Séance 2
Suite des exercices de balisage XML
Exploiter l'encodage

TD Antiquités numériques - 2ème partie

Université Lumière Lyon 2, site Berges du Rhône

Mardi 4/11/2025 - 12h00-14h00

exercice 2
Encodage TEI d’une carte postale

Consigne

Il vous faut transposer la première “modélisation” réalisée de l’exercice 1 dans le vocabulaire et la structure d'un fichier TEI

Rappel : structure d'un document TEI

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">

 <teiHeader>
              <!-- [ en-tête ]  -->
 </teiHeader>

 <text>

     <front>
              <!-- [ partie préliminaire ... ] -->
     </front>

     <body>
              <!-- [ corps du texte ... ]  -->
     </body>

     <back>
              <!--  [ partie annexe ...  ] -->
     </back>

  </text>

</TEI>
```

Remarques :

- **<!-- -- > :** commentaire dans le langage XML (non pris en compte, pratique pour annoter/documenter son code)
- <front> et <back> sont optionnels
- Les images peuvent être regroupées dans un élément <facsimile> ou être liées directement à une balise via l'attribut @facs

  Exemple :

```
<text>
    <pb facs="page1.png"/>
   <!-- texte de la page 1 ici -- >
      (...)
```

Rappel : structure XML ad hoc

```
<carte n="0010">
    <recto url="https://gallica.bnf.fr/ark:/12148/btv1b10582689t/f2.item">
        <legende>Le Grand Hôtel de Russie sur la Piazza del Populo - Via del
Babuino</legende>
    <verso url="https://gallica.bnf.fr/ark:/12148/btv1b10582689t/f2.item">

    <obliteration>
        <date>21/11/1914</date>
        <lieu>Rome</lieu>
        <cachet>Roma Ferrovia</cachet>
    </obliteration>

    <message>
        <p>Cher Toulet</p>
        <p>Vous savez qu'on n'a jamais le temps d'écrire à Paris. C'est pourquoi (...) </p>
        <p>Votre vieux dévoué</p>
        <p>Claude Debussy</p>
    </message>

    <destinataire>
             <p>Monsieur P. J. Toulet</p>
             <p>Chateau de la Rafette à Saint-Loubès</p>
             <p>France - Gironde.</p>
     </destinataire>

   </verso>
</carte>
```

Transposition en TEI (premiers niveaux)

Balise ad hoc	**Conseil**
<carte>	une section de type “carte”...
<recto> <verso>	une sous section
<legende>	liée à la notion d’illustration (<figure>)
<obliteration>	
<message>	un
<destinataire>	
…	

Prise en main de l'éditeur XML et premier en-tête TEI

1. Ouvrir le fichier '**tei-base.xml**'
2. Tapez un chevron ('**<**') à l'intérieur de la balise **<teiHeader>** : le plugin Scholarly XML vous suggère une balise TEI susceptible d'être insérée à cet endroit : **<fileStmt>**
3. Acceptez la suggestion avec la touche de tabulation
4. Continuez à remplir l'entête TEI en suivant les suggestions…
5. Tant qu'une balise est soulignée en rouge, c'est qu'il manque des balises enfants ou qu'une erreur a été faite. Lisez les descriptions du problème pour les résoudre.
6. Vous pouvez vous aider des guidelines de la TEI pour reproduire un TEI Header minimal valide (toutes les balises obligatoires sont renseignées).

Reformatez si nécessaire en XML à l'aide de la palette de commandes
(voir la cheatsheet de la séance 1)

Aides

- Dans VSCode, regardez la documentation du plugin Scholarly XML (animations)
  - `CTRL + barre d'espace` (faire apparaître les suggestions)
  - `CTRL + E` (entourer du texte avec une balise)
  - `Command Palette =>"Scholarly XML: Validate XML with associated RELAX NG schema."`
- <teiHeader> : https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-teiHeader.html

Encodage TEI de la carte postale

Il s’agit maintenant de baliser la carte postale en TEI.

Si vous souhaitez partir d’une base de départ, vous pouvezuvrez le fichier ‘**carte-tei-0.xml**’. Sinon vous pouvez continuer avec tei-base.xml (en l’enregistrant sous un nouveau nom).

Pour relire le texte

1. Cliquez sur l'icône **TEI Publisher** en haut à droite de l'écran 
2. Choisissez l'ODD **Dantiscus Letters** ou **TEI Publisher base**
   La prévisualisation vous permet de relire le texte directement (la mise en forme des balises dépend de l'ODD choisi… on y reviendra plus tard).

Corrigé 2

en tête TEI

```
<teiHeader>

    <fileDesc>

      <titleStmt>
        <title>Exercice d'édition numérique d'une carte postale</title>
      </titleStmt>

      <publicationStmt>
        <p>Document de formation.</p>
      </publicationStmt>

      <sourceDesc>
        <msDesc>
          <msIdentifier>
            <repository>Médiathèque intercommunale Pau-Pyrénées</repository>
          </msIdentifier>
          <msContents/>
          <physDesc/>
        </msDesc>
      </sourceDesc>

    </fileDesc>

    <profileDesc>

      <correspDesc>

        <correspAction type="sent">
          <persName/>
          <placeName/>
          <date/>
        </correspAction>

        <correspAction type="received">
          <persName/>
          <placeName/>
        </correspAction>

      </correspDesc>

    </profileDesc>

  </teiHeader>
```

Encodage du message (base de départ)

```
<text>
      <body>

         <div type="recto">
            <figure>
               <graphic url="https://gallica.bnf.fr/ark:/12148/btv1b10582689t"/>
               <!--
                  <figDesc> </figDesc>
                  <head> </head>
                -->
            </figure>

         </div>

         <div facs="https://gallica.bnf.fr/ark:/12148/btv1b10582689t/f2.item"
            type="verso">

            <div type="message">
               <!-- ... -->
            </div>

            <div type="destination">

               <p>
                  <!-- stamps -->
               </p>

               <p>
                  <address>
                     <!-- ... -->
                  </address>
               </p>

            </div>

         </div>

      </body>

   </text>
```

Commentaire

- De multiples encodages sont possibles !
- Modulaire et flexible, la TEI permet d'avoir à la fois :
  - Un modèle commun et déjà documenté
  - La possibilité de l'adapter aux spécificités de son projet
- Son modèle est évolutif ouvert à toutes les contributions
  - Si des balisent manquent de votre point de vue, vous pouvez les proposer, elles bénéficieront ainsi à d'autres projets
  - Il y a 2 mises à jour par an
  - Le premier support est la liste discussion. Voir la rubrique support pour vous aider dans votre apprentissage

Pour aller plus loin

- Regardez les définitions des éléments et y trouver des idées d'attributs intéressants à encoder
- Réfléchissez des encodages alternatifs ou complémentaires
  - insertion des débuts de ligne
  - utilisation de l'élément TEI/facsimile pour les images
  - enregistrement de la forme fautive et de sa correction
  - signalement des lieux
  - latitude et longitude
  - etc.
- Manuel d'encodage de correspondances et cartes postales

fichier complet

Voir ‘carte-tei-1.xml’

exercice 3

Customisez le schéma TEI pour votre projet...

Le but de cet exercice est de créer un schéma personnalisé pour contrôler le balisage de notre fichier ce carte postale.

Un schéma rend l'encodage plus facile et plus sûr :

- Seules les balises et attributs autorisés peuvent être saisies
- On peut y ajouter une documentation et des exemples spécifiques au projet
- Créer un schéma personnalisé est une bonne pratique qui permet une meilleure gestion de projet et un résultat final mieux maîtrisé

Nous allons partir d'un schéma créé par Lou Burnard. Celui ci est incomplet.
Nous devrons donc y ajouter les balises de notre modélisation manquantes...

Préparation

- Copiez le fichier **tei_cartes.rng** qui se trouve dans le dossier **exercices/schemas** à côté de votre fichier de travail.
- **tei_cartes.rng** correspond au schéma créé par Lou Burnard pour l'encodage des cartes postales et ne contient pas toutes les balises nécessaires par rapport à notre encodage.

Modifiez le lien vers le schéma dans le fichier TEI

- Dans le prologue XML, pointez sur ce même fichier `tei_cartes.rng`

```
<?xml-model href="tei_cartes.rng" type="application/xml"
schematypens="http://relaxng.org/ns/structure/1.0"?>
<?xml-model href="tei_cartes.rng" type="application/xml"
schematypens="http://purl.oclc.org/dsdl/schematron"?>
```

- Les balisent manquantes sont repérées par le plugin Scholarly XML qui contrôle la validité du fichier avec ce schéma.
- Cliquez sur le symbole "attention" dans la barre d'état pour voir les erreurs. Elles vous indiquent quelles balises sont manquantes : vous allez pouvoir les intégrer à l'aide d'un outil en ligne appelé "Roma"...

Découverte de l’outil ROMA

- Allez à l'adresse https://roma.tei-c.org/
- Cliquez sur "Télécharger un ODD"
- Dans le dossier `exercices/schemas`, choisissez `tei_cartes.odd` et cliquez sur Commencer
- Cliquez sur Personnaliser l'ODD
- Les balises s'affichent par ordre alphabétique.
  Les balises du schéma "cartes" sont cochées, le nom de leur module d'appartenance s'affiche à droite.

- [ ] addName
(additional name) contains an additional name component, such as a nickn… (namesdates)

- [x] address
(address) contient une adresse postale ou d'un autre type, par exemple l'adr… (core)

- [x] addrLine
(ligne d'adresse) contient une ligne d'adresse postale. (core)

- [ ] addSpan
(added span of text) marks the beginning of a longer sequence of text adde… (transcr)

- [ ] adminInfo
(administrative information) contains information about the present custod… (msdescription)

- [ ] affiliation
(affiliation) contains an informal description of a person's present or past af… (namesdates)

Consignes

- Vous remarquez que la sélection ne porte que sur un petit nombre de balises.
  Il est en effet préférable de partir d'un schéma minimal et d'ajouter des balises au fur et à mesure des besoins...
- Complétez les balises manquantes dans votre schéma

  Vous pouvez si vous le souhaitez en profiter pour ajouter la balise <lb/> (line beginning) pour marquer les passages à la ligne...

Export du schéma au format relaxNg

- Roma a chargé un fichier **.odd**, autrement dit le schéma exprimé en TEI.
- Mais les éditeurs XML ne connaissant pas ce format. ODD est un langage de spécification (construction) de schémas, pas de validation.
  Il faut l'exporter dans un format lisible par l'éditeur XML. Par exemple Relax NG compact (**.rnc**)
- Cliquez sur Télécharger et choisissez Relax NG
- Choisissez le dossier **exercices/ex-3-roma** et nommez le fichier **tei_cartes** pour remplacer l'ancien

Test du schéma

- Créez un nouveau fichier TEI lié à ce schéma et testez :
  - la suggestion des balises ;
  - la saisie de balises absentes du schéma…
- Avez-vous besoin de nouvelles balises ? Recommencez à l’ étape ODD puis ré-exportez votre schéma…
  - Le processus est itératif par nature.

Cours : introduction à XPATH

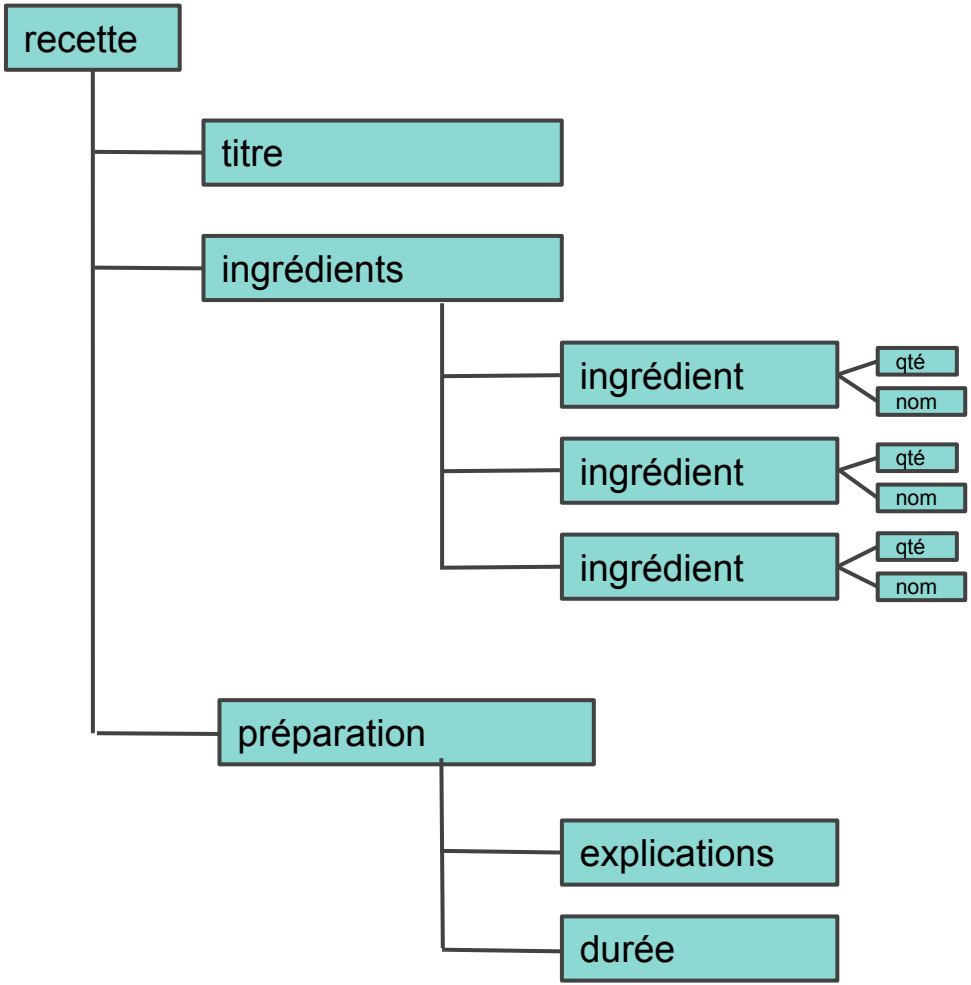
L’arbre XML et sa sérialisation

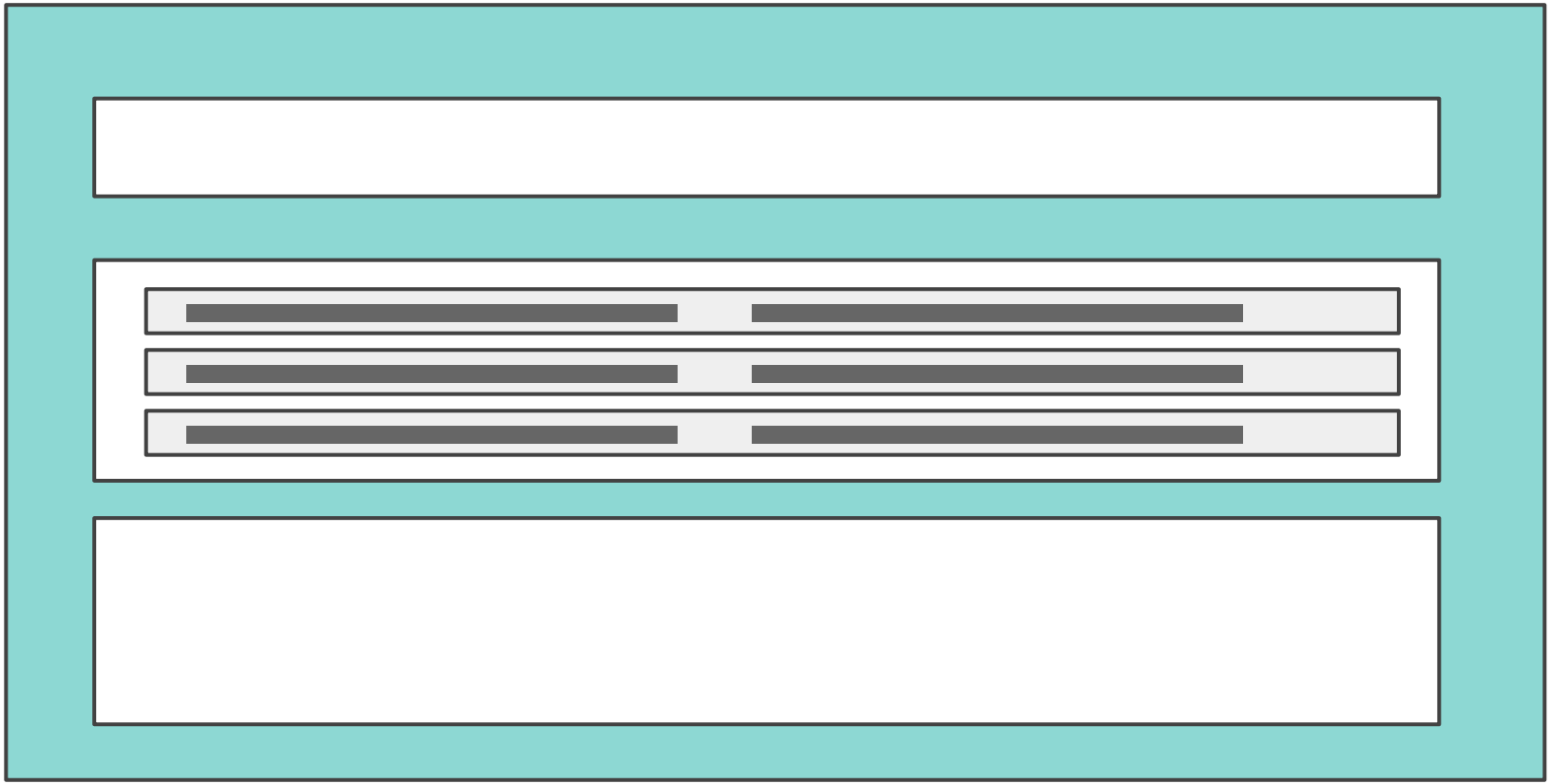
- Structurellement XML est un arbre
- La sérialisation est la représentation de l'arbre sous forme d'un flux de caractères
- Les propriétés de XML doivent être maintenues dans sa sérialisation :
    - Une racine unique (élément qui contient tous les autres éléments)
    - Imbrication des balises sans chevauchements
    - Tous les éléments sont fermés
    - Les attributs ont des valeurs entre guillemets (simples ou doubles)

```
<recette>
        <titre>Lait de poule</titre>
        <ingredients>
            <ingredient><qté>1</qté><nom>oeuf</nom></ingredient>
            <ingredient><qté>10 cl</qté><nom>de lait chaud</nom></ingredient>
            <ingredient><nom>sucre en poudre</nom></ingredient>
        </ingredients>
        <preparation>
                <explications>Verser le tout dans un verre à anse. Sucrer selon
                 son goût. Remuer et ajouter un peu de noix de muscade râpée.
                </explications>
                 <durée/>
         </preparation>
</recette>
```

[[lait de poule][[[1] [oeuf]]

[[10 cl][de lait chaud]]] [[Verser le tout dans un verre à anse. Sucrer selon son goût. Remuer et ajouter un peu de noix de muscade râpée.]]


recette
titre
ingrédients
ingrédient
qté
nom
ingrédient
qté
nom
ingrédient
qté
nom
préparation
explications
durée



Notion de validité

- Terme technique qui veut dire que le document n'utilise que certains éléments pré-déterminés dans un certain contexte.
    - Par exemple, dans un dictionnaire, les éléments doivent toujours apparaître dans un certain ordre.
- Ensemble de règles régissant le contexte d'apparition des balises en fonction d’un schéma donné.
- La validité est généralement contrôlée pendant l'édition.
- Un document peut être bien formé mais non valide.
- Un document mal formé ne peut pas être valide.

XPATH

- XPath = XML path language
- Il sert à identifier et localiser des fragments de documents XML.
- Il décrit des chemins à prendre pour aller de la racine ou d'un noeud "local" vers le noeud recherché en suivant les filiations...

Raccourcis XPATH

. élément courant

/ élément racine

.. parent

../.. grand-parent

x/ enfant

x// descendant

tei: namespace

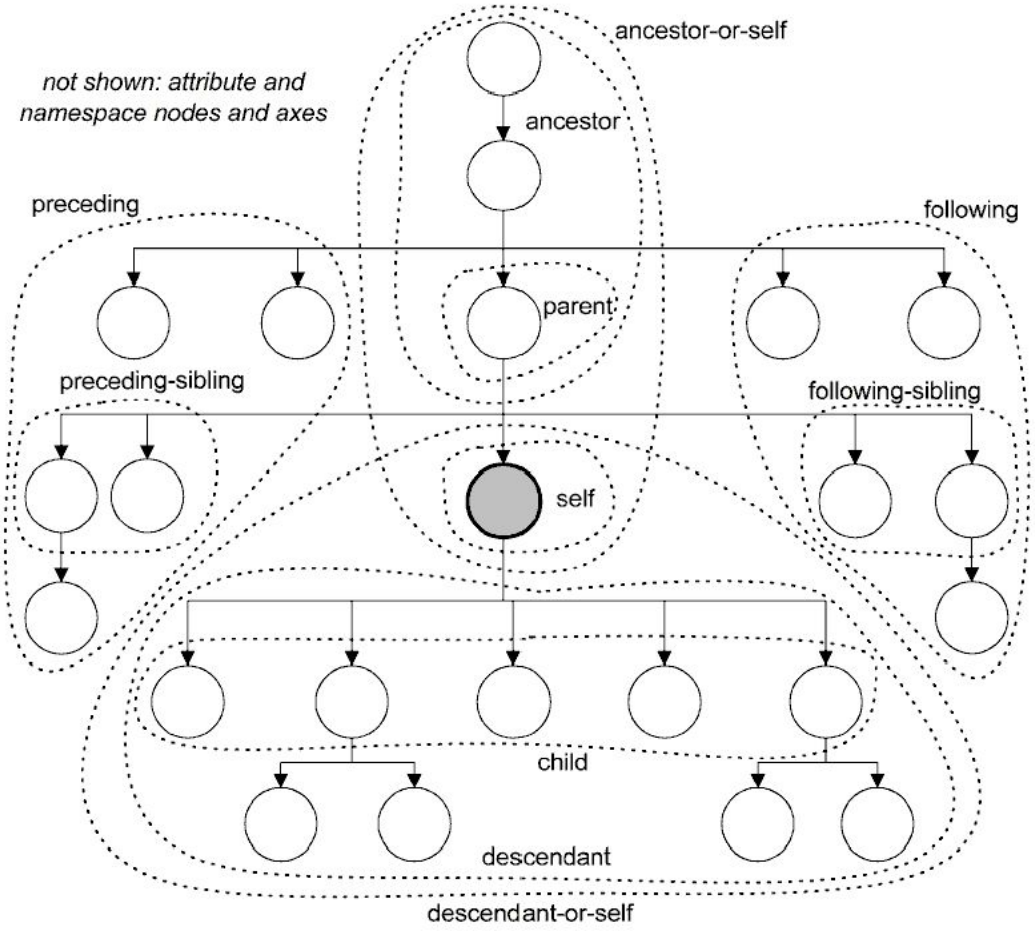
[ ] predicat

@ attribut

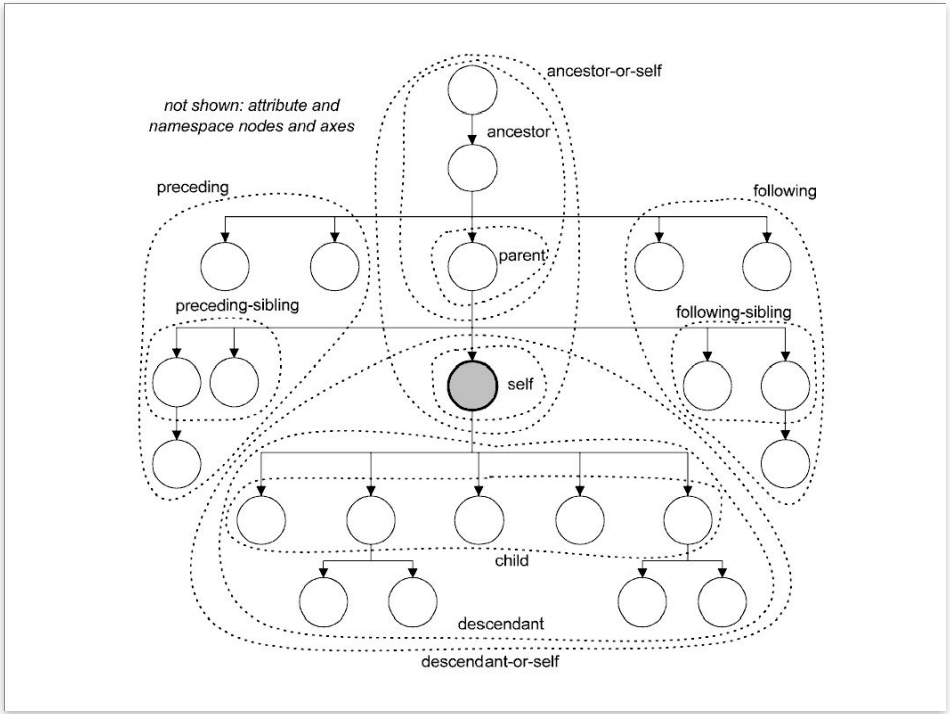
* n’importe quel element

not() sélection négative

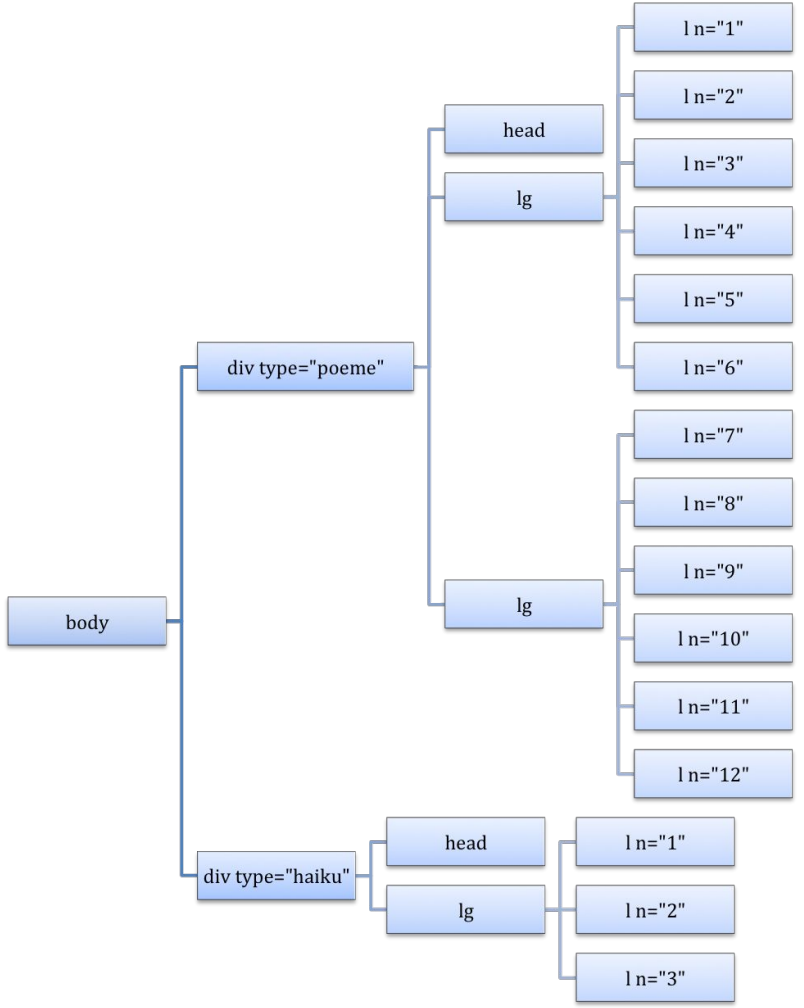
Axes XPath


not shown: attribute and
namespace nodes and axes
ancestor-or-self
ancestor
preceding
following
parent
preceding-sibling
following-sibling
self
child
descendant
descendant-or-self

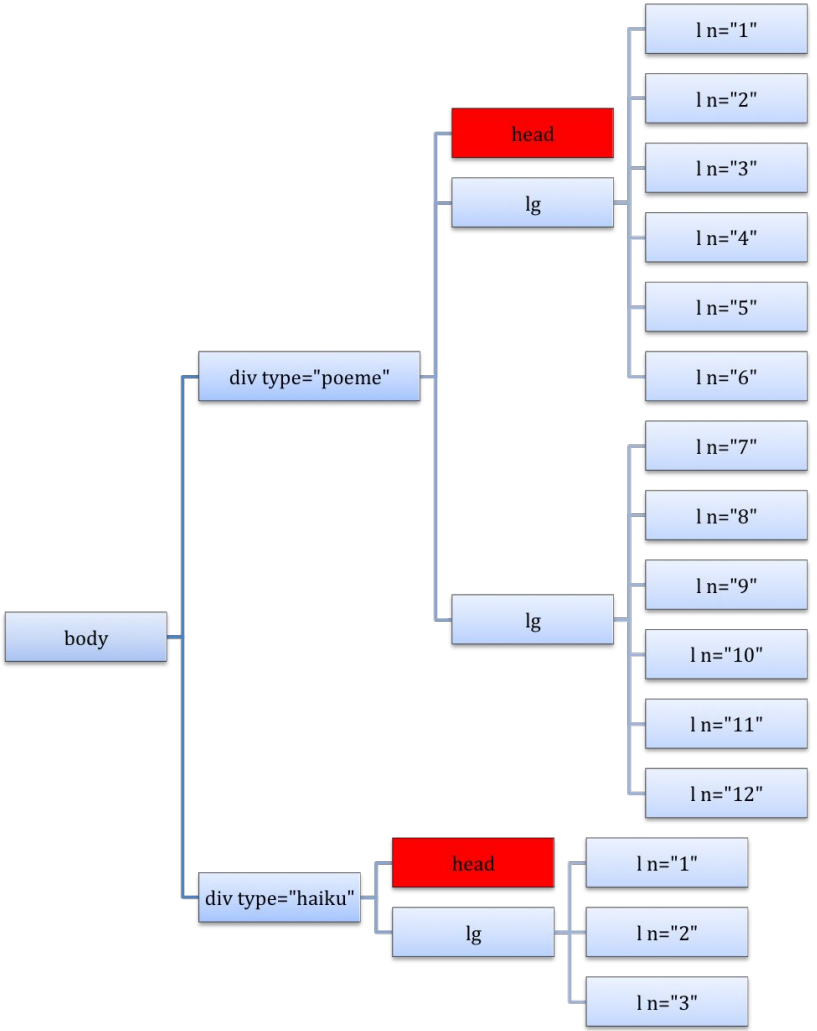
Axes XPATH


ancestor-or-self
not shown: attribute and namespace nodes and axes
ancestor
preceding
following
parent
preceding-sibling
following-sibling
self
child
descendant
descendant-or-self

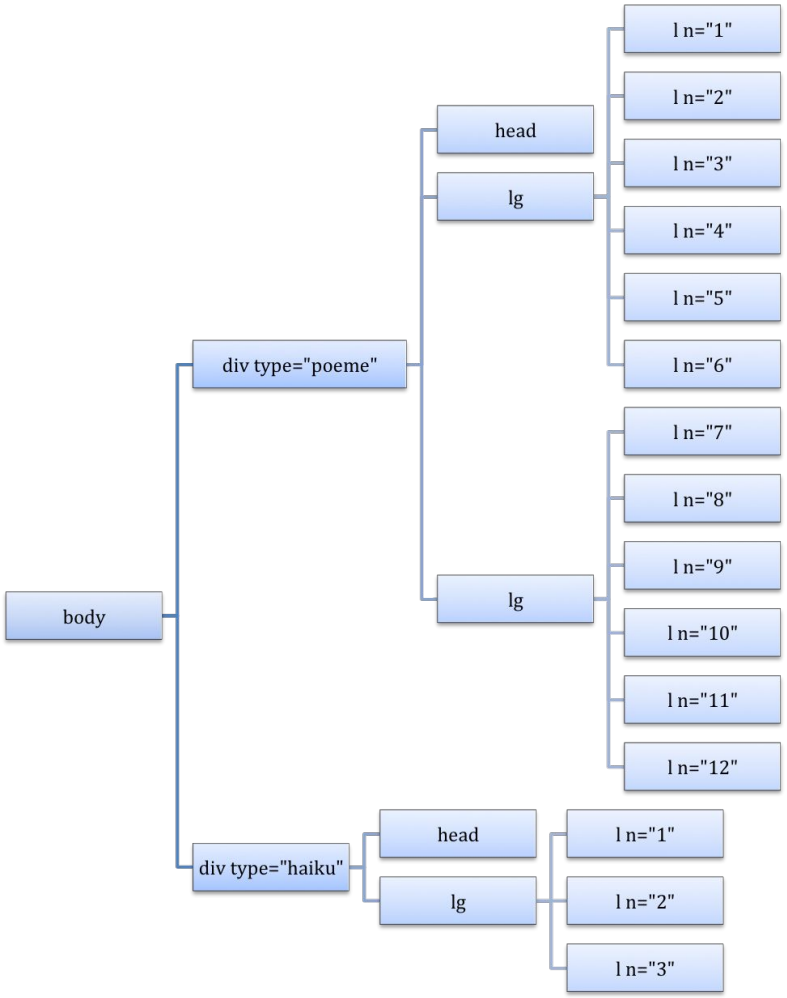
Titres de section ?



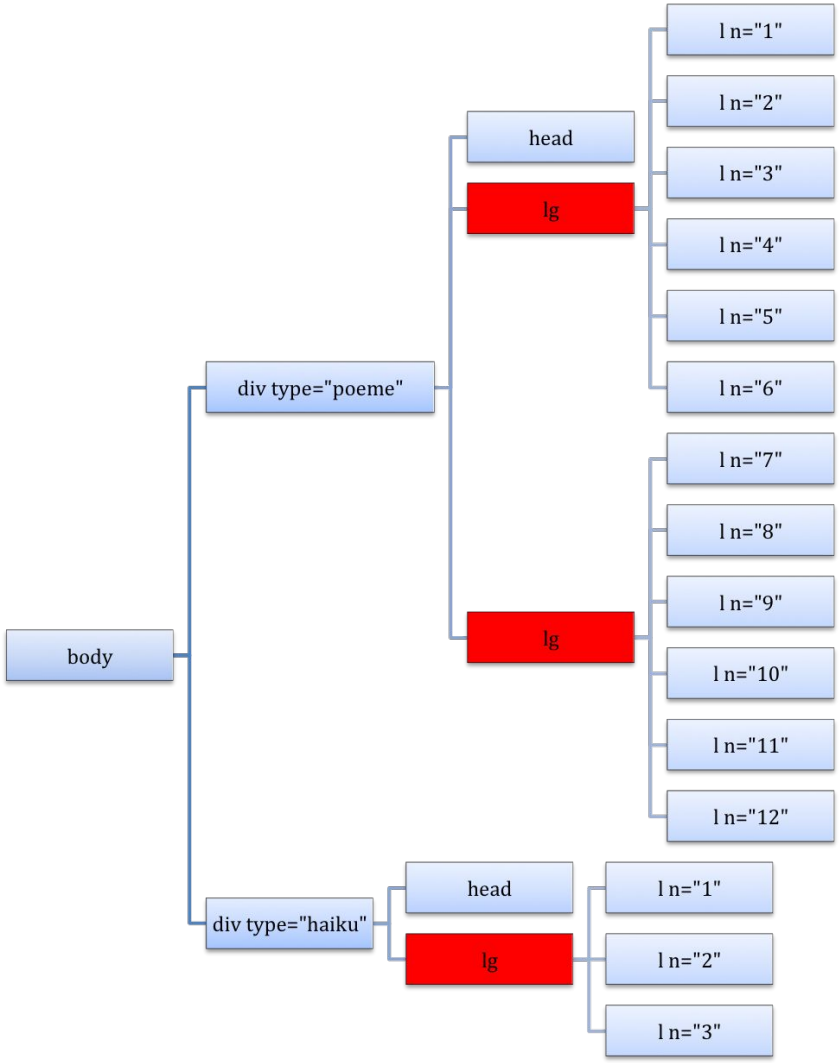
//body/div/head



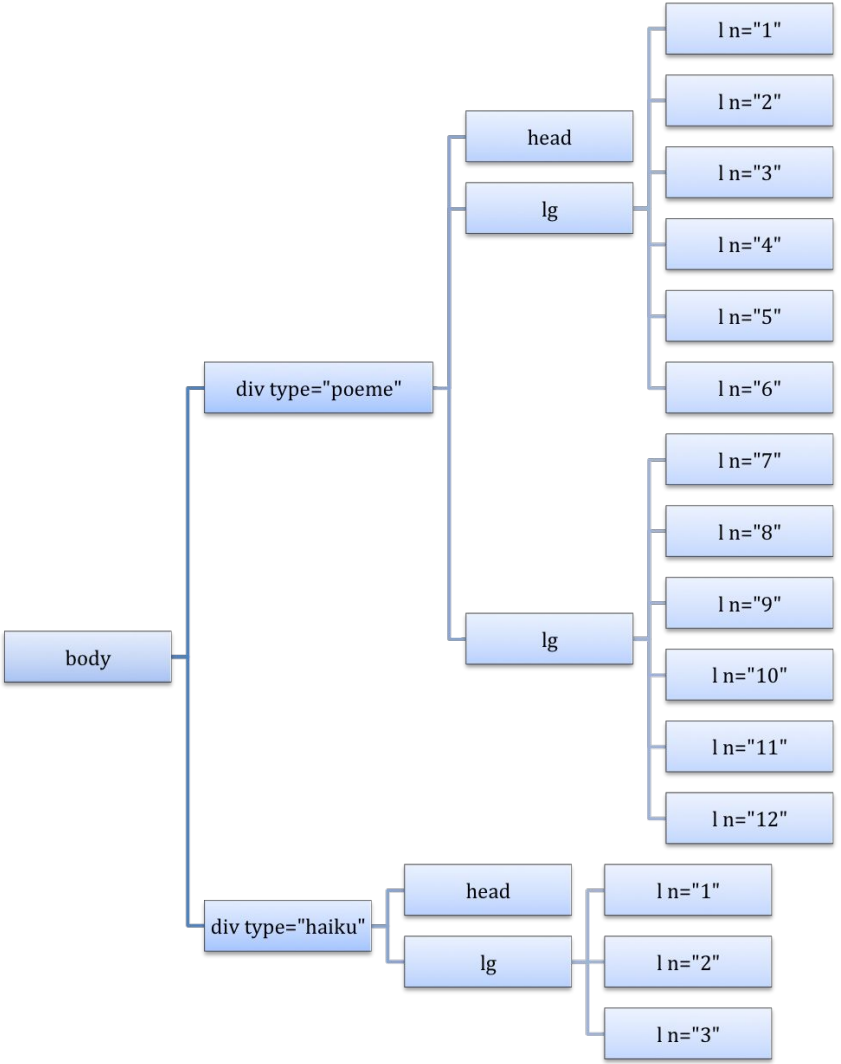
Strophes ?


body
div type="poeme"
head
lg
l n="1"
l n="2"
l n="3"
l n="4"
l n="5"
l n="6"
lg
l n="7"
l n="8"
l n="9"
l n="10"
l n="11"
l n="12"
div type="haiku"
head
lg
l n="1"
l n="2"
l n="3"

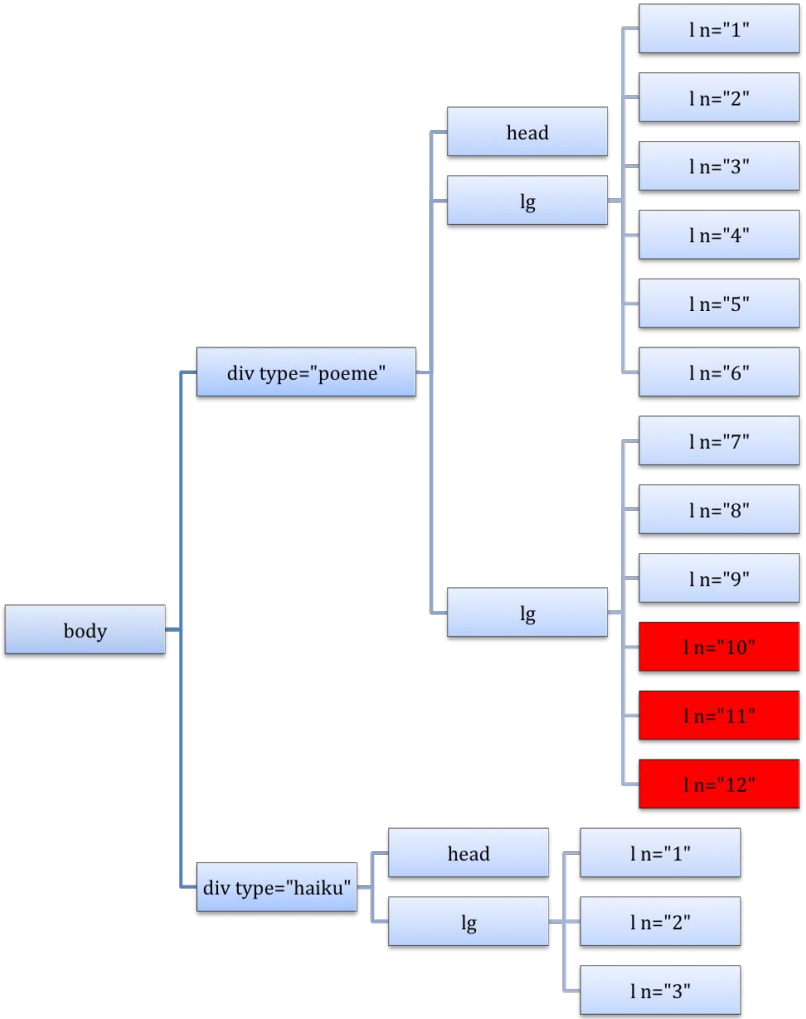
//body/div/lg


body
div type="poeme"
head
lg
l n="1"
l n="2"
l n="3"
l n="4"
l n="5"
l n="6"
lg
l n="7"
l n="8"
l n="9"
l n="10"
l n="11"
l n="12"
div type="haiku"
head
lg
l n="1"
l n="2"
l n="3"

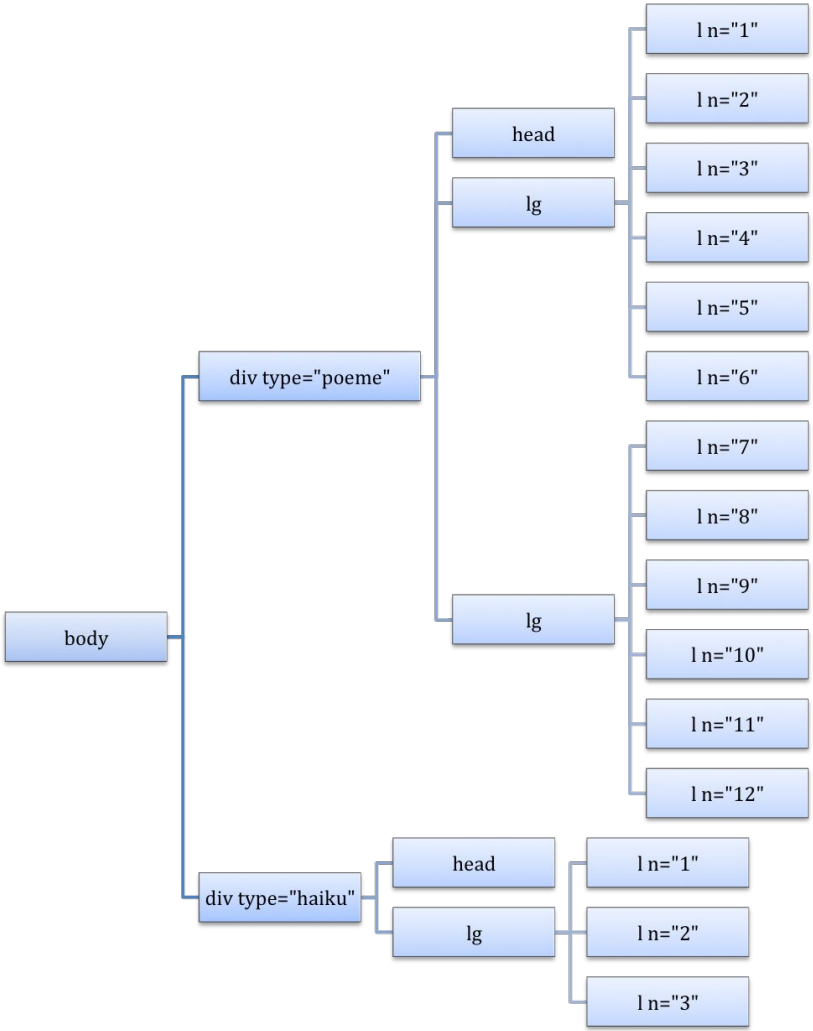
Tous les vers à partir du 9ème ?



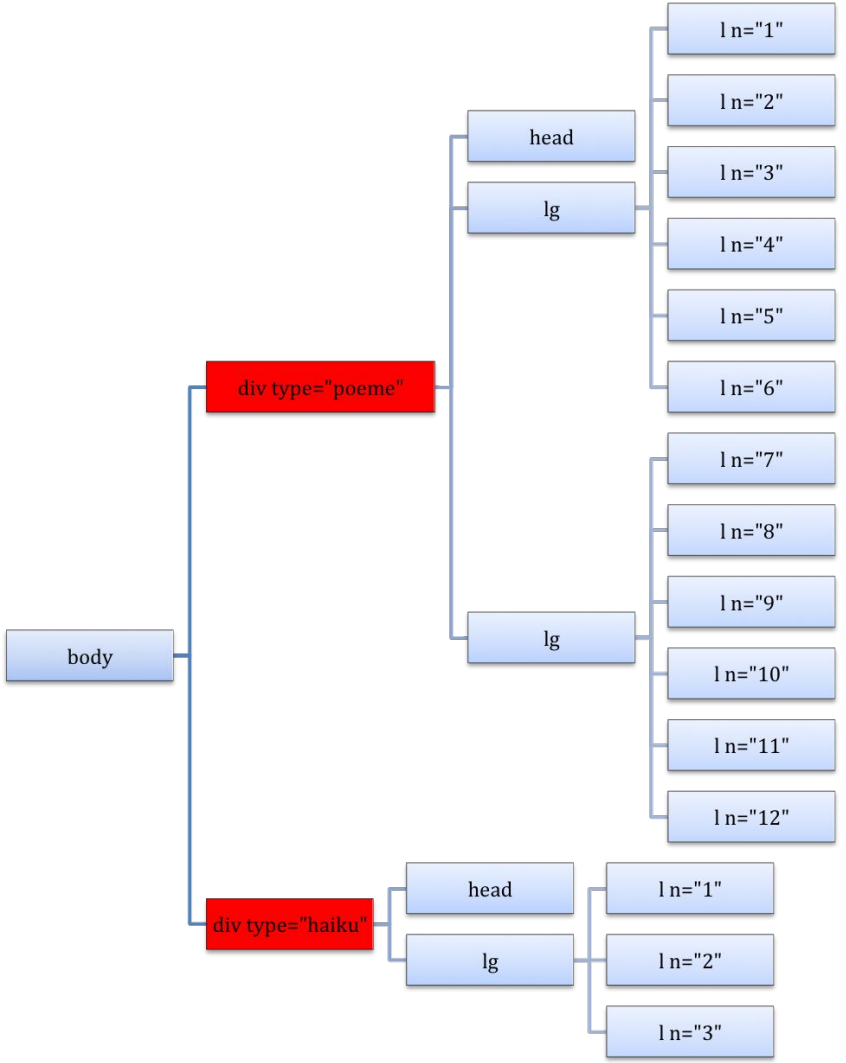
//l[@n > 9]


body
div type="poeme"
head
lg
l n="1"
l n="2"
l n="3"
l n="4"
l n="5"
l n="6"
lg
l n="7"
l n="8"
l n="9"
l n="10"
l n="11"
l n="12"
div type="haiku"
head
lg
l n="1"
l n="2"
l n="3"

Sections avec un type ?



//body/div/@type



XPATH

- un chemin de localisation retourne un node-set (ensemble non ordonné de noeuds)
- Les chemins peuvent être :
  - absolus : (/div/lg[1]/l)
  - relatifs (l/../../head)
- Syntaxe formelle : (axe::type[predicat])

  Exemple : child::div[contains(head, 'Chanson')]

XPATH : Pour aller plus loin

- https://docs.google.com/presentation/d/14HZwA3TmHf_y7RGNOOB4jLMTKMDWn2tFGo10I5OS0VI/edit?slide=id.i800#slide=id.i800
- https://newtfire.org/courses/tutorials/explainXPath.html
- http://dh.obdurodon.org/introduction-xpath.xhtml