

---

## 11 – Noms, dates, chiffres et abréviations

---

La TEI définit des éléments pour un grand nombre de types de données spéciales que l'on peut rencontrer presque partout dans des textes de toutes sortes. Ces types de données peuvent être d'un intérêt particulier dans tout un éventail de disciplines. Ils se réfèrent tous à des objets externes au texte lui-même (noms de personnes et de lieux, chiffres, dates). Ils posent toutefois des problèmes particuliers à beaucoup d'applications de traitement du langage naturel (NLP), à cause des formes variées sous lesquelles ils peuvent apparaître dans les textes. Les éléments décrits dans le présent chapitre, en rendant ces types de données explicites, facilitent le traitement des textes qui les contiennent.

### 11.1. Noms et chaînes de caractère de référence

Une « chaîne de référence » (*referring string*) est une expression qui se réfère à une personne, un endroit, un objet donné, etc. Deux éléments sont fournis pour marquer de telles chaînes :

`<rs>` contient une chaîne de référence ou un nom général ; parmi les attributs possibles, citons :

*type* indique plus spécifiquement l'objet auquel la chaîne se réfère. Des exemples de valeurs sont : *personne*, *endroit*, *navire*, *élément*, etc.

`<name>` contient un nom propre ou une proposition substantive ; parmi les attributs possibles, citons :

*type* indique le type d'objet qui est nommé par l'expression.

L'attribut *type* est employé pour distinguer (par exemple) entre des noms de personnes, d'endroits ou d'organisations, dans les cas où cela est possible :

```
<q>My dear <rs type=person>Mr. Bennet</rs>, </q>
said his lady to him one day, <q>have you heard
```

that <rs type=place>Netherfield Park</rs> is let  
at last?</q>

It being one of the principles of the  
<rs type=organization>Circumlocution Office</rs> never,  
on any account whatsoever, to give a straightforward answer,  
<rs type=person>Mr Barnacle</rs> said, <q>Possibly.</q>

Comme le montrent les exemples suivants, l'élément <rs> peut être employé pour toute référence à une personne, un endroit, etc., qui n'est pas forcément un nom propre ou une proposition substantive :

<q>My dear <rs type=person>Mr. Bennet</rs>,</q>  
said <rs type=person>his lady</rs> to him  
one day...

<q>Peu après son installation <rs type=lieu>rue Simon-  
Crubelier</rs>, <rs type=personne>Maurice Réol</rs>, qui  
était <rs type=metier>aide-rédacteur</rs> à la  
<rs type=organisation>CAMPA</rs> ...

L'élément <name>, au contraire, est prévu pour le cas spécial des chaînes de référence contenant uniquement des noms propres ; il peut être utilisé de la même façon que l'élément <rs>, ou imbriqué au sein de celui-ci si une chaîne de référence contient un mélange de noms communs et de noms propres.

Le simple fait de baliser un objet en tant que «nom» ne suffit généralement pas pour permettre le traitement automatique des noms de personnes afin d'obtenir les formes canoniques généralement requises à des fins de référence. Un nom tel qu'il apparaît dans le texte peut être orthographié de façon incohérente, ou être partiel ou flou. Qui plus est, des particules de noms tel que *van* ou *de la* peuvent ou non être incluses dans la forme de référence d'un nom. Ceci dépend de la langue et du pays de celui qui porte le nom en question.

Les attributs suivants sont également disponibles pour ces éléments et pour des éléments similaires, afin de surmonter ces difficultés :

*key* fournit un autre identifiant pour l'objet nommé, telle qu'une clé d'un enregistrement de base de données ;

*reg* donne une forme régularisée ou normalisée du nom utilisé.

L'attribut *key* peut être un moyen utile pour rassembler toutes les références se rapportant à la même personne ou au même emplacement éparpillés à travers un document :

```
<q>My dear <rs type=person key=BENM1>Mr. Bennet</rs>,
  </q> said <rs type=person key=BENM2>his lady</rs>
  to him one day, <q>have you heard that
  <rs type=place key=NETP1>Netherfield Park</rs>
  is let at last?</q>
```

Cette utilisation devrait être distinguée du cas de l'attribut *reg* (régularisation), qui permet de marquer la forme standard d'une chaîne de référence, comme ci-dessous :

```
<name type=person key=WADLM1 reg='de la Mare, Walter'>
  Walter de la Mare
</name>
was born at
<name key=Ch1 type=place>Charlton</name>, in
<name key=KT1 type=county>Kent</name>, in 1873.
```

On peut aussi baliser de façon plus détaillée les composants de noms propres, en utilisant le jeu de balises supplémentaires traitant les noms et les dates.

## 11.2. Dates et heures

Les balises suivantes permettent un codage plus détaillé des dates et de l'heure :

*<date>* contient une date dans n'importe quel format ; parmi les attributs possibles, citons :

*calendar* indique le système ou le calendrier auquel la date se rattache ;

*value* donne la valeur de la date sous une forme standard, habituellement aaaa-mm-jj ;

*<time>* contient une expression définissant une heure du jour dans n'importe quel format ; parmi les attributs possibles, citons :

*value* donne la valeur de l'heure sous une forme standard.

L'attribut *value* indique une forme normalisée pour la date ou l'heure, au moyen d'un format reconnu tel que celui qui est prescrit par la norme ISO 8601. Les dates ou les heures partielles (par exemple « 1990 », « septembre 1990 », « autour de midi ») peuvent habituellement être exprimées en omettant simplement une partie de la valeur donnée; ou bien, les dates ou les heures imprécises (par exemple « début août », « entre dix et douze heures ») peuvent être exprimées comme une plage de dates ou d'heures. Si l'une ou l'autre extrémité de la plage d'heure ou de date est connue avec certitude (par exemple, « avant 1230 », « quelques jours après Hallowe'en »), l'attribut *exact* peut être employé pour le préciser.

Exemples :

```
<date value='1980-02-21'>21 Feb 1980</date>
<date value='1990'>1990</date>
<date value='1990-09'>September 1990</date>
```

```
Given on the <date value='1977-06-12'>Twelfth Day of June
in the Year of Our Lord One Thousand Nine Hundred and
Seventy-seven of the Republic the Two Hundredth and first
and of the University the Eighty-Sixth.</date>
```

```
<l>pecially when it's nine below zero
<l>and <time value='15:00'>three o'clock in the afternoon</time>
```

```
<p>C'était une belle matinée de la <date value='1323-11'>fin
novembre</date> ...
```

### 11.3. Nombres

Les nombres peuvent être écrits en lettres ou en chiffres (vingt et un, XXI et 21) et leur présentation dépend de la langue (par exemple *5th* en anglais devient 5. en grec; *123,456.78* en anglais équivaut à 123.456,78 en français<sup>1</sup>). Dans des applications de traitement du langage naturel ou de traduction automatique, il est souvent utile de les différencier par rapport à d'autres parties

1. Toutefois les codes typographiques français recommandent aujourd'hui d'écrire plutôt 123 456,78 {NdT}.

plus « lexicales » du texte. Dans d'autres applications, la capacité d'enregistrer une valeur numérique en utilisant une notation standard est importante. L'élément `<num>` fournit cette possibilité :

`<num>` contient un chiffre, écrit dans n'importe quel format ; attributs possibles :

<i>type</i>	indique le type de valeur numérique ; les valeurs suggérées comprennent : <code>fraction</code> , <code>ordinal</code> (pour des chiffres ordinaux, par exemple « vingt et unième », pourcentage, et <code>cardinal</code> (un nombre absolu, par exemple « 21 », « 21,5 », etc.) ;
<i>value</i>	fournit la valeur du nombre dans un format dépendant de l'application.

Par exemple :

```
<num value='33'>xxxiii</num>
<num type=cardinal value='21'>twenty-one</num>
<num type=percentage value='10'>ten percent</num>
<num type=percentage value='10'>10%</num>
<num type=ordinal value='5'>5th</num>
```

#### 11.4. Les abréviations et leur développement

De même que les noms, les dates et les nombres les abréviations peuvent être transcrites telles quelles ou sous une forme développée ; elles peuvent être soit non-balisées, soit codées au moyen de l'élément suivant :

`<abbr>` contient une abréviation de tout genre ; parmi les attributs possibles, citons :

<i>expan</i>	donne le développement de l'abréviation ;
<i>type</i>	permet au codeur de classer l'abréviation selon une typologie adéquate ; exemples : <code>contraction</code> , <code>suspension</code> , <code>brevigraph</code> , <code>superscription</code> ou <code>acronym</code> ; l'attribut <i>type</i> peut aussi recevoir une valeur telle que <code>titre</code> (pour des titres d'adresse), <code>géographique</code> , <code>organisation</code> , etc., décrivant la nature de l'objet auquel on se réfère.

L'élément `<abbr>` est utile pour distinguer les éléments semi-lexicaux tels que des acronymes ou des termes de jargon :

```
We can sum up the above discussion as follows: the identity
of a <abbr>CC</abbr> is defined by that calibration of values
which motivates the elements of its <abbr>GSP</abbr>;
```

```
Every manufacturer of <abbr>3GL</abbr> or <abbr>4GL</abbr>
languages is currently nailing on <abbr>OOP</abbr> extensions
```

```
Le brevet <abbr>BBR</abbr> de justification des lignes a été
repris par la <abbr>CII</abbr> ...
```

L'attribut *type* peut être employé pour distinguer des types d'abréviation selon leur fonction, et l'attribut *expan* peut être employé pour fournir un développement :

```
Le <name><abbr type=titre expan='Docteur'>Dr.</abbr>
<abbr type=initiale expan='Jean'>J.</abbr>
Transen</name>
est le <abbr type=fonction expan='Président directeur
général'>Pdg</abbr> de la <abbr type=acronyme expan='Compagnie
des Compteurs Informatisés'>C2I</abbr>.
```

Cet élément est aussi particulièrement utile lorsqu'on transcrit des documents manuscrits contenant de fréquentes abréviations.

### 11.5. Adresses

L'élément `<address>` est employé pour noter une adresse postale de n'importe quelle sorte. Il contient un ou plusieurs éléments `<addrLine>`, un pour chaque ligne de l'adresse.

`<address>`

contient une adresse postale ou d'un autre type, par exemple l'adresse d'un éditeur, d'une organisation ou d'une personne ;

`<addrLine>`

contient une ligne d'une adresse postale ou autre.

Voici un exemple simple :

---

```
<address>
<addrLine>Rédaction des Cahiers GUTenberg</addrLine>
<addrLine>Irisa/Inria, campus de Beaulieu</addrLine>
<addrLine>F-35042 Rennes cedex</addrLine>
<addrLine>France</addrLine>
</address>
```

Les parties individuelles d'une adresse peuvent être subdivisées au moyen de l'élément `<name>` expliqué ci-dessus en section 11.1 «Noms et chaînes de référence».

```
<address>
<addrLine>Rédaction des Cahiers GUTenberg</addrLine>
<addrLine>Irisa/Inria, campus de Beaulieu</addrLine>
<addrLine><name type=code>F-35042</name>
      <name type=city>Rennes cedex</name>
<addrLine><name type=country>France</name></addrLine>
</address>
```