Concepts et modèle de données du logiciel *NotaBene* Spécification technique*

Nicolas Mazziotta

Version : `$Id: specification.tex 205 2010-04-09 14:09:46Z nmzi $`

Ce document définit les notions ainsi que la terminologie linguistique et technique mobilisés par le logiciel d'annotation *NotaBene* [1].

Le logiciel a été développé avec comme objectif de permettre l'annotation de textes anciens, la plupart du temps transmis par un *medium* concret. Les notions abordées ci-dessous devront être repensées s'il fallait les adapter aux ressources électroniques.

La première section traite la dimension linguistique et introduit les concepts qui lui sont spécifiques. La deuxième section est centrée sur la dimension informatique et présente la manière dont RDF, le modèle de données choisi, est employé pour exprimer les analyses linguistiques. La dernière section reprend l'ensemble des termes introduits dans le document et en fournit une définition.

1 Dimension linguistique

Il s'avère nécessaire de choisir et de définir précisément une terminologie qui rende possible la discussion, sans interférence de la part de modèles spécifiques. La terminologie suivante sera introduite : *analyse*, *catégorisation*, *constituant*, *délimitation*, *document*, *édition*, *groupement*, *liaison*, *mot*, *note*, *terme*, *texte*. Tous ces termes sont à entendre selon la définition donnée ci-dessous, à l'exclusion de toute autre acception, quel que soit le modèle linguistique dont il est issu.

Nous commencerons par présenter les unités de base de la description (1.1) avant de nous occuper de leur analyse (1.2).

1.1 Unités de base

Distinguons les unités maximales (1.1.1) de celles qui ne le sont pas (1.1.2).

1.1.1 Document/texte/édition

a. Texte et document. Il est commode, pour travailler sur les matériaux linguistiques anciens antérieurs au développement de l'informatique, de distinguer leur dimension

*Bien que centrée sur le logiciel *NotaBene*, cette spécification a été élaborée dans le cadre d'une concertation entre les membres du LASLA (L) et des projets *Khartês* (K) et *Ramsès* (R). L'objectif poursuivi est de rendre possible la mutualisation des ressources et des outils d'analyse linguistiques. Étaient présents : Stéphanie Gohy (R), Anne-Claude Honnay (R), Dominique Longrée (L), Nicolas Mazziotta (K), Stéphane Polis (R), Gérald Purnelle (L), Serge Rosmorduc (R), Laurent Simon (L) et Jean Winand (R).

1. Accessible à l'adresse https ://sourceforge.net/projects/notabene.

concrète de leur dimension *abstraite*. Cette distinction, parallèle à celle qui existe entre la matière et la substance chez Louis Hjelmslev (1968 : 74-77), permet principalement de départir tout ce qui est extralinguistique de ce qui est proprement linguistique. Nous nommons *texte* l'unité linguistique véhiculée par le *medium*. Le terme *document* est réservé à ce dernier.

b. Édition Nous savons cependant que le texte est une unité abstraite inaccessible. Le travail philologique permet de projeter cette unité dans un *medium* différent de celui qui l'a véhiculée et qui est adapté aux besoins des lecteurs qui veulent l'exploiter. Cette réalisation du texte est nommée *édition*. Dans le cadre de l'annotation électronique, l'édition consiste généralement en un fichier lisible par l'ordinateur ; voir sous 2.1 pour un exemple d'édition électronique.

1.1.2 Mot et constituant

a. Mot. Qu'on récuse ou qu'on admette la pertinence linguistique de la notion, la plupart des corpus informatisés fournissent des éditions qui livrent des unités assimilables à ce que la grammaire traditionnelle a identifié comme des *mots*. Nous ne nous risquerons pas à une définition précise du terme de manière générale. Dans le cadre défini ici-même, un mot est une unité linguistique signifiante dont les limites sont définies par un modèle d'analyse implicite ou explicite. Par exemple, il peut être décidé que le mot correspond à toute chaîne graphique séparée des autres chaînes par des « blancs » ou qu'il est assimilable au monème des théories fonctionnalistes. Le modèle employé peut également être intuitif : tout dépend du choix des annotateurs.

D'autre part, il n'est pas nécessaire que les mots correspondent à des unités intuitivement perçues comme telles. Suivant les intérêts qu'on leur porte, les signes de ponctuation peuvent être considérés comme des mots ou non.

Le mot est l'unité minimale pouvant être analysée. Il est possible qu'un mot ait une forme d'expression nulle, ce qui correspond au *signifiant zéro* présent dans les théories génératives ou fonctionnalistes.

b. Constituant. Les théories syntaxiques regroupent souvent les mots en groupes fonctionnant ensemble suivant le modèle de l'analyse en constituants immédiats. Ces groupes et les relations qu'ils entretiennent peuvent être analysés (voir le point 1.2.2, ci-dessous). Nous appelons *constituant* tout mot ou groupe de mot susceptible d'être analysé.

Ce terme n'est pas tributaire d'un quelconque modèle d'analyse. Aucune démarche n'est donc préconisée pour l'identification des constituants, leur découpage ou leur association. De cette manière, dans la phrase ci-dessous,

> Saichent tot cilh ki ces lettres verront que li cuens de Gelre a fait hommage le veske de Liege [...] (15 décembre 1236, Archives de l'État à Liège, Cathédrale Saint-Lambert)

rien n'empêche de grouper (voir 1.2.2 ci-dessous) *vesque* et *Liege* en un constituant

1.2 Analyses

Une *analyse* consiste en la mise en relation d'un constituant avec *quelque chose*. Il peut s'agir d'une notion définie dans le modèle d'analyse (1.2.1) ou d'un constituant. Ces relations entre constituants peuvent être de plusieurs types (1.2.2). L'analyse peut également rester vague et être formulée en langue naturelle (1.2.3).

1.2.1 Modèle et relations au modèle

Les constituants du texte peuvent être mis en relation avec un modèle formalisé.

a. Modèle et termes. De manière générale, l'analyse linguistique d'un texte se fait suivant un modèle prédéterminé ou construit de manière heuristique. Ce modèle peut être flou et implicite ou formalisé et explicite. Nous ne parlons ici que de ce dernier cas. Dans ce type de modèle, les notions et les relations entre les notions sont définies et identifiées par un nom spécifique et idéalement univoque, appelé *terme*. Par exemple, en grammaire générative, le terme P est défini par sa relation au schème de réécriture classique $P \rightarrow SN + SV$ [2]. Un autre exemple est la définition analytique des *morphèmes* et des *lexèmes* chez André Martinet, pour qui ils sont une sorte de *monème* [3].

b. Catégorisation. L'opération qui consiste à mettre en relation un terme (défini dans le modèle) avec un constituant (attesté dans le texte) est une opération de *catégorisation*. Ainsi, le fait de considérer que dans la phrase en ancien français mentionnée sous 1.1.2, le mot *lettres* est un représentant de la classe des substantif met en relation le terme *substantif* (défini en dehors du texte) et le mot *lettres*, attesté.

1.2.2 Relations intratextuelles

Les constituants d'un texte peuvent être groupés et la relation qu'ils entretiennent peut être spécifiée sémantiquement.

a. Groupement. Les constituants peuvent entrer en relation les uns avec les autres, et l'opération d'analyse qui consiste à repérer une relation entre deux d'entre eux est nommée *groupement*. Les groupements peuvent être de deux types : soit ils permettent de délimiter les unités, soit de les lier dans une autre relation.

b. Délimitation et liaison. Le mot a été défini ci-dessus (1.1.2) comme l'unité minimale pouvant être soumise à l'analyse. L'opération de *délimitation* consiste à grouper des mots et des constituants pour construire des constituants plus larges. Par exemple, dans la phrase en ancien français citée ci-dessus, les mots *ces* et *lettres* peuvent être groupés et ce groupement délimite un constituant, catégorisé par exemple comme un *syntagme nominal* dans le cadre de la grammaire générative. Une délimitation est donc un groupement qui définit un nouveau constituant. Il est important de noter que cette relation n'est pas nécessairement binaire : plus de deux constituants peuvent être groupés.

S'il ne s'agit pas d'une délimitation, la relation qui existe entre les constituants peut être vue comme un *lien* (par exemple de nature syntaxique). L'opération d'analyse qui consiste à déterminer quelle est la nature du lien qui unit deux constituants est nommée *liaison*. Ainsi, dans l'exemple donné ci-dessus, la relation syntaxique de *complément déterminatif du nom* pourrait qualifier le lien qui existe entre le constituant *de Gelre* et le mot *veske*. Nous appelons *liaison* les groupements qui ne définissent pas un nouveau constituant.

1.2.3 Notes

Certaines analyses sont indépendantes de tout modèle formalisé. Ce peut être le cas de notes explicatives ou descriptives de la matérialité du document, qui sont générale-

2. Voir Ruwet 1967 : 113-120.
3. Voir XXXXX

ment étrangères à la dimension linguistique (le texte). Par exemple, noter que « l'encre est pâlie par endroits » est une *note*.

2 Dimension informatique

Les données suivent le modèle du *Resource Description Framework* (RDF), selon les conventions exposées dans Klyne et Carroll 2004. L'expression des données peut être faite dans les formats RDF/XML (Beckett 2004), Notation 3 (Berners-Lee 2000) ou N-Triples (Grant et Beckett 2004 : §3). Nous ne présenterons pas ces formats. Les exemples ci-dessous suivent la syntaxe de RDF/XML.

La terminologie suivante sera introduite : annotation, cible espace de nommage, graphe, littéral, nœud, nœud vide, origine, propriété, ressource, triplet, typage, Uniform Resource Identifier.

2.1 Ressources

Nous nommons *ressources* les données auxquelles il est possible d'avoir accès par voie informatique.

a. Identification. La plupart des éléments introduits dans la section 1.1 ci-dessus peuvent être identifié à l'aide d'un *Uniform Resource Identifier* (Berners-Lee *et al.* 2005), dont la forme la plus courante est une « adresse internet » (URL). Le mécanisme d'identification `xml:id` (Marsh *et al.* 2005) est une manière courante d'associer un URI à un mot dans une édition. Par exemple, dans le fragment d'édition électronique au format XML suivant,

```
<w xml:id="w1">Saichent</w>
<w xml:id="w2">tot</w>
<w xml:id="w3">cilh</w>
<w xml:id="w4">ki</w>
<w xml:id="w5">ces</w>
<w xml:id="w6">lettres</w>
<w xml:id="w7">verront</w>
<w xml:id="w8">que</w>
<w xml:id="w9">li</w>
<w xml:id="w10">cuens</w>
<w xml:id="w11">de</w>
<w xml:id="w12">Gelre</w>
<w xml:id="w13">a</w>
<w xml:id="w14">fait</w>
<w xml:id="w15">hommage</w>
<w xml:id="w16">le</w>
<w xml:id="w17">veske</w>
<w xml:id="w18">de</w>
<w xml:id="w19">Liege</w>
[...]
```

la valeur de l'attribut `xml:id` identifie et distingue de manière univoque les mots les uns par rapport aux autres.

Nous appelons *identification* l'opération d'édition qui consiste à associer un mot, un constituant, une analyse, un terme, etc. à un URI. D'un point de vue philologique, cette opération est une opération d'*édition* (voir 1.1.1) et non d'*analyse* (voir 1.2), parce qu'elle vise à rendre le texte accessible par les machines.

b. Espace de nommage. Un espace de nommage (Bray *et al.* 1999) est un ensemble de noms identifiés par un URI. Idéalement, cet URI donne accès aux ressources correspondant à ces noms. En imaginant que l'édition ci-dessus soit accessible à l'adresse

http ://khartes.org/textes/texte1

l'URI http ://khartes.org/textes/texte1# identifie un espace de nommage qui permet de regrouper tous les individus identifiés.

c. Nœuds vides. Une fois transposées dans le modèle de données présenté ci-dessous, l'analyse des ressources peut générer des unités opératoires. Ces unités sont appelées *nœuds vides* (angl. *blank nodes*, Klyne et Carroll 2004 : §6.6). Nous verrons ci-dessous (sous 2.2.3) des exemples de ces nœuds.

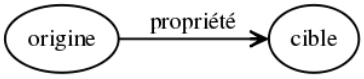
d. Littéraux. Certaines ressources ne sont pas identifiées ou identifiables par un URI, mais sont toujours mentionnées telles quelles. Concrètement, elles prennent la forme de chaînes, d'entiers, de booléens, etc. Tous ces types de données sont définis par la recommandation *XML-Schema Part 2 : Datatypes* (Biron et Malhotra 2004). Actuellement, seuls les littéraux sont utilisés ici. Ils ont la forme d'une chaîne unicode (encodage UTF-8) [4] associée à un code de langue et de région (ISO 639-2 1998).

2.2 Transposition des analyses linguistiques

Nous appelons *annotation* toute expression d'une analyse (telle que définie sous 1.2) suivant le modèle de donnée défini en 2.2.1 ci-dessous. La transposition des modèles d'analyse linguistique est abordée sous 2.2.2 et celle des principaux types d'analyse est décrite en 2.2.3.

2.2.1 Modèle de données

a. Sémantique générale. Le modèle RDF utilise le triplet comme unité fondamentale dans Klyne et Carroll 2004 : §3 Les triplets associent trois ressources identifiées dans une relation orientée qui va d'une *origine* à une *cible* (toutes deux une sorte de *nœud*). Cette relation porte le nom de *propriété* [5]. Schématiquement :

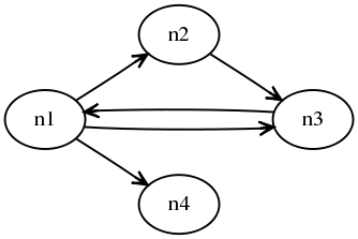


Chacun des trois éléments constitutifs d'un triplet peut prendre la forme d'une URI. Les propriétés ne peuvent être que des URI, mais les nœuds peuvent être des nœuds vides ou des littéraux. Les triplets définissent ainsi des relations binaires sémantiquement qualifiées.

Il n'y a pas de contrainte quant au nombre de propriétés qui peuvent être rattachées à une origine ou une cible, mais une propriété n'a qu'une et une seule origine et une et une seule cible. Pratiquement, un ensemble de triplets peut être considéré simultanément et permet de représenter un *graphe* ; par exemple, soit les nœuds *n1*, *n2*, *n3* et *n4*, unis par des propritétés non nommées ici par économie :

4. Voir http ://www.unicode.org.

5. *NotaBene* étant un logiciel destiné prioritairement à l'analyse linguistique, pour éviter toute ambiguïté, nous n'utilisons pas la terminologie standard et ne parlons pas de *sujet*, de *prédicat* ou d'*objet* à propos des éléments constituant un triplet.



b. Spécificité du modèle par rapport au modèle arborescent. Il est important de noter que les graphes se distinguent des arbres entre autres par les caractéristiques suivantes :

- la sémantique des propriétés liant les différents nœuds d'un graphe n'est pas prédéterminée, contrairement aux liens entre les nœuds d'un arbre, qui impliquent nécessairement un rapport d'*inclusion* ou de *précédence* ;
- de ce fait, il est possible de représenter un rapport d'inclusion ou de précédence dans un graphe, mais les propriétés impliquées doivent être sémantiquement déterminées de façon explicite ;
- de ce fait également, les contraintes portant sur la structure des arbres (un nœud ne peut avoir qu'un nœud-parent, ce dernier doit être distinct et les enfants d'un nœud sont nécessairement ordonnés) ne sont pas applicables aux graphes à moins d'être explicitement ajoutées.

En d'autres termes, le modèle de données du graphe est plus expressif, mais plus verbeux que le modèle arborescent.

c. Contrainte générale fondamentale. Si un seul constituant est présent dans un triplet, c'est l'origine qui pointe vers lui.

2.2.2 Systèmes terminologiques

Pour plus de facilité, les systèmes terminologiques (définitions des termes, voir 1.2.1 *supra*) sont exprimés dans le même modèle de données que les annotations, suivant une syntaxe spécifique nommée *OWL Web Ontology Language* (Bechhofer *et al.* 2004). Nous ne couvrirons pas entièrement ce sujet ici.

OWL permet de définir des *classes* (concepts) et des *propriétés* (relations entre concepts) ainsi que de leur attribuer un URI. Ces URI peuvent être employés dans les annotations. Par exemple, soit l'édition mentionnée sous 2.1, l'annotation suivante fait référence à une ontologie qui serait disponible à l'adresse http ://khartes.org/termes :

```
<?xml version="1.0"?>
 <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <rdf:Description rdf:about="http://khartes.org/textes/texte1#w6">
  <rdf:type rdf:resource="http://khartes.org/termes#Substantif">
 </rdf:RDF>
</rdf:Description>
```

les URI mentionnés comme valeurs des attributs `rdf:resource` correspondent : 1/ à l'adresse du mot annoté pour l'élément le plus englobant ; 2/ à l'adresse où se trouvent défini le terme employé pour l'élément englobé. La propriété `rdf:type` est définie au niveau de la spécification RDF.

2.2.3 Types d'annotations

Nous avons présenté ci-dessus deux grandes classes d'analyses : les catégorisations, qui mettent en relation le texte et le modèle (voir 1.2.1) et les groupements, qui mettent

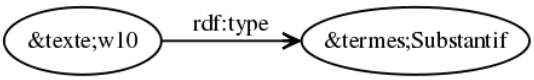
en relation les constituants du texte. Les groupements peuvent être des délimitations ou des liaisons (voir 1.2.2).

Ces trois types d'analyse sont transposés dans le modèle de données comme suit.

a. Catégorisation. L'exemple d'annotation donné sous 2.2.2 est un exemple de transposition de catégorisation. L'origine est un constituant (identifié par son URI), la propriété est définie par `rdf:type` (qui représente un URI) et la cible est un élément de terminologie (identifié également par son URI). L'URI peut être abrégé en utilisant le système des entités XML (Bray *et al.* 2006 : §4.2) pour réduire les espaces de nommage. Par exemple :

– http ://khartes.org/textes/texte1# peut être abrégé en `&text;` ;
– http ://khartes.org/termes# peut être abrégé en `&termes;`.

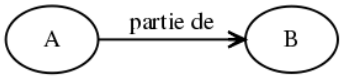
Schématiquement, le triplet est le suivant :



Bien entendu, d'autres propriétés peuvent être utilisées à la place de la propriété `rdf:type`.

b. Délimitation. À l'exception des mots (qui sont « donnés » par l'édition), les constituants sont toujours définis par rassemblement de constituants plus petits (ou *groupements*), jamais par subdivision.

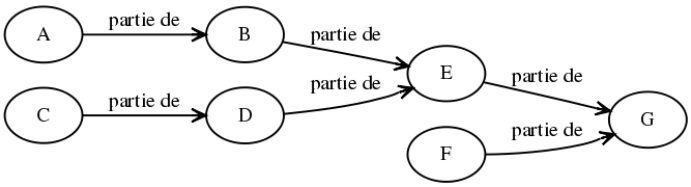
La délimitation de nouveaux constituants se fait en créant un nœud vide pour représenter le constituant plus englobant. Les constituants existant préalablement sont mis en relation de la manière schématisée ci-dessous :



où le plus petit constituant (A) est dans la position de l'origine, alors que le plus large (B) est dans la position de la cible. Cette contrainte est *exclusivement technique* et ne doit pas influer sur la démarche d'identification des constituants. B est nécessairement un *nœud vide*, qui peut éventuellement être catégorisé.

La propriété qui unit A et B est arbitrairement nommée « partie de » dans le cadre de cet exemple. Elle a une sémantique définie dans le modèle d'annotation comme fondatrice des constituants. Il est essentiel que la définition de ce type de propriété se fasse au niveau spécifique du modèle d'analyse et non au niveau de la présente spécification. En effet, il doit rester possible de définir des délimitations concurrentes et éventuellement incompatibles, distinguées par l'URI de la propriété mobilisée.

La délimitation des constituant est le plus souvent récursive (les noms des nœuds vides ci-dessous sont arbitraires) :



une traduction possible [6] de ce graphe en RDF/XML est la suivante (en admettant que les termes ont l'espace de nommage http ://khartes.org/termes#) :

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:termes="http://khartes.org/termes#">
 <rdf:Description rdf:nodeID="A">
  <termes:partie_de>
   <rdf:Description rdf:nodeID="B">
    <termes:partie_de>
     <rdf:Description rdf:nodeID="E">
      <termes:partie_de>
       <rdf:Description rdf:nodeID="G"/>
      </termes:partie_de>
     </rdf:Description>
    </termes:partie_de>
   </rdf:Description>
  </termes:partie_de>
 </rdf:Description>
 <rdf:Description rdf:nodeID="C">
  <termes:partie_de>
   <rdf:Description rdf:nodeID="D">
    <termes:partie_de rdf:resource="E"/>
   </rdf:Description>
  </termes:partie_de>
 </rdf:Description>
 <rdf:Description rdf:nodeID="F">
  <termes:partie_de rdf:resource="G"/>
 </rdf:Description>
</rdf:RDF>
```

c. Liaison. L'expression des liaisons (groupements qui ne délimitent pas un nouveau constituant) place également un constituant en position d'origine et un autre en position de cible, mais se distingue de celle des délimitations par deux caractéristiques :

- la cible n'est pas nécessairement un nœud vide ;
- la sémantique de la propriété reliant les deux nœuds est nécessairement disjointe d'une sémantique de délimitation.

3 Définitions

[Section en cours de rédaction]

● ***analyse*** (*ling.*) « toute information issue de l'étude d'un ou plusieurs constituants* et associée explicitement à ceux-ci. »

```
Analyse ::= Catégorisation || Groupement || Note
```

● ***annotation*** (*ling.*) « expression d'une analyse* sous la forme d'un graphe*. »

● ***catégorisation*** (*ling.*) « analyse* consistant à définir une relation entre, d'une part, un constituant* ou une liaison* et, d'autre part, un terme*. »

● ***cible*** (*inf.*) « troisième élément d'un triplet*. »

● ***constituant*** (*ling.*) « tout mot*, groupement* de mots, texte* ou subdivision du texte qui mérite d'être étudié. »

● ***délimitation*** (*ling.*) « groupement consistant à définir un nouveau constituant par regroupement d'autres constituants. Cette opération distingue le nouveau constituant défini de tous les autres constituants du texte. »

6. Le langage étant déclaratif et séquentiel, les triplets peuvent être déclarés dans n'importe quel ordre. Cette notation a été choisie parce qu'elle est la plus économique.

● ***document*** (*ling.*) « médium ayant permis la transmission d'un message sous une forme linguistique.

Remarque : en pratique, le document peut prendre la forme d'un papyrus, d'un parchemin, de tablettes gravées, etc. Le document est fondamentalement concret. »
● ***édition*** (*ling.*) « représentation du texte en permettant l'analyse*. »
● ***espace de nommage*** (*ling.*) « ensemble de noms identifiés par »
● ***graphe*** (*inf.*) « ensemble de triplets*. »

```
Graphe ::= Triplet+
```

● ***groupement*** (*ling.*) « analyse consistant à définir une relation entre deux constituants. »

```
Groupement ::= Délimitation || Liaison
```

● ***identification*** (*inf.*) « opération d'édition* associant un URI* à l'expression informatique d'un constituant*. »
● ***liaison*** (*ling.*) « groupement* ne définissant pas de nouveau constituant*. »
● ***mot*** (*ling.*) « unité linguistique minimale associant une forme d'expression à une forme de contenu. »
● ***note*** (*ling.*) « analyse* non formalisée associant un constituant à une glose de forme libre. »
● ***nœud*** (*inf.*) « désignation générique des origine* et cible* d'un triplet*. »
● ***nœud vide*** (*inf.*) « nœud non identifié par un URI, mais par un identifiant interne au graphe. »
● ***origine*** (*inf.*) « premier élément d'un triplet*. »
● ***propriété*** (*inf.*) « deuxième élément d'un triplet*. »
● ***ressource*** (*inf.*) « donnée accessible par voie informatique »
● ***terme*** (*ling.*) « désignation d'une notion linguistique relevant d'un modèle spécifique. »
● ***texte*** (*ling.*) « message linguistique transmis par un document. Un texte est fondamentalement abstrait. »
● ***triplet*** (*inf.*) « ensemble de trois éléments ordonnés représentant une relation orientée. »

```
Triplet ::= Origine Propriété Cible
Origine ::= URI || Noeud vide
Propriété ::= URI
Cible ::= URI || Noeud vide
```

● ***Uniform Resource Identifier (URI)*** (*inf.*) « référence de forme conventionnelle permettant d'identifier une ressource*. »

Références

Bechhofer, Sean, Van Harmelen, Frank, Hendler, Jim, Horrocks, Ian, McGuinness, Deborah L., Patel-Schneider, Peter F. et Stein, Lynn Andrea, éds (2004). *OWL Web Ontology Language Reference. Reference. W3C Recommendation 10 February 2004*, http ://www.w3.org/TR/2004/REC-owl-ref-20040210.

Beckett, Dave, éd. (2004). *RDF/XML Syntax Specification (Revised). W3C Recommendation 10 February 2004*, http ://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/.

Berners-Lee, Tim (2000). « Primer : Getting into RDF & Semantic Web using N3 », http ://www.w3.org/2000/10/swap/Primer.html.

Berners-Lee, Tim, Fielding, R. et Masinter, L. (2005). *Uniform Resource Identifier (URI) : Generic Syntax*, http ://www.ietf.org/rfc/rfc3986.txt.

Biron, Paul V. et Malhotra, Ashok, éds (2004). *XML Schema Part 2 : Datatypes Second Edition. W3C Recommendation 28 October 2004*, http ://www.w3.org/TR/2004/REC-xmlschema-2-20041028.

Bray, Tim, Hollander, Dave et Layman, Andrew, éds (1999). *Namespaces in XML*, http ://www.w3.org/TR/1999/REC-xml-names-19990114.

Bray, Tim, Paoli, Jean, Sperberg-McQueen, C. M., Maler, Eva, Yergeau, François et Cowan, John, éds (2006). *Extensible Markup Language (XML) 1.1 (Second Edition). W3C Recommendation 16 August 2006, edited in place 29 September 2006*, http ://www.w3.org/TR/2006/REC-xml11-20060816.

Grant, Jan et Beckett, Dave, éds (2004). *RDF Test Cases. W3C Recommendation 10 February 2004*, http ://www.w3.org/TR/2004/REC-rdf-testcases-20040210.

Hjelmslev, Louis (1968). *Prolégomènes à une théorie du langage*, Paris : Minuit (Arguments 35), traduit en français par Una Canger (original danois, 1943).

ISO 639-2 = *ISO 639-2 Registration Authority*, http ://www.loc.gov/standards/iso639-2/.

Klyne, Graham et Carroll, Jeremy J., éds (2004). *Resource Description Framework (RDF) : Concepts and Abstract Syntax W3C Recommendation 10 February 2004*, http ://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.

Marsh, Jonathan, Viellard, Daniel et Walsh, Norman, éds (2005). *xml :id Version 1.0. W3C Recommendation 9 September 2005*, http ://www.w3.org/TR/2005/REC-xml-id-20050909/.

Ruwet, Nicolas (1967). *Introduction à la grammaire générative*, Paris : Plon (Recherches en sciences humaines 22).