

Conversion d'inventaires MS WORD en données structurées

Vincent Bezençon
Fondation Jean Monnet

Point de départ

DOSSIER 36

LA CONFERENCE DE MESSINE

SOUS-DOSSIER 36/4

Comité intergouvernemental: La question des transports et travaux publics
Commission des Transports

Liste des pièces

- AMK 36/4/1 “La coopération européenne dans le domaine des transports”. 15.07.55.
- AMK 36/4/2 “Les modalités de la coopération européenne dans le domaine des transports”. 15.07.55.
- AMK 36/4/3 “Note de la délégation française. Transports”. 18.07.55. Document N°12.
- AMK 36/4/4 Commission des Transports et des Travaux Publics. Sommaire N°1. Réunion du 20 juillet 1955. 20.07.55. Document N°31.
- AMK 36/4/5 “Commission des Transports et des Travaux publics. Note de la délégation française. Programme de travail de la Commission”. 21.07.55. Document N° 36.
- AMK 36/4/6 “Commission des transports et des travaux publics. Groupe de travail infrastructure et investissements”. Note relative à l’infrastructure et aux investissements”. 5.08.55. Document N°130.

Objectif

```
2 <c level="otherlevel" otherlevel="dossier" id="AMK-036-004">
3   <did>
4     <unittitle>Comité intergouvernemental: La question des transports et travaux publics Commission des Transports</unittitle>
5     <unitid label="Cote">AMK 36/4</unitid>
6     <unitdatestructured>
7       <daterange>
8         <fromdate standarddate="1955-07-15">15 juillet 1955</fromdate>
9         <todate standarddate="1955-08-05">5 août 1955</todate>
10      </daterange>
11    </unitdatestructured>
12    <physdescstructured physdescstructuredtype="materialtype" coverage="whole">
13      <quantity>6</quantity>
14      <unittype>Pièces</unittype>
15    </physdescstructured>
16  </did>
17  <c level="otherlevel" otherlevel="pièce" id="AMK-036-004-001">
18    <did>
19      <unittitle>"La coopération européenne dans le domaine des transports".</unittitle>
20      <unitid label="Cote">AMK 36/4/1</unitid>
21      <unitdatestructured>
22        <datesingle standarddate="1955-07-15">15 juillet 1955</datesingle>
23      </unitdatestructured>
24      <physdescstructured physdescstructuredtype="materialtype" coverage="whole">
25        <quantity>1</quantity>
26        <unittype>Pièce</unittype>
27      </physdescstructured>
28    </did>
29  </c>
30  <c level="otherlevel" otherlevel="pièce" id="AMK-036-004-002">
31    <did>
32      <unittitle>"Les modalités de la coopération européenne dans le domaine destransports".</unittitle>
33      <unitid label="Cote">AMK 36/4/2</unitid>
34      <unitdatestructured>
35        <datesingle standarddate="1955-07-15">15 juillet 1955</datesingle>
36      </unitdatestructured>
37      <physdescstructured physdescstructuredtype="materialtype" coverage="whole">
38        <quantity>1</quantity>
39        <unittype>Pièce</unittype>
40      </physdescstructured>
41    </did>
42  </c>
```

Stratégie

MS
WORD

Reconnaissance de structure
(expressions régulières)

XML
simple

Feuille de
style XSL

XML
EAD3

DOSSIER 36	LA CONFERENCE DE MESSINE
	SOUS-DOSSIER 36/4
	<u>Comité intergouvernemental: La question des transports</u> <u>Commission des Transports</u>
	Liste des pièces
AMK 36/4/1	“La coopération européenne dans le domaine des tr
AMK 36/4/2	“Les modalités de la coopération européenne dans 15.07.55.
AMK 36/4/3	“Note de la délégation française. Transports”. 18.0
AMK 36/4/4	Commission des Transports et des Travaux Publics 20 juillet 1955. 20.07.55. Document N°31.
AMK 36/4/5	“Commission des Transports et des Travaux public

```
<ud>
  <niveau>dossier</niveau>
  <cote>AMK 36/4</cote>
  <intitule>Comité intergouvernemental: La question d
  <ud>
    <niveau>pièce</niveau>
    <cote>AMK 36/4/1</cote>
    <intitule>“La coopération européenne dans le do
    <dates>1955-07-15</dates>
  </ud>
  <ud>
    <niveau>pièce</niveau>
    <cote>AMK 36/4/2</cote>
    <intitule>“Les modalités de la coopération eur
    <dates>1955-07-15</dates>
  </ud>
  <ud>
    <niveau>pièce</niveau>
    <cote>AMK 36/4/3</cote>
    <intitule>“Note de la délégation française. Tra
    <dates>1955-07-18</dates>
    <notes>Document N°12.</notes>
  </ud>
</ud>
```

```
<c level="otherlevel" otherlevel="dossier" id="AMK-036-004">
  <did>
    <unittitle>Comité intergouvernemental: La questio
    <unitid label="Cote">AMK 36/4</unitid>
    <unitdatestructured>
      <daterange>
        <fromdate standarddate="1955-07-15">15 juil
        <todate standarddate="1955-08-05">5 août 1955
      </daterange>
    </unitdatestructured>
    <physdescstructured physdescstructuredtype="mater
      <quantity>6</quantity>
      <unittype>Pièces</unittype>
    </physdescstructured>
  </did>
  <c level="otherlevel" otherlevel="pièce" id="AMK-03
    <did>
      <unittitle>“La coopération européenne dans le d
      <unitid label="Cote">AMK 36/4/1</unitid>
      <unitdatestructured>
        <datesingle standarddate="1955-07-15">15 juil
      </unitdatestructured>
      <physdescstructured physdescstructuredtype="mat
        <quantity>1</quantity>
        <unittype>Pièce</unittype>
      </physdescstructured>
    </did>
  </c>
</c level="otherlevel" otherlevel="pièce" id="AMK-03
```

Expressions régulières ?

(regex de leur petit nom)

^ Start of string, or start of line in multi-line pattern	* 0 or more {3} Exactly 3	. Any character except new line (\n)
\A Start of string	+ 1 or more {3,} 3 or more	(a b) a or b
\$ End of string, or end of line in multi-line pattern	? 0 or 1 {3,5} 3, 4 or 5	(...) Group
\Z End of string	Add a ? to a quantifier to make it ungreedy.	(?...) Passive (non-capturing) group
\b Word boundary	Escape Sequences	[abc] Range (a or b or c)
\B Not word boundary	\ Escape following character	[^abc] Not (a or b or c)
\< Start of word	\Q Begin literal sequence	[a-q] Lower case letter from a to q
\> End of word	\E End literal sequence	[A-Q] Upper case letter from A to Q
Character Classes	"Escaping" is a way of treating characters which have a special meaning in regular expressions literally, rather than as special characters.	[0-7] Digit from 0 to 7
\c Control character	Common Metacharacters	\x Group/subpattern number "x"
\s White space	^ [. \$	Ranges are inclusive.
\S Not white space	{ * (\	Pattern Modifiers
\d Digit	+) ?	g Global match
\D Not digit	< >	i * Case-insensitive
\w Word	The escape character is usually \	m * Multiple lines
\W Not word	Special Characters	s * Treat string as single line
\x Hexadecimal digit	\n New line	x * Allow comments and whitespace in pattern
\O Octal digit		e * Evaluate replacement
POSIX		U * Ungreedy pattern
		* PCRE modifier

Expressions régulières ?

Un exemple

Repérer les dates comme ça :

24.05.82

Pour obtenir ça :

`<date>1982-05-24</date>`

On recherche :

`([0-9]{2})\.[0-9]{2}\.[0-9]{2}`

On remplace par :

`<date>19$3-$2-$1</date>`

Marche à suivre

10. Balise ud (+ en profiter pour corriger les lignes composées d'une seule cote)

```
(^AMF.*)  
<ud>$1</ud>
```

11. Balise cote

```
^<ud>(AMF \d+/\d+/\d+\.?(\w+)?\.\.?,'?~? ?\w?) ? : ?  
<ud><cote>$1</cote>
```

12. Corriger les chiffres mal ocr-isés dans les dates

```
(\d*\.\.? \d+ \.\.? \d*)1(\d*\.\.? \d*\.\.? \d*)  
$11$2  
  
(\d*\.\.? \d*\.\.? \d*)1(\d*\.\.? \d+ \.\.? \d*)  
$11$2  
  
(\d*\.\.? \d*\.\.? \d*)[o0](\d*\.\.? \d+ \.\.? \d*)  
$10$2  
  
(\d*\.\.? \d+ \.\.? \d*)[o0](\d*\.\.? \d*\.\.? \d*)  
$10$2
```

13. Mettre les dates au bon format

```
(\d{4}|\d{1,2})(\.\.|\.)(\d{1,2})(\.\.|\.)(\d{4}|\d{1,2})  
$5-$3-$1
```

ChatGPT

J'ai un texte contenant plusieurs dates écrites sous la forme : JJ.MM.AA (jour.mois.année). Par exemple : 24.05.82

J'aimerais transformer ces dates pour qu'elles se retrouvent sous la forme AAAA-MM-JJ, et encadrée par une balise XML <date>. Par exemple : <date>1982-05-24</date>

Donne-moi une expression régulière qui me permet de faire cette transformation.

Voici une expression régulière pour transformer des dates sous la forme JJ.MM.AA en <date>AAAA-MM-JJ</date> .

Expression régulière pour rechercher le format JJ.MM.AA :

```
regex
(\b\d{2})\.\(\d{2})\.\(\d{2})\b
```

Explication :

- `(\b\d{2})` : capture deux chiffres pour le jour en début de mot (`\b`).
- `\.` : reconnaît le point entre jour, mois, et année.
- `(\d{2})` : capture deux chiffres pour le mois.
- `\.\(\d{2})\b` : capture les deux chiffres de l'année en fin de mot (`\b`).

Expression régulière pour rechercher le format JJ.MM.AA :

```
regex
(\b\d{2})\.\(\d{2})\.\(\d{2})\b
```

Explication :

- `(\b\d{2})` : capture deux chiffres pour le jour en début de mot (`\b`).
- `\.` : reconnaît le point entre jour, mois, et année.
- `(\d{2})` : capture deux chiffres pour le mois.
- `\.\(\d{2})\b` : capture les deux chiffres de l'année en fin de mot (`\b`).

Format de remplacement :

```
replacement
<date>19$3-$2-$1</date>
```

Dans cet exemple, `$3` représente les deux derniers chiffres de l'année (`AA`). Si les dates se situent dans les années 1900, on préfixe `19` devant `$3`. Pour les années 2000, remplacez `19` par `20` selon vos besoins.

Se passer des Regex

Prompt:
Je possède un inventaire d'archive sous forme textuelle et j'aimerais le convertir sous la forme de données structurées en XML.

Voici un exemple d'une partie de cet inventaire :

INVENTAIRE TEXTUEL

DOSSIER 41 LE FONCTIONNEMENT D'EURATOM

SOUS-DOSSIER 41/4

Questions de financement, de budget d'Euratom

Liste des pièces

AMK 41/4/1 "Réflexions préliminaires sur le budget d'Euratom". 30.08.56.

AMK 41/4/2 "Note sur les aspects budgétaires d'Euratom", de Van Helmont. Manuscrit.

Voici comment les données doivent être structurées :

INVENTAIRE XML

```
<ud>
  <niveau>série</niveau>
  <cote>AMK 41</cote>
  <intitule>Le fonctionnement d'Euratom</intitule>
  <ud>
    <niveau>dossier</niveau>
    <cote>AMK 41/4</cote>
    <intitule>Questions de financement, de budget d'Euratom</intitule>
    <ud>
      <niveau>pièce</niveau>
      <cote>AMK 41/4/1</cote>
      <intitule>"Réflexions préliminaires sur le budget d'Euratom".</intitule>
      <dates>1956-08-30</dates>
    </ud>
    <ud>
      <niveau>pièce</niveau>
      <cote>AMK 41/4/2</cote>
      <intitule>"Note sur les aspects budgétaires d'Euratom", de Van Helmont.</intitule>
      <notes>Manuscrit.</notes>
    </ud>
  </ud>
</ud>
```

J'aimerais que tu convertisse le texte suivant, selon l'exemple donné

INVENTAIRE A CONVERTIR

[...]



Voici la conversion du texte en XML structuré, suivant le modèle donné :

```
xml
<ud>
  <niveau>série</niveau>
  <cote>AMK 36</cote>
  <intitule>La conférence de Messine</intitule>
  <ud>
    <niveau>dossier</niveau>
    <cote>AMK 36/4</cote>
    <intitule>Comité intergouvernemental: La question des transports et travaux public
    <contenu>Commission des Transports</contenu>
    <ud>
      <niveau>pièce</niveau>
      <cote>AMK 36/4/1</cote>
      <intitule>"La coopération européenne dans le domaine des transports".</intitul
      <dates>1955-07-15</dates>
    </ud>
    <ud>
      <niveau>pièce</niveau>
      <cote>AMK 36/4/2</cote>
      <intitule>"Les modalités de la coopération européenne dans le domaine des tran
      <dates>1955-07-15</dates>
    </ud>
  </ud>
  <ud>
    <niveau>pièce</niveau>
    <cote>AMK 36/4/3</cote>
    <intitule>"Note de la délégation française. Transports".</intitule>
    <dates>1955-07-18</dates>
    <notes>Document N°12.</notes>
  </ud>
</ud>
```

Allons encore plus loin



EAD3-TL-eng.pdf

PDF

On trouve dans le document joint la grammaire XML EAD3 pour les inventaires d'archives.

Sur la base de ce document, écris une version XML EAD3 de l'inventaire suivant :

DOSSIER 36 LA CONFERENCE DE MESSINE

SOUS-DOSSIER 36/4

Comité intergouvernemental: La question des transports et travaux publics

Commission des Transports

Liste des pièces

AMK 36/4/1 "La coopération européenne dans le domaine des transports". 15.07.55.

AMK 36/4/2 "Les modalités de la coopération européenne dans le domaine des transports".

15.07.55.

AMK 36/4/3 "Note de la délégation française. Transports".

18.07.55. Document N°12.

AMK 36/4/4 Commission des Transports et des Travaux Publics.

Sommaire N°1. Réunion du

20 juillet 1955. 20.07.55. Document N°31.

AMK 36/4/5 "Commission des Transports et des Travaux publics.

Note de la délégation

française. Programme de travail de la Commission". 21.07.55.

Document

N° 36.



Voici une version XML EAD3 pour l'inventaire que vous avez fourni, basée sur la grammaire XML EAD3 décrite dans le document :

```
xml
Copy code

<ead>
  <archdesc level="fonds">
    <did>
      <untititle>DOSSIER 36 LA CONFERENCE DE MESSINE</untititle>
      <unitid>36</unitid>
    </did>

    <dsc>
      <c level="series">
        <did>
          <untititle>Comité intergouvernemental: La question des transports et travaux pub
          <unitid>36/4</unitid>
        </did>

        <c level="file">
          <did>
            <untititle>Commission des Transports</untititle>
          </did>

          <c level="item">
            <did>
              <unitid>AMK 36/4/1</unitid>
              <untititle>La coopération européenne dans le domaine des transports</untititle>
              <unitdate normal="1955-07-15">15.07.55</unitdate>
            </did>
          </c>
        </c>
      </dsc>
    </archdesc>
  </ead>
```

```
<c level="item">
  <did>
    <unitid>AMK 36/4/3</unitid>
    <untititle>Note de la délégation française. Transports</untititle>
    <unitdate normal="1955-07-18">18.07.55</unitdate>
    <physdesc>Document N°12</physdesc>
  </did>
</c>
```

Il est temps de conclure

Merci :-)