Manipulation des formats des refbibs dans RONIE

**Du corpus GOLD pour éval à un extrait TRAIN pour grobid-trainer,
via Pub2TEI et raggréage des ponctuations**

1) Inventaire des tags présents pour conversion TEI

1.1) RSC

Encodage : **ISO-8859-1**
Noms de fichiers : /data/rsc/**DATE**/**ACR**/**DOCNAME**.xml

Exemple flatname : rsc_1990_FT_FT9908602163.xml

Structure des refbibs

```
└── #document 10
    └── article 20
        └── art-back 10
            ├── compoundgrp 10
            └── biblist 10
                └── citgroup 187
                    ├── journalcit 205
                    │   ├── citauth 725
                    │   │   ├── surname 725
                    │   │   └── fname 725
                    │   ├── title 205
                    │   ├── pages 205
                    │   │   ├── fpage 205
                    │   │   └── lpage 1
                    │   ├── YEAR 205
                    │   ├── volumeno 182
                    │   └── arttitle 1
                    ├── citation 30  @type="book|thesis|other"
                    │   ├── citauth 61
                    │   │   ├── fname 61
                    │   │   └── surname 61
                    │   ├── YEAR 27
                    │   ├── citpub 23  # éditeur
                    │   ├── title 23
                    │   ├── pubplace 22
                    │   ├── biblscope 18  #vérif attrs
                    │   ├── editor 17
                    │   ├── it 2
                    │   ├── arttitle 1
                    │   └── citext 1
                    └── citext 4
```

1.2) OUP

Encodage : ASCII, **UTF-8** ou non déclaré (ASCII ?)
Noms de fichiers : /data/oup/**NOMREVUE**/**ACRVOLISS**/**ACRVOLISS**xml/**DOCNAME**.xml
avec **DOCNAME** = **VOL-ISS-FPG**.xml
Exemple flatname : oup_Bioinformatics_1985-2010_v1-v26_bioinfo24_16_bioinfo24_16xml_btn308.xml

Structure des refbibs

```
(xpath '/article/back/ref-list')
└── ref-list 9
    ├── ref 257
    │   ├── label 161
    │   ├── nlm-citation 161
    │   │   ├── person-group 154
    │   │   │   ├── name 489
    │   │   │   │   ├── surname 489
    │   │   │   │   ├── given-names 489
    │   │   │   │   └── suffix 2
    │   │   │   └── etal 69
    │   │   ├── YEAR 161
    │   │   ├── source 157
    │   │   ├── volume 150
    │   │   ├── fpage 150
    │   │   ├── lpage 147
    │   │   ├── article-title 156
    │   │   │   ├── sup 4
    │   │   │   └── italic 3
    │   │   ├── publisher-name 10
    │   │   ├── publisher-loc 9
    │   │   ├── collab 6
    │   │   ├── comment 6
    │   │   └── uri 2
    │   ├── issue 5
    │   └── supplement 3
    │   citation 96
    │   ├── person-group 75
    │   │   ├── name 147
    │   │   │   ├── given-names 147
    │   │   │   └── surname 147
    │   │   └── etal 5
    │   ├── YEAR 96
    │   ├── source 92
    │   ├── volume 80
    │   ├── fpage 83
    │   ├── lpage 31
    │   ├── article-title 38
    │   ├── publisher-name 9
    │   ├── publisher-loc 4
    │   ├── bold 39
    │   └── italic 20
    └── title 7
```

sur 2 ex : OUP a conservé les ponctuations !

```
<ref id="rf10">
       <citation citation-type="journal">
              <label>10.</label>
              <person-group person-group-type="author"><name><surname>Pickering</surname>
<given-names>TG</given-names></name></person-group>: <article-title>Blood pressure
measurement and detection of hypertension</article-title>. <source>Lancet</source>
<year>1994</year>;<volume>344</volume>:<fpage>31</fpage>&#x2013;<lpage>35</lpage>.</citation>
       </ref>
```

mais parfois structure différente et sans `@citation-type` ni `<article-title>` :

```
      <ref id="GKH524C14">
        <label>14.</label>
        <citation>Teyssier,C., Belguise,K., Galtier,F., Cavailles,V. and Chalbos,D.
(<year>2003</year>) Receptor&#8208;interacting protein 140 binds c&#8208;Jun and inhibits
estradiol&#8208;induced activator protein&#8208;1 activity by reversing glucocorticoid
receptor&#8208;interacting protein 1 effect. <source>Mol. Endocrinol.</source>,
<volume>17</volume>, <fpage>287</fpage>&#8211;299.</citation>
      </ref>
```

=> Recensement des éléments <citation> qui n'ont pas d'attribut @citation-type

```
find -name "oup_*" -exec grep -Pzo "(?s)<citation[ >].*?</citation>" \{} \;
```

- sur 340 docs OUP dans ech_p+x_1078 je trouve 5952 refbibs <citation>
  - 1691 avec attribut "citation-type"
  - 4261 sans attribut

1.3) NATURE

Encodage : **UTF-8** ou non déclaré (ASCII ?)
Noms de fichiers : /data/nature/full/**ACR**/v**VOL**/n**ISSUE**/**DOCNAME**.xml

Exemple flatname : nature_full_NSMB_v11_n6_nsmb757.xml

Structure des refbibs (admirablement compacte)

```
(xpath '/article/bm/bibl')
└── bibl 19
    ├── bib 832
    │   └── reftxt 832      #préserve les séparateurs
    │       ├── refau 1827
    │       │   ├── snm 1827
    │       │   ├── fnm 1827
    │       │   └── suff 4
    │       ├── CD @year 816
    │       ├── vid 756
    │       ├── ppf 787
    │       ├── ppl 776
    │       ├── jtl 767
    │       ├── atl 765
    │       │   ├── i 87
    │       │   ├── super 16
    │       │   ├── sub 11
    │       │   └── sc 1
    │       ├── btl 66
    │       ├── i 160
    │       ├── newline 67
    │       ├── b 67
    │       │   ├── i 11
    │       │   └── sub 1
    │       └── url 1
    ├── bibgrp 29
    │   ├── bib 43
    │   │   └── reftxt 43
    │   │       ├── refau 43
    │   │       │   ├── snm 71
    │   │       │   └── fnm 70
    │   │       ├── CD @year 41
    │   │       ├── vid 41
    │   │       ├── ppf 41
    │   │       ├── ppl 41
    │   │       ├── jtl 41
    │   │       ├── atl 41
    │   │       │   ├── i 2
    │   │       │   └── sc 1
    │   │       ├── i 15
    │   │       └── Pub 1
    │   └── heading 29
    └── Pub 1
```

1.4) IOP

Encodage : **ISO-8859-1** ou non déclaré (ASCII ?)
Noms de fichiers : /data/iop/ISSN/DATE/ISS/DOCNAME.xml
ou /data/iop/ISSN/VOL/ISS/DOCNAME.xml

Exemple flatname : iop_1742-5468_2007_09_P09002_jstat7_09_p09002.xml

Structure des refbibs

```
(xpath '/article/back/references')
└── references 10
    ├── heading 10
    └── reference-list 10
        ├── journal-ref 294
        │   ├── authors 294
        │   │   ├── au 868
        │   │   │   ├── second-name 868
        │   │   │   └── first-names 867
        │   │   ├── others 51
        │   │   │   └── italic 51
        │   │   └── collaboration 3
        │   │       └── group 3
        │   ├── jnl-title 294
        │   ├── art-title 120
        │   │   ├── sup 25
        │   │   ├── sub 2
        │   │   └── italic 1
        │   ├── volume 289
        │   ├── pages 288
        │   ├── YEAR 293
        │   ├── crossref 213
        │   │   ├── cr_issn 361
        │   │   └── cr_doi 213
        │   ├── part 57
        │   ├── art-number 5
        │   ├── conf-title 5
        │   ├── misc-text 2
        │   │   └── webref 1
        │   └── issno 1
        ├── book-ref 18
        │   ├── authors 18
        │   │   └── au 31
        │   │       ├── first-names 31
        │   │       └── second-name 31
        │   ├── book-title 18
        │   ├── publication 18
        │   │   ├── place 18
        │   │   └── publisher 18
        │   ├── editors 5
        │   │   └── author 12
        │   │       ├── second-name 12
        │   │       └── first-names 12
        │   ├── art-title 5
        │   ├── volume 1
        │   ├── pages 4
        │   ├── YEAR 18
        │   ├── edition 2
        │   └── series 1
        │       └── volume 1
        ├── conf-ref 15
        │   ├── authors 15
        │   │   ├── au 21
        │   │   │   ├── second-name 21
        │   │   │   └── first-names 21
        │   │   └── others 10
        │   │       └── italic 10
        │   ├── conf-place 11
        │   ├── conf-title 15
        │   ├── YEAR 15
        │   ├── misc-text 10
        │   │   └── webref 10
        │   ├── art-title 10
        │   ├── pages 7
        │   ├── publication 5
        │   │   ├── publisher 5
        │   │   └── place 4
        │   ├── editors 2
        │   │   └── author 5
        │   │       ├── second-name 5
        │   │       └── first-names 5
        │   ├── volume 2
        │   └── series 1
        │       └── volume 1
        ├── misc-ref 10
        │   ├── authors 10
        │   │   └── au 16
        │   │       ├── second-name 16
        │   │       └── first-names 16
        │   ├── YEAR 9
        │   ├── art-title 7
        │   │   ├── italic 2
        │   │   └── sub 1
        │   │       └── italic 1
        │   ├── preprint-info 5
        │   │   ├── preprint 5
        │   │   └── art-number 1
        │   ├── links 5
        │   │   └── arxiv 5
        │   ├── misc-text 3
        │   │   └── italic 2
        │   ├── source 3
        │   └── thesis 2
        ├── multipart 3
        │   └── journal-ref 5
        │       ├── crossref 5
        │       │   ├── cr_issn 10
        │       │   └── cr_doi 5
        │       ├── volume 5
        │       ├── authors 5
        │       │   └── au 12
        │       │       ├── first-names 12
        │       │       └── second-name 12
        │       ├── pages 5
        │       ├── YEAR 5
        │       ├── jnl-title 5
        │       └── part 4
        └── ref-group 1
            └── journal-ref 2
                ├── volume 2
                ├── authors 2
                │   └── au 8
                │       ├── first-names 8
                │       └── second-name 8
                ├── pages 2
                ├── YEAR 2
                └── jnl-title 2
```

NB : une petite partie du corpus IOP adopte une toute autre structure, mais connue par ailleurs : le format nlm utilisé aussi chez OUP ou Pubmed

1.5) ELSEVIER

Encodage : **UTF-8** (toujours?) déclaré
Noms de fichiers : /data/elsevier/raw/IS[TY]DATE.../ISSN/VVOLIISS/DOCID/main.xml

Exemple flatname :
elsevier_raw_ISY19940590000053_0022510X_v111i2_0022510X9290077X_main.xml
Structure des refbibs

```
(xpath '/converted-article/tail')
└── tail 10
    └── ce:bibliography 10
        ├── ce:bibliography-sec 10
        │   └── ce:bib-reference 180
        │       ├── ce:label 180
        │       ├── sb:reference 165
        │       │   ├── sb:host 165
        │       │   │   ├── sb:pages 152
        │       │   │   │   ├── sb:first-page 152
        │       │   │   │   └── sb:last-page 81
        │       │   │   ├── sb:issue 135
        │       │   │   │   ├── sb:series 135
        │       │   │   │   │   ├── sb:title 135
        │       │   │   │   │   │   └── sb:maintitle 135
        │       │   │   │   │   └── sb:volume-nr 134
        │       │   │   │   ├── sb:DATE 135
        │       │   │   │   └── sb:issue-nr 22
        │       │   │   ├── sb:edited-book 20
        │       │   │   │   ├── sb:DATE 20
        │       │   │   │   ├── sb:title 15
        │       │   │   │   │   └── sb:maintitle 15
        │       │   │   │   ├── sb:publisher 13
        │       │   │   │   │   ├── sb:name 13
        │       │   │   │   │   └── sb:location 11
        │       │   │   │   ├── sb:editors 9
        │       │   │   │   │   └── sb:editor 10
        │       │   │   │   │       ├── ce:given-name 10
        │       │   │   │   │       └── ce:surname 10
        │       │   │   │   ├── sb:book-series 4
        │       │   │   │   │   └── sb:series 4
        │       │   │   │   │       ├── sb:volume-nr 4
        │       │   │   │   │       └── sb:title 4
        │       │   │   │   │           └ sb:maintitle 4
        │       │   │   │   └── sb:conference 2
        │       │   │   └── sb:book 10
        │       │   │       ├── sb:DATE 10
        │       │   │       ├── sb:publisher 8
        │       │   │       │   ├── sb:name 8
        │       │   │       │   └── sb:location 8
        │       │   │       └── sb:title 7
        │       │   │           └── sb:maintitle 7
        │       │   ├── sb:contribution 164
        │       │   │   ├── sb:authors 164
        │       │   │   │   ├── sb:author 407
        │       │   │   │   │   ├── ce:given-name 407
        │       │   │   │   │   ├── ce:surname 407
        │       │   │   │   │   └── ce:suffix 2
        │       │   │   │   ├── sb:et-al 18
        │       │   │   │   └── sb:collaboration 1
        │       │   │   └── sb:title 119
        │       │   │       ├── sb:maintitle 119
        │       │   │       │   ├── ce:inf 7
        │       │   │       │   └── ce:italic 6
        │       │   │       └── sb:subtitle 8
        │       │   └── sb:comment 8
        │       └── ce:other-ref 16
        │           └── ce:textref 16
        │               ├── ce:italic 2
        │               └── ce:small-caps 1
        └── ce:section-title 10
```

Les titres si host=sb:book peuvent être soit :

- exilés dans la contribution auparavant (le + souvent)
  - Ex : elsevier_raw_ISY19940130000106_03050483_v5i1_0305048377900214

```
<sb:reference><sb:contribution><sb:authors><sb:author><ce:given-name>G</ce:given-name> <ce:surname>Hadley</ce:surname>
</sb:author></sb:authors> <sb:title><sb:maintitle>Analysis of Inventory Systems</sb:maintitle></sb:title>
</sb:contribution>    <sb:host><sb:book><sb:date>1963</sb:date><sb:publisher><sb:name>Prentice-Hall</sb:name>
<sb:location>Englewood Cliffs</sb:location></sb:publisher> </sb:book></sb:host></sb:reference>
```

- sous le sb:book lui-même
  - Ex : elsevier_raw_ISY19940320000103_03788741_v2i4_S0378874180810166

```
<sb:book><sb:title><sb:maintitle>Principles of Tzeltal Plant Classification</sb:maintitle> </sb:title>
<sb:date>1974</sb:date> <sb:publisher><sb:name>Academic Press</sb:name> <sb:location>New York</sb:location>
</sb:publisher></sb:book>
```

Attention dans l'input éditeur on peut trouver côte à côte des refbibs structurées et non-structurées (ce:textref sans aucun élément de préfixe sb:)

```
<ce:bib-reference id="BIB2">
                <ce:label>[2]</ce:label>
                <ce:other-ref>
                    <ce:textref>
  Jacques J., Collet A., Wilen S.H., Enantiomers, Racemates and
  Resolution, John Wiley and Sons, New York, NY, 1981.
                </ce:textref>
                </ce:other-ref>
            </ce:bib-reference>
```

```
<ce:bib-reference id="BIB3">
    <ce:label>[3]</ce:label>
    <sb:reference>
        <sb:contribution>
            <sb:authors>
             <sb:author>
               <ce:given-name>W.M.</ce:given-name>
               <ce:surname>Pirkel</ce:surname>
             </sb:author>
             (...)
            </sb:authors>
        </sb:contribution>
        <sb:host>
            <sb:issue>
              <sb:series>
                <sb:title><sb:maintitle>J. Org. Chem.</sb:maintitle></sb:title>
                <sb:volume-nr>51</sb:volume-nr>
              </sb:series>
              <sb:date>1986</sb:date></sb:issue>
            <sb:pages><sb:first-page>4991</sb:first-page></sb:pages>
        </sb:host>
    </sb:reference>
</ce:bib-reference>
```

donc → pour avoir juste les bonnes bibs il faudrait ignorer ce qui n'a pas de namespace "sb:"

- Un cas rare : mise en forme <ce :...> trouvée comme simple note parmi des biblStruct normaux :

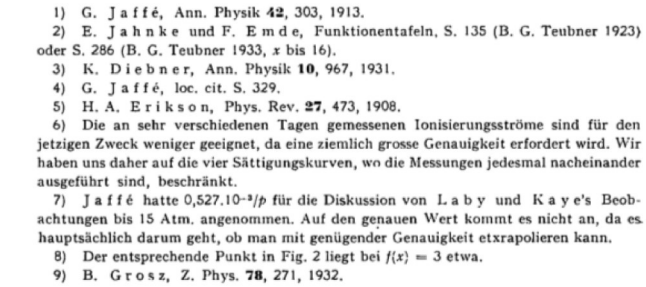
NATIF

elsevier_raw_ISY19940010000208_00318914_v2i1-12_S003189143590163X_main.xml

```
        <ce:bib-reference id="bib7">
          <ce:label>7)</ce:label>
          <ce:note>
            <ce:simple-para>Jaffé hatte 0.527.10<ce:sup loc="post">-3</ce:sup>/<ce:italic>p</ce:italic> für die Diskussion von Laby und Kaye's
Beobachtungen bis 15 Atm. angenommen. Auf den genauen Wert kommt es nicht an, da es hauptsächlich darum geht, ob man mit genügender Genauigkeit
etxrapolieren kann.</ce:simple-para>
          </ce:note>
        </ce:bib-reference>
        <ce:bib-reference id="bib8">
          <ce:label>8)</ce:label>
          <ce:note>
            <ce:simple-para>Der entsprechende Punkt in Fig. 2 liegt bei <ce:italic>f(x)</ce:italic> = 3 etwa.</ce:simple-para>
          </ce:note>
        </ce:bib-reference>
```

ORIGINAL PDF

elsevier_raw_ISY19940010000208_00318914_v2i1-12_S003189143590163X_main.pdf



1) G. Jaffé, Ann. Physik **42**, 303, 1913.
2) E. Jahnke und F. Emde, Funktionentafeln, S. 135 (B. G. Teubner 1923) oder S. 286 (B. G. Teubner 1933, x bis 16).
3) K. Diebner, Ann. Physik **10**, 967, 1931.
4) G. Jaffé, loc. cit. S. 329.
5) H. A. Erikson, Phys. Rev. **27**, 473, 1908.
6) Die an sehr verschiedenen Tagen gemessenen Ionisierungsströme sind für den jetzigen Zweck weniger geeignet, da eine ziemlich grosse Genauigkeit erfordert wird. Wir haben uns daher auf die vier Sättigungskurven, wo die Messungen jedesmal nacheinander ausgeführt sind, beschränkt.
7) Jaffé hatte $0{,}527.10^{-3}/p$ für die Diskussion von Laby und Kaye's Beobachtungen bis 15 Atm. angenommen. Auf den genauen Wert kommt es nicht an, da es hauptsächlich darum geht, ob man mit genügender Genauigkeit etxrapolieren kann.
8) Der entsprechende Punkt in Fig. 2 liegt bei $f(x) = 3$ etwa.
9) B. Grosz, Z. Phys. **78**, 271, 1932.

- Autre cas bizarre : la refbib est double avec comme seule différence le no de page

NATIF

elsevier_raw_ISY19940200000077_00098981_v60i3_0009898175900807_main.xml

```
<ce:bib-reference id="BIB7">
  <ce:label>7</ce:label>
  <sb:reference>
      <sb:contribution>
          <sb:authors>
              <sb:author>
                  <ce:given-name>A.J.</ce:given-name>
                  <ce:surname>Crowle</ce:surname>
              </sb:author>
          </sb:authors>
          <sb:title>
             <sb:maintitle>Immunodiffusion</sb:maintitle>
          </sb:title>
      </sb:contribution>
      <sb:host>
          <sb:edited-book>
              <sb:date>1973</sb:date>
              <sb:publisher>
                  <sb:name>Academic Press</sb:name>
           <sb:location>New York and London</sb:location>
              </sb:publisher>
          </sb:edited-book>
          <sb:pages>
              <sb:first-page>374</sb:first-page>
          </sb:pages>
      </sb:host>
  </sb:reference>
  <sb:reference>
      <sb:contribution>
          <sb:authors>
              <sb:author>
                  <ce:given-name>A.J.</ce:given-name>
                  <ce:surname>Crowle</ce:surname>
              </sb:author>
          </sb:authors>
          <sb:title>
             <sb:maintitle>Immunodiffusion</sb:maintitle>
          </sb:title>
      </sb:contribution>
      <sb:host>
          <sb:edited-book>
              <sb:date>1973</sb:date>
              <sb:publisher>
                  <sb:name>Academic Press</sb:name>
           <sb:location>New York and London</sb:location>
              </sb:publisher>
          </sb:edited-book>
          <sb:pages>
              <sb:first-page>398</sb:first-page>
          </sb:pages>
      </sb:host>
  </sb:reference>
</ce:bib-reference>
```

ORIGINAL PDF

elsevier_raw_ISY19940200000077_00098981_v60i3_0009898175900807_main.pdf

6 J. Kohn, in H. Peeters (ed.), Protides of the Biological Fluids, p. 124, 1958, Elsevier, Amsterdam
7 A.J. Crowle, Immunodiffusion, Academic Press, New York and London, 1973 p. 374 and p. 398
8 J. Krøll, Scand. J. Clin. Lab. Invest., 21 (1968) 187
9 N.H. Axelsen, J. Krøll and B. Weeke (eds), A Manual of Quantitative Immunoelectrophoresis — Methods and Applications, pp. 71—77, 1973, Universitetsforlaget, Oslo

---

- 3ème remarque

celui-ci à garder en tête, car majoritaire mais inattendu

La position du titre de monogr est souvent directement dans <sb:contribution> au lieu de <sb:host> comme en TEI avec monogr.

NATIF

elsevier_raw_ISY19940130000106_03050483_v5i1_0305048377900214_main.xml

```
<ce:bib-reference id="BIB4">
    <ce:label>4.</ce:label>
    <sb:reference>
        <sb:contribution>
            <sb:authors>
                <sb:author>
                    <ce:given-name>G</ce:given-name>
                    <ce:surname>Hadley</ce:surname>
                </sb:author>
                <sb:author>
                    <ce:given-name>TM</ce:given-name>
                    <ce:surname>Whitin</ce:surname>
                </sb:author>
            </sb:authors>
            <sb:title>
                <sb:maintitle>Analysis of Inventory Systems</sb:maintitle>
            </sb:title>
        </sb:contribution>
        <sb:host>
            <sb:book>
                <sb:date>1963</sb:date>
                <sb:publisher>
                    <sb:name>Prentice-Hall</sb:name>
                    <sb:location>Englewood Cliffs</sb:location>
                </sb:publisher>
            </sb:book>
        </sb:host>
    </sb:reference>
</ce:bib-reference>
```

ORIGINAL PDF

elsevier_raw_ISY19940130000106_03050483_v5i1_0305048377900214_main.pdf

4. Hadley G and Whitin TM (1963) *Analysis of Inventory Systems*. Prentice-Hall, Englewood Cliffs.
5. Hanssmann F (1962) *Operations Research in Production and Inventory Control*. Wiley, New York.

Idem comparé avec infos externes sur l'ouvrage

NATIF

elsevier_raw_ISY19940200000077_00098981_v60i3_0009898175900807_main.xml

```
<ce:bib-reference id="BIB18">
    <ce:label>18</ce:label>
    <sb:reference>
        <sb:contribution>
            <sb:authors>
                <sb:author>
                    <ce:given-name>&#x00D6;.</ce:given-name>
                    <ce:surname>Ouchterlony</ce:surname>
                </sb:author>
            </sb:authors>
        </sb:contribution>
        <sb:host>
            <sb:edited-book>
                <sb:edition>3rd edn</sb:edition>
                <sb:book-series>
                    <sb:editors>
                        <sb:editor>
                            <ce:given-name>P.</ce:given-name>
                            <ce:surname>Kallos</ce:surname>
                        </sb:editor>
                    </sb:editors>
                    <sb:series>
                        <sb:title>
                            <sb:maintitle>Progress in Allergy</sb:maintitle>
                        </sb:title>
                        <sb:volume-nr>Vol. V</sb:volume-nr>
```

```
                    </sb:series>
                </sb:book-series>
                <sb:date>1958</sb:date>
                <sb:publisher>
                    <sb:name>Karger</sb:name>
                    <sb:location>Basel and New York</sb:location>
                </sb:publisher>
            </sb:edited-book>
            <sb:pages>
                <sb:first-page>38</sb:first-page>
                <sb:last-page>48</sb:last-page>
            </sb:pages>
        </sb:host>
    </sb:reference>
</ce:bib-reference>
```

ORIGINAL PDF

17 W. Groc, M. Jendrey and W. Lahn, Clin. Chim. Acta, 54 (1974) 65
18 Ö. Ouchterlony, in P. Kallos (ed.), Progress in Allergy, Vol. V, pp. 38—48, 1958, Karger, Basel and New York
19 R.J. Wieme, Agar Gel Electrophoresis, p. 110, 1965, Elsevier, Amsterdam

Informations éditeur

cf. http://www.karger.com/Book/Toc/217983

Status: out of print, available online
Publication year: **1958**
Progress in Allergy
Vol. 5

Editor(s): **Kallós P. (Helsingborg)**
Get instant access *Order this title*

[Toc]

1 Diffusion-in-Gel Methods for Immunological Analysis (Part 1 of 2)
Ouchterlony Ö.
pp 1–36 of: Kallós P (ed): Progress in Allergy Volume 5. Progr Allergy. Basel/New York, Karger, 1958, vol 5, pp 1–78 (DOI:10.1159/000273347)

37 Diffusion-in-Gel Methods for Immunological Analysis (Part 2 of 2)
Ouchterlony Ö.
pp 37–78 of: Kallós P (ed): Progress in Allergy Volume 5. Progr Allergy. Basel/New York, Karger, 1958, vol 5, pp 1–78 (DOI:10.1159/000277641)

79 The Release of Histamine (Part 1 of 2)
Paton W.D.M.
pp 79–113 of: Kallós P (ed): Progress in Allergy Volume 5. Progr Allergy. Basel/New York, Karger, 1958, vol 5, pp 79-148 (DOI:10.1159/000277642)
(...)

=> pour tous ces cas problématiques, vérifier si solution dans feuilles Pub2TEI

- a priori pas vraiment : la feuille Elsevier originale passe toujours par la l.45 de la feuille Bibliography.xsl (`<xsl:template match="ce:bib-reference[sb:reference]">`) qui ne sait traiter que les articles de journaux standards.

- **ignorer** les cas *très* rares pour l'entraînement, *et tant que ça n'introduit apparement pas de biais,*

- les **recenser ici** pour l'avenir (enjeu d'alimentation des métas en TEI de l'API via docs natifs complets et les transformations qui seront issues de Pub2TEI)

2) Feuilles de styles Pub2TEI avant utilisation

```
/home/loth/refbib/tools/Pub2TEI/Stylesheets > git remote -v
origin https://github.com/kermitt2/Pub2TEI.git
```

Ce sont des feuilles XSLT 2.0

2.1) Vue générale

La feuille principale est `Publishers.xsl` :

- elle trie les cas de figure selon la balise racine observée dans l'input
- elle importe toutes les autres feuilles de 2 façons :
  - directement pour chaque « Publisher »
    - pour mapper la structure éditeur sur un **squelette TEI**

```
<TEI>
    <teiHeader>
        <fileDesc>...</fileDesc>
        <profileDesc>...<profileDesc>
    </teiHeader>
    <text>
        <front>...</front>
        <body>...</body>
        <back></back>
    </text>
</TEI>
```

    - via `Imports.xsl` pour chaque composant spécifique qu'on peut trouver (biblio, tables,...)

La logique est donc "input-guided" pour générer le squelette et "output-guided" pour remplir les sous-branches spécifiques.

- la structure d'import permet de faire tourner l'ensemble même si plein de détails particuliers restent à compléter

Méthode actuelle :

- Les cas traitant certaines valeurs d'attributs comme <article type="bidule"> couvrent actuellement les cas fréquents mais n'ont pas encore de cas prévu pour certaines valeurs rares
- la rencontre d'un élément inconnu est signalée par le message :

```
"Element inconnu: name: <xsl:value-of select="name()"/>
—     local-name: <xsl:value-of select="local-name()"/>
—     namespace-uri: <xsl:value-of select="namespace-uri()"/> -"
```

- ajout manuel du cas de figure lorsque observés dans les exemples

2.2) Bibliography.xsl

elsevier :
A parachever dans les détails l. 45 (`<xsl:template match="ce:bib-reference[sb:reference]">`)

Feuille :

```
<xsl:template match="sb:authors">
    <xsl:apply-templates/>
</xsl:template>
```

```
<xsl:template match="sb:author">
    <author>
        <xsl:apply-templates/>
    </author>
</xsl:template>
```

2.3) DTDs de chaque éditeur

Les sous-éléments (sous-dossiers ou fichiers annexes) en doublon (vus une 2è, 3è ... fois) sont notés en gras.

Pub2TEI/Samples/DTDs/art452/art452.dtd
Pub2TEI/Samples/DTDs/art452/catalog.xml
Pub2TEI/Samples/DTDs/art452/common110.ent
Pub2TEI/Samples/DTDs/art452/ESextra.ent
Pub2TEI/Samples/DTDs/art452/html
Pub2TEI/Samples/DTDs/art452/iso8879
Pub2TEI/Samples/DTDs/art452/iso9573-13
Pub2TEI/Samples/DTDs/art452/mathml
Pub2TEI/Samples/DTDs/art452/mathml2-mod-ES.dtd
Pub2TEI/Samples/DTDs/art452/mathml2-qname-1.mod
Pub2TEI/Samples/DTDs/art452/soextblx.dtd
Pub2TEI/Samples/DTDs/art452/tagbytag5-v1.pdf
Pub2TEI/Samples/DTDs/art501.dtd
Pub2TEI/Samples/DTDs/art501/catalog.txt
Pub2TEI/Samples/DTDs/art501/catalog.xml
doublon: "catalog.xml" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art501/common110.ent
doublon: "common110.ent" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art501/ESextra.ent
doublon: "ESextra.ent" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art501/esgrid.pdf
Pub2TEI/Samples/DTDs/art501/html
doublon: "html" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art501/iso8879
doublon: "iso8879" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art501/iso9573-13
doublon: "iso9573-13" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art501/mathml
doublon: "mathml" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art501/mathml2-mod-ES.dtd
doublon: "mathml2-mod-ES.dtd" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art501/mathml2-qname-1.mod
doublon: "mathml2-qname-1.mod" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art501/README.txt
Pub2TEI/Samples/DTDs/art501/relnote501.pdf
Pub2TEI/Samples/DTDs/art501/soextblx.dtd
doublon: "soextblx.dtd" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art502/art502.dtd
Pub2TEI/Samples/DTDs/art502/catalog
Pub2TEI/Samples/DTDs/art502/catalog.xml
doublon: "catalog.xml" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art502/common110_1.ent
Pub2TEI/Samples/DTDs/art502/ESextra.ent
doublon: "ESextra.ent" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art502/esgrid.pdf
doublon: "esgrid.pdf" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art502/html
doublon: "html" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art502/iso8879
doublon: "iso8879" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art502/iso9573-13
doublon: "iso9573-13" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art502/mathml
doublon: "mathml" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art502/mathml2-mod-ES.dtd
doublon: "mathml2-mod-ES.dtd" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art502/mathml2-qname-1.mod
doublon: "mathml2-qname-1.mod" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art502/README
Pub2TEI/Samples/DTDs/art502/relnote502.pdf
Pub2TEI/Samples/DTDs/art502/soextblx.dtd
doublon: "soextblx.dtd" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art510/art510.dtd
Pub2TEI/Samples/DTDs/art510/catalog
doublon: "catalog" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art510/catalog.xml
doublon: "catalog.xml" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art510/common115.ent
Pub2TEI/Samples/DTDs/art510/cookbook.xml
Pub2TEI/Samples/DTDs/art510/ESextra.ent
doublon: "ESextra.ent" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art510/esgrid.pdf
doublon: "esgrid.pdf" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art510/html
doublon: "html" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art510/index.html
Pub2TEI/Samples/DTDs/art510/iso8879
doublon: "iso8879" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art510/iso9573-13
doublon: "iso9573-13" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art510/mathml
doublon: "mathml" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art510/mathml2-mod-ES.dtd
doublon: "mathml2-mod-ES.dtd" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art510/mathml2-qname-1.mod
doublon: "mathml2-qname-1.mod" déjà vu ailleurs

Pub2TEI/Samples/DTDs/art510/README
doublon: "README" déjà vu ailleurs
Pub2TEI/Samples/DTDs/art510/relnote510.pdf
Pub2TEI/Samples/DTDs/art510/soextblx.dtd
doublon: "soextblx.dtd" déjà vu ailleurs
Pub2TEI/Samples/DTDs/A++V2.4/A++V2.4.dtd
Pub2TEI/Samples/DTDs/A++V2.4/A++V2.4JobSheetV2.4.1.dtd
Pub2TEI/Samples/DTDs/A++V2.4/A++V2.4JobSheetV2.4.dtd
Pub2TEI/Samples/DTDs/A++V2.4/Book.dtd
Pub2TEI/Samples/DTDs/A++V2.4/calstblx.dtd
Pub2TEI/Samples/DTDs/A++V2.4/Common.dtd
Pub2TEI/Samples/DTDs/A++V2.4/JobSheet.dtd
Pub2TEI/Samples/DTDs/A++V2.4/JobSheetV2.4.1.dtd
Pub2TEI/Samples/DTDs/A++V2.4/JobSheetV2.4.dtd
Pub2TEI/Samples/DTDs/A++V2.4/Journal.dtd
Pub2TEI/Samples/DTDs/A++V2.4/mathml2.dtd
Pub2TEI/Samples/DTDs/A++V2.4/mathml3.dtd
Pub2TEI/Samples/DTDs/book520/book520.dtd
Pub2TEI/Samples/DTDs/book520/common113.ent
Pub2TEI/Samples/DTDs/book520/ESextra.ent
doublon: "ESextra.ent" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book520/html
doublon: "html" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book520/index.html
doublon: "index.html" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book520/iso8879
doublon: "iso8879" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book520/iso9573-13
doublon: "iso9573-13" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book520/mathml
doublon: "mathml" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book520/mathml2-mod-ES.dtd
doublon: "mathml2-mod-ES.dtd" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book520/mathml2-qname-1.mod
doublon: "mathml2-qname-1.mod" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book520/soextblx.dtd
doublon: "soextblx.dtd" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book521/book521.dtd
Pub2TEI/Samples/DTDs/book521/common113.ent
doublon: "common113.ent" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book521/ESextra.ent
doublon: "ESextra.ent" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book521/html
doublon: "html" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book521/index.html
doublon: "index.html" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book521/iso8879
doublon: "iso8879" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book521/iso9573-13
doublon: "iso9573-13" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book521/mathml
doublon: "mathml" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book521/mathml2-mod-ES.dtd
doublon: "mathml2-mod-ES.dtd" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book521/mathml2-qname-1.mod
doublon: "mathml2-qname-1.mod" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book521/soextblx.dtd
doublon: "soextblx.dtd" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book530/book530.dtd
Pub2TEI/Samples/DTDs/book530/catalog
doublon: "catalog" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book530/catalog.xml
doublon: "catalog.xml" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book530/common116.ent
Pub2TEI/Samples/DTDs/book530/ESextra.ent
doublon: "ESextra.ent" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book530/index.html
doublon: "index.html" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book530/html
doublon: "html" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book530/index.html
doublon: "index.html" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book530/iso8879
doublon: "iso8879" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book530/iso9573-13
doublon: "iso9573-13" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book530/mathml
doublon: "mathml" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book530/mathml2-mod-ES.dtd
doublon: "mathml2-mod-ES.dtd" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book530/mathml2-qname-1.mod
doublon: "mathml2-qname-1.mod" déjà vu ailleurs
Pub2TEI/Samples/DTDs/book530/soextblx.dtd
doublon: "soextblx.dtd" déjà vu ailleurs
Pub2TEI/Samples/DTDs/contrast/book-item-project-1.1.xsd
Pub2TEI/Samples/DTDs/contrast/book-item-project-1.2.xsd
Pub2TEI/Samples/DTDs/contrast/book-item-project-1.3.xsd
Pub2TEI/Samples/DTDs/contrast/book-item-project-1.4.xsd

Pub2TEI/Samples/DTDs/contrast/book-item-project.xsd
Pub2TEI/Samples/DTDs/contrast/book-projects-1.1.xsd
Pub2TEI/Samples/DTDs/contrast/book-projects-1.2.xsd
Pub2TEI/Samples/DTDs/contrast/book-projects-1.3.xsd
Pub2TEI/Samples/DTDs/contrast/book-projects-1.4.xsd
Pub2TEI/Samples/DTDs/contrast/book-projects.xsd
Pub2TEI/Samples/DTDs/contrast/index.html
doublon: "index.html" déjà vu ailleurs
Pub2TEI/Samples/DTDs/contrast/issues.xsd
Pub2TEI/Samples/DTDs/contrast/items.xsd
Pub2TEI/Samples/DTDs/contrast/journal-item-issue.xsd
Pub2TEI/Samples/DTDs/entities/rsc_x.ent
Pub2TEI/Samples/DTDs/IOP/common.inc
Pub2TEI/Samples/DTDs/IOP/ent
Pub2TEI/Samples/DTDs/IOP/header.dtd
Pub2TEI/Samples/DTDs/IOP/html.inc
Pub2TEI/Samples/DTDs/IOP/img.inc
Pub2TEI/Samples/DTDs/IOP/iopv1_4_1.dtd
Pub2TEI/Samples/DTDs/IOP/iopv1_4_2.dtd
Pub2TEI/Samples/DTDs/IOP/iopv1_4_3.dtd
Pub2TEI/Samples/DTDs/IOP/iopv1_4_4.dtd
Pub2TEI/Samples/DTDs/IOP/iopv1_4_5.dtd
Pub2TEI/Samples/DTDs/IOP/iopv1_4_6.doc
Pub2TEI/Samples/DTDs/IOP/iopv1_4_6.dtd
Pub2TEI/Samples/DTDs/IOP/iopv1_4_7a.dtd
Pub2TEI/Samples/DTDs/IOP/iopv1_4_7.dtd
Pub2TEI/Samples/DTDs/IOP/iopv1_4.dtd
Pub2TEI/Samples/DTDs/IOP/iopv1_5_1.dtd
Pub2TEI/Samples/DTDs/IOP/iopv1_5_2a.dtd
Pub2TEI/Samples/DTDs/IOP/iopv1_5_2.dtd
Pub2TEI/Samples/DTDs/IOP/iopv1_6.dtd
Pub2TEI/Samples/DTDs/IOP/iopv1_7.dtd
Pub2TEI/Samples/DTDs/IOP/iopv1_8.dtd
Pub2TEI/Samples/DTDs/IOP/iso
Pub2TEI/Samples/DTDs/IOP/iso-grk3.ent
Pub2TEI/Samples/DTDs/IOP/iso.inc
Pub2TEI/Samples/DTDs/IOP/latexent.ent
Pub2TEI/Samples/DTDs/IOP/nlm-3.0
Pub2TEI/Samples/DTDs/IOP/reference_list.inc
Pub2TEI/Samples/DTDs/IOP/toc.dtd
Pub2TEI/Samples/DTDs/IOP/vol.dtd
Pub2TEI/Samples/DTDs/NPG_XML_Article/npgcateg.mod
Pub2TEI/Samples/DTDs/NPG_XML_Article/npgnonchars.ent
Pub2TEI/Samples/DTDs/NPG_XML_Article/NPG_XML_Article.dtd
Pub2TEI/Samples/DTDs/NPG_XML_Article/NPG_XML_Article.rlx
Pub2TEI/Samples/DTDs/NPG_XML_Article/old_XML_Isomath.dtd
Pub2TEI/Samples/DTDs/NPG_XML_Article/XML_Calstbl.dtd
Pub2TEI/Samples/DTDs/NPG_XML_Article/XML_entities
Pub2TEI/Samples/DTDs/NPG_XML_Article/XML_Isomath.dtd
Pub2TEI/Samples/DTDs/publishing2.3/articlemeta.ent
Pub2TEI/Samples/DTDs/publishing2.3/backmatter.ent
Pub2TEI/Samples/DTDs/publishing2.3/chars.ent
Pub2TEI/Samples/DTDs/publishing2.3/common.ent
Pub2TEI/Samples/DTDs/publishing2.3/default-classes.ent
Pub2TEI/Samples/DTDs/publishing2.3/default-mixes.ent
Pub2TEI/Samples/DTDs/publishing2.3/display.ent
Pub2TEI/Samples/DTDs/publishing2.3/format.ent
Pub2TEI/Samples/DTDs/publishing2.3/iso8879
doublon: "iso8879" déjà vu ailleurs
Pub2TEI/Samples/DTDs/publishing2.3/iso9573-13
doublon: "iso9573-13" déjà vu ailleurs
Pub2TEI/Samples/DTDs/publishing2.3/journalmeta.ent
Pub2TEI/Samples/DTDs/publishing2.3/journalpubcustom-classes.ent
Pub2TEI/Samples/DTDs/publishing2.3/journalpubcustom-mixes.ent
Pub2TEI/Samples/DTDs/publishing2.3/journalpubcustom-models.ent
Pub2TEI/Samples/DTDs/publishing2.3/journalpubcustom-modules.ent
Pub2TEI/Samples/DTDs/publishing2.3/journalpublishing.dtd
Pub2TEI/Samples/DTDs/publishing2.3/link.ent
Pub2TEI/Samples/DTDs/publishing2.3/list.ent
Pub2TEI/Samples/DTDs/publishing2.3/math.ent
Pub2TEI/Samples/DTDs/publishing2.3/mathml
doublon: "mathml" déjà vu ailleurs
Pub2TEI/Samples/DTDs/publishing2.3/mathml2.dtd
doublon: "mathml2.dtd" déjà vu ailleurs
Pub2TEI/Samples/DTDs/publishing2.3/mathml2-qname-1.mod
doublon: "mathml2-qname-1.mod" déjà vu ailleurs
Pub2TEI/Samples/DTDs/publishing2.3/mathmlsetup.ent
Pub2TEI/Samples/DTDs/publishing2.3/modules.ent
Pub2TEI/Samples/DTDs/publishing2.3/nlmcitation.ent
Pub2TEI/Samples/DTDs/publishing2.3/notat.ent
Pub2TEI/Samples/DTDs/publishing2.3/para.ent
Pub2TEI/Samples/DTDs/publishing2.3/phrase.ent
Pub2TEI/Samples/DTDs/publishing2.3/references.ent
Pub2TEI/Samples/DTDs/publishing2.3/section.ent
Pub2TEI/Samples/DTDs/publishing2.3/xhtml-inlstyle-1.mod
Pub2TEI/Samples/DTDs/publishing2.3/xhtml-table-1.mod

Pub2TEI/Samples/DTDs/publishing2.3/XHTMLtablesetup.ent
Pub2TEI/Samples/DTDs/publishing2.3/xmlchars
Pub2TEI/Samples/DTDs/publishing2.3/xmlspecchars.ent
Pub2TEI/Samples/DTDs/publishing/annotation3.ent
Pub2TEI/Samples/DTDs/publishing/articlemeta3.ent
Pub2TEI/Samples/DTDs/publishing/backmatter3.ent
Pub2TEI/Samples/DTDs/publishing/base-test3.dtd
Pub2TEI/Samples/DTDs/publishing/catalog.ent
Pub2TEI/Samples/DTDs/publishing/catalog-v3.xml
Pub2TEI/Samples/DTDs/publishing/chars3.ent
Pub2TEI/Samples/DTDs/publishing/common3.ent
Pub2TEI/Samples/DTDs/publishing/default-classes3.ent
Pub2TEI/Samples/DTDs/publishing/default-mixes3.ent
Pub2TEI/Samples/DTDs/publishing/display3.ent
Pub2TEI/Samples/DTDs/publishing/format3.ent
Pub2TEI/Samples/DTDs/publishing/funding3.ent
Pub2TEI/Samples/DTDs/publishing/htmltable.dtd
Pub2TEI/Samples/DTDs/publishing/iso8879
doublon: "iso8879" déjà vu ailleurs
Pub2TEI/Samples/DTDs/publishing/iso9573-13
doublon: "iso9573-13" déjà vu ailleurs
Pub2TEI/Samples/DTDs/publishing/journalmeta3.ent
Pub2TEI/Samples/DTDs/publishing/journalpubcustom-classes3.ent
Pub2TEI/Samples/DTDs/publishing/journalpubcustom-mixes3.ent
Pub2TEI/Samples/DTDs/publishing/journalpubcustom-models3.ent
Pub2TEI/Samples/DTDs/publishing/journalpubcustom-modules3.ent
Pub2TEI/Samples/DTDs/publishing/journalpublishing3.dtd
Pub2TEI/Samples/DTDs/publishing/link3.ent
Pub2TEI/Samples/DTDs/publishing/list3.ent
Pub2TEI/Samples/DTDs/publishing/math3.ent
Pub2TEI/Samples/DTDs/publishing/mathml
doublon: "mathml" déjà vu ailleurs
Pub2TEI/Samples/DTDs/publishing/mathml2.dtd
doublon: "mathml2.dtd" déjà vu ailleurs
Pub2TEI/Samples/DTDs/publishing/mathml2-qname-1.mod
doublon: "mathml2-qname-1.mod" déjà vu ailleurs
Pub2TEI/Samples/DTDs/publishing/mathmlsetup3.ent
Pub2TEI/Samples/DTDs/publishing/modules3.ent
Pub2TEI/Samples/DTDs/publishing/nlmcitation3.ent
Pub2TEI/Samples/DTDs/publishing/notat3.ent
Pub2TEI/Samples/DTDs/publishing/oasis-exchange.ent
Pub2TEI/Samples/DTDs/publishing/oasis-tablesetup3.ent
Pub2TEI/Samples/DTDs/publishing/para3.ent
Pub2TEI/Samples/DTDs/publishing/phrase3.ent
Pub2TEI/Samples/DTDs/publishing/references3.ent
Pub2TEI/Samples/DTDs/publishing/related-object3.ent
Pub2TEI/Samples/DTDs/publishing/section3.ent
Pub2TEI/Samples/DTDs/publishing/Smallsamples
Pub2TEI/Samples/DTDs/publishing/xhtml-inlstyle-1.mod
doublon: "xhtml-inlstyle-1.mod" déjà vu ailleurs
Pub2TEI/Samples/DTDs/publishing/xhtml-table-1.mod
doublon: "xhtml-table-1.mod" déjà vu ailleurs
Pub2TEI/Samples/DTDs/publishing/XHTMLtablesetup3.ent
Pub2TEI/Samples/DTDs/publishing/xmlchars
doublon: "xmlchars" déjà vu ailleurs
Pub2TEI/Samples/DTDs/publishing/xmlspecchars3.ent
Pub2TEI/Samples/DTDs/si1510/catalog
doublon: "catalog" déjà vu ailleurs
Pub2TEI/Samples/DTDs/si1510/catalog.xml
doublon: "catalog.xml" déjà vu ailleurs
Pub2TEI/Samples/DTDs/si1510/common113.ent
doublon: "common113.ent" déjà vu ailleurs
Pub2TEI/Samples/DTDs/si1510/ESextra.ent
doublon: "ESextra.ent" déjà vu ailleurs
Pub2TEI/Samples/DTDs/si1510/esgrid.pdf
doublon: "esgrid.pdf" déjà vu ailleurs
Pub2TEI/Samples/DTDs/si1510/html
doublon: "html" déjà vu ailleurs
Pub2TEI/Samples/DTDs/si1510/index.html
doublon: "index.html" déjà vu ailleurs
Pub2TEI/Samples/DTDs/si1510/iso8879
doublon: "iso8879" déjà vu ailleurs
Pub2TEI/Samples/DTDs/si1510/iso9573-13
doublon: "iso9573-13" déjà vu ailleurs
Pub2TEI/Samples/DTDs/si1510/mathml
doublon: "mathml" déjà vu ailleurs
Pub2TEI/Samples/DTDs/si1510/mathml2-mod-ES.dtd
doublon: "mathml2-mod-ES.dtd" déjà vu ailleurs
Pub2TEI/Samples/DTDs/si1510/mathml2-qname-1.mod
doublon: "mathml2-qname-1.mod" déjà vu ailleurs
Pub2TEI/Samples/DTDs/si510/si510.dtd
Pub2TEI/Samples/DTDs/si1510/soextblx.dtd
doublon: "soextblx.dtd" déjà vu ailleurs

2.4) État d'avancement par lot

On a mi-déc 2014 des feuilles éditeurs + ou - finalisées parmi les 5 RONIE :

- Elsevier est très complexe en entrée mais le xsl est presque complet
- NLM semble aussi bien débroussaillé
    - ok pour les articles de journaux, quelques corrections pour les autres
    - NB : ce NLM devrait normalement être compatible pour OUP
- RSC a un input un peu plus simple, et il est pratiquement bien traité
- IOP semble poser quelques pbs, la feuille de style est juste un stub
- Nature semble aussi en développement
    - certaines templates sont finies [TODO]
    - le début qui ne matche pas encore vraiment sur la racine <article>

+ des cas de figure de XML input hors RONIE, que je ne connais pas vraiment mais qui ont déjà l'air pas mal

- Springer, Sage, BMJ, EDPS

3) Feuilles Pub2TEI : résultats observés des transformations

On procède à la transformation sur 1000 documents *(echantillon-p+x-1078_RONIE)*

composition :
441 els
138 iop
120 nat
340 oup
39 rsc

1078 total

Commande de conversion :

```
saxonb-xslt -xsl:tools/Pub2TEI/Stylesheets/Publishers.xsl
            -s:ech/natifs/
            -o:ech/tei.xml/
```
==> 180 transformations failed (~17%)

3.1) résultats sur RSC

Novel conjugated, soluble polymers are very much in demand for research into their intriguing optical and electronic properties. We have here a very easy general synthesis of a novel type of polyacetylene where there is opportunity of varying both the nature of the polar substituents and skeletal hetero atom.[10]

Received, 10th October 1989; Com. 9/04357H

References

1 F. L. Klavetter and R. H. Grubbs, *J. Am. Chem. Soc.*, 1988, **110**, 807.
2 J. H. Edwards and J. H. Feast, *Polymer*, 1980, **21**, 595.
3 E. M. D. Gilan, J. G. Hamilton, O. N. D. Mackey. and J. J. Rooney, *J. Mol. Catal.*, 1988, **46**, 359.
4 I. Adams and W. Vogt, *Makromol. Chem., Rapid Commun.*, 1988, **9**, 327.
5 E. M. D. Gillan, J. G. Hamilton, and J. J. Rooney, *J. Mol. Catal.*, 1989, **50**, L23.
6 O. Diels and K. Alder, *Annalen.*, 1931, **490**, 243.
7 K. J. Ivin, 'Olefin Metathesis,' Academic Press, London, 1983.
8 H. E. Ardill, R. M. E. Greene, J. G. Hamilton, H. T. Ho, K. J. Ivin, G. Lapienis, G. M. McCann, and J. J. Rooney. ACS Symp. Ser. 286, 1985, 275.
9 J. G. Hamilton, K. J. Ivin, and J. J. Rooney, *Br. Polym. J.*, 1984, **16**, 21.
10 J. G. Hamilton and J. J. Rooney, Brit. Pat. 1989, No. 8914186-5.

document : `/data/rsc/1990/C3/C39900000119.*`

commandes préalables :

```
cat exemples_RONI_1513
/rsc_1990_C3_C39900000119.xml
| perl -pe 's!http://www.rsc.org/dtds /rscart36.dtd
!/home/loth/refbib/tools/
Pub2TEI/Samples/DTDs/rscart36.dtd!'
| saxonb-xslt -xsl:Publishers.xsl -s:-
```

extrait résultat :

```
  <p>9104357H  References  1 F. L. Klavetter and R. H. Grubbs,J. Am. Chem. Soc. 1988,110
807.  2 J. H. Edwards and J. H. Feast Polymer 1980 21 595.  3 E. M. D. Gilan J. G. Hamilton
O. N. D. Mackey. and J. J.  Rooney J. Mol. Catal. 1988 46 359.  4 I. Adams and W. Vogt
Makromol. Chem. Rapid Commun.  1988 9 327.  5 E. M. D. Gillan J. G. Hamilton and J. J. Rooney
I. Mol. Catal.  1989 50 L23.  6 O.Diels and K. Alder Annalen. 1931 490 243.  7 K. J. Ivin
'Olefin Metathesis,' Academic Press London 1983.  8 H. E. Ardill R. M. E. Greene J. G.
Hamilton H. T. Ho K. J.  Ivin G. Lapienis G. M. McCann and J. J. Rooney ACS Symp.  Ser. 286
1985 275.  9 J. G. Hamilton K. J. Ivin and J. J. Rooney Br. PolFm. J. 1984  16 21.  10 J. G.
Hamilton and J. J. Rooney Brit. Pat. 1989,No. 8914186-5.     </p>
         </div>
      </body>
      <back>
         <div type="references">
            <listBibl>
               <biblStruct type="article" xml:id="cit1">
                  <analytic>
                     <author>
                        <persName>
```

```
                        <forename type="first">F. L.</forename>
                        <surname>Klavetter</surname>
                     </persName>
                  </author>
                  <author>
                     <persName>
                        <forename type="first">R. H.</forename>
                        <surname>Grubbs</surname>
                     </persName>
                  </author>
               </analytic>
               <monogr>
                  <title level="j" type="main">J. Am. Chem. Soc.</title>
                  <imprint>
                     <date type="year">1988</date>
                     <biblScope unit="vol">110</biblScope>
                     <biblScope unit="page" from="807">807</biblScope>
                  </imprint>
               </monogr>
            </biblStruct>
```

Quelques bons points :

- le paragraphe précédent contient les références sous leur forme d'origine (mais sans les sauts de ligne qui seraient dans le pdftottext et avec des mini-scories OCR comme par ex. dans FT9928803587 Rev. Sci. → Reu. Sci.
- la TEI issue la feuille de LR a conservé l'ordre d'origine !
- mais ce dernier point n'est pas garanti pour tout fichier

Mais :

- la feuille de transformation saute les <citgroup> imbriqués (max observé : 2)
- **elle saute aussi les <citation type="book|thesis|other"> (15%)**
  - mais justement seuls ceux-ci ont leur ponctuation dans le natif
- donc elle ne prend que les <journalcit> sous un <citgroup>
- <biblScope unit="page" from="in the press">in the press</biblScope>
- quelques éléments inconnus
  - qualifier, inf, compoundgroup

DOCUMENT	contient le dernier <p> d'origine	ordre TEI
C39900000119	oui	ok
AP9912800217	oui (2 derniers <p>)	(initiales prénom) <=> (nom)
P29960001081	oui (2 derniers <p>)	ok
AN9931800171	oui	ok
CC9960001247	oui	ok
FT9928803587	oui mais « Reu. Sci. » dans le <p> contre « Rev. Sci. » dans la <cit>	ok
AI9953200365	oui, manque 1 ½ citation au début	(initiales prénom) <=> (nom)

cf. aussi ~/refbib/tools/Pub2TEI/Stylesheets/ex_trait_RSC

18 J. H. Bowie, *Mass Spectrom. Rev.*, 1990, **9**, 360 and 361.
19 M. B. Stringer, D. J. Underwood, J. H. Bowie, J. L. Holmes, A. A. Mommers and J. A. Szulejko, *Can. J. Chem.*, 1986, **64**, 764; G. J. Currie, J. H. Bowie, R. A. Massy-Westropp and G. W. Adams, *J. Chem. Soc., Perkin Trans. 2*, 1988, 403.
20 J. C. Sheldon, J. H. Bowie and D. E. Lewis, *Nouv. J. Chim.*, 1988, **12**, 269.

==> **Cas inhabituel : 2 bibs ligne 19**

Quatre cas non-reconnus à ajouter à la feuille Pub2TEI RSC

```
<citgroup id="cit17"> <citation type="other">
<citauth>
<fname>G. M.</fname>
<surname>Sheldrick</surname>
</citauth>, <title>SHELXTL</title>, <citpub>Siemens Analytical Instruments Inc.</citpub>, <pubplace>Madison, WI</pubplace>,
<year>1990</year>.</citation></citgroup>


<citgroup id="cit4">
<citation type="book">
<citauth>
<fname>E.</fname>
<surname>Deutsch</surname>
</citauth>, <citauth>
<fname>K.</fname>
<surname>Libson</surname>
</citauth> and <citauth>
<fname>J.-L.</fname>
<surname>Vanderheyden</surname>
</citauth>, <title>The Inorganic Chemistry of Rhenium and Technetium as related to Nuclear Medicine</title>, eds. <editor>M.
Niccolini</editor>, <editor>G. Bandoli</editor> and <editor>U. Mazzi</editor>, <citpub>Cortina Int.</citpub>,
<pubplace>Verona</pubplace>, <year>1990</year>, vol. 3, <biblscope>pp. 13&ndash;22</biblscope>.</citation>
</citgroup>

<citgroup id="cit17">
<citation type="thesis">
<citauth>
<fname>R.</fname>
<surname>Radaelli</surname>
</citauth>, Thesis in Physics, <citpub>University of Milan</citpub>, <year>1991</year>.</citation>
</citgroup>
```

```
<citgroup id="cit1">
  <citgroup id="cit1a">
    <journalcit>
      <citauth>
        <fname>W. B.</fname>
        <surname>Euler</surname>
      </citauth>
      (...)
      <title>Solid State Commun.</title>
      <year>1984</year>
      <volumeno>51</volumeno>
      <pages>
        <fpage>473</fpage>
      </pages>
    </journalcit>
  </citgroup>
  <citgroup id="cit1b">
    <journalcit>
      <citauth>
        <fname>C. R.</fname>
        <surname>Hauer</surname>
      </citauth>
      (...)
      <title>J. Am. Chem. Soc.</title>
      <year>1987</year>
      <volumeno>109</volumeno>
      <pages>
        <fpage>5760</fpage>
      </pages>
    </journalcit>
  </citgroup>
</citgroup>
```

1 (*a*) W. B. Euler and C. R. Hauer, *Solid State Commun.*, 1984, **51**, 473; (*b*) C. R. Hauer, G. S. King, E. L. McCool, W. B. Euler, J. D. Ferrara and W. D. Youngs, *J. Am. Chem. Soc.*, 1987, **109**, 5760.

Décompte global taux de conversion RSC

dans dossier natifs

```
ls | grep ^rsc | while read L ; do grep -Pazo "(?s)<art-back[ >].*</art-back>" < $L | grep
-Po "<citation|<journalcit" ; done | sort | uniq -c
```

1281 total dont 1151 <journalcit et 130 <citation

dans dossier tei.xml *(> 441 documents els convertis)*

```
ls | grep ^rsc | while read L ; do grep -Po "<biblStruct" $L ; done | sort | uniq -c
```

1213 <biblStruct

```
ls | grep ^rsc | while read L ; do grep -Pazo "(?s)<back[ >].*</back>" < $L |grep -o
"<biblStruct" ; done | sort | uniq -c
```

1173 <biblStruct

3.2) résultats OUP

NB : OUP est théoriquement du NLM

versions observées de la DTD archivearticle.dtd

- v2.3

Certains documents passent bien en <biblStruct>, par ex. :

- oup_American_Journal_of_Hypertension_1988-2010_v1-v23_ajh8_8_ajh8_8xml_8_8_790.xml

Exemples documents ne passant pas :

- oup_Bioinformatics_1985-2010_v1-v26_bioinfo21_6_bioinfo21_6xml_bti067.xml
- oup_Nucleic_Acids_Research_1974-2010_v1-v37_nar32_6_nar32_6xml_gkh524.xml

<bibl> pas pleinement structurées

LECTURE XML
WARN: aucun <biblStruct>, je tente les 25 <bibl>
ERR: aucune xbibs dans ce xml natif !

=> les <bibl> en présence ressembleraient presque au format d'entraînement Grobid mais les auteurs et le titre level= « m » sont absents :-/

exemple :

original

```
<ref id="GKG473C4"> <label>4.</label>
<citation>Pazin,M.J. and Kadonaga,J.T.
(<year>1997</year>) What&#8217;s up and down
with histone deacetylation and transcription?
<source>Cell</source>, <volume>89</volume>,
<fpage>325</fpage>&#8211;328.</citation>
</ref>
```

après conversion

```
<bibl xml:id="GKG473C4">Pazin,M.J. and
Kadonaga,J.T. (<date type="year">1997</date>)
What’s up and down with histone deacetylation
and transcription? <title
level="j">Cell</title>,
<biblScope unit="vol">89</biblScope>,
<biblScope unit="page"
from="325">325</biblScope>–328.</bibl>
```

Décompte global taux de conversion OUP

dans dossier natifs

```
ls | grep ^oup | while read L ; do grep -Po "<(?:nlm-)?citation" $L ; done | wc -l
```

11029

dans dossier tei.xml *(> 441 documents els convertis)*

```
ls | grep ^oup | while read L ; do grep -Po "<biblStruct" $L ; done | sort | uniq -c
```

6884 <biblStruct

```
ls | grep ^oup | while read L ; do grep -Pazo "(?s)<back[ >].*</back>" < $L |grep -o
"<biblStruct" ; done | sort | uniq -c
```

6543 <biblStruct

3.3) Nature

None =>Nature n'était pas traité dans leur projet (aucune feuille) : j'en fait une

3.4) IOP

None => la feuille est juste un stub : j'en fait une nouvelle

3.5) résultats Elsevier

Globalement bien converti :

- souci tout de même sur les citations autres qu'article de journaux
-

TEI

elsevier_raw_ISY19940010000208_00318914_v2i1-12_S003189143590163X_main.xml

```
            <biblStruct xml:id="bib7">
               <ce:note xmlns:ce="http://www.elsevier.com/xml/common/dtd">
                  <p>Jaffé hatte 0.527.10<hi rend="superscript">-3</hi>/<hi rend="italic">p</hi> für die Diskussion von
Laby und Kaye's Beobachtungen bis 15 Atm. angenommen. Auf den genauen Wert kommt es nicht an, da es hauptsächlich darum geht,
ob man mit genügender Genauigkeit etxrapolieren kann.</p>
               </ce:note>
            </biblStruct>
            <biblStruct xml:id="bib8">
               <ce:note xmlns:ce="http://www.elsevier.com/xml/common/dtd">
                  <p>Der entsprechende Punkt in Fig. 2 liegt bei <hi rend="italic">f(x)</hi> = 3 etwa.</p>
               </ce:note>
            </biblStruct>
```

Décompte global taux de conversion elsevier

dans dossier natifs

```
ls | grep ^els | while read L ; do grep -o "sb:reference" $L ; done | wc -l
```

16926

dans dossier tei.xml *(> 441 documents els convertis)*

```
ls | grep ^els | while read L ; do grep -Po "<biblStruct" $L ; done | sort | uniq -c
```

8709 <biblStruct

```
ls | grep ^els | while read L ; do grep -Pazo "(?s)<back[ >].*</back>" < $L |grep -o
"<biblStruct" ; done | sort | uniq -c
```

8234 <biblStruct

Par type d'oeuvre

dans dossier natifs

```
ls | grep ^els | while read L ; do grep -o "sb:reference" $L ; done | wc -l
```

16926

dossier tei.xml

```
ls | grep ^els | while read L ; do grep -o "title level=\".\"" $L ; done | sort | uniq -c
```

6148 title level="a"
442 title level="j"

4) Procédure ragreage.py : méthode et paramètres

4.1) Lecture formats

4.2) Identification zone intéressante du pdf

Principe tlmatches

```
# compile regex search tokens found in the xml
search_toks = set()
for st in XMLTEXTS:
	for tok in re.split(r"\W+", st):
		if is_searchable(tok):
			etok = re.escape(tok)
			search_toks.add(re.compile(r"\b%s\b" % etok))

print (search_toks, "<== Those are the searched tokens")

# pdflines matchcount array
tlmatches = []

for i, tline in enumerate(rawpdftxtlines):
	# new initial count for this line
	tlmatches.append(0)

	# filter out very short text lines
	if len(tline) <= 2:
		next

	# décompte
	for tok_xfrag_re in search_toks:
		if re.search(tok_xfrag_re, tline):
			tlmatches[i] += 1

	# log
	if debug >= 2:
		print("-"*20+"\n"+"line n. "+str(i)+" : "+tline)
		print("found %i known text fragments" % tlmatches[i] + "\n"+"-"*20)
```

Exemples de problèmes

Cas concret sur un oup qui cumule les différents obstacles :

```
python3 bin/ragreage.py
-x ech/tei.xml/oup_Geophysical_Journal_International_-2010_v1-v183_gji121_3_gji121_3xml_121-3-789.xml
```

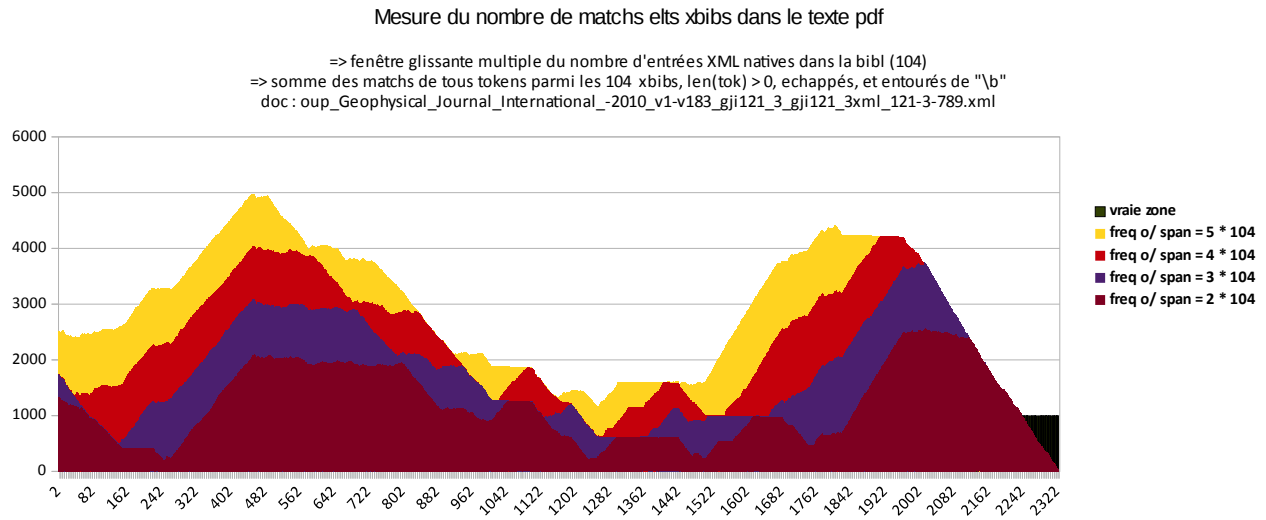
Certains tokens trouvé dans les xbibs sont des noms communs spécialisés qu'on peut retrouver partout dans le texte

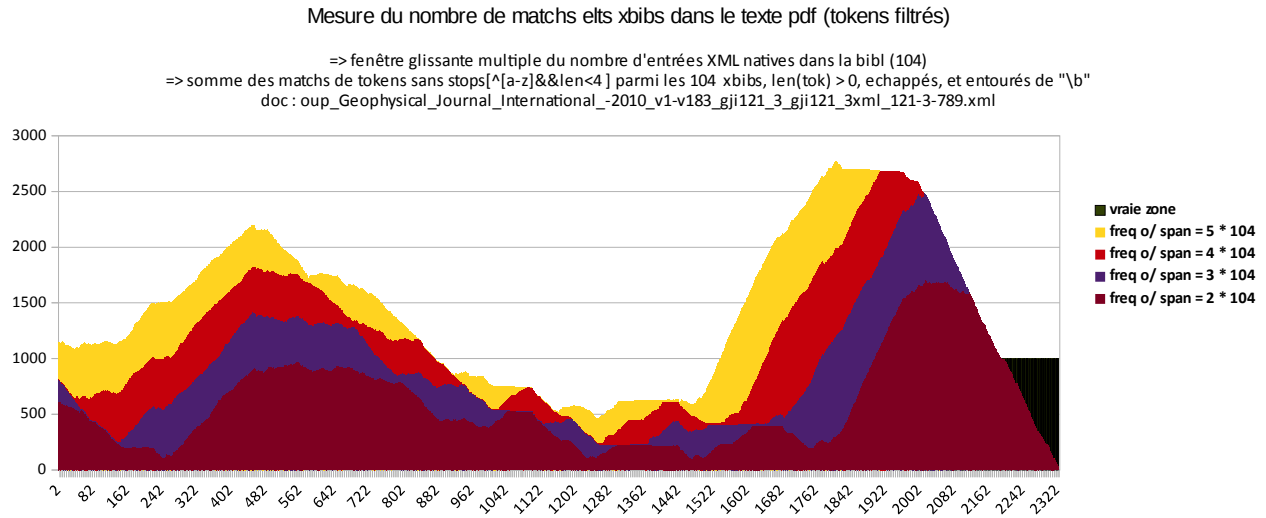
```
--------------------
line n. 1810 : The existing magnetization models of the oceanic lithosphere used for explaining the anomalous skewnesstemporal variations of the geomagnetic field, crustal tectonic
found 30 known text fragments
--------------------
```

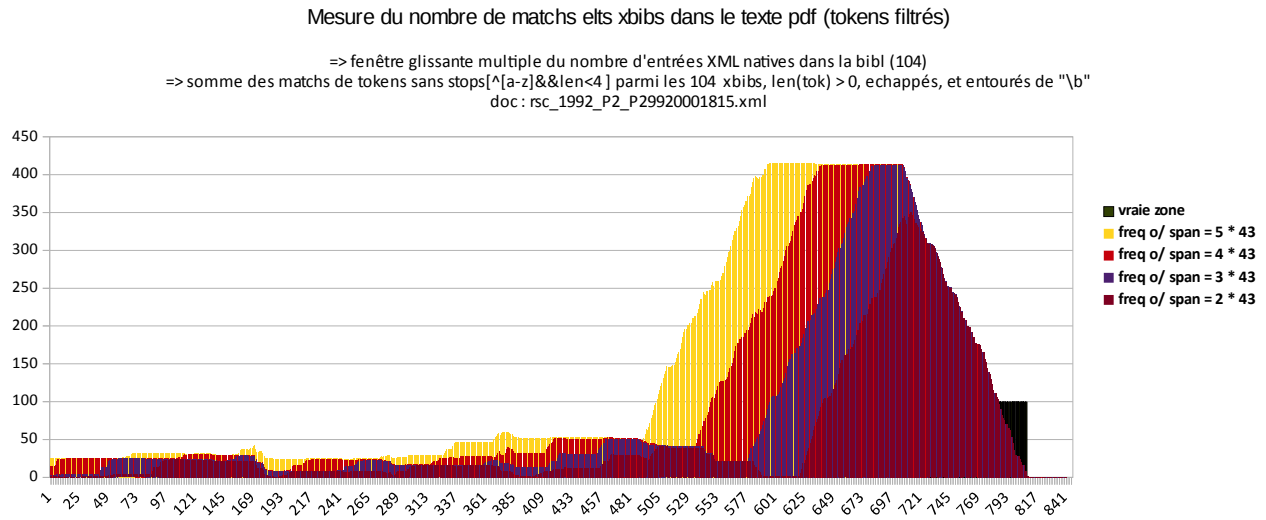
D'autres comme les noms d'auteurs et les dates peuvent se retrouver dans les appels de citation

```
--------------------
line n. 951 : variation of magnetization intensity with age observed by
found 15 known text fragments
--------------------
--------------------
line n. 952 : Bled & Petersen (1983), Furuta (1993). and Johnson &
found 12 known text fragments
--------------------
--------------------
line n. 953 : Pariso (1993) may obscure possible variations with spreading
found 11 known text fragments
--------------------
```

On regarde les décomptes sur une fenêtre de matchs…







5) Formats attendu pour l'entraînement grobid

Par modèle : balises et templates

5.1) Modèle « citations »

Balises utilisées

```
parsing doc 0354
(xpath '//*[local-name()='bibl']')
 └── bibl 4011
     ├── lb 326
     ├── date 3994
     │   ├── year 1
     │   └── month 1
     ├── author 3551
     │   └── lb 2
     ├── editor 121
     │   └── lb 1
     ├── orgName 85
     ├── title[@level="a"] 3316
     │   └── lb 1
     ├── title[@level="j"] 3049
     │   └── lb 2
     ├── title[@level="j"][@type="short"] 2
     ├── title[@level="m"] 662
     │   └── lb 4
     ├── title[@level="s"] 1
     ├── biblScope[@type="vol"] 3076
     ├── biblScope[@type="issue"] 323
     ├── biblScope[@type="chapter"] 4 => pourquoi pas ref[@type="chap" comme ici]
     ├── biblScope[@type="pp"] 3401
     ├── pubPlace 554
     ├── publisher 302
     ├── idno 18
     ├── idno[@type="vol"] 1
     ├── ptr[@type="web"] 13          => pourquoi pas ref[@type="url" comme ici]
     ├── note 107
     ├── note[@type="report"] 76
     ├── notes 13
     ├── ~~docAuthor~~ 2
     ├── ~~authors~~ 1
     └── ~~volumes~~ 1
```

`<lb>` indique les sauts de lignes

CRF Templates

TODO
cf. aussi https://github.com/kermitt2/grobid/wiki/Grobid-batch-quick-start#createtrainingfulltext

CORRESPONDANCES FORMAT <biblStruct> et format entraînement en sortie ragreage.py

PUB2TEI + ragreage => INPUT

```
7346 biblStruct
  6787 biblStruct/analytic
 19555 biblStruct/analytic/author
 10191 biblStruct/analytic/author/forename
  9321 biblStruct/analytic/author/persName
  9317 biblStruct/analytic/author/persName/forename
    43 biblStruct/analytic/author/persName/genName
  9321 biblStruct/analytic/author/persName/surname
    47 biblStruct/analytic/author/suffix
 10234 biblStruct/analytic/author/surname
    29 biblStruct/analytic/editor
    29 biblStruct/analytic/editor/persName
    29 biblStruct/analytic/editor/persName/forename
     1 biblStruct/analytic/editor/persName/genName
    29 biblStruct/analytic/editor/persName/surname
     4 biblStruct/analytic/name
  4452 biblStruct/analytic/title
   214 biblStruct/analytic/title/hi
     1 biblStruct/analytic/title/hi/hi
    12 biblStruct/analytic/title/subtitle
  2432 biblStruct/analytic/title/title
   344 biblStruct/analytic/title/title/hi
    10 biblStruct/analytic/translated-title
    10 biblStruct/analytic/translated-title/title
     3 biblStruct/idno
  7341 biblStruct/monogr
   773 biblStruct/monogr/author
   773 biblStruct/monogr/author/persName
   770 biblStruct/monogr/author/persName/forename
     2 biblStruct/monogr/author/persName/genName
   773 biblStruct/monogr/author/persName/surname
   284 biblStruct/monogr/editor
   284 biblStruct/monogr/editor/persName
   284 biblStruct/monogr/editor/persName/forename
   284 biblStruct/monogr/editor/persName/surname
  7341 biblStruct/monogr/imprint
 16776 biblStruct/monogr/imprint/biblScope
  6726 biblStruct/monogr/imprint/date
   489 biblStruct/monogr/imprint/publisher
   459 biblStruct/monogr/imprint/pubPlace
     5 biblStruct/monogr/meeting
  6672 biblStruct/monogr/title
    18 biblStruct/monogr/title/hi
```

OUTPUT => TRAIN

```
parsing doc 0354
(xpath '//*[local-name()='bibl']')
 └── bibl 4011
     ├── date 3994
     │   ├── year 1
     │   └── month 1
     ├── author 3551
     │   └── lb 2
     ├── biblScope[@type="pp"] 3401
     ├── title[@level="a"] 3316
     │   └── lb 1
     ├── biblScope[@type="vol"] 3076
     ├── title[@level="j"] 3049
     │   └── lb 2
     ├── title[@level="m"] 662
     │   └── lb 4
     ├── pubPlace 554
     ├── lb 326
     ├── biblScope[@type="issue"] 323
     ├── publisher 302
     ├── editor 121
     │   └── lb 1
     ├── note 107
     ├── orgName 85
     ├── note[@type="report"] 76
     ├── idno 18
     ├── notes 13
     ├── ptr[@type="web"] 13
     ├── biblScope[@type="chapter"] 4
     ├── docAuthor 2
     ├── title[@level="j"][@type="short"] 2
     ├── authors 1
     ├── volumes 1
     ├── idno[@type="vol"] 1
     └── title[@level="s"] 1
```

5.2) TODO liste à Noël 2014

training du modèle citation (et emplacements dans **~/refbib)**

- complétion `corpus/bibistex/03_POOLs_et_bases_PDFDOCS`
- ragréage `ragreage.py -x corpus/echantillon-p+x-$N_RONIE/tei.xml/$doc`
- transformations `tools/Pub2TEI`
- entraînement `tools/$G/grobid-trainer/resources/dataset/citation/corpus`
- lancement `tools/$G/grobid-service`
- stockage `istex@vi-istex-rc:~/mes_runs/res/`

5.3) Segmentation

Balises utilisées

```
(1078 fichiers .xml)
parsing doc 1078, match 001
(xpath '/')
 └── #document 1078
     └── tei 1078
         ├── text 1078
         │   ├── body 5039
         │   │   └── lb 170927
         │   ├── page 3560
         │   │   └── lb 3344
         │   ├── note 2381
         │   │   └── lb 2692
         │   ├── listBibl 1661
         │   │   └── lb 10017
         │   ├── front 1394
         │   │   └── lb 12313
         │   └── lb 191
         └── teiHeader 1078
             └── fileDesc 1078
```

PDF => createTrainingSegmentation* => { tei.xml + raw tokens.segmentation}

```
* java -Xmx1024m -jar ~/refbib/grobid/grobid-core/target/grobid-core-0.3.0.one-jar.jar
-gH ~/refbib/grobid/grobid-home/ -gP ~/refbib/tools/grobid/grobid-
home/config/grobid.properties -exe createTrainingSegmentation

# dure environ 0.75 s/doc
```

+ training du modèle « segmentation du doc » :
  - microechantillon qqs dizaines
  - transcriptions segments
  - alimentation modèle

5.4) Modèles grobid et configs

CASCADE	MODELE CRF	PRETRAINING EXE	TRAINING MVN TGT	TRAINING EXT	COMMANDE ragreage.py	FONCTION ragreage.py CORRESPONDANTE
1	fulltext	createTrainingFulltext	train_fulltext	.training.fulltext.tei.xml		
2	segmentation	createTrainingSegmentation	train_segmentation	.training.segmentation.tei.xml		find_bib_zone
3	reference-segmenter	createTrainingReferenceSegmentation	train_reference-segmentation	.referenceSegmenter.training.tei.xml		link_txt_bibs_with_xml()
4	citation	createTrainingFulltext	train_citation	.training.references.tei.xml	(par défaut)	TODO ignorer <lb> + 2 post-traitements : aut
5	name_citation	createTrainingFulltext	train_name_citation	.training.citations.authors.tei.xml		TODO ignorer tout sauf auteurs

5.5) Exemple de traitement préparatoire pour corpus d'apprentissage auto.

Document: corpus/s6-p+x-500/ELS500-s6.readme

```
BIBISTEX s6-500 echantillon complémentaire 500

shuf 03_POOLs_et_bases_PDFDOCS/par_lots/pools_dispos/els-BIBPOOL.tab | head -n 500 >
04_ECHANTILLONS/ech.500.els.pool.tab

# Aperçu de la représentativité
# -----------------------------
~/refbib/corpus/bibistex > cut -f9 04_ECHANTILLONS/ech.500.els.pool.tab | sort | uniq -c
:       3 [v1.0, v1.1]
:     341 [v1.2]
:     107 [v1.3]
:      37 [v1.4]
:      12 [v1.6, v1.7]

~/refbib/corpus/bibistex > cut -f7 04_ECHANTILLONS/ech.500.els.pool.tab | sort | uniq -c
:      11 [...-1959]
:     103 [1960-1979]
:     122 [1980-1989]
:     214 [1990-1999]
:      50 [2000-...]

~/refbib/corpus/bibistex > cut -f8 04_ECHANTILLONS/ech.500.els.pool.tab | sort | uniq -c | sort -rn | head
-n 30
:      28 MEDICINE, GENERAL & INTERNAL
:      14 PHYSICS/PHYSICS, MULTIDISCIPLINARY
:      11 CHEMISTRY, ORGANIC
:       9 PHYSICS, CONDENSED MATTER/CHEMISTRY, PHYSICAL
:       9 NEUROSCIENCES
:       8 BIOCHEMISTRY & MOLECULAR BIOLOGY/BIOPHYSICS
:       8 BIOCHEMISTRY & MOLECULAR BIOLOGY
:       6 PHYSICS, CONDENSED MATTER
:       6 CHEMISTRY, PHYSICAL
:       6 CHEMISTRY, ANALYTICAL/BIOCHEMICAL RESEARCH METHODS
:       5 POLYMER SCIENCE
:       5 PHYSICS, NUCLEAR
:       5 PHYSICS, CONDENSED MATTER/MATERIALS SCIENCE/MATERIALS SCIENCE, MULTIDISCIPLINARY
:       5 PHARMACOLOGY & PHARMACY
:       5 PEDIATRICS
:       5 NUCLEAR SCIENCE & TECHNOLOGY/INSTRUMENTS & INSTRUMENTATION/PHYSICS, PARTICLES &
FIELDS/SPECTROSCOPY
:       5 MATERIALS SCIENCE, CERAMICS/MATERIALS SCIENCE, MULTIDISCIPLINARY
:       5 DENTISTRY, ORAL SURGERY & MEDICINE
:       5 CHEMISTRY, INORGANIC & NUCLEAR/CHEMISTRY, ORGANIC
:       5 CHEMISTRY, INORGANIC & NUCLEAR
:       5 BIOCHEMISTRY & MOLECULAR BIOLOGY/PLANT SCIENCES
:       4 SOIL SCIENCE
:       4 PSYCHIATRY
:       4 IMMUNOLOGY
:       4 ENGINEERING, MECHANICAL/MECHANICS/ACOUSTICS
:       4 ENDOCRINOLOGY & METABOLISM/BIOCHEMISTRY & MOLECULAR BIOLOGY
:       4 ENDOCRINOLOGY & METABOLISM
:       4 BIOCHEMISTRY & MOLECULAR BIOLOGY/BIOPHYSICS/CELL BIOLOGY
:       3 SPECTROSCOPY
:       3 PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH/TROPICAL MEDICINE


# NATIF ~~> TEI
# --------------

# ajouts dtd articles elseviers seuls
mkdir bibidtd
find /home/loth/refbib/tools/Pub2TEI/Samples/DTDs/art* -maxdepth 1 -mindepth 1 | grep -v "README" | grep -v
"pdf$" | while read nodenam ; do cp -ru $nodenam bibidtd/. ; done
cd ech/natifs
for doc in `ls ../../bibidtd/` ; do ln -s ../../bibidtd/$doc ; done


# conversion TEI (~ 33docs/s)
mkdir tei
saxonb-xslt -xsl:/home/loth/refbib/tools/Pub2TEI/Stylesheets/Publishers.xsl -s:natifs -o:tei 2>errs.log
```

```
# PDF ~~> GBRAW
# --------------
# création flux raw tel que vu par grobid
mkdir trainers.refseg.praws
GB=/home/loth/refbib/grobid/
java -Xmx2048m -jar $GB/grobid-core/target/grobid-core-0.3.3-SNAPSHOT.one-jar.jar -gH $GB/grobid-home -gP
$GB/grobid-home/config/grobid.properties -dIn corpus/s6-p+x-500/ech/pdf  -dOut corpus/s6-p+x-
500/ech/trainers.refseg.praws/ -exe createTrainingReferenceSegmentation

# génération de la tei d'entraînement ragrée
mkdir trainers.refseg.ptei

# (GBRAW + TEI) => TRAINING.TEI
# =============================
# -a- ragreage
# ---
time for doc in `ls tei` ;
  do echo $doc ;
    bn=`basename -s ".xml" $doc` ;
    ragreage.py -m 'reference-segmenter' -x tei/$doc -t trainers.refseg.praws/
$bn.referenceSegmenter.training.rawtxt > trainers.refseg.ptei/$bn.refseg.tei.xml 2>> ragreage.refseg.log ;
    echo "ok: $bn" ;
  done
: real  2m24.981s         # NB: 2 ou 3 docs/s (juste align sans macrozone ni champs)

# -b- infos
# --- -----
# (rappel corpus actuel)
# 19 docs 416 bibl
#         --------
#            └─ dans grobid-trainer/res/data/refseg
#               au commit e848d0e du 2015-02-20 15h

# (notre corpus)
# diagnostic std:
cut -f2,3 checks.refseg.tab.ok | sort | uniq -c
:     68 0      0
:     84 0      1
:     91 1      0  # => qqch raté mais parfois pas gd chose
:    208 1      1  # => veut pas forcément dire que c réussi ms on essaye..

# si on prend que les 1 1 ...
grep -P "\t1\t1" < checks.refseg.tab | cut -f1 | sed 's!tei/!! ; s!\.xml!!' > refseg_train_ok.bn.ls

# ... combien ça fait de bibl "d'entraînement" ?
wc -l refseg_train_ok.bn.ls
: 208 refseg_train_ok.bn.ls          #  208 docs

for bn in `cat refseg_train_ok.bn.ls` ;
  do grep -o "<bibl>" trainers.refseg.ptei/$bn.tei.xml;
done | wc -l
: 3348                    # 3348 refbibs d'entraînement potentielles
                          # ----
                          #   └─ pas mal: alé les bløs!

# TRAINING.TEI => train grobid => MODEL
# -------------------------------------
# -a-  sous-ensemble ~ 5 x plus qu'actuellement
shuf refseg_train_ok.bn.ls | head -n 150 > refseg_train_ok.sample.refseg-test-2500.bn.ls

# >> ON VA RAJOUTER 150 docs (soit ~ 2500 refbibs d'entraînement) <<

# -b-   on initialise le dossier sous samp/
 coltrane_make_samp_dirs.sh refseg-test-2500 reference-segmenter
#-------------------------
: ok dirs created
: /home/loth/refbib/analyses/coltrane/samp/refseg-test-2500
: ├── data                          --------------------
: │   └── reference-segmenter
: │       ├── crfpp-templates
: │       │   └── reference-segmenter.template
: │       └── evaluation
: │           ├── 1.tei.xml
: │           ├── label_1.tei.xml
: │           └── label_2.tei.xml
: └── meta

# -c- /!\ fin   -manuelle- préparation
cd ~/refbib/analyses/coltrane/samp/refseg-test-2500
mkdir data/reference-segmenter/corpus

# -d- /!\ copie -manuelle- du corpus initial
export ORIG_DATASET="/home/loth/refbib/grobid/grobid-trainer/resources/dataset"
cp $ORIG_DATASET/reference-segmenter/corpus/*.tei.xml data/reference-segmenter/corpus/.
```

```
# -e- /!\ copie -manuelle- du nouveau corpus depuis le stock créé plus haut
export NEW_DATASET="/home/loth/refbib/corpus/s6-p+x-500/ech"
for bn in `cat $NEW_DATASET/refseg_train_ok.bn.ls` ; do echo $bn ; cp /home/loth/refbib/corpus/s6-p+x-
500/ech/trainers.refseg.ptei/$bn.refseg.tei.xml data/reference-segmenter/corpus/. ; done

# -f- entraînement proprement dit
coltrane_make_training.sh refseg-test-2500 reference-segmenter
```

6) Grobid-service

6.1) Premier test sur 292k docs els

Lancé grobid-service sur 292k documents le 09/01/2014 au soir

```
echantilloneur.pl \
--base par_lots/bases_quotas/els-PDFDOCS-path_type_period_vers_wostheme_issn_lot.tab \
--pool par_lots/pools_dispos/els-BIBPOOL.tab \
-n 300000 > ech.300k.els.tab

# il y en a 292075
mv ech.300k.els.tab ech.292k.els.tab

# représentativité par périodes
cut -f7 ech.292k.els.tab | sort | uniq -c
              23271 [...-1959]
              52257 [1960-1979]
              70845 [1980-1989]
             117454 [1990-1999]
              28248 [2000-...]
cut -f9 ech.292k.els.tab | sort | uniq -c
                    1 [data]
               2215 [v1.0, v1.1]
             188085 [v1.2]
              73343 [v1.3]
              23630 [v1.4]
               4801 [v1.6, v1.7]

# récupération sur serveur
tar -cz ech.292k.els.tab > ech.292k.els.tab.tgz
rsync -v ech.292k.els.tab.tgz $VM:~/.
```

```
# on y va et on prépare 5 morceaux
ssh $VM
tar -xzf ech.292k.els.tab.tgz
split -n 5 pdfpaths_292k.ls
wc -l xa*
               58403 xaa
               58413 xab
               58403 xac
               58479 xad
               58377 xae
              292075 total

# lancement service
cd toolz/grobid/grobid-service/
nohup mvn jetty:run-war &
cd ~

cat > lance_nomliste.sh
#! /bin/bash
for doc in `cat $1` ; do tgt=`echo $doc | sed 's/\.pdf$/.tei.xml/' | tr -s "/
()," "_" | sed 's/_data_
//'` ; curl -v -include --form input=@$doc 127.0.0.1:8080/processReferences >
stockage/resultats_292k_
gb_2015-01-09/$tgt ; done
```

Au retour :

- seuls 3059 sont bien passés
- 230636 passent en générant la même sortie d'erreur :

```
HTTP/1.1 100 Continue

HTTP/1.1 500 Internal Server Error
Content-Type: text/html; charset=iso-8859-1
Content-Length: 1427
Server: Jetty(6.1.10)

<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=ISO-8859-1"/>
<title>Error 500 Internal Server Error</title>
</head>
<body><h2>HTTP ERROR: 500</h2><pre>Internal Server Error</pre>
<p>RequestURI=/processReferences</p><p><i><small><a href="http://jetty.mortbay.org/">Powered
by Jetty://</a></small></i></p><br/>
<br/>
(...)
<br/>
<br/>
<br/>
<br/>

</body>
</html>
```

- plus 7 malformés qui correspondent à 3 ½ au départ (nom coupé => 2 pseudo fichiers) :
  1. `elsevier_r`
  2. `aw_ISY19940100000036_01486195_v42i2_0148619590900258_main.tei.xml`
  3. `elsevier_raw_IS`

4. Y19940770000054_00399140_v31i10sP1_0039914084802039_main.tei.xm
5. elsevier_raw_IST19960020000148_0012365X_
6. v162i1-3_0012365X9500319R_main.tei.xml
7. elsevier_raw_IST19980010000041_01419331_v22i6_S0141933198000866_m

– et donc enfin 58376 qui ne sont pas du tout passés

=> retours bugs et enseignements tirés

6.2) Exemple de lancement standard

Document : Montée en charge grobid

```
echantilloneur.pl --base par_lots/bases_quotas/els-PDFDOCS-path_type_period_vers_wostheme_issn_lot.tab
--pool par_lots/pools_dispos/els-BIBPOOL.tab  -n 300000 > ech.300k.els.tab
mv ech.300k.els.tab ech.292k.els.tab

# représentativité par périodes
cut -f7 ech.292k.els.tab | sort | uniq -c            cut -f9 ech.292k.els.tab | sort | uniq -c
          23271 [...-1959]                                     1 [data]
          52257 [1960-1979]                                 2215 [v1.0, v1.1]
          70845 [1980-1989]                               188085 [v1.2]
         117454 [1990-1999]                                73343 [v1.3]
          28248 [2000-...]                                 23630 [v1.4]
                                                            4801 [v1.6, v1.7]

# récupération sur serveur
tar -cz ech.292k.els.tab > ech.292k.els.tab.tgz
rsync -v ech.292k.els.tab.tgz $VM:~/.

# on y va et on prépare 5 morceaux
ssh $VM
tar -xzf ech.292k.els.tab.tgz
split -n 5 pdfpaths_292k.ls
wc -l xa*
           58403 xaa
           58413 xab
           58403 xac
           58479 xad
           58377 xae
          292075 total

# lancement service en nohup  # # # # # # # # # # # # # # # # # # # # # 1/2
cd toolz/grobid/grobid-service/
nohup mvn jetty:run-war &
cd ~

cat > lance_nomliste.sh
#! /bin/bash
outdir="resultats_292k_gb_2015-01-26"
total_nom=`wc -l $1` ;
compteur=0 ;
for doc in `cat $1` ;
        do compteur=$((compteur+1)) ;
        tgt=`echo $doc | sed 's/\.pdf$/.tei.xml/ ; s!/data/!!' | tr "/ (()," "_"`

        # appel du service  # # # # # # # # # # # # # # # # # # # # # 2/2
        curl --form input=@$doc 127.0.0.1:8080/processReferences > stockage/$outdir/$tgt ;
        # log dans la sortie ou dans nohup
        if mille=`echo $compteur | grep 000$` ;
                then echo -e "\n\n\n----------------------------------------------------\n  grobid-service
en cours : $compteur / $total_nom\n----------------------------------------------------\n\n\n" ;
        fi ;
done


tests résultats anormaux
------------------------
ls resultats_292k_gb_2015-01-26/ | grep -v "tei.xml" >> nom_fic_etranges-resultats_292k_gb_2015-01-26.ls

choix sous-ensemble à évaluer
-----------------------------
shuf xmlpaths_292k.ll | head -n 50000 | grep -Po "(?<= )/data.*xml$" | tar -czf stockage/50kxml.tgz -T -

tar -ztvf stockage/50kxml.tgz > extrait_IN_50k.tempo


correspondance GOLD <=> OUTPUT
```

```
-----------------------------
cat extrait_IN_50k.tempo | grep -Po "(?<= )/data.*$" | tr "/ ()," "_" |  sed 's/_data_// ;
s/\.xml/.tei.xml/'` > extrait_OUT_50k.ls

cd resultats_292k_gb/
tar czf ../extrait_OUT_50kxml.tgz -T ../extrait_OUT_50k.lsa
```

6.3) Lancement sur grande VM (4 mars)

Document : infos de mise en place

Voilà les caractéristiques pour la nouvelle VM qui abriterait Grobid :

PORT
====
Le port utilisé par défaut est en fait le 8080... Il serait interrogé depuis l'autre VM (client) vi-istex-rc.

SSH
===
clef publique : ssh-rsa (...)

PAQUETS
=======

Les 3 paquets absolument nécessaires pour lancer grobid :

  * la "openjdk-6-jdk" (ou une autre JDK >= 6) pour le lancement java
  * "maven" pour le build java
  * "git" pour les mises à jours d'application

Et voilà 8 paquets supplémentaires qui sont utiles:
  * utilitaires courants : "tree" et "unzip"
  * pour traitements XML:
    - "libxml2", "libxml2-utils"
    - "libxml2-dev" (pour les librairies perl "XML::LibXML" et python "lxml")
    - "zlib1g-dev" (pour la libraire perl "XML::LibXML")
  * pour diagnostics PDF:
    * "pdftk"
    * "poppler-utils" (pour avoir pdfinfo)

FICHIERS TEMPORAIRES
====================

J'ai regardé le service tourner et il y a bien 2 fichiers temporaires par traitement qui sont créés dans /tmp   (pas ~/tmp mais carrément à la racine)

Forme des noms du fichier créé:
MIME5626318907410971349.tmp
MIME8408513991555266075.tmp

et
origin5937360724333085108pdf
origin7200305170316414924pdf
…

Ce sont 2 fichiers qui ont une taille typique des fichiers pdf : entre 100k et 15M…

Etc