de *nouveaux types* d'entités (par exemple, civilisations, fonctions) ;
  2. l'inclusion d'expressions ne contenant pas de nom propre. Nous étendons donc la définition habituelle des entités nommées, centrée sur les noms propres, à des *expressions construites autour de noms communs*.

- Les types retenus correspondent à ce qui paraît intéressant pour une base de connaissances dans les domaines concernés par les textes (typiquement presse quotidienne). Les entités annotées ont vocation à être ultérieurement reliées par des relations, qui seront elles aussi à reconnaître dans les textes.
- Lorsque l'annotation porte sur des données transcrites de l'oral, on peut être amené à annoter des segments contenant des hésitations, des reprises etc. Il convient dans ce cas de les intégrer à l'entité à laquelle elles appartiennent (mon vol est le numéro <prod.serv> *AF euh euh euh 347* </prod.serv>).

1.1.2 Sources

Ce travail est en cohérence avec la définition des entités nommées d'ESTER2 [Ester2] et les discussions du groupe de travail qui a suivi cette campagne d'évaluation. En ce qui concerne la définition des entités nommées, nous avons utilisé celle de la thèse de Maud Ehrmann [Ehrmann, 2008]. La définition des entités temporelles repose sur TIMEX3 [Timex] tandis que celle des noms propres est basée sur la taxinomie de la thèse de Mickaël Tran [Tran, 2006]. Après avoir défini tous ces types d'entités nommées, nous nous sommes ap-

puyés sur la hiérarchie des entités nommées étendues proposée par Satoshi Sekine [Sekine, 2004, Nadeau and Sekine, 2007].

1.2 Principes

1.2.1 Types et sous-types

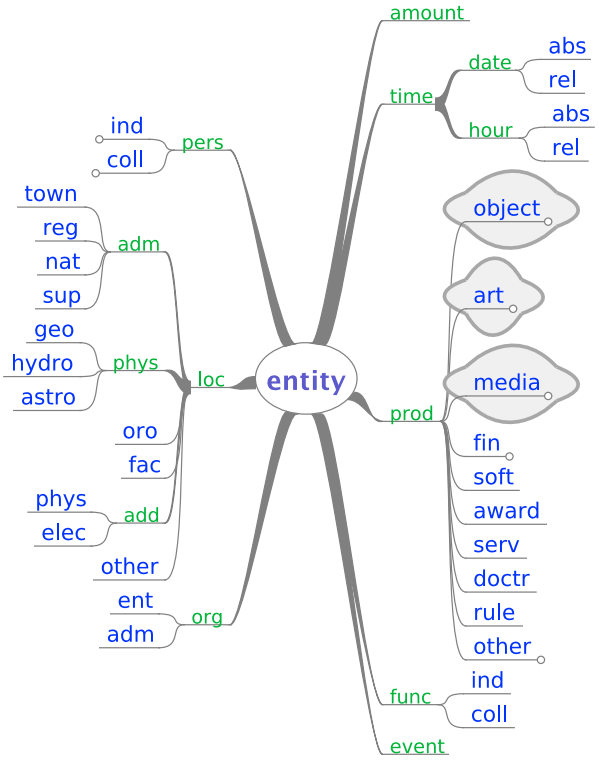


FIG. 1.1: Hiérarchie des entités

Les types d'entités sont organisés en hiérarchie (figure 1.1). Les sous-types sont notés en reprenant le nom du surtype auquel on ajoute un point et un nom identifiant le sous-type. Par exemple, <loc> a un sous-type qui est <loc.adm>.

Définition 1 (Étiquetage au plus spécifique)**.** L'annotation utilise uniquement les feuilles de la hiérarchie des types d'entités.

Définition 2 (Unknown)**.** On prévoit dans chaque classe une possibilité d'annotation .unknown (.unk) : on ne sait pas dire quelle sous-classe convient.

l'Atlantide *loc.adm.town, loc.adm.nat, loc.phys.geo ?*

l'
<loc.unk>
<name> Atlantide </name>
</loc.unk>

Le sous-type .unk sert à noter une entité dont on ne connaît pas le type. Le cas d'une entité ambiguë est traité différemment (par exemple, *Yves Rocher* est aussi bien une personne qu'une entreprise), voir la section 1.3.

Note : On pourrait être tenté dans certains cas d'annoter une entité par un type qui n'est pas une feuille de la hiérarchie (par exemple <loc>), ce qui est interdit. Le sous-type .unk (ici, <loc.unk>) apporte alors une solution.

Un système peut annoter <loc.unk> s'il ne sait pas dire plus précisément le sous-type. Dans l'évaluation, si un système annote avec un sous-type inconnu (par exemple, <loc.unk>), il aura cependant moins de points que s'il annotait au niveau spécifique attendu. Les annotateurs doivent éviter au maximum d'utiliser les sous-types <*.unk>. Toutes les annotations .unk mises par les annotateurs seront revues et discutées avant la livraison du corpus annoté de référence.

Définition 3 (Oth)**.** Une sous-classe .other est définie dans certaines classes de la hiérarchie des types d'entités. Elle est employée lorsqu'aucune des autres sous-classes indiquées ne convient. .other

le
<prod.oth>
<name> Monopoly </name>
</prod.oth>

Les sous-classes <*.other> correspondent à des manques dans la hiérarchie des types d'entités. Ces manquent proviennent du choix fait par les concepteurs de cette hiérarchie de limiter la prolifération des types d'entités dans des zones où un grand nombre de types peu fréquents aurait dû être défini.

Le sous-type .other doit être utilisé là où c'est pertinent par les annotateurs dans l'annotation de référence. Ce sous-type est alors celui qui doit être trouvé par les systèmes.

Définition 4 (Amb)**.** Une entité nommée peut se voir attribuer 2 ou plus types dans les cas d'ambiguité. Ce cas de figure est traité à la section1.3

1.2.2 Composants

Une entité est formée d'un ou plusieurs composants, ainsi éventuellement que des parties non annotées.

Tous les mots à l'intérieur d'une entité nommée doivent être étiquetés avec un composant, à l'exception de :

- certains articles et mots de liaison ;
- <name> est optionnel s'il est le seul composant à l'intérieur d'une entité ; tous les autres composants, même s'ils sont seuls dans l'entité, doivent être annotés ;

  le pompier
  le <func.ind> <kind> pompier </kind> </func.ind>

- il n'est pas nécessaire de décomposer un <event> en composants. En revanche, les types et sous-types à l'intérieur d'un <event> doivent être annotés.

Sur les parties non annotées, voir la section 1.2.2.6.

1.2.2.1 Typage des composants

Les composants sont typés (figure 1.2).

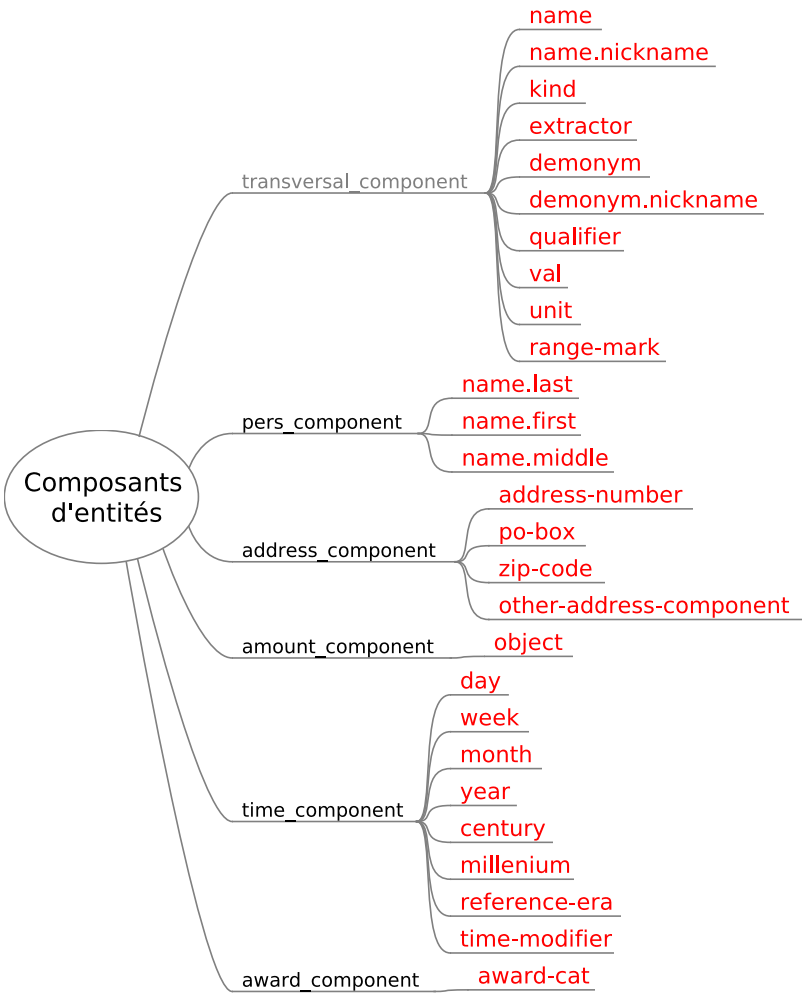


FIG. 1.2: Composants. Les noms de groupes de composants (ex : address_component) ne sont pas utlisés dans l'annotation, ils sont seulement présents pour structurer la description des composants.

Définition 5 (Imbrication d'entités)**.** Une entité peut servir de composant dans une autre entité [1]. La hiérarchie des types de composants inclut donc la hiérarchie des entités (bas de la figure 1.2).

1.2.2.2 Exemples de découpage d'entités en composants

Par exemple, un nom de rue (<loc.oro>) peut se composer d'un genre (<kind>) et d'un nom (<name>).

1. A priori, nous excluons les imbrications récursives : une entité ne peut pas être composant d'une entité du même sous-type. Par exemple, un <loc.adm.town> peut être un composant d'un <loc.add> ; en revanche, un <loc.add> ne peut pas être un composant d'un <loc.add>.

rue de Vaugirard

<loc.oro>

<kind> rue </kind>

de

<name> Vaugirard </name>

</loc.oro>

La figure 1.3 montre des cas typiques de découpage d'entités en composants.

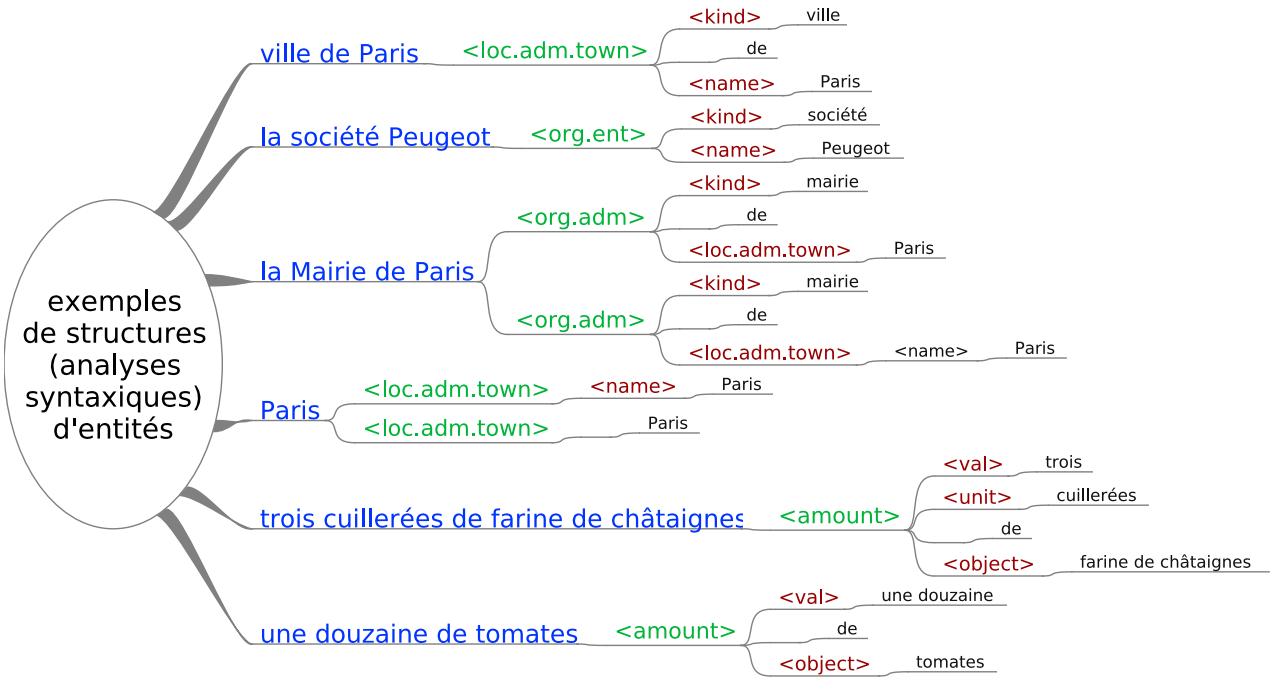


FIG. 1.3: Exemples d'entités.

1.2.2.3 Composants transversaux

Il existe de nombreux composants spécifiques à des classes particulières. Plusieurs composants se retrouvent néanmoins dans plusieurs classes :

name

<name> est le nom de l'entité :

Paris

<loc.adm.town>

<name> Paris </name>

</loc.adm.town>

kind

<kind> Partie d'une expression d'entité nommée qui désigne un hyperonyme (genre proche) de l'entité :

le maire de Paris

le <func.ind>

<kind> maire </kind>
de
<loc.adm.town> Paris </loc.adm.town>
</func.ind>

De ce fait, <kind> est une autre façon de noter un sous-type, que l'on n'a pas inclus « en dur » dans la hiérarchie des types d'entités.

<qualifier> Différents adjectifs qualificatifs peuvent apparaître dans une entité. Ils sont alors étiquetés comme des <qualifier>.

le socialiste Bertrand Delanoë

Le <pers.ind>
<qualifier> socialiste </qualifier>
<name.first> Bertrand </name.first>
<name.last> Delanoë </name.last>
</pers.ind>

<extractor> permet de spécifier une partie (un individu) d'un ensemble :

le cinquième Beatles

Le <pers.ind>
<extractor>cinquième</extractor>
<name> Beatles </name>
</pers.ind>

un des pompiers

<func.ind>
<extractor> un </extractor>
des
<kind> pompiers </kind>
</func.ind>

<demonym> spécifie l'origine géographique d'une entité sous forme d'un gentilé [2]. Nous distinguons ce type de composant spécifique car il apporte une information importante en rapport avec un lieu.

les mairies françaises

les <org.adm>
<kind> mairies </kind>
<demonym> françaises </demonym>
</org.adm>

le ministre américain

le <func.ind>

2. Un gentilé est le nom donné à un habitant d'un lieu (ville, région, pays, etc). Par exemple, *parisien*, *alsacienne*, *maltais*, *européen*.

<kind> ministre </kind>
<demonym> américain </demonym>
</func.ind>

<demonym.nickname> utilise une forme alternative désignant l'origine géographique d'une entité : demonym.nickname

les mairies franchouillardes
les <org.adm>
<kind> mairies </kind>
<demonym.nickname> franchouillardes </demonym.nickname>
</org.adm>

les froggies
les <pers.coll>
<demonym.nickname> froggies </demonym.nickname>
</pers.coll>

1.2.2.4 Annotation optionnelle du nom

Nous adoptons le principe suivant :

Affirmation 6 (Annotation implicite des noms). Une entité se compose souvent d'un genre (<kind>) et d'un nom (<name>). Si seul le nom est présent, il n'est pas nécessaire de l'indiquer : il est alors implicite.

le mont Ventoux — *kind + name, tout annoter*
le
<loc.phys.geo>
<kind> mont </kind>
<name> Ventoux </name>
</loc.phys.geo>

le Ventoux — *name seul, annotation facultative*
le
<loc.phys.geo>
<name> Ventoux </name>
</loc.phys.geo>

le Ventoux — *name seul, annotation facultative*
le
<loc.phys.geo> Ventoux </loc.phys.geo>

les pompiers de Paris — *kind + name, tout annoter*
les
<func.coll>

<kind> pompiers </kind>
de
<loc.adm.town> Paris </loc.adm.town>
</func.coll>

les pompiers *kind seul, annotation obligatoire*

les
<func.coll>
<kind> pompiers </kind>
</func.coll>

la 118 *name seul, annotation facultative*

la<loc.oro>
<name> 118 </name>
</loc.oro>

l'Etna *name seul, annotation facultative*

l'<loc.phys.geo> Etna </loc.phys.geo>

1.2.2.5 Annotation optionnelle du contenu de certains composants

Affirmation 7 (Annotation optionnelle du contenu d'un composant <name>, <object> ou <unit>). Un composant peut correspondre à une entité existante. Il peut par exemple être reconnu comme un nom de personne, de ville, etc. On a alors imbrication d'étiquettes (voir exemple plus bas). Idéalement, on aimerait que ces types soient annotés, mais les contraintes de temps et de coût nous ont amenés à décider d'abandonner ce niveau de détail. Cette optionalité concerne les composants des entités <loc.oro> (section 2.4.4), <loc.add.*> (section 2.4.6), <prod.*> (section 2.5), <amount> (section 2.6), etc.

Dans l'exemple ci-dessous, *Atlantique* peut donc être balisé comme un <name> qui est un <loc.phys.hydro>, ou simplement comme un <name>.

le TGV Atlantique *annotation riche*

le
<prod.serv>
<prod.object>TGV</prod.object>
<name>
<loc.phys.hydro>Atlantique</loc.phys.hydro>
</name>
</prod.serv>

le TGV Atlantique *annotation simplifiée*

le
<prod.serv>

```
        <prod.object>TGV</prod.object>
        <name>Atlantique</name>
    </prod.serv>
```

place de l'Abbé Georges Hénocque *annotation riche*

```
    <loc.oro>
        <kind> place </kind>
        de l'
        <name><pers.ind>
            <title>Abbé</title>
            <name.first> Georges </name.first>
            <name.last> Hénocque </name.last>
        </pers.ind></name>
    </loc.oro>
```

place de l'Abbé Georges Hénocque *annotation simplifiée*

```
    <loc.oro>
        <kind> place </kind>
        de l'
        <name>Abbé Georges Hénocque</name>
    </loc.oro>
```

trois régiments

```
    <amount>
        <val> trois </val>
        <unit>
            <pers.coll> régiments </pers.coll>
        </unit>
    </amount>
```

trois régiments

```
    <amount>
        <val> trois </val>
        <unit> régiments </unit>
    </amount>
```

1.2.2.6 Densité d'annotation dans une entité

Le contenu d'une entité doit être entièrement annoté en composants, sauf éventuellement les prépositions, comme dans *maire de Paris* :

<func.ind>
<kind> maire </kind>
de
<loc.adm.town> Paris </loc.adm.town>
</func.ind>

Attention, dans *De Gaulle*, *De* fait partie du nom de famille :

<pers.ind>
<name.last> De Gaulle </name.last>
</pers.ind>

1.2.2.7 Cas particulier du composant title

Un composant en général ne contient pas de type/sous-type sauf bien entendu <name> et <object> dont le type recouvre le composant. Le composant <title> est spécial et peut inclure un type <func.*> :

Sa Majesté le roi Mohamed VI

<pers.ind>
<title>
Sa Majesté le
<func.ind>roi</func.ind> annotation imbriquée dans un composant : optionnelle
</title>
<name.first>Mohamed</name.first>
<qualifier>VI</qualifier>
</pers.ind>

Une autre particularité de ce composant concerne la coordination. Nous regroupons sous un <pers.coll> les coordinations de personne qui se font au niveau du <title>. Il y a ainsi une différence forte avec ce qui est décrit dans la section sur les coordinations (cf. section 1.2.6.4).

Monsieur et Madame Chirac

<pers.coll>
<title>Monsieur</title>
et
<title>Madame</title>
<name.last> Chirac </name.last>
</pers.coll>

1.2.3 Étendue d'une entité nommée

Une entité nommée doit-elle inclure tout le syntagme nominal dont elle est la tête ? Non, cela donnerait des expressions trop longues et trop complexes.

Affirmation 8. Sont exclues des entités nommées :
- les propositions relatives ;
- les propositions subordonnées ;
- les incises : si une incise coupe une entité en deux, on annote chaque partie de l'entité séparément (les relations seront traitées ultérieurement).

1.2.4 Empan des expressions temporelles et des montants

Nous retenons les principes de limite d'expressions temporelles et des montants suivants :

- Borne gauche : on prend tout ce qui précède et qui permet de former un groupe nominal ou un groupe prépositionnel, y compris, pour les dates, les déterminants [3] (contrairement donc à toutes les autres entités, dont les déterminants sont exclus). Les prépositions font donc partie intégrante de la date ou de la durée, ainsi que les adjectifs préposés. Exemples (pour simplifier, l'annotation des composants n'est pas indiquée) :

  <time.date.rel> ce beau lundi </time.date.rel>

  <time.date.rel> après plusieurs mois </time.date.rel>

  <time.date.rel> au bout de de d' une demi-journée </time.date.rel>

  <time.date.abs> le 23 janvier </time.date.abs>

  <time.hour.abs> à 10 heures </time.hour.abs>

  <amount> pendant 2 heures </amount>

  la préparation <amount> en moins d' un mois </amount>

  les <amount> 2 heures </amount> que nous venons de passer

- Borne droite : entrent dans la partie droite,
  - les compléments du nom formés à partir d'une expression **temporelle** (ce qui exclut les événements). Ainsi, on aurait *<time.date.rel> le dimanche de la semaine dernière </time.date.rel>* mais *<time.date.rel> au troisième jour </time.date.rel> de sa visite*.
  - les adjectifs :

    <amount> une semaine entière </amount> une est une valeur <val>

    <time.date.rel> le trimestre prochain </time.date.rel>
  - les adverbes et locutions adverbiales

    <time.date.rel> cinq ans plus tard </time.date.rel>

    <amount> à un jour près </amount>

3. Car les déterminants participent en général à la spécification de la date. Voir aussi la section 2.7.1.1.

1.2.5 Imbrication d'entités

Les seules imbrications obligatoires d'entités sont :
- la métonymie (*la France demande l'arrêt des hostilités*)
- les cas comme *maire de Paris* (loc.adm.town dans func.ind) ou *ministre des affaires étrangères* (org.adm dans func.ind), etc., ou *dans la Grèce Antique* (loc dans time)

L'imbrication d'entité dans <object> est optionnelle

15 pompiers

<amount> <val> 15 </val> <object> pompiers </object> </amount>

=> il n'est pas utile ici d'annoter "pompiers" en <func.coll>

Cas particulier : Normalement les imbrications récursives sont proscrites. On fera toutefois une exception pour les organes internes d'une entreprise, administration, etc. :

haut commissariat des nations unies *on imbrique une <org.adm> dans une <org.adm> :*

<org.adm>
<kind> haut commissariat </kind>
des
<org.adm> nations unies </org.adm>
</org.adm>

1.2.6 Constructions particulières

1.2.6.1 Durée

Les durées sont représentées par des quantités <amount>. Voir les sections 2.6.3 et 2.6.4.

1.2.6.2 Dislocation/ellipse

Affirmation 9. Lorsqu'on fait face à une dislocation, chaque partie est annotée séparément.

la distance en kilomètres c'est trois

la distance en
<amount>
<unit> kilomètres </unit>
</amount>
c'est
<amount>
<val> trois </val>
</amount>

1.2.6.3 Les disfluences

Affirmation 10. Les phénomènes typiques de l'oral (hésitations, mots coupés etc.) doivent être intégrés dans l'entité s'ils coupent strictement l'entité et n'apportent pas de correction à l'élément.

Hésitation :

j'ai une Peugeot euh 107

j'ai une
<prod.object>
<org.ent> Peugeot </org.ent>
euh
<name> 107 </name>
</prod.object>

Affirmation 11. S'il s'agit d'une reformulation, d'une correction, il y a autant d'entités que de reformulations.

Correction :

j'ai une Renault euh Peugeot 107

j'ai une
<prod.object>
<org.ent> Renault </org.ent>
</prod.object>
euh
<prod.object>
<org.ent> Peugeot </org.ent>
<name> 107 </name>
</prod.object>

Hésitation et correction :

tout au long euh euh de la semaine euh non de l'année

<time.date.rel>
<time-modifier> tout au long </time-modifier>
euh euh de la
<kind> semaine </kind>
</time.date.rel>
euh non
<time.date.rel>
de l'
<kind> année </kind>
</time.date.rel>

Répétition :

Jean, Jean-Marie Le Pen

<pers.ind>
<name_first> Jean </name_first>
</pers.ind>
,
<pers.ind>
<name_first> Jean-Marie </name_first>
<name_last> Le Pen </name_last>
</pers.ind>

1.2.6.4 Coordinations

Affirmation 12 (Coordination). Lorsque plusieurs entités se suivent avec mise en facteur d'une partie commune, on annote séparément chaque entité, bien qu'elle soit incomplète. La coordination permet de décider du type commun des entités coordonnées ; on surannote donc chaque entité avec ce type, bien que son expression ne contienne pas nécessairement de genre (<kind>) correspondant.

vallées de la Lorraine , de l' Alsace , de la Bourgogne , de la Champagne

<loc.phys.geo>
<kind> vallées </kind>
de la
<name> Lorraine </name>
</loc.phys.geo>
,
<loc.phys.geo>de l' <name>Alsace</name> </loc.phys.geo>
,
<loc.phys.geo>de la <name> Bourgogne</name> </loc.phys.geo>
,
<loc.phys.geo>de la <name> Champagne</name></loc.phys.geo>

ministres de l' industrie , de l' énergie , des transports ferroviaires

<func.coll>
<kind> ministres </kind>
de l'
<org.adm><name> industrie </name></org.adm>
</func.coll>
,
<func.coll> *la coordination nous dit qu'on a un <func.coll>*
de l'
<org.adm><name> énergie</name></org.adm>
</func.coll>
,
<func.coll>

des
<org.adm><name> transports ferroviaires</name></org.adm>
</func.coll>

Affirmation 13 (Coordination de quantités : valeurs et range-mark). Pour les quantités (<amount>), si la coordination porte sur la valeur (<val>), on considère qu'il y a un intervalle ou un ensemble de valeurs, ce qui se note avec <range-mark> (voir la section 2.6, *intervalle de valeurs*), et l'ensemble coordonné est intégré dans une seule entité de type <amount>.

Affirmation 14 (Coordination de quantités : unités ou objets). Pour les quantités (<amount>), si la coordination porte sur l'unité (<unit>) ou l'objet (<object>), on considère qu'on a plusieurs entités distinctes.

des centaines de téléphones , d' ordinateurs , d' immenses rouleaux de câble

<amount>
<val> des centaines </val>
de
<object> téléphones </object>
</amount>
,
<amount>
d'
<object> ordinateurs </object>
</amount>
,
<amount>
d'
<object> immenses rouleaux de câble </object>
</amount>

Le cas suivant comporte une coordination entre deux entités complètes, sans mise en facteur, et ne pose donc pas de problème particulier :

il y a 3 minutes ou il y a 5 minutes

<time.hour.rel>
<time-modifier> il y a </time-modifier>
<amount>
<val> 3 </val>
<unit> minutes </unit>
</amount>
</time.hour.rel>
ou<time.hour.rel>
<time-modifier> il y a </time-modifier>
<amount>
<val> 5 </val>
<unit> minutes </unit>
</amount>
</time.hour.rel>

1.2.6.5 Élaboration

Affirmation 15 (Élaboration). Lorsqu'on a à la fois un acronyme ou un sigle et une définition, explication, expansion, élaboration de ce nom, on a affaire à deux entités distinctes.

Agipi association d' assurés pour la prévoyance , la dépendance et l' épargne-retraite

<org.ent>
<name> Agipi </name>
</org.ent>
<org.ent>
<name> association d' assurés pour la prévoyance , la dépendance
et l' épargne-retraite</name>
</org.ent>

1.3 Ambiguïtés et autres difficultés

Voir Leech & Wilson, EAGLES, `http://www.ilc.cnr.it/EAGLES/pub/eagles/corpora/annotate.ps.gz`, sur le principe d'annoter les ambiguïtés pour les catégories morphosyntaxiques (mais ils ne proposent pas de notation). Voir ce qu'il y a dans la TEI.

1.3.1 Observation

Certaines occurrences d'entités nommées ont un type qui peut rester ambigu, même en contexte.

Exemple : *Yves Rocher* peut être aussi bien personne qu'entreprise dans :

<?? >Yves Rocher</ ?? > va s'installer à Vannes

1.3.2 Prise de position

La référence pourra intégrer des annotations qui reflètent cette situation. Une même entité pourra donc être annotée par plusieurs types. Pour représenter concrètement cette annotation multiple, on fournit comme valeur d'un attribut *class-set* l'ensemble des types entre lesquels l'entité est ambiguë, séparés par des espaces. Cet attribut est placé dans un élément dont le type est *entity*. Les annotateurs doivent prendre en compte le contexte d'occurrence de l'entité pour désambiguïser. C'est seulement lorsque le contexte ne permet pas de désambiguïser l'entité qu'il faut utiliser une annotation d'ambiguïté [4].

Paris

<entity class-set="loc.adm.town org.adm">
<name> Paris </name>
</entity>

Lorsqu'une ambiguïté possible est levée par le contexte, on ne note pas l'ambiguïté.

4. Cette possibilité a été offerte aux annotateurs mais aucun n'en a fait usage dans le corpus Quaero.

> Cette réponse s'est doublée d'une réplique publicitaire conçue par l'agence Australie. (Le Monde, 28 avril 2010)

Ici, le contexte permet de déterminer que Australie est une organisation (<org.ent>) et non un pays (<loc.adm.nat>).

conçue par l'

<org.ent>

<kind> agence </kind>

<name> Australie </name>

</org.ent> .

Yves Rocher va s'installer à vannes.

<entity class-set="pers.ind org.ent">

<name> Yves Rocher </name>

</entity>.

va s'installer à Vannes

1.3.3 Ambiguïté entre titre et fonction : la frontière

Affirmation 16. Les titres et fonctions sont distingués. Un titre entre dans une entité de type <pers> alors qu'une fonction n'est jamais intégrée dans l'entité de type <pers> mais dans l'entité de type <func>.

Le Dr. Duboc, ancien chef de service à la Pitié-Salpêtrière

Le <pers.ind>

<title> Dr. </title>

<name.last> Duboc </name.last>

</pers.ind>

le général De Gaulle

le

<pers.ind>

<title> Général </title>

<name.last> De Gaulle </name.last>

</pers.ind>

le maire Delanoë

le

<func.ind><kind> maire </kind></func.ind>

<pers.ind><name.last> Delanoë </name.last></pers.ind>

Bertrand Delanoë, le maire de Paris

<pers.ind>

<name.first> Bertrand </name.first>
<name.last> Delanoë </name.last>
</pers.ind>
,
le <func.ind>
<kind> maire </kind>
de
<loc.adm.town> Paris </loc.adm.town>
</func.ind>

Affirmation 17. Un titre isolé (<title>) ne peut en aucun cas constituer à lui seul une entité de type <pers>.

Par exemple, *un général passa* ne doit pas être annoté.

1.3.4 Ambiguïté entre prod.media et org.ent

Il arrive que le nom d'un media (prod.media) soit aussi le nom d'une entreprise (org.ent) : *France Inter*, *le Monde*, etc. Pour faire la distinction entre les deux, le test suivant peut marcher : on fait précéder l'expression à annoter de *le média* ou *la société*.

Exemples :

- [le média] *Le Monde publie un article*
  (et non [la société] *Le Monde publie un article*) => <prod.media>
- *Bienvenue sur* [le média] *France Inter*
  (et non *Bienvenue sur* [la société] *France Inter*) => <prod.media>
- Et donc : [la société] *Libération a licencié trois journalistes*
  (et non [le média] *Libération a licencié trois journalistes*) = <org.ent>
- [la société] *Le Monde est victime d'une ingérence politique*
  (et non [le média] *Le Monde est victime d'une ingérence politique*)

1.4 Métonymie

Dans une métonymie (ou conversion ?), une catégorie d'entité est employée pour désigner une autre catégorie d'entité. On annote la catégorie à laquelle appartient intrinsèquement l'entité mentionnée. On surannote celle-ci avec la catégorie que désigne l'expression dans le contexte.

une déclaration du Quai d'Orsay
une déclaration du
<org.adm>
<loc.oro>
<kind>Quai</kind>
d'
<name>Orsay</name>
</loc.oro>

</org.adm>

la rue de Grenelle a réagi à cette déclaration

la
<org.adm>
<loc.oro>
<kind>rue</kind>
de
<name>Grenelle</name>
</loc.oro>
</org.adm>
à cette déclaration

l'Élysée a déclaré...

l' <org.adm>
<loc.fac>
<name> Élysée </name>
</loc.fac>
</org.adm>
a déclaré...

dans la Grèce Antique *réfère à la période temporelle*

dans la
<time.date.rel>
<loc.adm.nat>Grèce</loc.adm.nat>
<reference-era>Antique</reference-era></time.date.rel>la Grèce Antique a produit réfère à la civilisation grecque, donc
pers.coll
la
<pers.coll>
<time.date.rel>
<loc.adm.nat>Grèce</loc.adm.nat>
<reference-era>Antique</reference-era></time.date.rel>
</pers.coll>
a produit

La métonymie sur les lieux n'est, en revanche, pas annotée à l'intérieur d'un <func.*> :

le maire de Paris

le
<func.ind>
<kind> maire </kind>
de
<loc.adm.town> Paris </loc.adm.town>
</func.ind>

1.5 Noms propres et noms communs

– Du point de vue de leur reconnaissance, les noms propres (les Mayas) ont l'avantage de la majuscule.
– Cependant, du point de vue de leur sens et de leur intérêt en extraction d'information, les seconds (la civilisation maya, la civilisation précolombienne) sont aussi intéressants que les premiers.

Nous incluons donc des entités exprimées par des noms communs (mais pas *tous* les noms communs !).

Exemples :
le socialiste, le lacanien

la civilisation précolombienne

la <pers.coll>
<kind> civilisation </kind>
<name> précolombienne </name>
</pers.coll>

Chapitre 2

Entités

Ce chapitre décrit tour à tour chaque type d'entité et l'illustre par des exemples.

2.1 Personnes : pers

2.1.1 Définitions

On distingue deux sous-types de personnes, les *individus* et les *groupes* :

pers.ind

pers.coll

pers.ind Individu.

pers.coll On décide de n'annoter en <pers.coll> que les cas suivants :

- présence d'un nom propre (*les Beatles*)
- présence d'un gentilé (composant *demonym*) (*les victimes françaises de l'accident*, *les voyageurs chinois*, etc.), civilisations (*la civilisation Maya*).

On n'annote pas comme pers.coll les autres cas :

- le monde ouvrier, *les êtres humains*, *les blessés*, etc.

Pers.ind et pers.coll sont des sous-types, et pas des composants :

pers.ind *is-a* pers
name.first *part-of* pers.ind

Les personnes ou groupes de personnes désignées par leur métier sont annotées comme fonction (func) et non comme personne (pers) : voir la section 2.2.

Les personnages imaginaires (par exemple, *Asterix*) comme les icônes religieuses (*Dieu*) sont des personnes. Si l'on ne sait pas si leur nom est un nom (name.last) ou un prénom (name.first) ou un surnom (name.nickname), on l'annote comme name [1]. Dans certains contextes, il s'agit du titre de la série d'une œuvre, donc d'un prod.art (*Uderzo est le créateur de la BD Astérix*) (voir la section 2.5.2).

2.1.2 Composants

Composants possibles pour pers.ind et pers.coll : name, qualifier (VI), kind (civilisation), name.first, name.last, name.middle (pour les noms anglo-saxons), name.nickname [2],

1. Les personnages de Molière (*Alceste, Tartuffe,* etc.) ont des prénoms (composant <name.first>) et non un nom indéfini (composant <name>).

2. Le composant <name.nickname> ne doit pas être utilisé pour un style de langue familier (par exemple, *pute*, qui est un <kind> du type <func.ind>).

particle, title, demonym, other.

Sa Majesté le roi Mohamed VI

```
<pers.ind>
    <title>
        Sa Majesté le
        <func.ind>roi</func.ind>    annotation imbriquée dans un
        composant : optionnelle
    </title>
    <name.first>Mohamed</name.first>
    <qualifier>VI</qualifier>
</pers.ind>
```

Son Altesse Royale le prince Rainier

```
<pers.ind>
    <title>
        Son Altesse Royale le
        <func.ind>prince</func.ind>
    </title>
    <name.first>Rainier</name.first>
</pers.ind>
```

Sans le reste du titre (title), on a simplement affaire à une fonction (func.ind) :

le roi Mohamed VI

```
le
<func.ind>roi</func.ind>
<pers.ind>
    <name.first>Mohamed</name.first>
    <qualifier>VI</qualifier>
</pers.ind>
```

2.1.3 Exemples

2.1.3.1 pers.ind

Bertrand Delanoë

```
<pers.ind>
    <name.first> Bertrand </name.first>
    <name.last> Delanoë </name.last>
</pers.ind>
```

le socialiste Bertrand Delanoë

```
Le <pers.ind>
```

<qualifier> socialiste </qualifier>
<name.first> Bertrand </name.first>
<name.last> Delanoë </name.last>
</pers.ind>

le cinquième Beatles

le <pers.ind>
<extractor>cinquième</extractor> extractor
<name> Beatles </name> name
</pers.ind>

le français s'est classé quatrième

le <pers.ind>
<demonym> français </demonym> demonym
</pers.ind>
s'est classé quatrième

Asterix et Obelix ont rapporté un sanglier

<pers.ind>
<name> Asterix </name>
</pers.ind>
et
<pers.ind>
<name> Obelix </name>
</pers.ind>
ont rapporté un sanglier

2.1.3.2 pers.coll

les Beatles

les
<pers.coll>
<name> Beatles </name>
</pers.coll>

Monsieur et Madame Chirac

<pers.coll>
<title>Monsieur</title>
et
<title>Madame</title>
<name.last> Chirac </name.last>
</pers.coll>

les Français[3] ensemble de personnes caractérisé par un démonyme

les <pers.coll>
<demonym> Français </demonym>
</pers.coll>

la diaspora argentine s'est réunie à Paris ensemble de personnes caractérisé par un démonyme

la <pers.coll>
<kind> diaspora </kind>
<demonym> argentine </demonym>
</pers.coll>
s'est réunie à
<loc.adm.town>Paris</loc.adm.town>

l'immigration arménienne s'est offusquée de ensemble de personnes caractérisé par un démonyme

l' <pers.coll>
<kind> immigration </kind>
<demonym> arménienne </demonym>
</pers.coll>
s'est offusquée de

Note : *diaspora*, *immigration* peuvent désigner aussi bien un processus que l'ensemble de personnes résultant de ce processus. Dans le premier cas, les expressions correspondantes ne sont pas annotées (*l'immigration arménienne s'est amplifiée*). Dans le second, elles sont annotées en tant que pers.coll.

Note : *3 personnes* n'est pas un pers.coll, car *personnes* n'est ni un nom propre, ni un démonyme. Il est annoté comme une quantité (voir la section 2.6.2 page 52). Même chose pour *1300 enfants*, *21 blessés*.

Note : *les êtres humains* n'est pas annoté du tout, car ce n'est ni une personne (ni nom propre ni démonyme), ni une quantité (*les* est indéfini). *Le monde ouvrier* n'est pas annoté pour la même raison.

Ne pas annoter les groupes de personnes *jeunes, bébés, séniors,* sauf s'ils sont suivis d'un démonyme :

les bébés français

les
<pers.coll>
<kind> bébés </kind>
<demonym> français </demonym>
</pers.coll>

3. Le Belge aime la bière : généralité qui représente l'ensemble de la population, doc pers.coll. Le Belge est arrivé troisième : individu singulier, donc pers.ind.

2.2 Fonctions : func

2.2.1 Définitions

Métier (*un pompier*), fonction (*les chefs d'État*, *Miss Italie*) , rôle social (*des opposants*), etc., d'une personne (func.ind) ou d'un ensemble de personnes (func.coll).

On distingue donc deux sous-types de fonctions, les *individus* et les *groupes* :

func.ind

func.coll

func.ind peut contenir kind, loc, org

func.coll peut contenir kind, loc, org

Une fonction n'est pas la même chose qu'une personne. Une personne peut avoir une fonction (le préfet Érignac), dans ce cas on annote séparément la fonction et la personne. Le lien entre les deux sera établi ultérieurement lorsque l'on s'attaquera aux relations entre entités. Un <func.*> ne contient donc normalement pas de <pers>.

Règle pour distinguer pers et func :

1. Si on a un *nom de personne*, alors *pers* (y compris *le* <*qualifier*>*socialiste*</*qualifier*> *Bertrand Delanoë*) ; sinon
2. Si on a un nom de *fonction*, alors *func* (y compris *les* <*kind*>*pompiers*</*kind*> <*demonym*>*finlandais*</*demonym*>) ; sinon
3. Si on a un *démonyme*, alors *pers* (*les finlandais*, *la diaspora finlandaise*, *les jeunes finlandais*).

Dans une entité de type <func>, on peut avoir un <kind> seul, sans <name> associé.

Portée d'une fonction : Le func inclut toujours l'organisation ou le lieu qui y est rattaché, s'il est présent dans la phrase. Dans ce cas de figure [4], on n'utilise pas le composant <name> ; en revanche, on intègre directement le type <loc.adm.*> :

l'ambassadeur de Turquie en France

l' <func.ind>
<kind> ambassadeur </kind>
de
<loc.adm.nat> Turquie </loc.adm.nat>
en
<loc.adm.nat> France </loc.adm.nat>
</func.ind>

ministre de la chambre des représentants de l'état

<func.ind>
<kind>ministre</kind>
de la
<org.adm>

4. Attention ! Seul le type <func.*> est concerné par cette absence d'utilisation du composant <name> ; dans tous les autres cas, on utilise d'abord le composant <name>, puis de manière optionnelle, une sous-annotation au moyen du type <loc.adm.*> (par exemple en section 2.5.9).

<org.adm>chambre des représentants</org.adm>
de l'
<org.adm>état</org.adm>
</org.adm>
</func.ind>

La spécialisation d'un métier fait partie du terme qu'on annote. Par exemple, on annote entièrement chacune des expressions suivantes :

loueur de salle, *vendeur de chaussettes*, *représentant de l'OLP à Paris*, *un chirurgien spécialiste des yeux*

En ce qui concerne les fonctions « d'expert » ou de « spécialiste » (*spécialiste de la chasse*, *experts en explosifs*), on peut considérer qu'il s'agit de métiers, donc on les annote en <func>. Dans ce cas, comme dit plus haut, l'ensemble de l'expression (en incluant la spécialité) est à annoter.

On inclut également des rôles ou fonctions qui, sans être complètement des métiers, s'en rapprochent :

les SDF, *un chômeur, des détenus*

2.2.2 Exemples

le maire de Paris

le <func.ind>
<kind> maire </kind>
de
<loc.adm.town> Paris </loc.adm.town>
</func.ind>

les maires de France

les <func.coll>
<kind> maires </kind>
de
<loc.adm.nat> France </loc.adm.nat>
</func.coll>

le chercheur CNRS

le <func.ind>
<kind>chercheur</kind>
<org.ent>CNRS</org.ent>
</func.ind>

le préfet Érignac

le
<func.ind><kind> préfet </kind> </func.ind>
<pers.ind><name.last> Érignac </name.last></pers.ind>

le préfet est parti manger

le
<func.ind><kind> préfet </kind></func.ind>
est parti manger

le ministre des affaires étrangères

le <func.ind>
<kind> ministre </kind>
des
<org.adm> affaires étrangères </org.adm>
</func.ind>

le ministre américain des affaires étrangères

le <func.ind>
<kind> ministre </kind>
<demonym> américain </demonym>
des
<org.adm> affaires étrangères </org.adm>
</func.ind>

un journaliste britannique

<func.ind>
<kind> journaliste </kind>
<demonym> britannique </demonym>
</func.ind>

l'ancien maire de Paris

l'<func.ind>
<qualifier> ancien </qualifier>
<kind> maire </kind>
de
<loc.adm.town> Paris </loc.adm.town>
</func.ind>

les pompiers

les
<func.coll>
<kind> pompiers </kind>
</func.coll>

je voudrais être footballeur

je voudrais être
<func.ind>

<kind> footballeur </kind>
</func.ind>

les pompiers de Paris

les
<func.coll>
<kind> pompiers </kind>
de
<loc.adm.town> Paris </loc.adm.town>
</func.coll>

président de la république

<func.ind>
<kind> président </kind>
de la
<org.adm>
<kind> république </kind>
</org.adm>
</func.ind>

président de la République islamique du Pakistan

<func.ind>
<kind> président </kind>
de la
<org.adm>
<kind> République islamique </kind>
du
<loc.adm.nat> Pakistan </loc.adm.nat>
</org.adm>
</func.ind>

l'un des pompiers

l'
<func.ind>
<extractor>un</extractor>
des
<kind> pompiers </kind>
</func.ind>

ex Miss Italie

<func.ind>
<qualifier> ex </qualifier>
<kind> Miss </kind>
<org.adm>
<name> Italie </name>
</org.adm>
</func.ind>

2.3 Organisations : org

On ne définit pas de type organisation <org> général, mais directement des sous-types <org.ent>, <org.adm>. Une organisation peut être exprimée par un genre (<kind>) et des précisions (<name>, <demonym>, <loc>...) :

la police française

la <org.ent> org.ent

<kind> police </kind>
<demonym> française </demonym>

</org.ent>

l'hôpital Lariboisière

l'<org.ent>

<kind>hôpital</kind>
<name> Lariboisière </name>

</org.ent>

En revanche, un nom commun employé seul (la police, la cour de cassation, l'hôpital, l'entreprise) est annoté uniquement si contextuellement il suffit à nommer ou identifier l'organisation :

Peugeot a annoncé un plan de licenciement. L'entreprise a subi des pertes...

... L' <org.ent><kind> entreprise </kind></org.ent> ...

2.3.1 Entreprise : org.ent

2.3.1.1 Définitions

Une organisation <org.ent> commercialise des produits ou rend des services non uniquement administratifs. On inclut aussi bien des organisations privées et publiques, y compris hôpitaux, écoles, universités, partis politiques, syndicats, police, gendarmerie, cour de cassation, etc. On exclut les organisations à caractère principalement administratif (<org.adm>, voir ci-dessous la section 2.3.2). Voir la section 1.3.4 pour la distinction avec prod.media.

2.3.1.2 Exemples

La société Peugeot

la <org.ent>

<kind> société </kind>
<name> Peugeot </name>

</org.ent>

je travaille chez Peugeot

je travaille chez
<org.ent>

<name> Peugeot </name>

</org.ent>

l'UNESCO

l'<org.ent><name>UNESCO</name></org.ent>

l'hôpital de la Pitié Salpêtrière

l'<org.ent>
<kind> hôpital </kind>
de la
<name> Pitié-Salpêtrière </name>
</org.ent>

la Pitié Salpêtrière

la <org.ent>
<name> Pitié-Salpêtrière </name>
</org.ent>

l'hôpital d'instruction des armées du Val-de-Grâce

l'<org.ent>
<kind> hôpital d'instruction des armées </kind>
du
<name> Val-de-Grâce </name>
</org.ent>

le parti socialiste *« parti » fait partie du nom de ce parti (PS)*

l'<org.ent>
<name>parti socialiste</name>
</org.ent>

le parti Europe Écologie *« parti » ne fait pas partie du nom de ce parti (EE)*

le <org.ent>
<kind> parti </kind>
<name>Europe Écologie</name>
</org.ent>

le syndicat FSU

le <org.ent>
<kind> syndicat </kind>
<name>FSU</name>
</org.ent>

le syndicat national de la magistrature

le <org.ent>
<name>syndicat national de la magistrature</name>
</org.ent>

2.3.2 Administration org.adm

2.3.2.1 Définitions

Une organisation <org.adm> joue un rôle principalement administratif. Elle relève souvent d'un découpage administratif et/ou géographique. Cela inclut les mairies, préfectures, ministères, diocèses, etc. org.adm

2.3.2.2 Exemples

la Mairie de Paris

la <org.adm>
<kind> mairie </kind>
de
<loc.adm.town> Paris </loc.adm.town>
</org.adm>

le diocèse de Blois

le <org.adm>
<kind> diocèse </kind>
de
<loc.adm.town> Blois </loc.adm.town>
</org.adm>

le ministère des affaires étrangères

le <org.adm>
<kind> ministère </kind>
des
<name> affaires étrangères </name>
</org.adm>

les forces alliées

les <org.adm>
<kind> forces </kind>
<qualifier> alliées </qualifier>
</org.adm>

2.4 Localisations : loc

2.4.1 Définitions

Localisations, lieux, entités spatiales.

On distingue les localisations administratives <loc.adm.*>, géographiques <log.phys.*>, les voies <loc.oro>, les bâtiments <loc.fac>, les adresses <loc.add.*>. D'autres localisations ne font partie d'aucun de ces types, et sont regroupées dans <loc.other>.

Un <loc> peut contenir la désignation du type de localisation (<kind>) en plus du nom de cette localisation (<name>). Conformément à la règle 6 page 7. Les points cardinaux doivent être annotés avec le composant <qualifier> (au *sud* d'Israël)

2.4.2 Lieux administratifs : loc.adm

Un <loc.adm> désigne une portion de territoire dont le contour est géopolitique. Les sous-types de <loc.adm> correspondent à différentes granularités de portions de territoire. Pour simplifier, ces différents niveaux de granularité sont à gros grain. Dans le même niveau (par exemple <loc.adm.town>), on peut trouver des entités qui sont en réalité de tailles différentes (ville, quartier, lieu-dit) et qui peuvent être imbriquées l'une dans l'autre.

quartier, ville <loc.adm.town> ville, village, hameau, lieu-dit (*Bézenet*), commune ; partie de ville : quartier, arrondissement (de Paris, Lyon, Marseille), etc.

Paris, Orsay, Blois, Rungis, Essertines-en-Donzy, La Bolline, Le Poil, Le Pont Colbert

Paris

<loc.adm.town> Paris </loc.adm.town>

La Bolline

<loc.adm.town> La Bolline </loc.adm.town>

Val de Crüye

<loc.adm.town> Val de Crüye </loc.adm.town>

Maison Blanche [5]

<loc.adm.town> Maison Blanche </loc.adm.town>

le 13e arrondissement

```
<loc.adm.town>
    <name>13e</name>
    <kind>arrondissement</kind>
</loc.adm.town>
```

la ville de Paris

```
la <loc.adm.town>
    <kind> ville </kind>
    de
    <name> Paris </name>
</loc.adm.town>
```

the Big Apple

```
the
    <loc.adm.town>
        <name.nickname> Big Apple </name.nickname>
    </loc.adm.town>
```

la ville rose

```
la
    <loc.adm.town>
```

<name.nickname> ville rose </name.nickname>
</loc.adm.town>

région <loc.adm.reg> découpage interne à un État[6]. Inclut les régions (administratives et traditionnelles), départements, comtés, länder, arrondissements départementaux, cantons suisses, inclut aussi les associations de communes (communautés de communes, communautés urbaines), les cantons français (bien que certains soient des parties de villes), etc. :

la Bretagne, les Alpes-Maritimes, la Côte d'Azur, le littoral breton, le Baden-Württemberg, le Lancashire, l'arrondissement de Palaiseau,le Pays basque espagnol, le Pays basque français, le canton de Schwytz, le Valais, la Courly, la CAPS

le Pays basque espagnol

le <loc.adm.reg> Pays basque espagnol </loc.adm.reg>

la CAPS

la <loc.adm.reg> CAPS </loc.adm.reg>

au sud d'Israël

<loc.adm.reg>
<qualifier> au sud </qualifier>
d'
<name>
<loc.adm.nat> Israël </loc.adm.nat>
</name>
</loc.adm.reg>

national <loc.adm.nat> loc.adm.nat

la France, le Royaume-Uni, les États-Unis, Andorre, Monaco

le Royaume-Uni

le <loc.adm.nat> Royaume-Uni </loc.adm.nat>

supranational <loc.adm.sup> région du monde, continent, etc. : loc.adm.sup

le Moyen Orient, le Pays basque, la Catalogne, le Commonwealth, l'Afrique subsaharienne, le Nord, le Sud[7]

le Pays basque

le <loc.adm.sup> Pays basque </loc.adm.sup>

la région de l' Atlas

la
<loc.adm.sup>
<kind> région </kind>
de l'
<name>Atlas</name>
</loc.adm.sup>

5. Il s'agit d'un quartier de Paris.

6. Monaco est soit un État, soit une ville, mais pas une région. La bande de Gaza est une région d'Israël et doit donc être annotée avec le type <loc.adm.reg>.

7. Dans le sens *les pays du Sud*. Dans d'autres contextes, *le Sud* pourrait désigner d'autres lieux géographiques (*le Sud de la France*).

2.4.3 Lieux physiques : loc.phys

Lieux physiques terrestres : loc.phys.geo « La classe des géonymes [8] est formée de noms donnés aux espaces géographiques naturels, tels que les déserts, les montagnes, les massifs montagneux, les glaciers, les plaines, les gouffres, les plateaux, les vallées, les volcans, les canyons, etc. »

l'Etna

l'<loc.phys.geo> Etna </loc.phys.geo>

le désert de Gobi

le
<loc.phys.geo>
<kind> désert </kind>
de
<name> Gobi </name>
</loc.phys.geo>

Lieux physiques aquatiques : loc.phys.hydro « Les hydronymes [9] renvoient à des noms d'étendue d'eau [10]. Il peut s'agir par exemple de rivières, de fleuves, d'étangs, de marais, de lacs, de mers, d'océans, de courants marins, de canaux, de sources, etc. »

la Seine

la <loc.phys.hydro> Seine </loc.phys.hydro>

le canal Saint-Martin

le
<loc.phys.hydro>
<kind> canal </kind>
<name> Saint-Martin </name>
</loc.phys.hydro>

Lieux physiques astronomiques : loc.phys.astro Les astronymes sont les planètes, les étoiles, les galaxies, etc., et leurs parties.

la Lune

la <loc.phys.astro> Lune </loc.phys.astro>

la mer de la tranquillité [11]

le
<loc.phys.astro>
<kind> mer </kind>
de la
<name> tranquillité </name>
</loc.phys.astro>

8. Définition tirée de la thèse de Mickaël Tran, Université de Tours, 2006, p. 84
9. Définition tirée de la thèse de Mickaël Tran, Université de Tours, 2006, p. 84
10. Nous ajoutons : « ou de cours d'eau ».
11. Faudrait-il considérer qu'il s'agit d'un <loc.phys.hydro> sur la Lune ?

2.4.4 Voies : loc.oro

Les oronymes (loc.oro) sont les rues, places, routes, autoroutes, etc.

Les oronymes ne peuvent pas être rangés sous les lieux physiques ou administratifs. En effet, ils sont souvent artificiels mais quelquefois naturels, et ne sont pas nécessairement « géopolitiques » : cela justifie de les ranger hors des loc.adm et des loc.phys.

Affirmation 18. Un oronyme se compose souvent d'un genre <kind> et d'un nom <name> (*l'autoroute A6*, *la rue de Vaugirard*). Le nom peut être reconnu comme un nom de personne, de ville, etc., auquel cas son contenu est étiqueté par ce type (voir l'exemple de la *ligne de Sceaux* plus bas). Idéalement, on aimerait que ces types soient annotés, mais les contraintes de temps et de coût nous ont amenés à décider d'abandonner ce niveau de détail.

l'autoroute A6

l'<loc.oro> loc.oro

<kind> autoroute </kind>
<name> A6 </name>

</loc.oro>

la nationale 118

la <loc.oro>

<kind> nationale </kind>
<name> 118 </name>

</loc.oro>

rue des Glycines

<loc.oro>

<kind> rue </kind>
des
<name> Glycines </name>

</loc.oro>

place de l'Abbé Georges Hénocque

<loc.oro>

<kind> place </kind>
de l'
<name><pers.ind>

<title>Abbé</title>
<name.first> Georges </name.first>
<name.last> Hénocque </name.last>

</pers.ind></name>

</loc.oro>

rue de Vaugirard(le fait que Vaugirard ait été un hameau peut être pris en compte ou pas selon ce que décident les annotateurs)

<loc.oro>

<kind> rue </kind>
de
<name> Vaugirard </name>
</loc.oro>

l'A6

l'<loc.oro>
<name> A6 </name>
</loc.oro>

la 118

la <loc.oro>
<name> 118 </name>
</loc.oro>

le triangle de Rocquencourt

le <loc.oro>
<kind> triangle </kind>
de
<name>
<loc.adm.town> Rocquencourt </loc.adm.town>
</name>
</loc.oro>

2.4.5 Bâtiments : loc.fac

Les bâtiments nommés (gare, musée...) et leur extension (stade, campus, université, camping...), nommée également.

Un loc.fac est souvent une vue physique d'un org.

la gare de Rungis

<loc.fac>
<kind> gare </kind>
de
<name> Rungis </name>
</loc.fac>

la gare Saint-Germain Grande Ceinture

la <loc.fac>
<kind> gare </kind>
<name> Saint-Germain Grande Ceinture </name>
</loc.fac>

l'ancienne gare de Rungis

l' <loc.fac>

ancienne
<kind> gare </kind>
de
<name> Rungis </name>
</loc.fac>

la gare Saint-Germain Grande Ceinture
la <loc.fac>
<kind> gare </kind>
<name> Saint-Germain Grande Ceinture </name>
</loc.fac>

le palais de l'Élysée
le <loc.fac>
<kind> palais </kind>
de l'
<name> Élysée </name>
</loc.fac>

l'Élysée
l' <loc.fac>
<name> Élysée </name>
</loc.fac>

2.4.6 Adresses : loc.add

Cette entité se décompose en deux sous-sous-types : loc.add.phys et loc.add.elec

2.4.6.1 loc.add.phys

Une adresse, par opposition à un loc.oro, est un « point » (« point » éventuellement un peu large, comme une place) dans l'espace, par exemple un point dans une voie (rue).

9 place de Rungis
<loc.add.phys> loc.add.phys
<address-number> 9 </address-number>
<loc.oro>
<kind> place </kind>
de
<name><loc.adm.town> Rungis </loc.adm.town></name>
</loc.oro>
</loc.add.phys>

J'habite 15 rue de Vaugirard escalier 2

J' habite
<loc.add.phys>
<address-number> 15 </address-number>
<loc.oro>
<kind> rue </kind>
de
<name> Vaugirard </name>
</loc.oro>
<other-address-component> escalier 2 </other-address-component>
</loc.add.phys>

2.4.6.2 loc.add.elec

Ce sous-sous-type désigne les coordonnées électroniques : numéro de téléphone, télécopie, url, adresse de messagerie électronique, fréquence radio, identifiants de réseaux sociaux (Facebook, Twitter) ou d'outils de communication par Internet (Skype), etc.

mon numéro est le 01 69 85 80 02

mon numéro est le
<loc.add.elec>
<name>01 69 85 80 02</name>
</loc.add.elec>

mon identifiant skype est jean.dupont

mon identifiant skype est
<loc.add.elec>
<name>jean.dupont</name>
</loc.add.elec>

Radio Bleue sur 98.8 MHz

<prod.media>Radio Bleue</prod.media>
sur
<loc.add.elec>
<name>98.8 MHz</name>
</loc.add.elec>

suivez-moi sur Twitter à @leguidedannotation

suivez-moi sur
<prod.soft>
<name> Twitter </name>
</prod.soft>
à
<loc.add.elec>
<name> @leguidedannotation </name>
</loc.add.elec>

Sites Internet. On trouve beaucoup de cas particuliers :
- si l'on fait référence à l'accès au site c'est un <lod.add.elec> :
  retrouvez cet article sur lemonde.fr
- si l'on fait référence au site dans sa globalité, c'est un <prod.media> :
  interview à retrouver sur lemonde.fr
  mediapart.fr indique que Eric Woerth a bien touché 50000 euros
- si l'on fait référence à l'entreprise qui édite le site c'est un <org.ent> :
  Sarkozy dénonce mediapart.fr

Les adresses de site (*www.radio-france.fr*) sont annotées comme <loc.add.elec>. Par contre, *le site internet Radio France* ne constitue pas une entité nommée en soi (annoter ici uniquement *Radio France*, en prod.media).

2.5 Productions humaines : prod

Cette catégorie est proche de la catégorie PRODUCT de Sekine.

Nous considérons que le nombre de sous-catégories de PRODUCT chez Sekine est trop grand et serait trop lourd pour l'étiquetage de référence. Nous distinguons quelques catégories et regroupons les autres en quelques grandes catégories.

2.5.1 Marque d'objet : prod.object

Il s'agit principalement d'objets manufacturés. Les prod.objects ont les sous-types suivants. *Pour réduire la charge d'annotation, ces sous-types n'ont pas à être annotés.*

véhicule la navette Endeavour, la Peugeot 107 (mais pas navette, voiture)

aliment Perrier, Pink Lady, Label 5, Big Mac (mais pas limonade, pomme, whisky, hamburger, buche de Noel, Christmas pudding)

vêtement et autres accessoires : Levi's, Rolex (mais pas jean, montre)

médicament Lovenox, Xantax, Nurofen (mais pas antibiotique, paracétamol, ibuprofène)

arme Colt, Winchester 94, AK 47

Exemples

J'ai piloté la navette Endeavour

j'ai piloté la

<prod.object> prod.object

<kind>navette</kind>
<name>Endeavour</name>

</prod.object>

La société Peugeot commercialise des 107

la

<org.ent>

<kind>société</kind>
<name>Peugeot</name>

</org.ent>
commercialise des
<prod.object><name>107</name></prod.object>

L'AK 47 est le fusil le plus répandu dans le monde
L'
<prod.object>
<name>AK 47</name>
</prod.object>
est le fusil le plus répandu dans le monde

La Kalachnikov est le fusil le plus répandu dans le monde
La
<prod.object>
<name>Kalachnikov</name>
</prod.object>
est le fusil le plus répandu dans le monde

J'ai vendu ma vieille R21
<prod.object>
<name>R21</name>
</prod.object>

Par contre, lorsque le nom contient une entité d'un type différent, alors ce type doit être annoté.

la Renault 21
<prod.object>
<org.ent>Renault</org.ent>
<name>21</name>
</prod.object>

J'ai conduit une Renault Espace
j'ai conduit une
<prod.object>
<org.ent><name>Renault</name></org.ent>
<name>Espace</name>
</prod.object>

J'ai acheté une Peugeot 107
j'ai acheté une
<prod.object>
<org.ent><name>Peugeot</name></org.ent>
<name>107</name>
</prod.object>

2.5.2 Œuvre artistique : prod.art

Cette catégorie inclut les peintures, sculptures, photographies, œuvres de musique (dont opéra), pièces de théâtre, danse (et autres spectacles), films (y compris téléfilms), livres, etc.

Note : nous n'introduisons pas de différence entre une œuvre (*la Mégère apprivoisée*) et le fait de jouer cette œuvre (*je suis allé voir la Mégère apprivoisée*).

Exemples

Le Baiser de Klimt

prod.art

```
<prod.art>
    <name>Le Baiser</name>
</prod.art>
de
<pers.ind>
    <name.last>Klimt</name.last>
</pers.ind>
```

Le Malade Imaginaire

```
<prod.art>
    <name>Le Malade Imaginaire</name>
</prod.art>
```

le Guernica de Picasso

```
le
<prod.art>
    <name>Guernica</name>
</prod.art>
de
<pers.ind>
    <name.last>Picasso</name.last>
</pers.ind>
```

2.5.3 Production médiatique : prod.media

Tout ce qui est diffusé dans la presse, à la radio ou à la télévision : journaux, magazines, émissions, catalogues de vente, etc. Voir la section 1.3.4 pour la distinction avec org.ent.

Exemples

Mickaël Vendetta veut quitter La Ferme des Célébrités

<pers.ind>
<name.first>Mickaël</name.first>
<name.last>Vendetta</name.last>
</pers.ind>
veut quitter
<prod.media>
<name>La Ferme des Célébrités</name>
</prod.media>

prod.media

Le Monde a publié un article sur 1984 de George Orwell.

<prod.media>
<name>Le Monde</name>
</prod.media>
a publié un article sur
<prod.art>
<name>1984</name>
</prod.art>
de
<pers.ind>
<name.first>George</name.first>
<name.last>Orwell</name.last>
</pers.ind>

Sont exclus les produits rangés dans les œuvres d'art : films, téléfilms, etc.

2.5.4 Produits financiers : prod.fin

Les produits bancaires (*PEP*, *CEL*) sont des produits financiers. Les indices boursiers (*CAC 40*, *NASDAQ*, etc.) et les monnaies (l'Euro, le Dollar, etc.) sont annotés comme des produits financiers dans le cas où on évoque leur évolution à la bourse.

Ex : *Le CAC 40 a perdu 5%*, *Le Dollar a été réevalué*, etc.

Les dispositifs de rémunération (*le SMIC*, *RSA*, allocations diverses, etc., y compris allocations de recherche) sont annotés comme des produits financiers.

Ex : *Le RMI ne suffit pas pour payer un loyer.*

la gérante du fonds DWS signature court terme chez DWS Investments en France

la
<func.ind>gérante</func.ind>
du
<prod.fin>
<kind>fonds</kind>
<name>DWS signature court terme</name>

</prod.fin>
chez
<org.ent>DWS Investments</org.ent>
en
<loc.adm.nat>France</loc.adm.nat>

2.5.5 Logiciels : prod.soft

Firefox 3.5, World of Warcraft, Facebook, Twitter, MS-DOS, Windows 2000, Vista, etc.

le système d'exploitation Unix

le
<prod.soft> prod.soft
<kind>système d'exploitation</kind>
<name>Unix</name>
</prod.soft>

Firefox 3.5

<prod.soft>
<name>Firefox 3.5</name>
</prod.soft>

Linux Mandriva 2010.0 [12]

<prod.soft>
<name>Linux</name>
<name>Mandriva 2010.0</name>
</prod.soft>

Windows Vista

<prod.soft>
<name>Windows Vista</name>
</prod.soft>

2.5.6 Prix divers : prod.award

Le Prix Nobel de chimie, la palme d'or, l'Ours d'argent, le Renaudot, le César de la meilleure actrice.

2.5.6.1 Composants d'un prix

Le nom du prix est étiqueté par <name>. On inclut dedans les éléments indiquant le niveau du prix : or (médaille d'*or*), argent, 1er (*1er* accessit), 2e, etc.

Le composant <award-cat> permet de préciser la catégorie du prix, lorsque celui-ci en comporte plusieurs (Prix Nobel de *chimie*, César de la *meilleure actrice*).

12. Dans cet exemple, <name>Linux</name> est peu satisfaisant : on serait tenté de le mettre en <kind>, mais c'est un nom propre.

2.5.6.2 Exemples

Le Prix Nobel de chimie

le
<prod.award>
<name>Prix Nobel</name>
de
<award-cat>chimie</award-cat>
</prod.award>

la palme d'or

la
<prod.award>
<name>palme d'or</name>
</prod.award>

l'Ours d'argent

l'
<prod.award>
<name>Ours d'argent</name>
</prod.award>

le Renaudot

le
<prod.award>
<name>Renaudot</name>
</prod.award>

le César de la meilleure actrice

le
<prod.award>
<name>César</name>
de la
<award-cat>meilleure actrice</award-cat>
</prod.award>

2.5.7 Ligne de transport : prod.serv (service)

Ligne de train, d'avion...

Exemples

le RER B

le
<prod.serv>
<kind>RER</kind>
<name>B</name>
</prod.serv>

le TGV 822

le
<prod.serv>
<prod.object>TGV</prod.object>
<name>822</name>
</prod.serv>

le vol AF 2334

le
<prod.serv>
<kind>vol</kind>
<name>AF 2334</name>
</prod.serv>

le TGV Paris-Lyon

le
<prod.serv>
<prod.object>TGV</prod.object>
<name>
<loc.adm.town>Paris</loc.adm.town>
-
<loc.adm.town>Lyon</loc.adm.town>
</name>
</prod.serv>

le TGV Paris-Lyon

le
<prod.serv>
<prod.object>TGV</prod.object>
<name>Paris-Lyon</name> [13]
</prod.serv>

13. Rappel : on admet que pour des questions de coût d'annotation, on peut décider de ne pas annoter au-delà du composant <name>. Néanmoins, l'annotation idéale devrait inclure les types <loc.adm.town> sur chacune des villes.

le TGV Atlantique

```
le
<prod.serv>
    <prod.object>TGV</prod.object>
    <name>
        <loc.phys.hydro>Atlantique</loc.phys.hydro>
    </name>
</prod.serv>
```

le TGV Atlantique

```
le
<prod.serv>
    <prod.object>TGV</prod.object>
    <name>Atlantique</name>
</prod.serv>
```

2.5.8 Doctrine

Le socialisme, le communisme, le bouddhisme, le protestantisme...

Exemples

le socialisme est un un type d'organisation sociale...

```
le
<prod.doctr>
    <name>socialisme</name>
</prod.doctr>
```

2.5.9 Rule

Loi, décret, accord, traité, types de contrats (*CDD*, *CDI*)...

Exemples

le traité de Versailles

```
le
<prod.rule>
    <kind>traité</kind>
    de
    <name><loc.adm.town>Versailles</loc.adm.town></name>
</prod.rule>
```

le dernier plan de paix international

```
le
<prod.rule>
```

<qualifier> dernier </qualifier>
<kind> plan de paix </kind>
<qualifier> international </qualifier>
</prod.rule>

Nous rappelons que le <loc.adm.town> est facultatif (cf. section 1.2.2.4).

2.5.10 Productions autres : prod.other

Les productions humaines qui ne sont pas prévues dans les catégories explicites ci-dessus.

le Monopoly
le
<prod.other> prod.other
<name> Monopoly </name>
</prod.other>

On considère également comme prod.other les mythes et mythologies (*Mythologie grecque*).

2.6 Quantités : amount

2.6.1 Définitions : quantités autres que durées

Une quantité est composée d'une valeur assortie éventuellement d'une unité de mesure, suivie éventuellement d'un objet.

- La valeur doit être numérique (*trois kilomètres*, *une dizaine de kilomètres*, *une douzaine de tomates*, *une demi livre*). Les quantificateurs ne font pas partie des valeurs (tous les pensionnaires, plusieurs articles, quelques comptes, de nombreux chars, beaucoup de gens).
  - Néanmoins, lorsqu'une unité (plutôt qu'un objet) est présente et précédée d'un quantificateur (*quelques mètres*, *plusieurs jours*), on a une quantité que l'on annote.
- La valeur numérique peut être précise (douze) ou non (une douzaine). Douzaine, million entrent dans les valeurs numériques.
- Les unités sont d'une part les unités de mesure standards [14] ; une unité peut aussi être un contenant qui contient un objet mentionné dans l'expression (*un cube de bouillon*, *une charette de foin*).
- Un objet assorti d'une valeur numérique ou d'une spécification de partie est donc annoté (*deux tomates*, *une boîte de tomates*, *une demi tomate*) ainsi que des entités d'un autre type (*les 22 bleus*).

Proposition 19. *En revanche, un objet unique (*une tomate*,* la charette*), typiquement précédé de l'article un ou le, sans spécification de partie non plus, n'est pas annoté.*

14. Voir par exemple Le Système international d'unités[BIPM 2006, p. 14] : « Les grandeurs de base utilisées dans le SI sont la longueur, la masse, le temps, le courant électrique, la température thermodynamique, la quantité de matière et l'intensité lumineuse ». Les unités de base du SI « sont le mètre, le kilogramme, la seconde, l'ampère, le kelvin, la mole et la candela » (ibid.). De nombreuses autres unités existent, elles en sont dérivées ou pas.

Proposition 20. *De même, un simple pluriel (*les tomates, les êtres humains*,* les syndicats*,* les chômeurs*) ne spécifie pas une quantité, et n'est pas annoté non plus.*

Affirmation 21. Un objet est souvent une entité définie par ailleurs (*trois pompier*s, *22 bleus*), et possède donc un type. Idéalement, on aimerait que ces types soient annotés (comme précision supplémentaire, nécessairement imbriquée sous le composant *object* ; voir le deuxième exemple d'annotation pour *trois pompiers* ci-dessous), mais les contraintes de temps et de coût nous ont amenés à décider d'abandonner ce niveau de détail.

trois pompiers — *valeur + objet*

```
<amount>
    <val> trois </val>
    <object> pompiers </object>
</amount>
```

trois pompiers — *valeur + objet, annoté*

```
<amount>
    <val> trois </val>
    <object> <func.coll>pompiers</func.coll> </object>
</amount>
```

22 bleus — *valeur + objet*

```
<amount>
    <val> 22 </val>
    <object> bleus </object>
</amount>
```

22 bleus — *valeur + objet, annoté*

```
<amount>
    <val> 22 </val>
    <object> <pers.coll>bleus</pers.coll> </object>
</amount>
```

quelques mètres — *quantificateur + unité, annoté*

```
<amount>
    <val> quelques </val>
    <unit> mètres </unit>
</amount>
```

quelques pompiers — *quantificateur + objet, pas un amount*

```
quelques
<func.coll>
    <kind> pompiers </kind>
</func.coll>
```

Affirmation 22. Les résultats sportifs ne sont pas annotés (*Nadal a battu Grosjean en 3 sets 7/6 7/5 6/4*).

2.6.2 Exemples : quantités autres que durées

trois kilomètres *valeur + unité*

```
<amount>
    <val> trois </val>
    <unit> kilomètres </unit>
</amount>
```

unit

une dizaine de kilomètres

```
<amount>
    <val> une dizaine </val>
    de
    <unit> kilomètres </unit>
</amount>
```

trois régiments valeur + unité qui est par ailleurs une entité désignant un ensemble

```
<amount>
    <val> trois </val>
    <unit> <pers.coll>régiments</pers.coll> </unit>
</amount>
```

une douzaine de tomates *valeur + objet*

```
<amount>
    <val> une douzaine </val>
    de
    <object> tomates </object>
</amount>
```

quatre milliers de tomates

```
<amount>
    <val>quatre milliers</val>
    de
    <object>tomates</object>
</amount>
```

une tomate *un + objet, on n'annote pas*

```
une tomate
```

une boîte de tomates *spécification de partie, on annote*

```
<amount>
    <val> une </val>
    <unit> boîte </unit>
    de
    <object>tomates</object>
```

</amount>

trois kilos de tomates *valeur + unité + objet*

<amount>
<val> trois </val>
<unit> kilos </unit>
de
<object>tomates</object>
</amount>

trois cuillerées de farine de châtaignes *unité = contenant*

<amount>
<val> trois </val>
<unit> cuillerées </unit>
de
<object>farine de châtaignes</object>
</amount>

un car de CRS

<amount>
<val> un </val>
<unit> car </unit>
de
<object>CRS</object>
</amount>

deux tomates de 300 g *deux entités annotées séparément*

<amount>
<val>deux</val>
<object>tomates</object>
</amount>
de
<amount>
<val>300</val>
<unit>g</unit>
</amount>

une tomate de 300 g *un seul objet entier, on ne l'annote pas → une seule entité*

une tomate de
<amount>
<val>300</val>
<unit>g</unit>

</amount>

deux boîtes de tomates concassées de 300 g *deux entités*

<amount>
<val>deux</val>
<unit>boîtes</unit>
de
<object>tomates concassées</object>
</amount>
de
<amount>
<val>300</val>
<unit>g</unit>
</amount>

une boîte de tomates concassées de 400g [15]

spécification de partie, on annote→ deux entités

<amount>
<val> une </val>
<unit> boîte </unit>
de
<object>tomates concassées</object>
</amount>
de
<amount>
<val> 400 </val>
<unit> g </unit>
</amount>

deux cubes [16] *de bouillon de poule* *unité particulière*

<amount>
<val> deux </val>
<unit> cubes </unit>
de
<object>bouillon de poule</object>
</amount>

deux bouillons-cubes de volaille [17] *pas une unité*

<amount>

15. http://www.marmiton.org/Recettes/Recette_poulet-aux-olives-de-cess-facon-tajine_92816.aspx
16. L'unité non conventionnelle cube indique le volume de bouillon de poule à utiliser dans la recette.
17. bouillon-cube ne désigne pas une unité de mesure pour une quantité de volaille : ce n'est pas un contenant pour une quantité de volaille.

<amount>
<val> deux </val>
<object>bouillons-cubes de volaille</object>
</amount>

environ trois kilomètres *modifieur de valeur*

<amount>
<qualifier>environ</qualifier>
<val>trois</val>
<unit>kilomètres</unit>
</amount>

entre deux et trois kilomètres *intervalle de valeurs ; voir l'affirmation 13*

<amount>
<range-mark>entre</range-mark>
<val>deux</val>
<range-mark>et</range-mark>
<val>trois</val>
<unit>kilomètres</unit>
</amount>

de trois à cinq kilos

<amount>
<range-mark>de</range-mark>
<val>trois</val>
<range-mark>à</range-mark>
<val>cinq</val>
<unit>kilos</unit>
</amount>

la distance en kilomètres c'est trois *dislocation : deux <amount>*

la distance en
<amount>
<unit> kilomètres </unit>
</amount>
c'est
<amount>
<val> trois </val>
</amount>

3 personnes *(ni un nom propre, ni un démonyme : voir la section 2.1.3.2)*

<amount>

<val> 3 </val>
<object> personnes </object>
</amount>

1300 enfants *(ni un nom propre, ni un démonyme : voir la section 2.1.3.2)*
<amount>
<val> 1300 </val>
<object> enfants </object>
</amount>

21 blessés *(ni un nom propre, ni un démonyme : voir la section 2.1.3.2)*
<amount>
<val> 21 </val>
<object> blessés </object>
</amount>

j'ai 10 ans *les âges sont des quantités*
j'ai
<amount>
<val> 10 </val>
<unit> ans </unit>
</amount>

il a quitté la maison à 12 ans
il a quitté la maison à
<amount>
<val> 12 </val>
<unit> ans </unit>
</amount>

2.6.3 Définitions : durées

Les durées sont annotées comme les autres quantités.

2.6.4 Exemples : durées

la réunion a eu lieu tous les lundis pendant trois ans
la réunion a eu lieu
<time.date.abs>
<time-modifier>tous les</time-modifier>
<week> lundis </week>
</time.date.abs>
<amount>

<qualifier>pendant</qualifier>
<val>trois</val>
<unit>ans</unit>
</amount>

la réunion va durer entre 3 et 5 heures

la réunion va durer
<amount>
<range-mark>entre</range-mark>
<val>3</val>
<range-mark>et</range-mark>
<val>5</val>
<unit>heures</unit>
</amount>

une semaine entière

<amount>
<val> une </val>
<unit> semaine </unit>
<qualifier> entière </qualifier>
</amount>

2.6.5 Principes généraux pour les amount

Étendue de l'objet En général, on ne prend pas les adjectifs et compléments dans <object>. Si nécessaire, on se limite à un seul adjectif ou complément.

150 salariés américains => on peut prendre *salarié américains* comme <object>
10 employés de l'ambassade américaine => on ne prend que *employé* dans <object>

Élision de l'objet On met le complément d'objet dans <object> même quand l'objet est implicite.

La boite comporte 10 couverts en bois et 5 en métal

La boite comporte
<amount>
<val> 10 </val>
<object> couverts en bois </object>
</amount>
et
<amount>
<val> 5 </val>
<object> en métal </object>
</amount>

Amount et pourcentage

55 pour cent des contrôleurs de la sncf

<amount>
<val> 55 </val> <unit> pour cent </unit>
des
<object> contrôleurs </object>
</amount>
de la
<org.ent> sncf </org.ent>

Valeurs possibles des range-mark Les conjonctions de coordination (*et*, *ou*), les prépositions contractées (*au*) ou pas (*à*). Les valeurs reliées par un OU inclusif sont regroupées au sein d'un type <amount> alors que les valeurs reliées par un OU exclusif sont séparées en deux types <amount> :

4 ou 5 blessés

<amount>
<val> 4 </val>
<range-mark> ou </range-mark>
<val> 5 </val>
<object> blessés </object>
</amount>

une heure ou une demi-heure ou 3 ou 20 minutes

<amount>
<val> une </val>
<unit> heure </unit>
</amount>
ou
<amount>
<val> une </val>
<unit> demi-heure </unit>
</amount>
ou
<amount>
<val> 3 </val>
<range-mark> ou </range-mark>
<val> 20 </val>
<unit> minutes </unit>
</amount>

Voir aussi la section 1.2.6.4 sur les coordinations.

Les fractions : <val> + <unit>, ou seulement <val> ? les entités comme *3/4*, *1/2*, etc. sont annotées comme valeur uniquement.

Exception : *pourcent*, % sont considérés comme <unit>

Empan des amount

– *X objets sur Y*

60 fédérations départementales sur 95

<amount>
<val> 60 </val>
<unit> fédérations départementales </unit>
</amount>
sur
<amount> <val> 95 </val> </amount>

– *X des Y objets*
La première valeur est un <extractor>

une des trois filles du président

<amount>
<extractor> une </extractor>
des
<val> trois </val>
<object> filles </object>
</amount>

Valeur sans spécification d'unité Faut-il annoter une quantité si l'unité n'est pas référencée ?
Ex : *j'en ai vu 2 3*
Si le contexte nous donne l'unité alors oui.

2.7 Point dans le temps : time

Nous distinguons dates (time.date) et heures (time.hour). À l'intérieur de ces deux types, nous distinguons également temps absolu (.abs) et temps relatif (.rel). De même que dans les quantités (amount), le composant range-mark peut être utilisé.

Rappel : les propositions relatives ou subordonnées ne peuvent pas entrer dans une entité (voir la section 1.2.4), notamment dans le composant time-modifier.

2.7.1 Date : time.date

Une date peut être composée d'un jour de semaine (*lundi...dimanche*) d'un jour (*1...31*), d'un mois (*janvier....décembre*), d'une saison (*printemps, été, automne, hiver*) [18], d'une année (*nombre*) et d'une référence temporelle (avant/après JC).

Un time.date peut aussi inclure, en plus de la date, une référence à une partie de la journée (*demain soir*, *demain midi*, *demain à 15 heures*, *demain dans la soirée*). Si date et partie de la journée peuvent être intervertis, sans les modifier, en conservant le sens (*demain*

18. Les saisons seules sont annotées comme <time.date.rel> ; en revanche, elles sont annotées <time.date.abs> si elles sont combinées avec un spécifieur temporel : été 2009.

à 15 heures : à 15 heures demain, mais pas *demain soir : soir demain*), on annote deux entités séparées. Sinon on annote une seule entité qui inclut le tout. Voir plus bas l'exemple *hier matin*.

2.7.1.1 Date absolue : time.date.abs

Les dates absolues et précises sont annotées en time.date.abs. Les composants suivants peuvent être utilisés : jour de la semaine (<week>), numéro du jour (<day>), mois (<month>), année (<year>), siècle (<century>), millénaire (<millenium>), ère (<reference-era>).

lundi 25 janvier 2010

<time.date.abs> — time.date.abs
- <week> lundi </week> — *day of the week* — week
- <day> 25 </day> — day
- <month> janvier </month> — month
- <year> 2010 </year> — year

</time.date.abs>

Les dates absolues mais imprécises sont également annotées en time.date.abs. Leur point de référence dans le calendrier n'est pas connu avec certitude mais leurs composants sont absolus.

rendez-vous mardi

rendez-vous
<time.date.rel>
- <week> mardi </week>

</time.date.rel>

Cependant, dès qu'un marqueur explicite de relativité (par exemple, *prochain*) est spécifié (*rendez-vous mardi prochain*), on a une date relative (voir plus bas la section 2.7.1.2) Le déterminant (*le*, *la*, *les*, *l'*) est inclus dans l'entité uniquement s'il a une fonction équivalente à une préposition (*à*, *en*) (*un rendez-vous le 23 janvier* est parallèle à *un rendez-vous à dix heures*) :

le 25 janvier

<time.date.abs>
- <time-modifier>le</time-modifier> — time-modifier
- <day> 25 </day>
- <month> janvier </month>

</time.date.abs>

vers le 25 janvier

<time.date.abs>
- <time-modifier>vers le</time-modifier>
- <day> 25 </day>
- <month> janvier </month>

</time.date.abs>

en 300 avant JC

<time.date.abs>
<time-modifier>en</time-modifier>
<year> 300 </year>
<time-modifier>avant JC </time-modifier>
</time.date.abs>

le mois de janvier

mois *est une catégorie de date ;* le *n'a pas la fonction d'une préposition*
le
<time.date.abs>
<kind>mois</kind>
de
<month> janvier </month>
</time.date.abs>

l'année 1900

l'
<time.date.abs>
<kind>année</kind>
<year> 1900 </year>
</time.date.abs>

le 21e siècle

le
<time.date.abs>
<century>21e siècle</century>
</time.date.abs>

le troisième millénaire

le
<time.date.abs>
<millenium>troisième millénaire</millenium>
</time.date.abs>

le second empire

le
<time.date.abs>
<reference-era>second empire</reference-era>
</time.date.abs>

Mais on n'annote pas *ère professionnelle* dans *la cinquième joueuse de l' ère professionnelle*.

2.7.1.2 Date relative : time.date.rel

Une date relative est formée autour d'un terme intrinsèquement relatif (*maintenant*, *aujourd'hui*, *hier*, *demain*) ou est marquée comme telle par un mot comme *ce*, *prochain*, etc. Une date relative peut inclure des composants de dates absolues (*lundi*, *janvier*), qui sont alors annotés comme habituellement pour ces composants.

Les termes comme *demain*, *maintenant*, *en ce moment*, *ce soir*, *hier matin*, mais aussi le *jour de Noël*, le *nouvel an* etc. sont annotés comme des dates relatives. On utilise le composant <name> pour les mots comme *maintenant*, *moment*, *soir*, etc. On utilise le composant <kind> pour les types de dates comme *jour*, *mois*, *année*, etc.

Par contre des expressions comme le *jour de sa naissance* ou encore il est né le *jour de l'élection de Mitterrand* ne sont pas annotées car elles nécessitent des raisonnements plus poussés pour déterminer la date absolue à laquelle elles correspondent (par exemple, l'élection de François Mitterrand le 21 mai 1981).

lundi prochain

time.date.rel

```
<time.date.rel>
    <week> lundi </week>
    <time-modifier>prochain</time-modifier>
</time.date.rel>
```

hier matin

```
<time.date.rel>
    <name> hier </name>
    <time-modifier>matin</time-modifier>
</time.date.rel>
```

hier matin *annotation optionnelle du contenu du composant time-modifier*

```
<time.date.rel>
    <name> hier </name>
    <time-modifier>
        <time.hour.rel><name>matin</name></time.hour.rel>
    </time-modifier>
</time.date.rel>
```

Il y a quinze ans, rares étaient les utilisateurs de téléphones portables, maintenant tout le monde en a un.
ici, maintenant a une valeur plus large qu'une journée

```
<time.date.rel>
    <name> maintenant </name>
</time.date.rel>
```

cette semaine

```
<time.date.rel>
```

<time-modifier>cette</time-modifier>
<kind>semaine</kind>
</time.date.rel>

tout au long de la semaine
<time.date.rel>
<time-modifier>tout au long</time-modifier>
de la
<kind>semaine</kind>
</time.date.rel>

Il y a quinze ans , rares étaient les utilisateurs de téléphones portables , maintenant tout le monde en a un .
<time.date.rel>
<time-modifier> Il y a </time-modifier>
<amount>
<val> quinze </val>
<unit> ans </unit>
</amount>
</time.date.rel>
, rares étaient les utilisateurs de téléphones portables ,
<time.date.rel>
<name> maintenant </name>
</time.date.rel>
tout le monde en a un .

2.7.2 Heure : time.hour

Ce qui est inférieur à la journée (24 heures) est annoté comme heure (time.hour). Une heure peut être composée d'une valeur, d'une unité [19] ou d'un name et éventuellement d'une référence temporelle (*de l'après-midi*, *AM*, *PM*).

Les périodes de la journée (*matin*, *soir*) sont inférieures à la journée, donc annotées comme time.hour.

2.7.2.1 Heure absolue : time.hour.abs

Les heures absolues sont annotées time.hour.abs.

quinze heures trente
<time.hour.abs>

19. La signification de la « valeur » est différente de celle qu'elle a dans une quantité. Dans une heure, la « valeur » est plutôt de type ordinal : *3 heures* signifie *la troisième heure*.

<val> quinze </val>
<unit> heures </unit>
<val> trente </val>
</time.hour.abs>

Les heures absolues mais imprécises sont également annotées en time.hour.abs. Leur point de référence dans le calendrier n'est pas connu avec certitude mais leurs composants sont absolus.

aux alentours de quinze heures trente

<time.hour.abs>
<time-modifier>aux alentours de</time-modifier>
<val> quinze </val>
<unit> heures </unit>
<val> trente </val>
</time.hour.abs>

3 heures du matin

<time.hour.abs>
<val> 3 </val>
<unit> heures </unit>
<time-modifier>du matin </time-modifier>
</time.hour.abs>

Les valeurs temporelles comme *et demie* dans *une heure et demie* sont annotées comme valeur.

une heure et demie

<time.hour.abs>
<val> une </val>
<unit> heure </unit>
<val> et demie </val>
</time.hour.abs>

une heure moins dix

<time.hour.abs>
<val> une </val>
<unit> heure </unit>
<val> moins dix </val>
</time.hour.abs>

2.7.2.2 Heure relative : time.hour.rel

Les heures qui ne peuvent être déterminées qu'avec un raisonnement s'ancrant sur la connaissance de l'heure de l'énonciation sont annotées en time.hour.rel ; pour les heures relatives, il est important d'utiliser le type <amount> autour des composants <val> et <unit> car on ne sait pas nécessairement faire la distinction entre durée et montant :

la réunion commence dans deux heures

```
la réunion commence
<time.hour.rel>                                          time.hour.rel
    <time-modifier>dans</time-modifier>
    <amount>
        <val> deux </val>
        <unit> heures </unit>
    </amount>
</time.hour.rel>
```

ce matin

```
<time.hour.rel>
    <time-modifier>ce</time-modifier>
    <name> matin </name>
</time.hour.rel>
```

2.7.3 Intervalles de dates / imprécision

Le lien entre les deux entités est remis à plus tard.

La réunion commencera entre 3 et 5 heures

```
la réunion commencera
    <time.hour.abs>
        <range-mark>entre</range-mark>
        <val> 3 </val>
        <range-mark>et</range-mark>
        <val> 5 </val>
        <unit> heures </unit>
    </time.hour.abs>
```

2.8 Événements : event

Il est gênant dans l'annotation de ne pas pouvoir annoter les événements (par exemple, *le 23e congrès de la CFDT*). Leur absence fait se demander comment annoter, avec le guide actuel, des entités qui sont des événements. On introduit donc une définition souple des événements.

2.8.1 Définitions

On définit un seul type d'événement.

event événement

Les composants possibles d'un événement sont toutes les entités et tous les composants qui existent par ailleurs dans le guide. Toutes ces entités et ces événements doivent être annotés.

2.8.2 Exemples

Le Tour de France

Le

<event>

Tour de

<loc.adm.nat>France</loc.adm.nat>

</event>

le 44e congrès de la CFDT

le

<event>

44e congrès de la

<org.ent>CFDT</org.ent>

</event>

la coupe du monde

la

<event> coupe du monde </event>

la ligue des champions

la

<event> ligue des champions </event>

2.9 Entités non annotées

Nous listons ici des expressions qui n'ont pas à être annotées, car elles ne rentrent dans aucune des catégories définies jusqu'à présent.

- Les passages en langues étrangères
- Les noms de maladies (*SIDA*, *Grippe A*, etc.)
- Les phénomènes psychologiques (*complexe d'Œdipe*, *syndrome de Stockholm*, etc.)
- Les termes scientifiques ne pouvant se ramener à un « produit » (*ADN*, etc.)
- Les filières d'enseignement (*Staps*, *DEUG*, etc.)
- Les contrats particuliers (*le contrat Coca-Cola/Danone*, etc.)
  - mais dans *le contrat Coca-Cola*, l'entité *Coca-Cola* doit être annotée (prod.object).

- Les affaires politiques et/ou judiciaires (*Watergate*, *Monica-gate*, *affaire Dickinson*, etc.). Optionnel : peuvent dans certains cas entrer dans la catégorie <event> si les annotateurs les considèrent comme telles.
- Les phénomènes climatiques (*le Mistral*, *la tempête Xinthya*, etc.). Optionnel : peuvent dans certains cas entrer dans la catégorie <event> si les annotateurs les considèrent comme tels.
- Les phénomènes sociaux (*l'immigration arménienne*[20], etc.). Optionnel : Peuvent dans certains cas entrer dans la catégorie <event> si les annotateurs les considèrent comme tels.

REMARQUE : Dans certains cas, il faut quand même annoter les composants de ces expressions. Exemples :

- on n'annote pas *syndrome de Stockholm* mais il faut annoter *Stockholm* (<loc.adm.town>)
- on n'annote pas *complexe d'Oedipe* mais il faut annoter *Oedipe* (<pers.ind>)
- etc.

20. Cette expression est néanmoins annotée si elle désigne un groupe de personnes plutôt qu'un processus, voir la section 2.1.3.2.

Chapitre 3

Guide rapide

3.1 Types et sous-types

Cette section liste l'ensemble des sous-types associés à chaque type d'entité.

3.1.1 pers

pers.ind Une seule personne (*François Mitterrand*).

pers.coll Un ensemble de personnes y compris les groupes musicaux (les *français*, les *Beatles*, la *Mano Negra*).

3.1.2 func

func.ind Une fonction ou un métier (*maire* de Paris, le *pompier*).

func.coll Un ensemble de personnes d'une même corporation (les *maires* de France, les *pompiers* de Paris).

3.1.3 org

org.ent Organisation qui commercialise des produits ou qui rend des services (la *société Peugeot*, la *Pitié-Salpêtrière*).

org.adm Organisation qui joue un rôle principalement administratif (le *Ministère des Affaires Étrangères*).

3.1.4 loc

3.1.4.1 loc.adm

Ce sous-type se décompose en quatre sous sous-types.

loc.adm.town Quartier, arrondissement, lieu-dit, hameau, village, ville, etc. (*Paris*, le *Val de Crüye*).

loc.adm.reg Cantons, communautés de communes, départements, régions, Länder, etc. (les *Bouches du Rhône*, le *Pays-Basque espagnol*).

loc.adm.nat Pays (*France*).

loc.adm.sup Régions du monde, continent (*Pays-Basque*, *Maghreb*).

3.1.4.2 loc.phys

Ce sous-type se décompose en trois sous sous-types.

loc.phys.geo Montagnes, plaines, plateaux, grottes, volcans, canyons (*gouffre de Padirac*, le *mont Ventoux*).

loc.phys.hydro Océans, mers, rivières, fleuves, étangs, marais (la *Seine*, le *lac Paladru*).

loc.phys.astro Planètes, étoiles, galaxies et leurs parties (la *Lune*, la *mer de la Tranquillité*).

3.1.4.3 loc.oro

Désigne les route, autoroute, rue, avenue, place, etc. (l' *autoroute A6*).

3.1.4.4 loc.fac

Désigne les bâtiments (le *Palais de l'Élysée*).

3.1.4.5 loc.add

Ce sous-type se décompose en deux sous sous-types.

loc.add.phys Désigne les adresses physiques (*LIMSI-CNRS, Bâtiment 508, BP133, 91403 Orsay Cedex*).

loc.add.elec Désigne les coordonnées électroniques (numéros de téléphone, de télécopie, URL, adresse de messagerie électronique, identifiants de réseaux sociaux ou d'outil de communication par Internet, etc., *http ://www.limsi.fr/*, *01-69-85-80-00*).

3.1.5 prod

prod.object Produit manufacturé correspondant aux sous-types suivants (non annotés) : véhicules, aliment, vêtements, médicament, armes (la *navette Endeavour*, le *Big Mac*, une *Rolex*, le *Nurofen*, l' *AK47*).

prod.art Peintures, sculptures, pièces de théâtre, photographies, œuvres de musique, films, livres, etc. (*Le Baiser* de Klimt, *Le malade imaginaire*, le *Guernica*).

prod.media Journaux, magazines, émissions, catalogues de vente, etc. (*Le Figaro*, *Le sept à huit*, *La ferme célébrités*).

prod.fin Produits financiers. (*PEP*, *CEL*, *DWS signature court terme*).

prod.soft Logiciels. (*Linux Mandriva 2010.0*, *Mac OS X*, *Système d'exploitation Unix*, *MSN Messenger*, *World of Warcraft*, *Tetris*).

prod.award Récompenses universitaires ou professionnelles, prix, diplômes, etc. (*Prix Nobel de Chimie*, *Palme d'Or*, le *Renaudot*, le *César de la meilleure actrice*, *Doctorat de physique nucléaire*, le *Brevet d'études professionnelles Métiers de la mode et des industries connexes*, le *Grand chelem*, la *cuillère en bois*, le *Bouclier de Brennus*).

prod.serv Lignes de transports. (le *RER A*, le *TGV 822*, le *vol AF 2334*, le *TGV Paris - Lyon*, le *TGV Atlantique*).

prod.doctr Doctrines politiques, philosophiques, religieuses, sectaires. (le *socialisme*, le *bouddhisme theravâda*, le *structuralisme*, le *mitterrandisme*, le *minimalisme*, la *scientologie*).

prod.rule Lois, décrets, accords, etc. (le *Traité de Versailles*, le *49.3*, la *Constitution*, l' *Arrêté du 12 novembre 2009 portant organisation du service de l'aviation civile de Saint-Pierre-et-Miquelon*, la *Loi Hadopi 2*, la *Loi Scellier*).

prod.other Les autres ENEs de type `prod` qui ne correspondent à aucun des sous-types précédents. (le *Monopoly*, le *passe Navigo*).

3.1.6 amount

Rassemble les quantités, montants et durées (*trois kilomètres*, *environ trois kilomètres*, *deux bouillon-cubes de volaille*, la réunion va durer *entre 3 et 5 heures*, il y a *3 ans*).

3.1.7 time

3.1.7.1 time.date

Ce sous-type se décompose en deux sous sous-types.

time.date.abs Une date absolue (*lundi 25 janvier 2010*).

time.date.rel Une date relative (*lundi prochain*, *demain*).

3.1.7.2 time.hour

Ce sous-type se décompose en deux sous sous-types.

time.hour.abs Une heure absolue (*15 heures 30*).

time.hour.rel Une heure relative (dans *une demi-heure*).

3.1.8 event

Un événement (*Le Tour de France*).

3.2 Composants

Cette section définit chaque composant. Des exemples d'usage sont déjà fournis dans le chapitre sur les entités.

3.2.1 Composants transverses

3.2.1.1 name

Partie correspondant directement à l'entité.
le mont *Ventoux*

```
le
  <loc.phys.geo>
      <kind> mont </kind>
      <name> Ventoux </name>
  </loc.phys.geo>
```

Dans cet exemple, *Ventoux* est l'entité annotée.

name est aussi employé si l'on a un nom (par exemple : *Asterix*) pour lequel on ne sait pas choisir entre nom (name.last), prénom (name.first), surnom (name.nickname) (voir la section 3.2.2.1).

3.2.1.2 name.nickname

Surnom, diminutif et acronyme (*tonton* dans un document sur François Mitterrand, *Sarko*, *DSK*, *PPDA*, *ville rose*).

3.2.1.3 kind

Partie d'une expression d'entité nommée qui désigne un hyperonyme (genre proche) de l'entité :
le maire de Paris

```
le <func.ind>
      <kind> maire </kind>
      de
      <loc.adm.town> Paris </loc.adm.town>
  </func.ind>
```

Dans cet exemple, *maire* est un genre de *fonction* (`<func.ind>`).

3.2.1.4 extractor

Permet de spécifier une partie (un individu) d'un ensemble :
le cinquième Beatles

Le <pers.ind>

<extractor>cinquième</extractor>
<name> Beatles </name>

</pers.ind>

Dans cet exemple, *cinquième* est une partie de l'ensemble *Beatles*.

<func.ind>

<extractor>un</extractor>
des
<kind> pompiers </kind>

</func.ind>

Dans cet exemple, *un* est une partie de l'ensemble *pompiers*.

Ce composant s'applique uniquement dans les cas qui portent sur des fonctions ou des personnes collectives (`<func.coll>` ou `<pers.coll>`). L'existence de ce composant transforme le sous-type collectif en sous-type individuel (respectivement `<func.ind>` et `<pers.ind>`).

3.2.1.5 demonym

Un démonyme spécifie l'origine géographique d'une entité sous forme d'un gentilé.
les mairies *françaises*

les <org.adm>

<kind> mairies </kind>
<demonym> françaises </demonym>

</org.adm>

Un démonyme peut s'appliquer à une personne, à une fonction, à une organisation, à un montant, etc.

demonym.nickname Un cas particulier de démonyme utilise une forme alternative pour désigner l'origine géographique d'une entité.
les froggies

les <pers.coll>

<demonym.nickname> froggies </demonym.nickname>

</pers.coll>

3.2.1.6 qualifier

Un qualifieur spécifie une entité sous forme d'un adjectif qualificatif.
le socialiste Bertrand Delanoë

```
Le <pers.ind>
    <qualifier> socialiste </qualifier>
    <name.first> Bertrand </name.first>
    <name.last> Delanoë </name.last>
</pers.ind>
```

3.2.1.7 val

Ce composant spécifie la valeur numérique d'une quantité ou d'une heure.
trois kilos

```
<amount>
    <val> trois </val>
    <unit> kilos </unit>
</amount>
```

trois heures

```
<time.hour.abs>
    <val> trois </val>
    <unit> heures </unit>
</time.hour.abs>
```

3.2.1.8 unit

Ce composant spécifie l'unité de mesure (unité standard, contenant [1], élément de date [2]) d'une quantité ou d'une heure (trois *kilos*, trois *heures*).
trois kilos

```
<amount>
    <val> trois </val>
    <unit> kilos </unit>
</amount>
```

trois heures

```
<time.hour.abs>
    <val> trois </val>
    <unit> heures </unit>
</time.hour.abs>
```

1. Un contenant est un élément qui peut contenir d'autres éléments : un *régiment* contient des hommes, un *car* de CRS contient des CRS, etc.
2. Année, mois, semaine, jour, heure, etc.

3.2.1.9 range-mark

Ce composant marque un intervalle entre deux valeurs, deux dates, deux heures.
entre deux et trois kilomètres

```
<amount>
    <range-mark>entre</range-mark>
    <val>deux</val>
    <range-mark>et</range-mark>
    <val>trois</val>
    <unit>kilomètres</unit>
</amount>
```

3.2.2 Composants employés dans chaque classe

3.2.2.1 Composants de personnes

Si l'on a un nom (par exemple : *Asterix*) pour lequel on ne sait pas choisir entre nom (name.last), prénom (name.first), surnom (name.nickname), on l'annote simplement comme name.

name.last Nom de famille (Samuel L. *Jackson*)

name.first Prénom (*Samuel* L. Jackson)

name.middle Enièmes prénoms ou initiales de ces prénoms, pour les noms anglo-saxons (Samuel *L.* Jackson).

title Titre ou désignateur de personne (*Monsieur* Chirac, *Son Altesse*).

3.2.2.2 Composants d'adresses

address-number Numéro de rue (au sens large y compris bis, ter, etc : *3 ter* rue des Lilas)

po-box Boîte postale (LIMSI-CNRS *BP133* 91403 Orsay Cedex)

zip-code Code postal (LIMSI-CNRS BP133 *91403* Orsay Cedex)

other-address-component Bâtiment, escalier, etc. (LIMSI-CNRS BP133 91403 Orsay *Cedex*)

3.2.2.3 Composants de quantités

Les composants val et unit sont transverses car ils apparaissent aussi dans les heures.

val Chiffre (*deux* litres)

unit Unité standard (deux *litres*), contenant (quatre *cars* de CRS), élément de date (3 *ans*)

object Objet (deux *cartons*, deux *bouillons-cubes de volaille*)

3.2.2.4 Composants de dates

day Numéro du jour dans le mois (mardi *30* mars 2010)

week Nom du jour (*mardi* 30 mars 2010)

month Nom du mois (mardi 30 *mars* 2010)

year Année (mardi 30 mars *2010*)

century Siècle (le *XXI^{ème} siècle*)

millenium Millénaire (le *troisième millénaire*)

reference-era Ère, période (l' *âge de fer*, le *second empire*)

3.2.2.5 Composants de date et heure

time-modifier Tout ce qui précise la spécification d'une date ou d'une heure, par exemple par rapport à un point du calendrier ou à un moment de la journée (en 300 *avant JC*, 3 heures *du matin*, trois heures *moins dix*, *vers* trois heures *moins dix*, *après* trois heures).

3.2.2.6 Composants d'heures

Les composants val et unit sont transverses car ils apparaissent aussi dans les quantités.

val Chiffre (*deux* minutes)

unit Unité (deux minutes)

3.2.2.7 Composants de prix

award-cat Catégorie de récompense (le Prix Nobel de *Chimie*, le César de la *Meilleure actrice*).

Bibliographie

BIPM 2006. *Le Système international d'unités (SI)*. Bureau international des poids et mesures. Organisation intergouvernementale de la Convention du Mètre, Sèvres, 8e édition edition, 2006. `http://www.bipm.org/utils/common/pdf/si_brochure_8.pdf`.

Maud Ehrmann. *Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*. PhD thesis, Université Paris 7 - Denis Diderot, 2008.

Ester2. *Annotation Entités Nommées, dates, heures et montants*. ESTER2, 2009.

Olivier Galibert, Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum, and Ludovic Quintard. Structured and extended named entity evaluation in automatic speech transcriptions. In *Proc IJCNLP*, Chiang Mai (Thailand), November 2011.

Cyril Grouin, Olivier Galibert, Sophie Rosset, Ludovic Quintard, and Pierre Zweigenbaum. Mesures d'évaluation pour entités nommées structurées. In *Évaluation des méthodes d'Extraction de Connaissances dans les Données (EvalECD 2011)*, Brest, January 2011a.

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard. Proposal for an extension of traditional named entities : From guidelines to evaluation, an overview. In *Proceedings of the Fifth Linguistic Annotation Workshop (LAW-V)*, pages 92–100, Portland, Oregon, 2011b. Association for Computational Linguistics.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticæ Investigationes*, 30(1), 2007.

Sophie Rosset, Cyril Grouin, Olivier Galibert, Pierre Zweigenbaum, Karën Fort, and Ludovic Quintard. Proposition pour une extension de la définition des entités nommées : de la définition à l'évaluation. In Jean-Yves Antoine, Nathalie Friburger, Denis Maurel, and Damien Nouvel, editors, *Journée Atala 2011 : Reconnaissance d'Entités Nommées Nouvelles Frontières & Nouvelles Approches*, Paris (France), June 2011. ATALA. `http://tln.li.univ-tours.fr/Tln_Colloques/Tln_REN2011/Rosset-ATALA-20juin2011.pdf`.

Satoshi Sekine. Definition, dictionaries and tagger of extended named entity hierarchy. In *LREC'04*, Lisbon, Portugal, 2004.

Timex. *Guidelines for Temporal Expression Annotation for English for TempEval 2010*. TimeML Working Group, 2009. URL `http://www.timeml.org/tempeval2/tempeval2-trial/guidelines/timex3guidelines-072009.pdf`.

Mickaël Tran. *Prolexbase. Un dictionnaire relationnel multilingue de noms propres : conception implantation et gestion en ligne*. PhD thesis, Université François Rabelais de Tours, 2006.

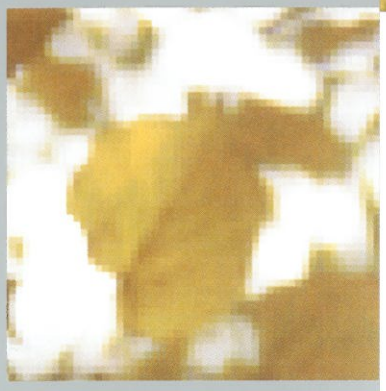
Notes et Documents LIMSI N^{o} : 2011-04
Septembre 2011

Auteurs (Authors) : S. Rosset, C. Grouin, P. Zweigenbaum

Titre : Entités nommées structurées : guide d'annotation Quaero

Title: Structured Named Entities : Quaero Annotation Guidelines

Nombre de pages (Number of pages) : 82



Résumé : *Ce document présente les principes et spécifications des entités nommées structurées définies dans le projet Quaero. Dans le projet, ces spécifications ont servi à annoter des corpus de presse audio et de presse ancienne, pour un total de trois millions de mots. Le travail de définition et de structuration qui a abouti à ce guide, ainsi que les mesures d'accord inter-annotateurs, sont décrits dans Grouin* et al. *(2011b). Une réflexion concernant les métriques pour l'évaluation de systèmes de détection d'entités nommées structurées, en lien avec des tâches finalisées, est proposée dans Grouin* et al. *(2011a). Les problématiques liées à l'évaluation sur des transcriptions automatiques de parole sont abordées dans Galibert* et al. *(2011). Enfin, un récapitulatif de l'ensemble du travail est présenté dans Rosset* et al. *(2011).*

Mots clés : Entités nommées, Types d'entités, Composants d'entités, entités nommées structurées, Projet Quaero, Annotation de corpus

Abstract : *This document presents the principles and specifications of the structured named entities defined in the Quaero project. In the project, these specifications were used to annotate broadcast news and old press corpora, for a total of three million words. The definition and structuring work which led to these guidelines, along with inter-annotator agreement measures, were described in Grouin* et al. *(2011b). A discussion about metrics for evaluating structured named entity detection systems, in relation to finalized tasks, is proposed in Grouin* et al. *(2011a). The issues linked to the evaluation of automatic speech transcripts are addressed in Galibert* et al. *(2011). Finally, a general overview of the overall work is presented in Rosset* et al. *(2011).*

Key words: Named Entities, Entity Types, Entity Components, Structured Named Entities, Quaero Project, Corpus Annotation

Tél. : + 33 (0) 1 69 85 80 80 - Fax : + 33 (0) 1 69 85 80 88 - Toile : http://www.limsi.fr