Informatique pour le TAL 2 (SL02244X)
XML

Franck Sajous/CLLE-ERSS

29 mars 2012

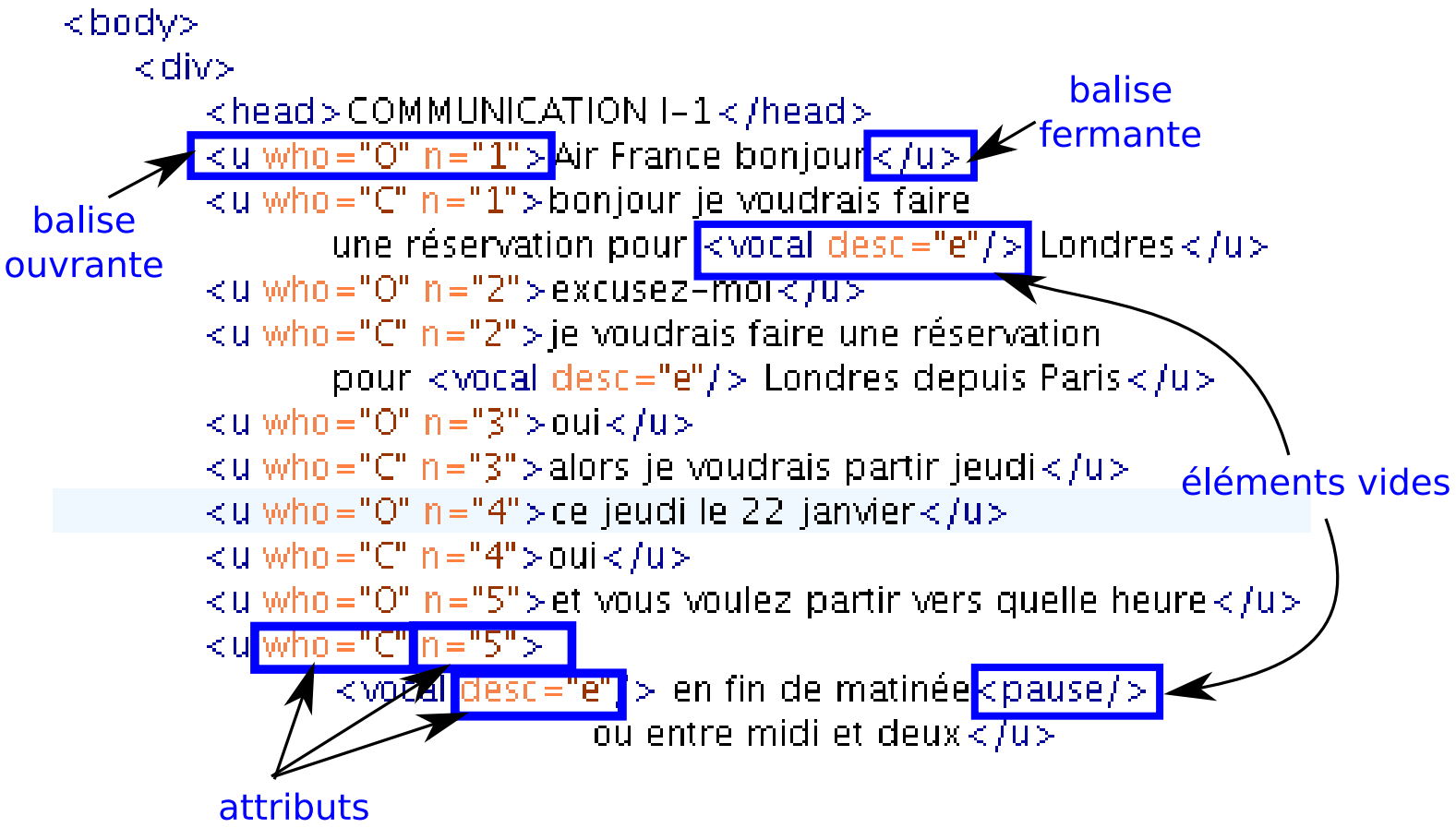
http://w3.erss.univ-tlse2.fr/membre/fsajous/

Description rapide

XML : eXtensible Markup Language

- XML est un langage à balises (comme HTML) ;
- ses balises ne sont pas prédéfinies (pas comme HTML) : c'est l'auteur d'un document qui crée ses balises ;
- sémantique des balises pas nécessairement transparente : elle dépend d'une convention entre auteur et utilisateur ;
- XML sert juste à représenter des données ayant une structure *arborescente* ;
- utilisé en TAL pour représenter des corpus annotés, des lexiques, etc., mais sert de support d'échange de données dans nombreux autres domaines.

1er exemple : corpus AirFrance



Commentaires

La balise est arbitraire. . .

Ici, la balise <u> signifie *utterance*. Dans un autre document, elle pourrait signifier *underline*, *unresolved anaphora*, *ungrammatical*, etc.

De même, un tour de parole pourrait être matérialisé par <speech>, <sequence>, etc.

Human readable VS program readable

L'exemple précédent est « lisible » par l'humain. Certains sont générés automatiquement ou semi-automatiquement et inexploitables « à l'oeil nu ».

```
<div type="level2">
  <head>III.3. Émission et réception</head>
  <p>Quel que soit le média de transmission, un émetteur convertit l'information
     en signal électrique, optique ou radioélectrique adapté au média, en le
     modulant et en l'amplifiant. </p>
  <p>La technique de ces fonctions d'interface est donc très dépendante du média,
     de la fréquence d'utilisation, et surtout de la puissance nécessaire pour
     compenser les pertes de propagation.</p>
  <p>Dans un canal de transmission hertzien, le signal porté par l'onde
     radioélectrique est atténué par la perte d'espace, les absorptions
```

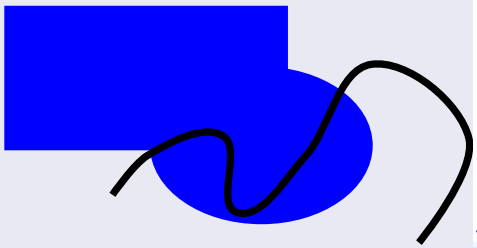
annotation

```
</unit>
<unit id="markmacr_1254503981">
  <metadata>
    <author>markmacr</author>
    <creation-date>1254503981</creation-date>
  </metadata>
  <characterisation>
    <type>paragraph</type>
    <featureSet/>
  </characterisation>
  <positioning>
    <start>
      <singlePosition index="708"/>
    </start>
    <end>
      <singlePosition index="1054"/>
    </end>
  </positioning>
</unit>
```

« Tout domaine » : pas seulement du texte annoté

SVG : Scalar Vector Graphics

Exemple de dessin vectoriel



```
<g inkscape:label="Calque 1" inkscape:groupmode="layer" id="layer1">
  <rect
    style="opacity:1;fill:#0000ff;fill-opacity:1;"
    id="rect8482" width="362.85715" height="180" x="68.571426" y="103.79076"/>
  <path sodipodi:type="arc"
    style="opacity:1;fill:#0000ff;"
    id="path8484" sodipodi:cx="401.42856" sodipodi:cy="282.36218" sodipodi:rx="141.42857"
    sodipodi:ry="98.571426"
    d="M 542.85713,282.36218 A 141.42857,98.571426 0 1 1 259.99998,282.36218 A 141.42857,98.571426 0 1 1
  <text xml:space="preserve" style="font-size:41.59119797px;font-style:normal;">
      <tspan sodipodi:role="line" id="tspan8488" x="82.810081" y="71.954323">Exemple de dessin vectoriel</tspan>
  </text>
</g>
```

GPX : GPS exchange format

```
<gpx>
  <wpt lat="39.921055008" lon="3.054223107">
    <ele>12.863281</ele>
    <time>2005-05-16T11:49:06Z</time>
    <name>Cala Sant Vicenç - Mallorca</name>
    <sym>City</sym>
  </wpt>
</gpx>
```

Autre exemple : flux RSS

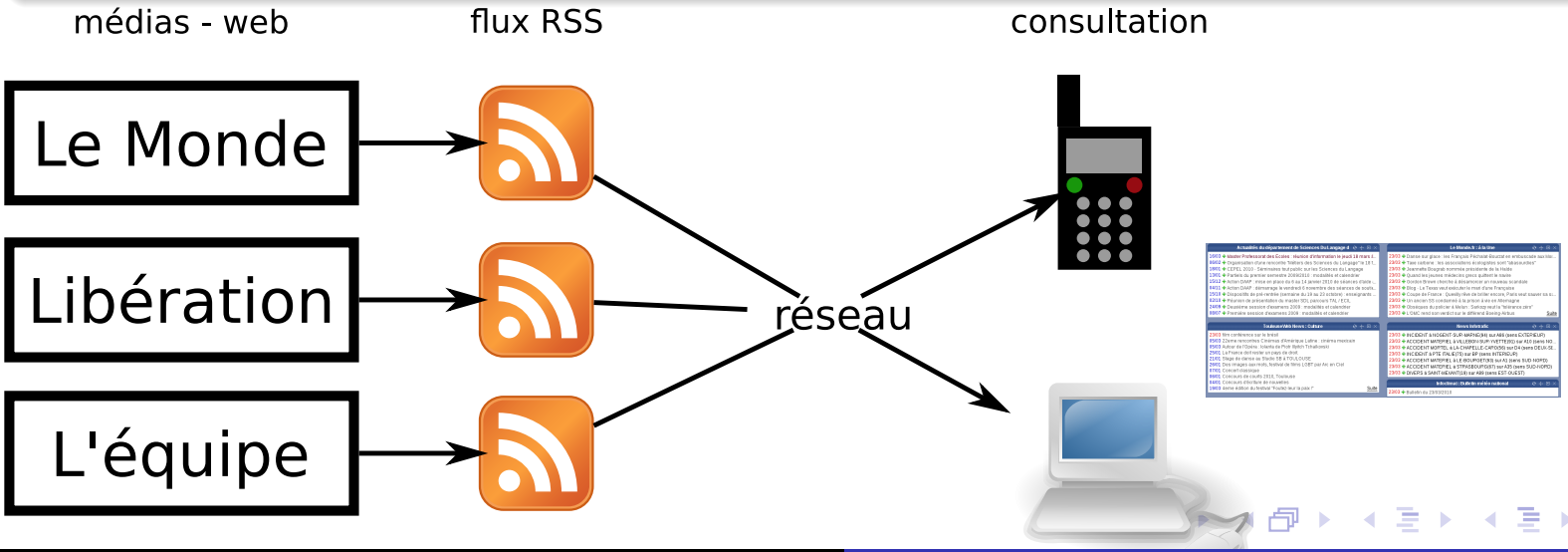
RSS

- format basé sur XML pour la « syndication » de contenu web ;
- plusieurs entités d'une même composante métier publient des résumés d'information selon le même format ;
- application (exemple) : web mobile (internet sur téléphone portable).

Autre exemple : flux RSS

RSS

- format basé sur XML pour la « syndication » de contenu web ;
- plusieurs entités d'une même composante métier publient des résumés d'information selon le même format ;
- application (exemple) : web mobile (internet sur téléphone portable).



RSS : format

Flux « Libération »

```
<channel>
  <title>Libération - La une</title>
  <link>http://www.liberation.fr/</link>
  <description>Actualites</description>
  <language>fr-fr</language>
  <pubDate>Wed, 17 Mar 2010 15:22:43 GMT</pubDate>
  <lastBuildDate>Wed, 17 Mar 2010 15:22:43 GMT</lastBuildDate>
  <ttl>2</ttl>
  <item>
    <title>Fabius veut une union à gauche «au-delà des régionales»</title>
    <link>http://rss.feedsportal.com/c/32268/f/438243/s/989a5cf/l/0L0S
    <description>... Et aussi: Boutin «horrifiée» par le débat sur l'identité nat
      Conseil d'Etat se prononce sur un dossier chaud dans la campagne als
      veut une minute de silence dans les bureaux de vote.</description>
    <category domain="">Politiques</category>
    <pubDate>Wed, 17 Mar 2010 15:22:34 GMT</pubDate>
    <guid>http://rss.feedsportal.com/c/32268/f/438243/s/989a5cf/l/0L0
  </item>
  <item>
    <title>Policier tué en Seine-et Marne: le gardé à vue poursuivi en Espagn
    <link>http://rss.feedsportal.com/c/32268/f/438243/s/989ad1e/l/0L0
    <description>Le membre présumé de l'ETA, interpellé mardi après la fu
      Dammarie-les-Lys, fait l'objet de poursuites judiciaires en Espagne pc
      de violences urbaines.
    <category domain="">Société</category>
    <pubDate>Wed, 17 Mar 2010 15:07:45 GMT</pubDate>
    <guid>http://rss.feedsportal.com/c/32268/f/438243/s/989ad1e/l/0L
  </item>
```

Flux « Le Monde »

```
<channel>
  <title>Le Monde.fr : à la Une</title>
  <link>http://www.lemonde.fr</link>
  <description>Toute l'actualité au moment de la connexion</description>
  <language>en</language>
  <copyright>Copyright Le Monde.fr</copyright>
  <pubDate>Wed, 17 Mar 2010 15:23:22 GMT</pubDate>
  <lastBuildDate>Wed, 17 Mar 2010 15:23:22 GMT</lastBuildDate>
  <ttl>2</ttl>
  <item>
    <title>DSK sceptique sur la création d'un fonds monétaire européen</title>
    <link>http://www.lemonde.fr/europe/article/2010/03/17/dsk-sceptique-su
    <description>Le directeur général du FMI estime que cette idée est "une distr
      rapport aux problèmes budgétaires urgents du moment que doit régler la
      Grèce</description>
    <pubDate>Wed, 17 Mar 2010 15:18:59 GMT</pubDate>
    <guid isPermaLink="false">http://www.lemonde.fr/europe/article/2010/03/
  </item>
  <item>
    <title>Italie : quatre banques envoyées en correctionnelle pour fraude présur
    <link>http://www.lemonde.fr/europe/article/2010/03/17/italie-quatre-ban
    <description>Ces banques auraient notamment caché à la commune de Milar
      présentés par les produits financiers dérivés qu'elles ont émis dans le cadr
      restructuration de la dette de la municipalité.</description>
    <pubDate>Wed, 17 Mar 2010 15:14:13 GMT</pubDate>
    <guid isPermaLink="false">http://www.lemonde.fr/europe/article/2010/03/
  </item>
```

RSS : format

Flux « Libération »

```
<channel>
  <title>Libération - La une</title>
  <link>http://www.liberation.fr/</link>
  <description>Actualites</description>
  <language>fr-fr</language>
  <pubDate>Wed, 17 Mar 2010 15:22:43 GMT</pubDate>
  <lastBuildDate>Wed, 17 Mar 2010 15:22:43 GMT</lastBuildDate>
  <ttl>2</ttl>
  <item>
    <title>Fabius veut une union à gauche «au-delà des régionales»</title>
    <link>http://rss.feedsportal.com/c/32268/f/438243/s/989a5cf/l/0L0S
    <description>... Et aussi: Boutin «horrifiée» par le débat sur l'identité nat
      Conseil d'Etat se prononce sur un dossier chaud dans la campagne als
      veut une minute de silence dans les bureaux de vote.</description>
    <category domain="">Politiques</category>
    <pubDate>Wed, 17 Mar 2010 15:22:34 GMT</pubDate>
    <guid>http://rss.feedsportal.com/c/32268/f/438243/s/989a5cf/l/0L0
  </item>
  <item>
    <title>Policier tué en Seine-et Marne: le gardé à vue poursuivi en Espagn
    <link>http://rss.feedsportal.com/c/32268/f/438243/s/989ad1e/l/0L0
    <description>Le membre présumé de l'ETA, interpellé mardi après la fu
      Dammarie-les-Lys, fait l'objet de poursuites judiciaires en Espagne pc
      de violences urbaines.
    <category domain="">Société</category>
    <pubDate>Wed, 17 Mar 2010 15:07:45 GMT</pubDate>
    <guid>http://rss.feedsportal.com/c/32268/f/438243/s/989ad1e/l/0L
  </item>
```

Flux « Le Monde »

```
<channel>
  <title>Le Monde.fr : à la Une</title>
  <link>http://www.lemonde.fr</link>
  <description>Toute l'actualité au moment de la connexion</description>
  <language>en</language>
  <copyright>Copyright Le Monde.fr</copyright>
  <pubDate>Wed, 17 Mar 2010 15:23:22 GMT</pubDate>
  <lastBuildDate>Wed, 17 Mar 2010 15:23:22 GMT</lastBuildDate>
  <ttl>2</ttl>
  <item>
    <title>DSK sceptique sur la création d'un fonds monétaire européen</title>
    <link>http://www.lemonde.fr/europe/article/2010/03/17/dsk-sceptique-su
    <description>Le directeur général du FMI estime que cette idée est "une distr
      rapport aux problèmes budgétaires urgents du moment que doit régler la
      Grèce</description>
    <pubDate>Wed, 17 Mar 2010 15:18:59 GMT</pubDate>
    <guid isPermaLink="false">http://www.lemonde.fr/europe/article/2010/03/
  </item>
  <item>
    <title>Italie : quatre banques envoyées en correctionnelle pour fraude présur
    <link>http://www.lemonde.fr/europe/article/2010/03/17/italie-quatre-ban
    <description>Ces banques auraient notamment caché à la commune de Milar
      présentés par les produits financiers dérivés qu'elles ont émis dans le cadr
      restructuration de la dette de la municipalité.</description>
    <pubDate>Wed, 17 Mar 2010 15:14:13 GMT</pubDate>
    <guid isPermaLink="false">http://www.lemonde.fr/europe/article/2010/03/
  </item>
```

sources différentes, contenu différent, mais même format

RSS : agrégateurs

Agrégateur en ligne : http://rssnewsbox.com

Flux (sources)

- http://w3.sc-du-langage.univ-tlse2.fr/blog/rss.php
- http://www.toulouseweb.com/news-CULT-culture.rss
- http://www.infotrafic.fr/rss_infotrafic.php
- http://www.lemonde.fr/rss/une.xml
- http://www.infoclimat.fr/rss/flux_rss.opml

Cas d'utilisation

- annotation de corpus : à partir d'un texte, production d'un document enrichi
  - découpage structurel (sections, paragraphes, etc.) ;
  - annotations automatiques ou manuelles : (repérage d'entités nommées, marquage d'anaphores, etc.) ;
- « sorties d'outils », résultats d'analyses (morpho-syntaxiques, syntaxiques, etc.)
- données formalisées : lexiques, bases de données, etc.
- métadonnées.

Existence de standards

Pas de balises prédéfinies : précisions

C'est l'auteur qui définit ses balises. . .
. . . mais pour un usage donné, il existe souvent des standards.

- méta-données : Dublin Core, TEI, RDF
- encodage de corpus/lexiques : TEI $\rightarrow$ structure (paragraphes, sections, titres, etc.), entités (dates, mesures, personnes, adresses, etc.), poésie (vers, analyse métrique, etc.), oral retranscrit (tours de parole, pauses, etc.), dictionnaires, etc.
- annotation d'expressions temporelles et d'événements : TimeML (Time Markup Language) ;
- ontologies : OWL (Web Ontologies language)

Morphalou (ATILF) : lexique de formes fléchies

http://www.cnrtl.fr/lexiques/morphalou/

```
<lexicalEntry id="abaissable_1">
    <formSet>
        <lemmatizedForm>
            <orthography>abaissable</orthography>
            <grammaticalCategory>adjective</grammaticalCategory>
        </lemmatizedForm>
        <inflectedForm>
            <orthography>abaissable</orthography>
            <grammaticalNumber>singular</grammaticalNumber>
            <grammaticalGender>masculine</grammaticalGender>
        </inflectedForm>
        <inflectedForm>
            <orthography>abaissables</orthography>
            <grammaticalNumber>plural</grammaticalNumber>
            <grammaticalGender>masculine</grammaticalGender>
        </inflectedForm>
    </formSet>
    <originatingEntry target="TLF">ABAISSABLE, adj.</originatingEntry>
</lexicalEntry>
<lexicalEntry id="abaissant_1">
    <formSet>
        <lemmatizedForm>
            <orthography>abaissant</orthography>
            <grammaticalCategory>adjective</grammaticalCategory>
        </lemmatizedForm>
        <inflectedForm>
            <orthography>abaissant</orthography>
            <grammaticalNumber>singular</grammaticalNumber>
            <grammaticalGender>masculine</grammaticalGender>
        </inflectedForm>
        <inflectedForm>
            <orthography>abaissants</orthography>
            <grammaticalNumber>plural</grammaticalNumber>
            <grammaticalGender>masculine</grammaticalGender>
        </inflectedForm>
    </formSet>
    <originatingEntry target="TLF">ABAISSANT, ANTE, adj.</originatingEntry>
</lexicalEntry>
```

Bonne formation

Un document est *bien formé* si :

- en-tête correct : `<?xml version='1.0' encoding='UTF-8'?>`

Bonne formation

Un document est *bien formé* si :

- en-tête correct : `<?xml version='1.0' encoding='UTF-8'?>`
- à chaque balise ouvrante correspond une balise fermante ;

Bonne formation

Un document est *bien formé* si :

- en-tête correct : `<?xml version='1.0' encoding='UTF-8'?>`
- à chaque balise ouvrante correspond une balise fermante ;
- balises vides terminées par un / : `<br/>`, `<pause/>`

Bonne formation

Un document est *bien formé* si :

- en-tête correct : `<?xml version='1.0' encoding='UTF-8'?>`
- à chaque balise ouvrante correspond une balise fermante ;
- balises vides terminées par un / : `<br/>`, `<pause/>`
- attributs possibles pour les balises ouvrantes et vides, pas pour les fermantes : `<u who='bob'> ... </u>`, `<vocal desc='e'/>`
  ~~`<u>...</u who='bob'>`~~

Bonne formation

Un document est *bien formé* si :

- en-tête correct : `<?xml version='1.0' encoding='UTF-8'?>`
- à chaque balise ouvrante correspond une balise fermante ;
- balises vides terminées par un / : `<br/>`, `<pause/>`
- attributs possibles pour les balises ouvrantes et vides, pas pour les fermantes : `<u who='bob'> ... </u>`, `<vocal desc='e'/>` ~~`<u>...</u who='bob'>`~~
- deux attributs d'une balise donnée ne peuvent avoir le même nom : ~~`<u who='bob' who='max'>...</u>`~~

Bonne formation

Un document est *bien formé* si :

- en-tête correct : `<?xml version='1.0' encoding='UTF-8'?>`
- à chaque balise ouvrante correspond une balise fermante ;
- balises vides terminées par un / : `<br/>`, `<pause/>`
- attributs possibles pour les balises ouvrantes et vides, pas pour les fermantes : `<u who='bob'> ... </u>`, `<vocal desc='e'/>` ~~`<u>...</u who='bob'>`~~
- deux attributs d'une balise donnée ne peuvent avoir le même nom : ~~`<u who='bob' who='max'>...</u>`~~
- les valeurs des attributs sont encadrées par des quotes (simples : ' ou doubles : ").

Bonne formation

Un document est *bien formé* si :

- en-tête correct : `<?xml version='1.0' encoding='UTF-8'?>`
- à chaque balise ouvrante correspond une balise fermante ;
- balises vides terminées par un / : `<br/>`, `<pause/>`
- attributs possibles pour les balises ouvrantes et vides, pas pour les fermantes : `<u who='bob'> ... </u>`, `<vocal desc='e'/>` ~~`<u>...</u who='bob'>`~~
- deux attributs d'une balise donnée ne peuvent avoir le même nom : ~~`<u who='bob' who='max'>...</u>`~~
- les valeurs des attributs sont encadrées par des quotes (simples : ' ou doubles : ").
- pas de chevauchment des balises, imbrication possible (structure arborescente) : ~~`<p><b></p></b>`~~. Correct : `<p><b></b></p>`, `<div><div></div></div>`

Bonne formation

Un document est *bien formé* si :

- en-tête correct : `<?xml version='1.0' encoding='UTF-8'?>`
- à chaque balise ouvrante correspond une balise fermante ;
- balises vides terminées par un / : `<br/>`, `<pause/>`
- attributs possibles pour les balises ouvrantes et vides, pas pour les fermantes : `<u who='bob'> ... </u>`, `<vocal desc='e'/>` ~~`<u>...</u who='bob'>`~~
- deux attributs d'une balise donnée ne peuvent avoir le même nom : ~~`<u who='bob' who='max'>...</u>`~~
- les valeurs des attributs sont encadrées par des quotes (simples : ' ou doubles : ").
- pas de chevauchment des balises, imbrication possible (structure arborescente) : ~~`<p><b></p></b>`~~. Correct : `<p><b></b></p>`, `<div><div></div></div>`
- un et un seul élément racine (balise englobante)

Bonne formation

Un document est *bien formé* si :

- en-tête correct : `<?xml version='1.0' encoding='UTF-8'?>`
- à chaque balise ouvrante correspond une balise fermante ;
- balises vides terminées par un / : `<br/>`, `<pause/>`
- attributs possibles pour les balises ouvrantes et vides, pas pour les fermantes : `<u who='bob'> ... </u>`, `<vocal desc='e'/>` ~~`<u>...</u who='bob'>`~~
- deux attributs d'une balise donnée ne peuvent avoir le même nom : ~~`<u who='bob' who='max'>...</u>`~~
- les valeurs des attributs sont encadrées par des quotes (simples : ' ou doubles : ").
- pas de chevauchment des balises, imbrication possible (structure arborescente) : ~~`<p><b></p></b>`~~. Correct : `<p><b></b></p>`, `<div><div></div></div>`
- un et un seul élément racine (balise englobante)

Un document DOIT être bien formé pour être manipulé par un programme.

Structure arborescente

```
<body>
  <div type="level1">
    <head>I. Généralités</head>
    <div type="level2">
      <head>I.1. Étymologie</head>
      <p>Le mot télécommunications vient du préfixe grec tele- (XXXX-), signifiant
        loin, et du latin communicare, signifiant partager.</p>
    </div>
    <div type="level2">
      <head>I.2. Définition</head>
      <p>Les télécommunications (abrév. fam. télécoms), sont considérées comme des
        technologies et techniques appliquées et non comme une science. </p>
      <p>On entend par télécommunications toute transmission, émission et réception à
        distance, de signes, de signaux, d'écrits, d'images, de sons ou de
        renseignements de toutes natures, par fil électrique, radioélectricité,
        liaison optique, ou autres systèmes électromagnétiques. </p>
    </div>
  </div>
  <div type="level1">
    ...
  </div>
  ...
</body>
```