INU3011 Documents structurés

Cours 3

XML et caractères spéciaux

Plan

- Quelques points logistiques
- Préparation au Quiz 1
  - Exercices en classe
- Entités prédéfinies et entités caractères
  - Retour sur la lecture *Les entités en XML*
- oXygen et caractères spéciaux
- Émojis

Logistique

- Formation des équipes : c'est le temps d'y penser !
  - Me faire part bientôt par courriel de la personne avec qui vous travaillerez, ou…
  - Si vous souhaitez travailler en solo
- Rappels :
  - Documents malformés dans Firefox
  - Ouvrir avec… *vs* Envoyer vers… aux labos

Mini-quiz 1 (1/2)

- À faire entre demain et jeudi (incl.)
  - Durée 55 minutes
  - 16 questions
    - Entre 1 et 5 points
    - Quatre questions valent 5 points
    - Total : 50 points
  - Individuel (engagement sur l'honneur)
  - Pondération : 10%

Mini-quiz 1 (2/2)

- Portée : Tout ce qui a été vu jusqu'à aujourd'hui avant le début du cours
  - ***Incluant*** les exercices auto-évalués en ligne et sous forme de quiz StudiUM donnés la semaine dernière
- N'inclut pas le cours d'aujourd'hui

Exercices

Arbres inversés et unités de contenu

Exemple 1

<MEMO><CONFIDENTIEL/><AUTEUR>Julien</AUTEUR><DESTINATAIRES>
<NOM>Viateur</NOM></DESTINATAIRES><SUJET>
Vacances</SUJET><CORPS><PAR
TYPE="normal">Ça change pas le monde, <EM>sauf que</EM>... On peut écrire directement la plupart des caractères, comme ' " : ? > ! /, mais pas & ni <.
</PAR></CORPS></MEMO>

Dossier d'exemples 008-arbres
Fichier ex-arbre-1.xml

Même document, après indentation automatique dans oXygen

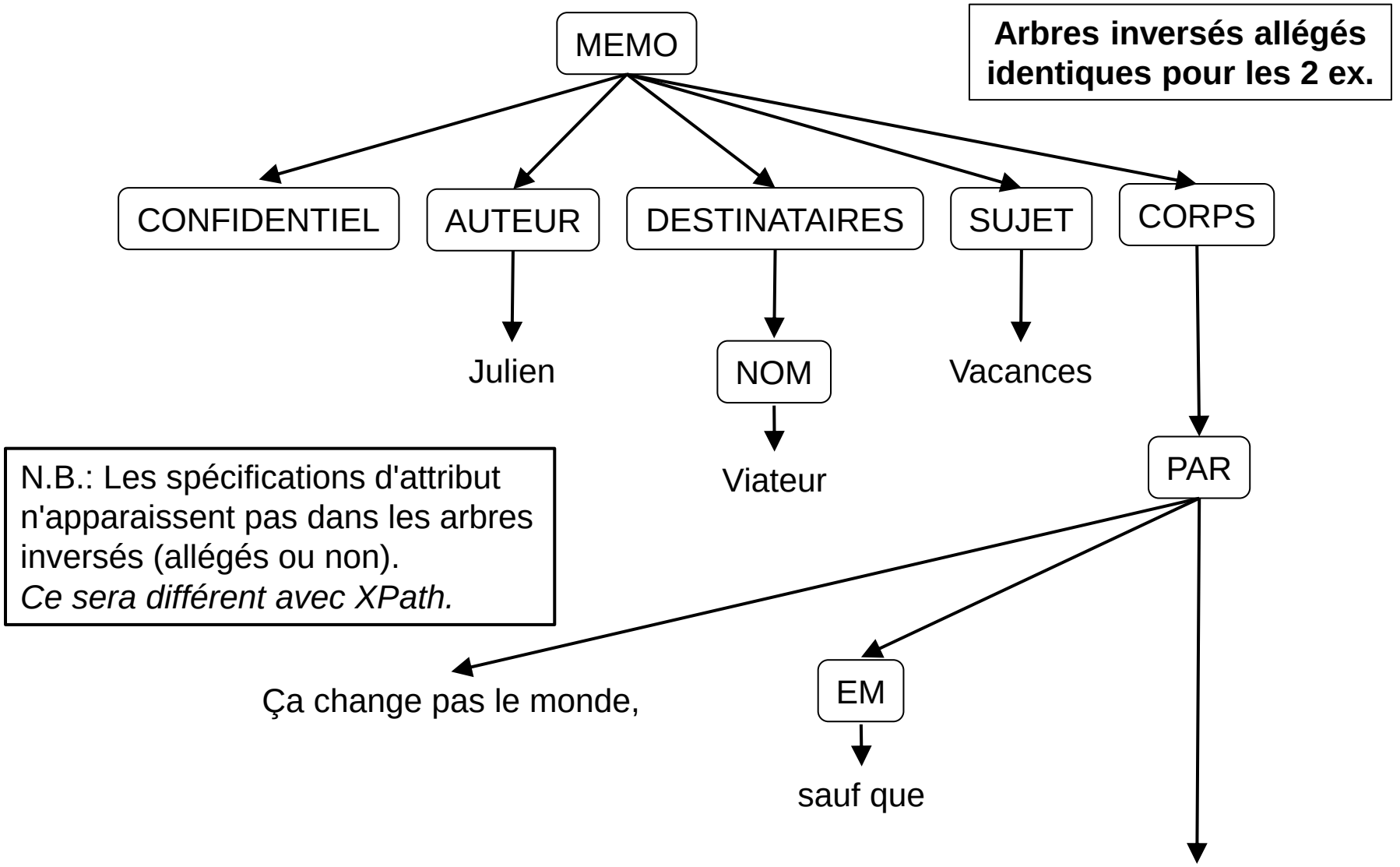
```
<MEMO>
  <CONFIDENTIEL/>
  <AUTEUR>Julien</AUTEUR>
  <DESTINATAIRES>
    <NOM>Viateur</NOM>
  </DESTINATAIRES>
  <SUJET> Vacances</SUJET>
  <CORPS>
    <PAR TYPE="normal">Ça change pas le monde, <EM>sauf que</EM>... On peut écrire
      directement la plupart des caractères, comme ' " : ? > ! /, mais pas &amp;amp; ni &amp;lt;.
    </PAR>
  </CORPS>
</MEMO>
```

Dossier d'exemples 008-arbres
Fichier ex-arbre-2.xml

Note

- Dans un navigateur Web, qui *interprète* le document XML, les deux documents précédents produisent *exactement le même affichage !*

```
−<MEMO>
   <CONFIDENTIEL/>
   <AUTEUR>Julien</AUTEUR>
  −<DESTINATAIRES>
      <NOM>Viateur</NOM>
   </DESTINATAIRES>
   <SUJET> Vacances</SUJET>
  −<CORPS>
    −<PAR TYPE="normal">
        Ça change pas le monde,
        <EM>sauf que</EM>
        ... On peut écrire directement la plupart des caractères, comme ' " : ? > ! /, mais pas & ni <.
      </PAR>
   </CORPS>
 </MEMO>
```

… On peut écrire directement la plupart des caractères, comme ' " : ? > ! /, mais pas & ni <.

Exercices (1/3)

Répondre pour `ex-arbre-1.xml` puis pour `ex-arbre-2.xml`

- Identifier contenu textuel direct de MEMO
- Nombre d'unités de contenu dans MEMO
- MEMO est conteneur/mixte/textuel/vide?
- Identifier contenu textuel direct de CORPS
- Nombre d'unités de contenu dans CORPS
- CORPS est conteneur/mixte/textuel/vide?

Exercices (2/3)

Répondre pour `ex-arbre-1.xml` puis pour `ex-arbre-2.xml`

- Identifier contenu textuel direct de PAR
- Identifier contenu textuel de PAR
- Nombre d'unités de contenu dans PAR
- PAR est conteneur/mixte/textuel/vide?
- Identifier tous les éléments-conteneurs
- Les arbres allégé et non allégé ont-ils le même nombre de nœuds ?

Exercices (3/3)

Répondre pour l'arbre inversé allégé (qui est le même pour les deux exemples)

- Identifier les frères aîné et cadet de EM
- Identifier les ancêtres de EM qui ne sont pas son parent
- Nombre de descendants de PAR
- Nombre de frères de AUTEUR

Entités prédéfinies et entités caractère

Problème un (1/2)

- En XML, "<" indique **toujours** le début d'une balise
- Ce caractère ne peut donc pas être inscrit directement dans le contenu textuel d'un élément **ni** dans une valeur d'attribut
- Que faire si on *veut* l'inscrire ?

Problème un (2/2)

- Par convention, un *appel d'entité prédéfinie*, de forme &*nom*; permet de représenter **comme texte** un caractère, même s'il est normalement "défendu"

  Exemple: **`<`** ↔ **`<`**

- Du coup, l'& est elle aussi un caractère "défendu" dans du contenu textuel ou dans une valeur d'attribut, puisque réservée aux appels d'entité

Entités prédéfinies (1/2)

- En fait, il y a cinq entités prédéfinies :

Appel	Nom de l'entité	Caractère représenté
&lt;	lower than	<
&gt;	greater than	>
&amp;	ampersand (en français, esperluette)	&
&apos;	apostrophe	'
&quot;	quotation mark	"

Ex.: B&G ↔ B&G

Entités prédéfinies (2/2)

- N.B.: Les deux seules entités prédéfinies vraiment essentielles sont & et <
- En effet:

  ' " > *peuvent* être inscrits *directement* comme contenu textuel et dans une valeur d'attribut
- Comme à peu près tous les autres caractères, incluant: / ! : - _ $ % \ ? [ ]…
- ... et les lettres accentuées (é, ç, È, etc.)

L’esperluette (&)

- L’esperluette (&) ne doit survenir *que dans un appel d’entité*
  - L’utiliser dans n’importe quel autre contexte constitue une erreur de bien-formé. Ex.:

… **`recherche & développement`** …

Erreur : document mal formé

Problème deux

- On n'est pas toujours en mesure de saisir directement au clavier les 149 186 caractères Unicode existants
- Comment représenter *facilement* n'importe quel caractère dans un document XML, même si on n'est pas capable de le taper directement ?

Entités caractères

- Une des deux formes suivantes:

  &#*nnnnn*;

  &#x*hhhhh*;

  – où *nnnnn* et *hhhhh* représentent le numéro Unicode du caractère voulu:

nnnnn	numéro Unicode en décimal
hhhhh	numéro Unicode en hexadécimal

  – Ex.: € et € désignent tous deux le symbole de l'euro (€)

Exemples d'appel d'entité

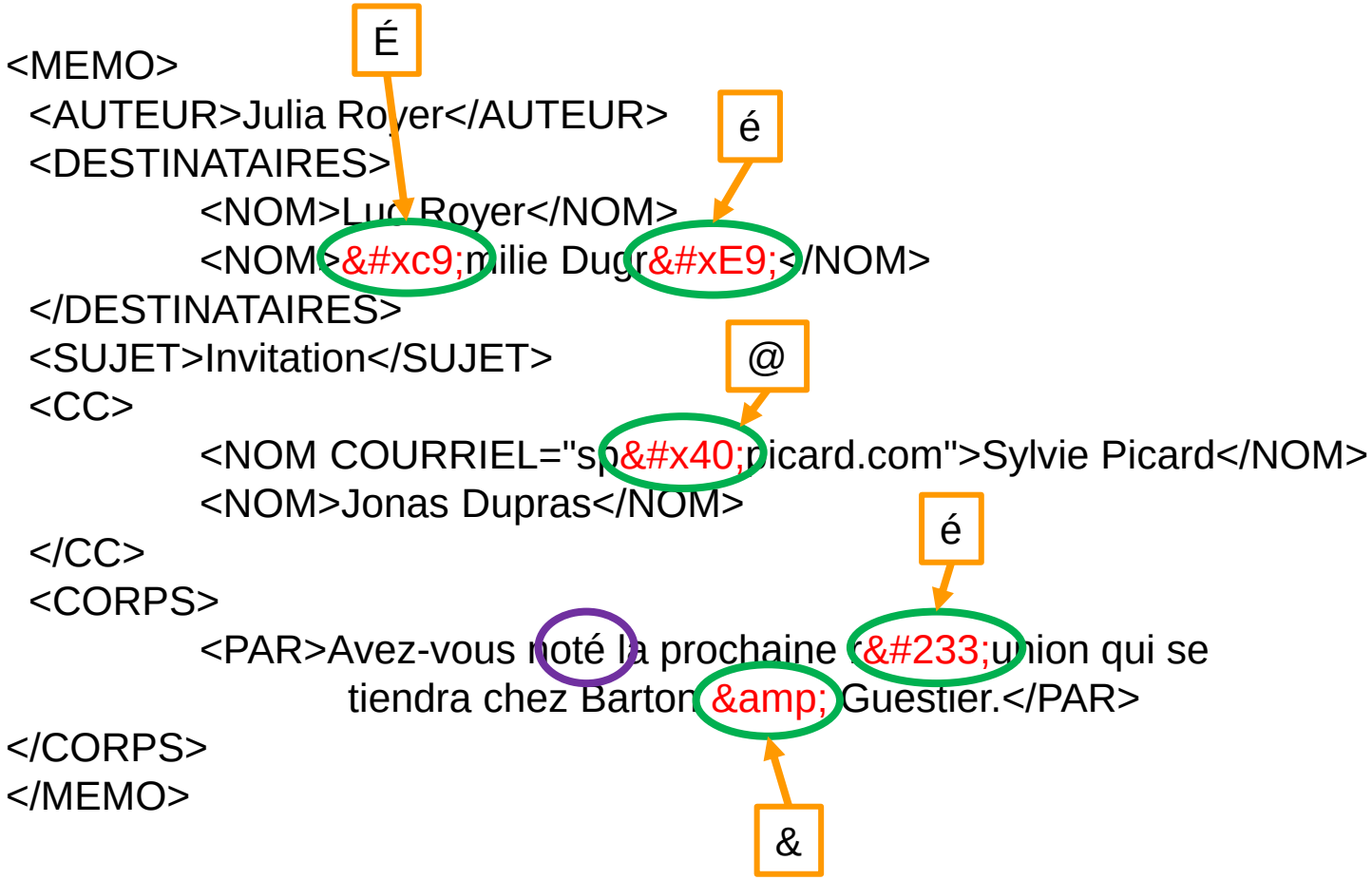
```
<MEMO>
  <AUTEUR>Julia Royer</AUTEUR>
  <DESTINATAIRES>
            <NOM>Luc Royer</NOM>
            <NOM>&#xc9;milie Dugr&#xE9;</NOM>
  </DESTINATAIRES>
  <SUJET>Invitation</SUJET>
  <CC>
            <NOM COURRIEL="sp&#x40;picard.com">Sylvie Picard</NOM>
            <NOM>Jonas Dupras</NOM>
  </CC>
  <CORPS>
            <PAR>Avez-vous noté la prochaine r&#233;union qui se
                  tiendra chez Barton &amp;amp; Guestier.</PAR>
</CORPS>
</MEMO>
```

Dossier 009-entites
Fichier MEMO.xml

Exemples d'appel d'entité

```
<MEMO>
  <AUTEUR>Julia Royer</AUTEUR>
  <DESTINATAIRES>
            <NOM>Luc Royer</NOM>
            <NOM>&#xc9;milie Dugr&#xE9;</NOM>
  </DESTINATAIRES>
  <SUJET>Invitation</SUJET>
  <CC>
            <NOM COURRIEL="sp&#x40;picard.com">Sylvie Picard</NOM>
            <NOM>Jonas Dupras</NOM>
  </CC>
  <CORPS>
            <PAR>Avez-vous noté la prochaine r&#233;union qui se
                  tiendra chez Barton &amp; Guestier.</PAR>
</CORPS>
</MEMO>
```

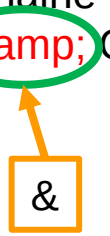

Exemples d'appel d'entité



Exemples d'appel d'entité

```
<MEMO>
  <AUTEUR>Julia Royer</AUTEUR>
  <DESTINATAIRES>
            <NOM>Luc Royer</NOM>
            <NOM>&#xc9;milie Dugr&#xE9;</NOM>
  </DESTINATAIRES>
  <SUJET>Invitation</SUJET>
  <CC>
            <NOM COURRIEL="sp&#x40;picard.com">Sylvie Picard</NOM>
            <NOM>Jonas Dupras</NOM>
  </CC>
  <CORPS>
            <PAR>Avez-vous noté la prochaine r&#233;union qui se
                  tiendra chez Barton &amp; Guestier.</PAR>
</CORPS>
</MEMO>
```

Dans cet exemple
De tous ces appels d'entité,
le seul obligatoire est le &
Tous les autres caractères
peuvent être inscrits directement
dans le document…



Exemples d'appel d'entité

```
<MEMO>
  <AUTEUR>Julia Royer</AUTEUR>
  <DESTINATAIRES>
            <NOM>Luc Royer</NOM>
            <NOM>Émilie Dugré</NOM>
  </DESTINATAIRES>
  <SUJET>Invitation</SUJET>
  <CC>
            <NOM COURRIEL="sp@picard.com">Sylvie Picard</NOM>
            <NOM>Jonas Dupras</NOM>
  </CC>
  <CORPS>
            <PAR>Avez-vous noté la prochaine réunion qui se
                  tiendra chez Barton &amp;amp; Guestier.</PAR>
</CORPS>
</MEMO>
```

&

Dans cet exemple
De tous ces appels d'entité,
le seul obligatoire est le &
Tous les autres caractères
peuvent être inscrits directement
dans le document…

Dossier 009-entites
Fichier MEMO2.xml

Faites afficher la source
pour voir la différence !

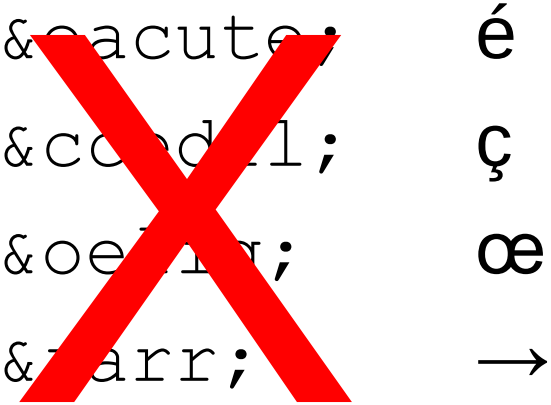
Pas dans les noms XML !

- Un appel d'entité (prédéfinie, caractère) ne peut *pas* se trouver dans un :
  - nom d'élément
  - nom d'attribut
- Ex.: `<prénom>Lise</prénom>`
  - ne peut **pas** s'écrire :

Dossier 009-entites
Fichier mal-forme.xml

Notez bien

- Les entités HTML pour caractères accentués et spéciaux ne peuvent *pas* être utilisées en XML*, p.ex. :



é	é
ç	ç
œ	œ
→	→

Mais évidemment, tous ces caractères peuvent être inscrits directement dans un document XML

*À moins de les *déclarer*, ce qui n'est pas vu dans le Premier Tour d'horizon.

Interprétation du XML par les navigateurs

- Dernière section du texte "Les entités en XML"

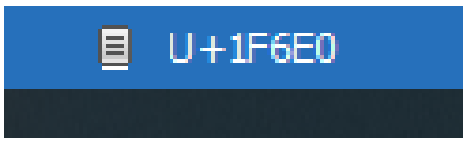
oXygen et caractères spéciaux
(1/7)

- Option générale à régler

  Options → Préférences → Éditeur → Ouvrir

  Activer la prise en charges des caractères spéciaux

- Fonction utile
  - En tout temps, en bas à droite, le numéro Unicode du caractère à droite du curseur est indiqué (en hexadécimal); ex. :

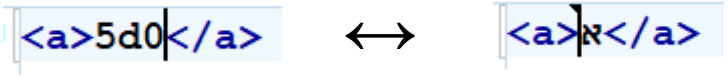


oXygen et caractères spéciaux
(2/7)

- Éditer → Insérer à partir de la table de caractères…
  - Bloc Unicode
  - Onglet "Détails"
  - Recherche par nom
  - Choix d'insertion comme :
    - caractère
    - appel d'entité caractère (numéro décimal)
    - appel d'entité caractère (numéro hexadécimal)

oXygen et caractères spéciaux
(3/7)

- Aussi :
  - Document → Source → Convertir la séquence hexadécimale (jusqu'à 4 chiffres hexadécimaux) en caractères
    - Maj+Ctrl+X
  - Exemple :



oXygen et caractères spéciaux
(4/7)

- On peut aussi insérer un appel d'entité (prédéfinie ou caractère) en le tapant directement, en commençant par l'&
  - On peut alors choisir une entité prédéfinie dans un menu déroulant
    - Le point-virgule final est mis automatiquement
  - Pour un appel de caractère, il faut inscrire soi-même le point-virgule final

oXygen et caractères spéciaux
(5/7)

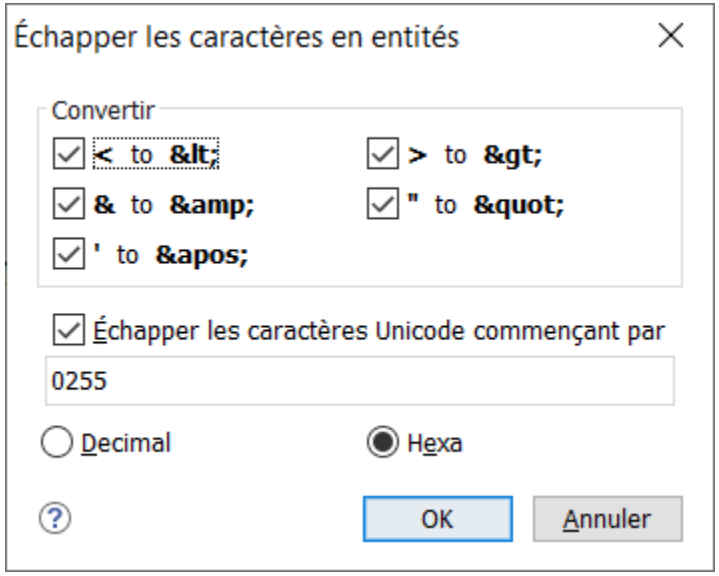
- Copier-coller à partir d'une autre source (Web, PDF, Word)
  - Fonctionne habituellement très bien
  - Y compris pour les émojis

Exemples :

或者透過郵寄或電話的方式投訴

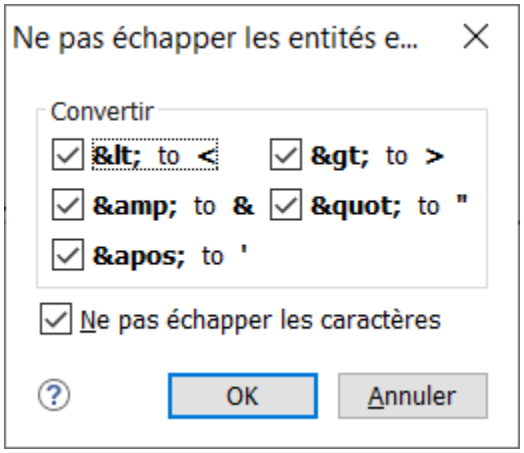
oXygen et caractères spéciaux
(6/7)

- On peut passer de caractères à appels de caractères :
  - Sélectionner le passage à traiter
  - Document → Source → Échapper la sélection



oXygen et caractères spéciaux
(7/7)

- On peut passer d'appels de caractères à caractères :
  - Sélectionner le passage à traiter
  - Document → Source → Ne pas échapper la sélection



Émojis (1/2)

- Les émojis sont des cas spéciaux
  - Notamment car domaine très dynamique
  - Certains émojis sont composés de caractères distincts pouvant être "combinés"

    Ex.:  (1 car.) = 1F469 + 200D + 1F37C

     (2 car.) = 1F469 + 1F37C

    Dossier 009-entites - Fichiers commençant par "emoji"
    Faire afficher la source !

- Les polices de caractères disponibles sur notre appareil sont parfois la limite

Émojis (2/2)

- Point positif : les différents acteurs commerciaux se font un point d'honneur de réagir vite à l'ajout de nouveaux émojis
- Sources à consulter :
  - Full Emoji List, v13.1

    https://unicode.org/emoji/charts/full-emoji-list.html
  - Emojipedia

    https://emojipedia.org/

StudiUM

