INU3011 Documents structurés

Cours 2

Historique de XML

Structuration de l'information en XML

Nouvelle licence

- La licence pour oXygen version 25 est maintenant disponible à partir de la page StudiUM du cours

Historique de XML

SGML = ISO/IEC 8879:1986

- Standard Generalized Markup Language
- Plus complexe que XML; légèrement plus puissant
- Développé à partir de GML (1969), un produit d'IBM
- GML = auteurs **G**oldfarb, **M**osher, **L**orie !
- Développement motivé par la publication dans le domaine du droit

L'élan "CALS"

- Le succès de SGML a au départ été grandement favorisé par l'initiative CALS (*Continuous Acquisition and Life-Cycle Support*) du *Dept. of Defense* états-unien
- But : rationaliser toute la documentation technique de l'équipement de défense
- Ce mouvement s'est transmis à l'industrie, augmentant ainsi la popularité de SGML

L'élan "publication technique et savante" (1/3)

- En partie poussée par le mouvement CALS, l'industrie de la publication commerciale (surtout technique et savante) embrasse SGML grâce aux efforts de deux groupes :
  - Association of American Publishers (AAP)
  - Le groupe "Davenport"

L'élan "publication technique et savante" (2/3)

- Plusieurs fruits de leurs efforts sont encore visibles aujourd'hui
  - Le modèle DocBook pour la documentation technique
  - Le modèle JATS (Journal Article Tag Suite) pour les articles scientifiques

L'élan "publication technique et savante" (3/3)

  - La vague des IETM (Interactive Electronic Technical Manual)
  - HyTime (ISO/IEC 10744:1997) pour l'hypermédia
  - DITA (Darwin Information Typing Architecture) par OASIS Open (2018)
- La rencontre avec le courant LaTeX pour la publication scientifique a donné MathML

L'élan "humanités numériques"

- L'avènement de la *Text Encoding Initiative* (TEI) en 1987 a taillé une place de choix à SGML dans les applications informatiques (naissantes) du domaine des arts et des sciences sociales
- La TEI a joint le courant XML en 2002
- La TEI est aujourd'hui présente dans à peu près tous les projets d'human. num.

L'élan "Web"

- HTML 1 : 1989 (texte, hyperliens)
  - Inventé par Tim Berners-Lee, reconnu comme créateur du Web, base HTML sur SGML
- HTML 2 : 1994 (images, formulaires)
- HTML 3 : 1996 (son, applets)
- HTML 4 : 1998 (vidéo, CSS)
- HTML 4.01: 1999
- XHTML : 2000 → devient du XML

L'élan "Web sémantique" (1/2)

- Au tournant de l'an 2000, l'idée du Web sémantique voit le jour et est développée sur la base de XML, qui est en plein essor dans de nombreux domaines
- L'idée de coder *toute* information numérique en XML fait son chemin : c'est la prolifération du XML "orienté-données" (s'oppose à "orienté-document")

L'élan "Web sémantique" (2/2)

- La majorité des fichiers XML que l'on trouve dans les fichiers système de nos ordinateurs, tablettes et téléphones est du XML "orienté-données"
  - Par exemple, ceux dans **`C:\ProgramData`**
- Pour ces utilisations, XML tend aujourd'hui à être remplacé par JSON ou YAML (plus légers mais moins lisibles que XML)

HTML (1/2)

- HTML = HyperText Markup Language
- Langage de base du Web
  - Permet de créer des *liens hypertextuels* entre les documents
- Pas un format unique, mais une *famille*
  - Il existe de nombreuses versions de HTML
  - La plupart basées sur SGML

HTML (2/2)

- Une des plus utilisées : HTML 4.01
- Il existe aussi (depuis 2000) une version du HTML normalisée par ISO/IEC
  - Norme internationale ISO/IEC 15445:2000
- Dernière version officielle : HTML 5.2
  - Recommandation W3C 14 décembre 2017
  - N'est plus basée ni sur SGML, ni sur XML

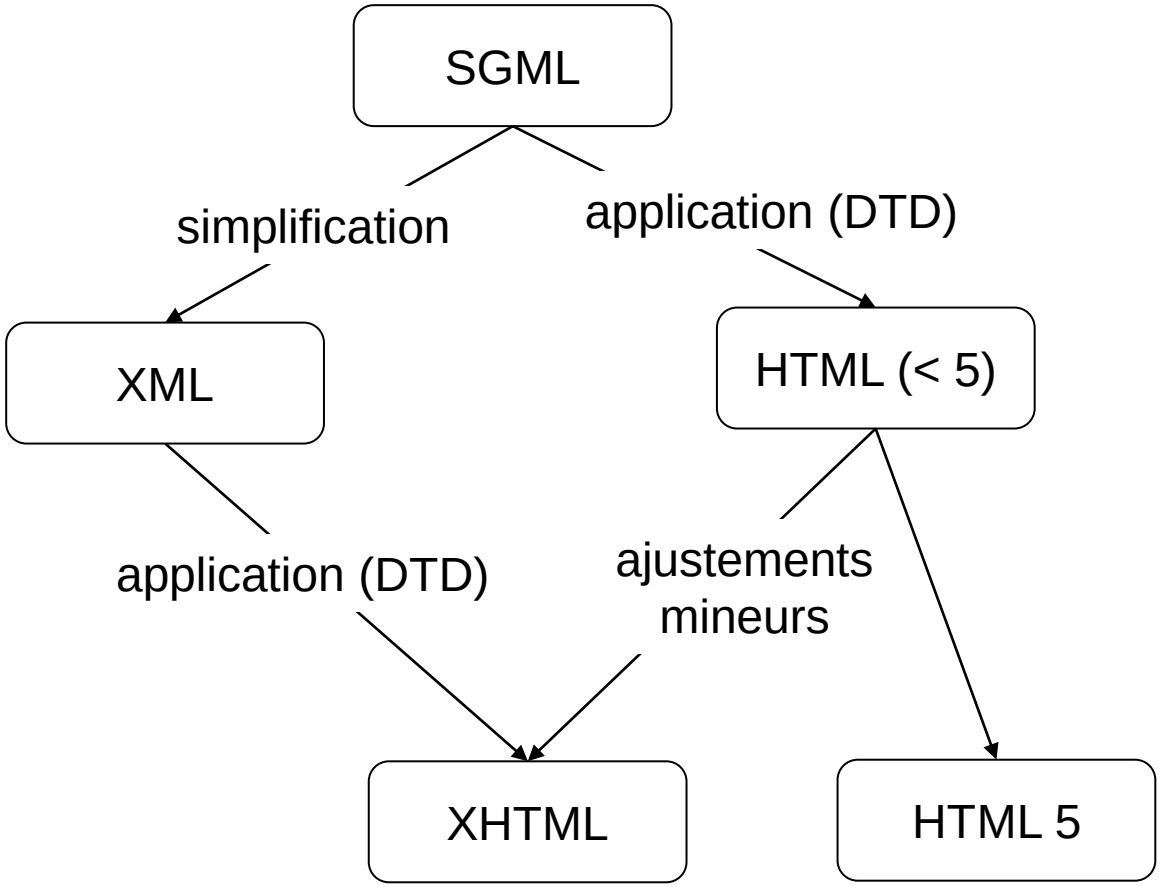
XHTML (1/2)

- Extensible Hypertext Markup Language
- Version qui a succédé à HTML 4.01 au sein du W3C
- Essentiellement, c’est une réécriture de HTML 4.01 basée sur *XML*, plutôt que *SGML*

XHTML (2/2)

- XHTML 1.0 (2e édition): Extensible HyperText Markup Language
  - Recommandation W3C janvier 2000
  - "A Reformulation of HTML 4 in XML 1.0"
- XHTML 1.1: Module-based XHTML
  - Recommandation W3C mai 2001
  - Introduit la possibilité d'extensions modulaires (par exemple, XForms pour les formulaires)
- XHTML est, comme HTML 4, maintenant supplanté (*superseded*) par HTML5

Liens entre SGML, XML et HTML



Quelques modèles XML « documentaires » très utilisés

- TEI (Text Encoding Initiative) https://www.tei-c.org/
- DocBook https://www.docbook.org/
- EAD (Encoded Archival Description) https://www.loc.gov/ead/
- Journal Article Tag Suite https://dtd.nlm.nih.gov/
- MARC21 https://www.loc.gov/standards/marcxml/

Formats bureautiques

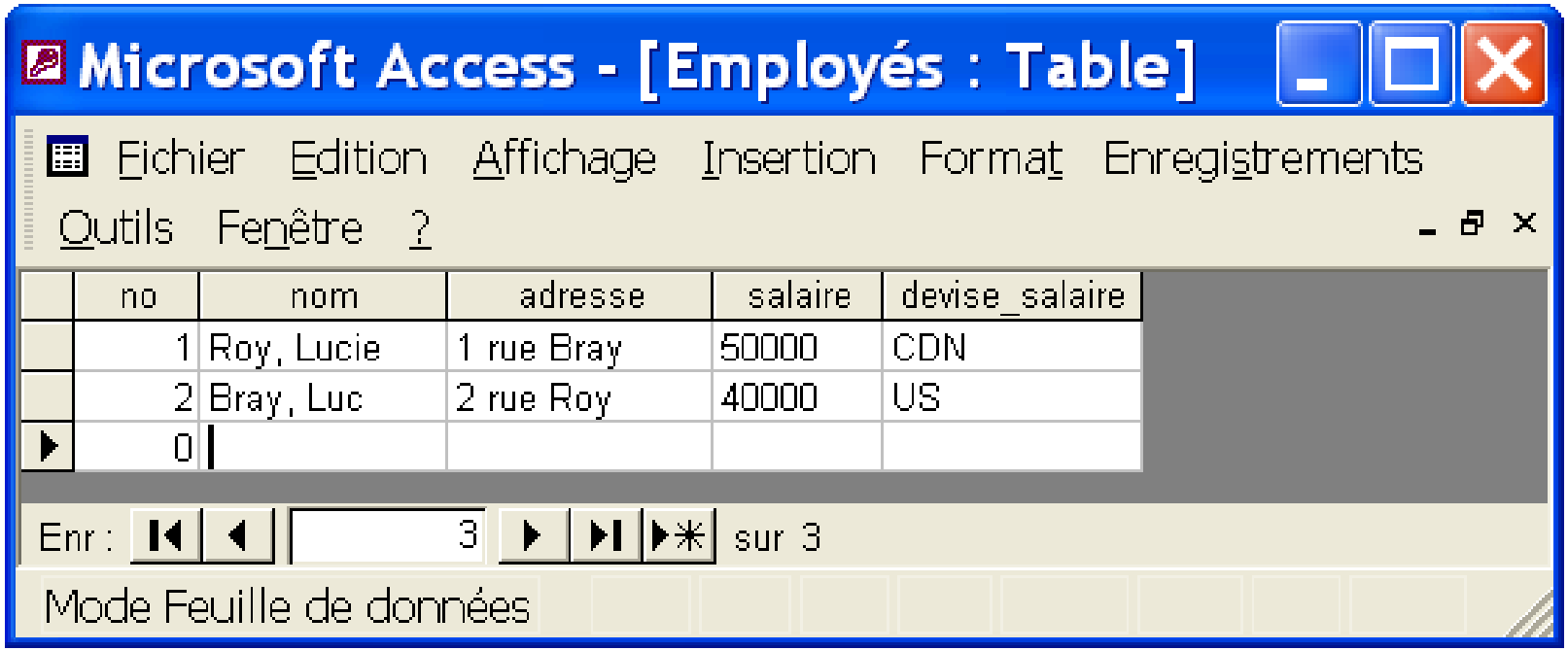
- Office Open XML File Formats (OOXML)
  - ISO/IEC 29500:2008 Office Open XML File Formats
  - Filiation avec Microsoft Office
- Open Document Format for Office Applications (OpenDocument ou ODF)
  - ISO/IEC 26300:2006 Open Document Format for Office Applications (OpenDocument) v1.0
  - Filiation avec le format natif de OpenOffice

N.B.: Ces deux normes ISO sont disponibles gratuitement au <https://standards.iso.org/ittf/PubliclyAvailableStandards/>

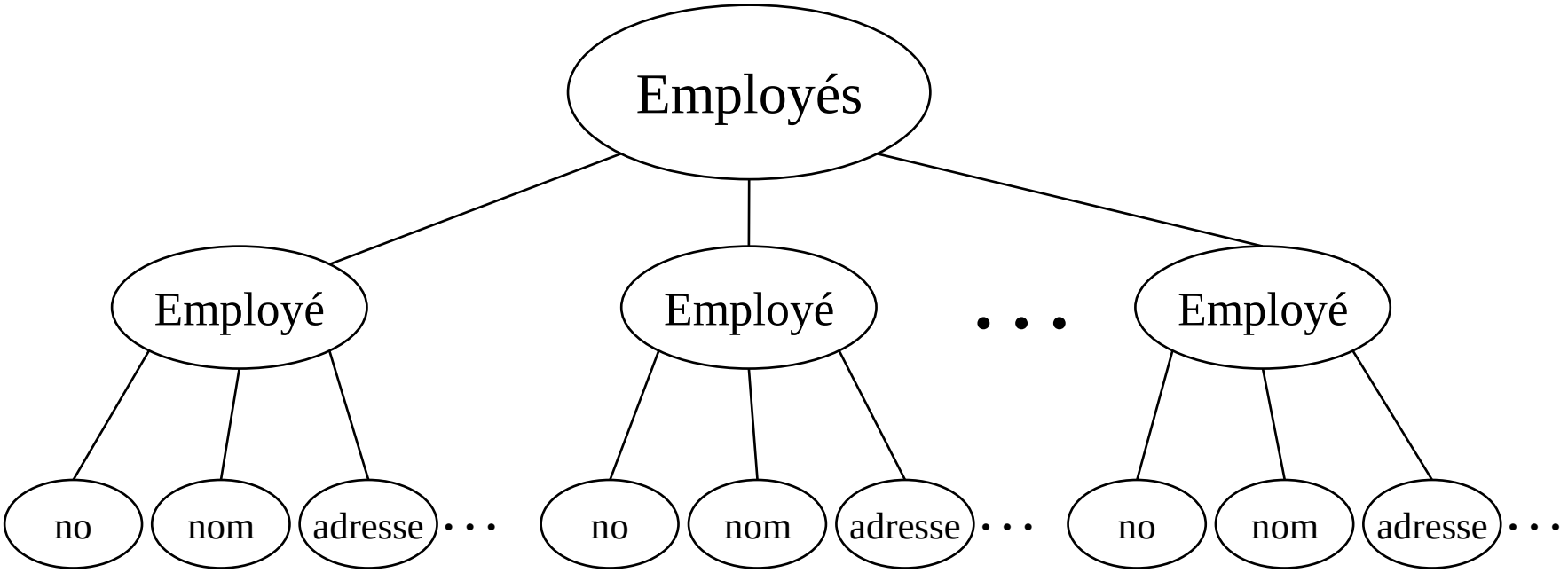
Comment l'information est-elle structurée en XML ?

- Deux comparaisons :
  - Bases de données
  - Traitement de texte

Comparaison 1:
Base de données



Vue hiérarchique des mêmes données



En XML…

```
<Employés>
        <Employé>
                <no>1</no>
                <nom>Roy, Lucie</nom>
                <adresse>1 rue Bray</adresse>
                ...
        </Employé>
        <Employé>
                <no>2</no>
                <nom>Bray, Luc</nom>
                <adresse>2 rue Roy</adresse>
                ...
        </Employé>
...
</Employés>
```


Comparaison 2: Traitement de texte

Mes dernières vacances

J’ai mangé de la dinde, beaucoup de dinde, et j’ai bu du café, beaucoup de café !

Le Goinfre

Pour indiquer que c’est un titre

On sélectionne…

Mes dernières vacances

J’ai mangé de la dinde, beaucoup de dinde, et j’ai bu du café, beaucoup de café !

Le Goinfre

Pour indiquer que c’est un titre

Puis on applique le style voulu…

Mes dernières vacances

J’ai mangé de la dinde, beaucoup de dinde, et j’ai bu du café, beaucoup de café !

Le Goinfre

Pour indiquer l’emphase

On sélectionne…

Mes dernières vacances

J’ai mangé de la dinde, beaucoup de dinde, et j’ai bu du café, beaucoup de café !

Le Goinfre

Pour indiquer l’emphase

Puis, on applique le style voulu…

Mes dernières vacances

J’ai mangé de la dinde, ***beaucoup*** de dinde, et j’ai bu du café, beaucoup de café !

Le Goinfre

Pour indiquer l’emphase

Même chose pour l’autre mot « beaucoup »…

Mes dernières vacances

J’ai mangé de la dinde, ***beaucoup*** de dinde, et j’ai bu du café, ***beaucoup*** de café !

Le Goinfre

Pour la signature

On sélectionne…

Mes dernières vacances

J’ai mangé de la dinde, ***beaucoup*** de dinde, et j’ai bu du café, ***beaucoup*** de café !

Le Goinfre

Pour la signature

Puis on applique le style voulu…

Mes dernières vacances

J’ai mangé de la dinde, ***beaucoup*** de dinde, et j’ai bu du café, ***beaucoup*** de café !

Le Goinfre

En XML… (1/3)

- Au lieu de sélectionner et d'appliquer un style, l'auteur « marque » les bouts de texte avec des *balises* avant et après…

<titre>Mes dernières vacances</titre>

J'ai mangé de la dinde,
<emphase>beaucoup</emphase> de dinde, et j'ai bu du café, <emphase>beaucoup</emphase> de café !

<signature>Le Goinfre</signature>

Note: Ceci n'est pas un document XML bien formé, car il manque un élément-document.

En XML… (2/3)

- Une « application XML », créée indépendamment des documents, s’occupe au moment voulu **d’interpréter** les balises, p.ex. les **traduire en styles de présentation** appropriés

En XML… (3/3)

- Le travail est divisé en deux :
  - création du contenu et de la structure des documents (auteurs, contributeurs, etc.)
  - conception de l'application informatique interprète (graphistes, informaticiens, etc.)

N.B.: L'application peut être une simple feuille de style (CSS ou XSLT) !

Unités de contenu - exemple

```
<historiette>
  <para>
    <personne>Dracula</personne> alla en France. Là,
    il rencontra <personne>Barbe-Bleue</personne>.
  </para>
</historiette>
```

```
<historiette>↲
¤¤<para>↲
¤¤¤¤<personne>Dracula</personne>¤alla¤en¤France.¤Là,↲
¤¤¤¤il¤rencontra¤<personne>Barbe-Bleue</personne>.↲
¤¤</para>↲
</historiette>
```

Élément `historiette`

```
<historiette>↲
¤¤<para>↲
¤¤¤¤<personne>Dracula</personne>¤alla¤en¤France.¤Là,↲
¤¤¤¤il¤rencontra¤<personne>Barbe-Bleue</personne>.↲
¤¤</para>↲
</historiette>
```

Élément `historiette`

```
<historiette>↲
¤¤<para>↲
¤¤¤¤<personne>Dracula</personne>¤alla¤en¤France.¤Là,↲
¤¤¤¤il¤rencontra¤<personne>Barbe-Bleue</personne>.↲
¤¤</para>↲
</historiette>
```

Unités de contenu de l'élément `historiette` :

1. La chaîne ↲¤¤

Élément `historiette`

```
<historiette>↲
¤¤<para>↲
¤¤¤¤<personne>Dracula</personne>¤alla¤en¤France.¤Là,↲
¤¤¤¤il¤rencontra¤<personne>Barbe-Bleue</personne>.↲
¤¤</para>↲
</historiette>
```

Unités de contenu de l'élément `historiette` :

1. La chaîne ↲¤¤
2. Le sous-élément `para`

Élément `historiette`

```
<historiette>↲
¤¤<para>↲
¤¤¤¤<personne>Dracula</personne>¤alla¤en¤France.¤Là,↲
¤¤¤¤il¤rencontra¤<personne>Barbe-Bleue</personne>.↲
¤¤</para>↲
</historiette>
```

Unités de contenu de l'élément `historiette` :

1. La chaîne ↲¤¤
2. Le sous-élément `para`
3. La chaîne ↲

Élément `historiette`

```
<historiette>↲
¤¤<para>↲
¤¤¤¤<personne>Dracula</personne>¤alla¤en¤France.¤Là,↲
¤¤¤¤il¤rencontra¤<personne>Barbe-Bleue</personne>.↲
¤¤</para>↲
</historiette>
```

Unités de contenu de l'élément `historiette` :

1. La chaîne ↲¤¤
2. Le sous-élément `para`
3. La chaîne ↲

N.B.: L'élément `historiette` *est donc un élément* ***conteneur***

Élément para

```
<historiette>↲
¤¤<para>↲
¤¤¤¤<personne>Dracula</personne>¤alla¤en¤France.¤Là,↲
¤¤¤¤il¤rencontra¤<personne>Barbe-Bleue</personne>.↲
¤¤</para>↲
</historiette>
```

Élément `para`

```
<historiette>↲
¤¤<para>↲
¤¤¤¤<personne>Dracula</personne>¤alla¤en¤France.¤Là,↲
¤¤¤¤il¤rencontra¤<personne>Barbe-Bleue</personne>.↲
¤¤</para>↲
</historiette>
```

Unités de contenu de l'élément `para` :

1. La chaîne ↲¤¤¤¤

Élément `para`

```
<historiette>↲
¤¤<para>↲
¤¤¤¤<personne>Dracula</personne>¤alla¤en¤France.¤Là,↲
¤¤¤¤il¤rencontra¤<personne>Barbe-Bleue</personne>.↲
¤¤</para>↲
</historiette>
```

Unités de contenu de l'élément `para` :

1. La chaîne ↲¤¤¤¤
2. Un sous-élément `personne`

Élément para

```
<historiette>↲
¤¤<para>↲
¤¤¤¤<personne>Dracula</personne>¤alla¤en¤France.¤Là,↲
¤¤¤¤il¤rencontra¤<personne>Barbe-Bleue</personne>.↲
¤¤</para>↲
</historiette>
```

Unités de contenu de l'élément para **:**

1. La chaîne ↲¤¤¤¤
2. Un sous-élément personne
3. La chaîne ¤alla¤en¤France.¤Là,↲¤¤¤¤il¤rencontra¤

Élément para

```
<historiette>↲
¤¤<para>↲
¤¤¤¤<personne>Dracula</personne>¤alla¤en¤France.¤Là,↲
¤¤¤¤il¤rencontra¤<personne>Barbe-Bleue</personne>.↲
¤¤</para>↲
</historiette>
```

Unités de contenu de l'élément para **:**

1. La chaîne ↲¤¤¤¤
2. Un sous-élément personne
3. La chaîne ¤alla¤en¤France.¤Là,↲¤¤¤¤il¤rencontra¤
4. Un sous-élément personne

Élément para

```
<historiette>↲
¤¤<para>↲
¤¤¤¤<personne>Dracula</personne>¤alla¤en¤France.¤Là,↲
¤¤¤¤il¤rencontra¤<personne>Barbe-Bleue</personne>.↲
¤¤</para>↲
</historiette>
```

Unités de contenu de l'élément para **:**

1. La chaîne ↲¤¤¤¤
2. Un sous-élément personne
3. La chaîne ¤alla¤en¤France.¤Là,↲¤¤¤¤il¤rencontra¤
4. Un sous-élément personne
5. La chaîne .↲¤¤

Élément para

```
<historiette>↲
¤¤<para>↲
¤¤¤¤<personne>Dracula</personne>¤alla¤en¤France.¤Là,↲
¤¤¤¤il¤rencontra¤<personne>Barbe-Bleue</personne>.↲
¤¤</para>↲
</historiette>
```

Unités de contenu de l'élément `para` **:**

1. La chaîne `↲¤¤¤¤`
2. Un sous-élément `personne`
3. La chaîne `¤alla¤en¤France.¤Là,↲¤¤¤¤il¤rencontra¤`
4. Un sous-élément `personne`
5. La chaîne `.↲¤¤`

N.B.: L'élément `para` *est donc un élément* ***mixte***

Premier élément `personne`

```
<historiette>↲
¤¤<para>↲
¤¤¤¤<personne>Dracula</personne>¤alla¤en¤France.¤Là,↲
¤¤¤¤il¤rencontra¤<personne>Barbe-Bleue</personne>.↲
¤¤</para>↲
</historiette>
```

Premier élément `personne`

```
<historiette>↲
¤¤<para>↲
¤¤¤¤<personne>Dracula</personne>¤alla¤en¤France.¤Là,↲
¤¤¤¤il¤rencontra¤<personne>Barbe-Bleue</personne>.↲
¤¤</para>↲
</historiette>
```

Unités de contenu de l'élément :

1. La chaîne `Dracula`

Premier élément `personne`

```
<historiette>↲
¤¤<para>↲
¤¤¤¤<personne>Dracula</personne>¤alla¤en¤France.¤Là,↲
¤¤¤¤il¤rencontra¤<personne>Barbe-Bleue</personne>.↲
¤¤</para>↲
</historiette>
```

Unités de contenu de l'élément :

1. La chaîne `Dracula`

*N.B.: Cet élément est donc un élément **textuel***

Second élément `personne`

```
<historiette>↲
¤¤<para>↲
¤¤¤¤<personne>Dracula</personne>¤alla¤en¤France.¤Là,↲
¤¤¤¤il¤rencontra¤<personne>Barbe-Bleue</personne>.↲
¤¤</para>↲
</historiette>
```

Second élément `personne`

```
<historiette>↲
¤¤<para>↲
¤¤¤¤<personne>Dracula</personne>¤alla¤en¤France.¤Là,↲
¤¤¤¤il¤rencontra¤<personne>Barbe-Bleue</personne>.↲
¤¤</para>↲
</historiette>
```

Unités de contenu de l'élément :

1. La chaîne `Barbe-Bleue`

Second élément `personne`

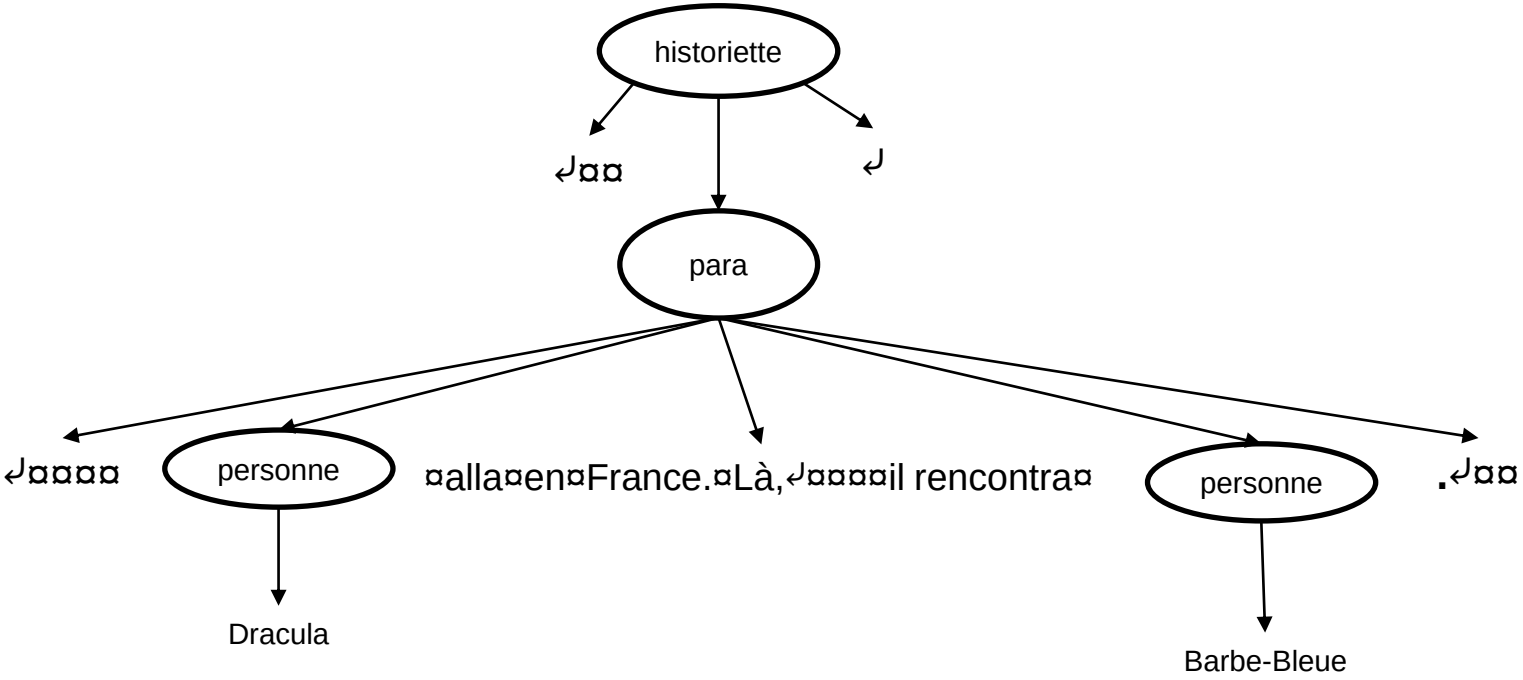
```
<historiette>↲
¤¤<para>↲
¤¤¤¤<personne>Dracula</personne>¤alla¤en¤France.¤Là,↲
¤¤¤¤il¤rencontra¤<personne>Barbe-Bleue</personne>.↲
¤¤</para>↲
</historiette>
```

Unités de contenu de l'élément :

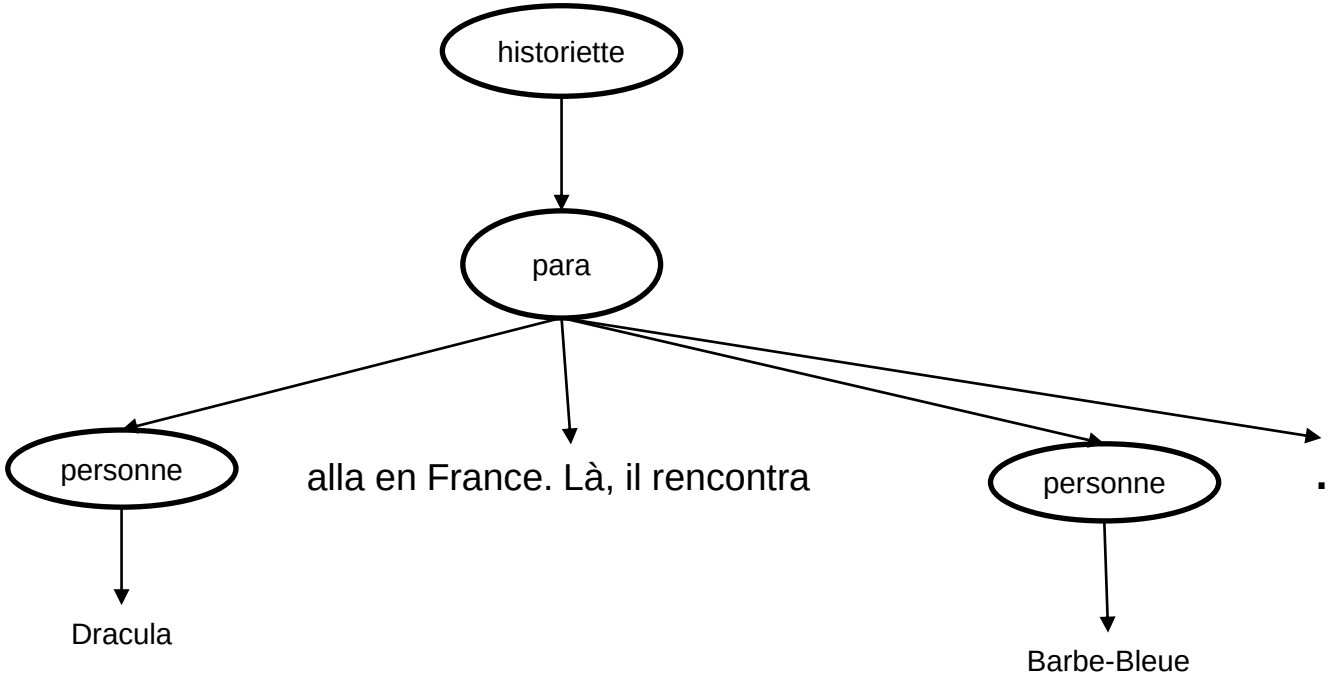
1. La chaîne `Barbe-Bleue`

*N.B.: Cet élément est donc aussi un élément **textuel***

Arbre inversé



Arbre inversé allégé



StudiUM

