

# 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

```
<body>
  <div>
    <head>COMMUNICATION I-1</head>
    <u who="O" n="1">Air France bonjour</u>
    <u who="C" n="1">bonjour je voudrais faire
      une réservation pour <vocal desc="e"/> Londres</u>
    <u who="O" n="2">excusez-moi</u>
    <u who="C" n="2">je voudrais faire une réservation
      pour <vocal desc="e"/> Londres depuis Paris</u>
    <u who="O" n="3">oui</u>
    <u who="C" n="3">alors je voudrais partir jeudi</u>
    <u who="O" n="4">ce jeudi le 22 janvier</u>
    <u who="C" n="4">oui</u>
    <u who="O" n="5">et vous voulez partir vers quelle heure</u>
    <u who="C" n="5">
      <vocal desc="e"> en fin de matinée <pause/>
      ou entre midi et deux</u>
  </div>
</body>
```

Diagram annotations:

- balise ouvrante**: points to the opening tag `<u who="O" n="1">`.
- balise fermante**: points to the closing tag `</u>`.
- attributs**: points to the attributes `who="O" n="1"`.
- éléments vides**: points to the empty element `<vocal desc="e"/>`.
- pause**: points to the `<pause/>` element.

# 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 inexploitable « 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
```



```
</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>
<unit id="markmacr_1254503776">
```

## « 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

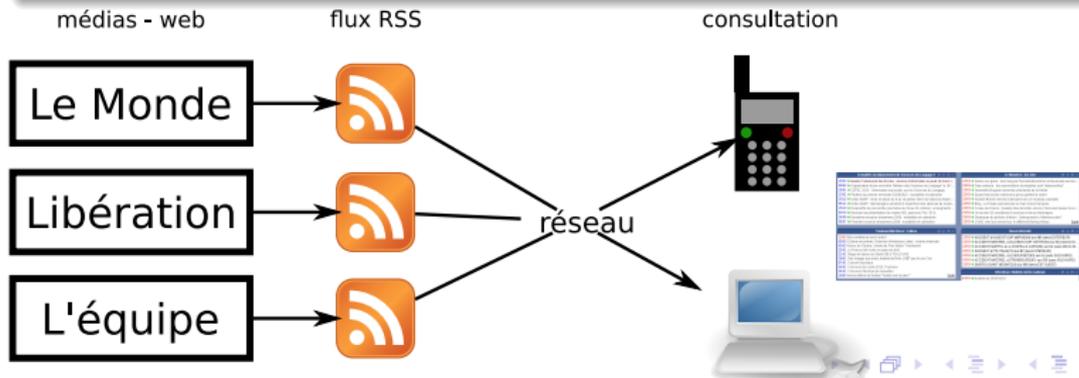
### RSS

- 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/0L05
<description>... Et aussi: Boutin «horrrifée» 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 Espagne
<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 pt
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/0L0
</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-s
<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 Mila
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/0L00</description>... Et aussi: Boutin «horrrifié» 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/0L00</item>
<item>
<title>Policier tué en Seine-et-Marne: le gardé à vue poursuivi en Espagne</link>http://rss.feedsportal.com/c/32268/f/438243/s/989ad1e/l/0L00</description>Le membre présumé de l'ETA, interpellé mardi après la fu
Dammarie-les-Lys, fait l'objet de poursuites judiciaires en Espagne pct
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/0L00</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-s</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 Milaz
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>

<p><b>Actualités du département de Sciences Du Langage d</b>    </p> <p>16/03  Master Professorat des Ecoles : réunion d'information le jeudi 18 mars é...</p> <p>06/02  Organisation d'une rencontre "Métiers des Sciences du Langage" le 18 f...</p> <p>18/01  CEPEL 2010 - Séminaires tout public sur les Sciences du Langage</p> <p>13/01  Partiels du premier semestre 2009/2010 : modalités et calendrier</p> <p>15/12  Action DAAP : mise en place du 6 au 14 janvier 2010 de séances d'aide t...</p> <p>04/11  Action DAAP : démarrage le vendredi 6 novembre des séances de soutie...</p> <p>15/10  Dispositifs de pré-rentree (semaine du 19 au 23 octobre) : enseignants ...</p> <p>02/10  Réunion de présentation du master SDL parcours TAL / ECIL</p> <p>24/09  Deuxième session d'examens 2009 : modalités et calendrier</p> <p>08/07  Première session d'examens 2009 : modalités et calendrier</p> <p style="text-align: right;"><a href="#">Suite</a></p>	<p><b>Le Monde.fr : à la Une</b>    </p> <p>23/03  Danse sur glace : les Français Péchalat-Bourzat en ambassade aux Mor...</p> <p>23/03  Taxe carbone : les associations écologistes sont "abasourdies"</p> <p>23/03  Jeannette Bougrab nommée présidente de la Halde</p> <p>23/03  Quand les jeunes médecins grecs quittent le navire</p> <p>23/03  Gordon Brown cherche à désamorcer un nouveau scandale</p> <p>23/03  Blog - Le Texas veut exécuter le mari d'une Française</p> <p>23/03  Coupe de France : Quevilly rêve de briller encore, Paris veut sauver sa st...</p> <p>23/03  Un ancien SS condamné à la prison à vie en Allemagne</p> <p>23/03  Obsèques du policier à Melun : Sarkozy veut la "tolérance zéro"</p> <p>23/03  L'OMC rend son verdict sur le différend Boeing-Airbus</p> <p style="text-align: right;"><a href="#">Suite</a></p>
<p><b>ToulouseWeb News : Culture</b>    </p> <p>23/03 film conférence sur le Brésil</p> <p>05/03 22eme rencontres Cinémas d'Amérique Latine : cinéma mexicain</p> <p>05/03 Autour de l'Opéra : Iolanta de Piotr Illyitch Tchaïkovski</p> <p>25/01 La France doit rester un pays de droit.</p> <p>21/01 Stage de danse au Studio SB à TOULOUSE</p> <p>26/01 Des images aux mots, festival de films LGBT par Arc en Ciel</p> <p>07/01 Concert classique</p> <p>06/01 Concours de courts 2010, Toulouse</p> <p>04/01 Concours d'écriture de nouvelles</p> <p>19/03 4eme édition du festival "Foutez-leur la pak !"</p> <p style="text-align: right;"><a href="#">Suite</a></p>	<p><b>News Infotrafic</b>    </p> <p>23/03  INCIDENT à NOGENT-SUR-MARNE(94) sur A86 (sens EXTERIEUR)</p> <p>23/03  ACCIDENT MATERIEL à VILLEBON-SUR-YVETTE(91) sur A10 (sens NO...</p> <p>23/03  ACCIDENT MORTEL à LA-CHAPELLE-CAPO(56) sur D4 (sens DEUX-SE...</p> <p>23/03  INCIDENT à PTE ITALIE(75) sur BP (sens INTERIEUR)</p> <p>23/03  ACCIDENT MATERIEL à LE-BOURGET(93) sur A1 (sens SUD-NORD)</p> <p>23/03  ACCIDENT MATERIEL à STRASBOURG(67) sur A35 (sens SUD-NORD)</p> <p>23/03  DIVERS à SAINT-MEXANT(19) sur A89 (sens EST-OUEST)</p> <p><b>Infoclimat : Bulletin météo national</b>    </p> <p>23/03  Bulletin du 23/03/2010</p>

## 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.infotrafic.fr/rss_infotrafic.php)
- <http://www.lemonde.fr/rss/une.xml>
- [http://www.infoclimat.fr/rss/flux\\_rss.opml](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 → 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 chevauchement 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 chevauchement 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 chevauchement 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- (0000-), 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>

```

