

Règles d'annotation de tableaux de données

Romane Guari

Sommaire

Introduction.....	3
La structure de l'ontologie	5
Annotation de l'en-tête.....	6
Distinction colonnes numériques et symboliques	6
Balise qc.....	6
Balise sc	6
Exemple	6
Annotation de la légende	8
Balise rc	8
Exemple	8
Annotation des lignes.....	9
Balise ri	9
Balise ai.....	9
Balise aii.....	9
Exemple	9
Application.....	12

Introduction

Ce document propose une méthode d'annotation sémantique des tableaux de données issus d'article scientifique à l'aide d'une ontologie. Cette méthode est destinée à être rendue automatique. Elle consiste à ajouter des balises HTML/XML dans le document contenant les tableaux de données. Pour un document donné, les balises liées à un tableau sont incluses dans une division du document HTML/XML ce qui est indiqué par la présence d'une balise div. Cette méthode d'annotation est exploitable pour tous les tableaux de données au format suivant et ayant un code HTML/XML associé similaire à cet exemple:

Headcell1	Headcell2	Headcell3	Headcell4
Cell 11	Cell12	Cell13	Cell14
Cell 21	Cell22	Cell23	Cell24

Table 1 : a caption with infoCaption

```
<div>
  <span class= "captions">
    <span id="cap001">
      <span class="label"> Table 1</span>
      : a caption with an infoCaption
    </span>
  </span>
  <table>
    <thead>
      <tr>
        <th>
          Headcell1
        </th>
        <th>
          Headcell2
        </th>
        <th>
          Headcell3
        </th>
        <th>
          Headcell4
        </th>
      </tr>
    </thead>
    <tbody>
      <tr>
        <td>
          Cell11
        </td>
        <td>
          Cell12
        </td>
        <td>
          Cell13
        </td>
        <td>
```

```

Cell14
</td>
</tr>
<tr>
<td>
Cell21
</td>
<td>
Cell22
</td>
<td>
Cell23
</td>
<td>
Cell24
</td>
</tr>
</tbody>
</table>
</div>

```

Notons que pour pouvoir être annoté par la méthode présentée, un tableau doit avoir une seule ligne de cellules d'en-tête et ne doit pas posséder de lignes ou de colonnes se divisant dans la suite du tableau.

Dans un premier temps, il est important d'expliquer la structure de l'ontologie utilisée pour appliquer la méthode. Ensuite, les différentes balises utilisées pour l'annotation seront expliquées avant de terminer par une application sur un exemple dans le domaine des emballages bio-sourcés avec l'utilisation de l'ontologie [TRANSMAT V115](#).

La structure de l'ontologie

La structure de l'ontologie repose sur la structure des schémas de base de données et est organisée en concepts numériques, concepts symboliques et relations. Pour le traitement de documents dans le domaine des emballages bio-sourcés, l'ontologie TRANSMAT disponible sur @Web est utilisée.

Les concepts symboliques sont décrits par un nom, une liste de synonymes pour ce nom et une taxonomie des valeurs possibles. Par exemple, le concept symbolique «Produits alimentaires» a une taxonomie contenant des valeurs telles que «Viande fraîche» ou «Cuisses de grenouilles».

Les concepts numériques sont décrits par un nom, une liste de synonymes pour ce nom et un ensemble d'unités dans lequel le concept peut être exprimé et éventuellement une plage numérique. Le concept numérique «Teneur en eau» comprend un ensemble d'unités comprenant le pourcentage avec une restriction numérique de [0.0,100.0].

Les relations sont décrites par le nom de la relation et sa signature. La signature d'une relation est composée d'un concept résultat unique et un ou plusieurs concepts d'accès. On remarque qu'un concept résultat d'une relation est un concept numérique. La signature de la relation «Matrix properties Thickness» de l'ontologie possède «Thickness» comme concept résultat et «Matrix» comme concept d'accès.

Cette structure de l'ontologie se retrouve dans la méthode d'annotation. En effet, les balises sont définies en fonction des types de concepts identifiés dans le tableau de données.

Annotation de l'en-tête

Pour annoter l'en-tête du tableau, deux balises sont utilisées. Elles permettent d'annoter les concepts identifiés dans une colonne. Il s'agit de la balise qc pour les concepts numériques et sc pour les concepts symboliques. Ces balises HTML/XML se placent autour du contenu de la cellule d'en-tête de la colonne que l'on annote. Cependant, ces balises ne dépendent pas uniquement du contenu de la cellule d'en-tête mais également de toutes les autres cellules de la colonne.

Distinction colonnes numériques et symboliques

La balise qc est utilisé pour l'annotation des concepts quantités et est donc associé à une colonne dite numérique, contrairement à la balise sc destinée à l'annotation d'une colonne représentant un concept symbolique et qui sera donc associé à une colonne dite symbolique.

Afin de déterminer si une colonne est symbolique ou numérique, on examine le contenu de chacune des cellules la composant. On appelle indice numérique, un nombre, et un indice symbolique, un mot. Une cellule est numérique si l'on note la présence d'un nombre en notation scientifique ou une unité dans son contenu. Autrement, on compare le nombre d'indices symboliques et numériques :

- si le nombre d'indices numériques est supérieur au nombre d'indices symboliques, la cellule est considérée comme numérique
- si le nombre d'indices numériques est inférieur au nombre d'indices symboliques, la cellule est considérée comme symbolique
- si le nombre d'indices numériques est égal au nombre d'indices symboliques, la cellule est considérée de type inconnu.

Une fois que toutes les cellules ont été classifiées, la colonne est elle-même classifiée en sachant que:

- si le nombre de cellules numériques est supérieur ou égal au nombre de cellules symboliques alors la colonne est numérique
- si le nombre de cellules numériques est inférieur au nombre de cellules symboliques alors la colonne est symbolique

Balise qc

La balise qc a un attribut «type» pour spécifier le concept numérique de l'ontologie représenté dans le contenu de toutes les cellules incluant la cellule d'en-tête de la colonne et un attribut facultatif «unit» désignant l'unité pouvant apparaître dans la colonne. Elle possède également un autre attribut facultatif «exponent» auquel on associe, s'il existe, l'exposant des valeurs contenu dans la colonne. Tous ces attributs peuvent être vides si aucun concept de l'ontologie n'est identifié dans la colonne.

Balise sc

La balise sc a un attribut «type» pour spécifier le concept symbolique de l'ontologie représenté dans le contenu de toutes les cellules y compris la cellule d'en-tête de la colonne. De la même manière que pour la balise qc, cet attribut peut être vide si aucun concept de l'ontologie n'est identifié dans la colonne.

Exemple

```
<thead>
  <tr>
    <th>
      <sc type="Symbolic_concept">
        Headcell1
      </sc>
```

```

</th>
<th>
    <qc type="Quantity_concept1">
    Headcell2
    </qc>
</th>
<th>
    <qc type="Quantity_concept2" unit="conceptUnit">
    Headcell3
    </qc>
</th>
<th>
    <qc type="Quantity_concept2" unit="conceptUnit">
    Headcell4
    </qc>
</th>
</tr>
</thead>

```

Annotation de la légende

Balise rc

Cette balise rc est utilisée pour annoter le ou les concepts relations de l'ontologie identifiés dans le tableau. Elle doit être placée après la balise qui indique l'id de la légende. Cette balise a un attribut «type» pour spécifier les concepts relation de l'ontologie représentés dans le contenu du tableau et sa légende. Notons qu'un tableau peut contenir plusieurs relations. Néanmoins, si une relation est représentée plusieurs fois dans le tableau de données, une seule balise rc est nécessaire. Il peut également n'en contenir aucune, en particulier dans le cas où le tableau ne contient que des colonnes symboliques annotées par la balise sc ou alors dans le cas où aucun concept numérique annoté est un concept résultat d'une relation. En effet, un concept relation ne doit pas être annoté si le concept de résultat n'a pas été annoté sur l'en-tête d'une des colonnes du tableau.

Exemple

```
<span class="captions">
  <span id="cap001">
    <rc type="Relation_concept1"></rc>
    <rc type="Relation_concept2"></rc>
    <span class="label"> Table 1</span>
    : a caption with an infoCaption
  </span>
</span>
```


Annotation des lignes

Au sein des lignes des tableaux de données, on ajoute trois balises afin d'annoter les instances des concepts de relation qui ont été identifiés. Ces trois balises sont les suivantes: ri, ai et aii.

Balise ri

Cette balise est utilisée pour annoter une instance de relation. Elle est imbriquée dans les balises <tr> indiquant le début d'une ligne dans un tableau de données et placée avant les balises <td> utilisées pour la déclaration d'une cellule. La balise possède deux attributs:

- un attribut «type» prenant comme valeur un concept de relation qui a été préalablement identifié dans le tableau lors de l'annotation faite sur la légende
- un attribut «id», associé à un entier positif et utilisé pour identifier de manière unique un concept de relation dans une ligne. Cet attribut permet de traiter le cas où un concept relation possède plusieurs instances dans une même ligne.

Les balises ai et aii sont imbriquées dans les balises ri, elles permettent d'annoter des instances de concepts symboliques ou numériques liés à la relation.

Balise ai

Une balise ai est utilisée pour identifier une instance de concept, qu'il soit numérique ou symbolique, présente dans une cellule du tableau. Cette balise possède un attribut «type» associé à un terme de l'ontologie. Ce terme doit appartenir à la taxonomie du concept annoté par la balise qc ou sc de la colonne dans laquelle se trouve la cellule. Un autre attribut «id» est utilisé pour identifier de manière unique une instance d'un concept dans une ligne. Cette balise se trouve à deux endroits sur la même ligne. D'une part, elle est imbriquée dans les balises ri en sachant que le terme du concept annoté par la balise ai doit appartenir à la signature de la relation donnée par la balise ri dans laquelle il est imbriqué. D'autre part, cette balise est imbriquée dans les balises td où se trouve la valeur nécessaire à l'instanciation du concept. Dans le cas où la cellule du tableau appartient à une colonne numérique, cette deuxième apparition de la balise ai se fait uniquement autour de la valeur numérique. Par exemple, si la case du tableau contient la valeur « 2.34±0.07f », la balise ai entoure uniquement « 2.34±0.07 ».

Balise aii

Une balise aii possède également un attribut «type» associé à un terme de l'ontologie appartenant à la taxonomie d'un concept numérique ou symbolique. Contrairement à la balise ai, une balise aii apparaît uniquement imbriquée dans les balises ri. Cette balise est utilisée pour annoter une instance implicite d'un concept présent dans la signature de la relation annotée par la balise ri. Cette instance peut se trouver dans la légende, dans l'en-tête ou encore dans les notes du tableau. Il s'agit donc d'un terme d'un concept qui appartient à la relation annotée par la balise ri mais n'étant pas représenté dans la ligne.

Notons que parmi les balises ai ou aii imbriquées dans une balise ri annotant un concept relation, on ne doit retrouver que des termes appartenant à la taxonomie d'un concept quantité ou symbolique qui appartient à la signature du concept relation. De plus, chaque concept possède au maximum une seule instanciation, c'est-à-dire, au maximum une seule balise soit ai soit aii avec un terme de sa taxonomie.

Exemple

```
<tbody>
  <tr>
    <ri type="Relation_concept1" id="0">
      <ai type="infoCaption"> </ai>
```

```

                <ai type="Symbolic_concept" id="0"> </ai>
                <ai type="Quantity_concept1" id="0"> </ai>
            </ri>
            <ri type=" Relation_concept2" id="1">
                <ai type=" Symbolic_concept " id="0"> </ai>
                <ai type=" Quantity_concept2" id="1" > </ai>
            </ri>
            <ri type=" Relation_concept2" id="2">
                <ai type=" Symbolic_concept " id="0"> </ai>
                <ai type=" Quantity_concept2" id="2"> </ai>
            </ri>
        <td>
            <ai type=" Symbolic_concept" id="0">
                Cell11
            </ai>
        </td>
        <td>
            <ai type=" Quantity_concept1" id="0">
                Cell12
            </ai>
        </td>
        <td>
            <ai type="Quantity_concept2" id="1">
                Cell13
            </ai>
        </td>
        <td>
            <ai type="Quantity_concept2" id="2">
                Cell14
            </ai>
        </td>
    </tr>
    <tr>
        <ri type=" Relation_concept1" id="0">
            <aii type="infoCaption"> </aii>
            <ai type="Symbolic_concept" id="0"> </ai>
            <ai type="Quantity_concept1" id="0"> </ai>
        </ri>
        <ri type=" Relation_concept2" id="1">
            <ai type=" term_Symbolic_concept " id="0"> </ai>
            <ai type=" term_Quantity_concept2" id="1" > </ai>
        </ri>
        <ri type=" Relation_concept2" id="2">
            <ai type=" Symbolic_concept " id="0"> </ai>
            <ai type=" Quantity_concept2" id="2"> </ai>
        </ri>
    <td>
        <ai type=" Symbolic_concept" id="0">
            Cell21
        </ai>
    </td>
    <td>
        <ai type=" Quantity_concept1" id="0">
            Cell22
        </ai>
    </td>
    <td>

```

```

        <ai type=" Quantity_concept2" id="1">
        Cell23
        </ai>
    </td>
    <td>
        <ai type=" Quantity_concept2" id="2">
        Cell24
        </ai>
    </td>
</tr>
</tbody>

```

L'exemple a été annoté selon la méthode présentée ci-dessus et en utilisant l'ontologie TRANSMAT V115. Ce tableau est tiré de l'article «Barrier properties of chitosan coated polyethylene»¹ Mia Kurek, Mario Ščetar, Andree Voilley, Kata Galić, Frédéric Debeaufort, Journal of Membrane Science, Volumes 403–404, 2012, Pages 162-168. ISSN 0376-7388.

Sample	WVP $\times 10^{-13}$ (g/m s Pa) Δ RH 70%	WVP $\times 10^{-13}$ (g/m s Pa) Δ RH 45%	WVP $\times 10^{-13}$ (g/m s Pa) Δ RH 33%
PE	4.62 \pm 0.73f	5.55 \pm 0.23f	7.72 \pm 2.58f
CS coated PE	12.37 \pm 1.14f	6.67 \pm 0.23f	7.88 \pm 2.39f
PECSEinv	9.14 \pm 1.09f	6.41 \pm 3.28f	2.85 \pm 0.34f
CSA	4161.31 \pm 656.17a,b	2199.80 \pm 1048.33d,e	25.71 \pm 2.20f
CSE	4100.77 \pm 588.88a,b	2884.37 \pm 346.43b,c,d,e	38.71 \pm 2.61f
CSAGLY	5410.08 \pm 1543.67a	1905.39 \pm 149.64e	26.14 \pm 1.24f
CSEGLY	3481.46 \pm 343.88b,c,d	2635.38 \pm 414.28c,d,e	105.17 \pm 6.57f

Different letters (a–f) indicate significant differences between formulations ($p < 0.05$).

¹ <https://doi.org/10.1016/j.memsci.2012.02.037>


```

</td>
</tr>
<tr>
<ri type="H2O Permeability_relation" id="0">
  <ai type="PE chitosan coated samples" id="0"> </ai>
  <ai type="H2O Permeability" id="0"> </ai>
  <aII type="Temperature"> </aII>
  <aII type="Relative_Humidity"> </aII>
</ri>
<ri type="H2O Permeability_relation" id="1">
  <ai type="PE chitosan coated samples" id="0"> </ai>
  <ai type="H2O Permeability" id="1"> </ai>
  <aII type="Temperature"> </aII>
  <aII type="Relative_Humidity"> </aII>
</ri>
<ri type="H2O Permeability_relation" id="2">
  <ai type="PE chitosan coated samples" id="0"> </ai>
  <ai type="H2O Permeability" id="2"> </ai>
  <aII type="Temperature"> </aII>
  <aII type="Relative_Humidity"> </aII>
</ri>
<td class="align-left">
  <ai type="PE chitosan coated samples" id="0">
    CS coated PE
  </ai>
</td>
<td class="align-char">
  <ai type="H2O Permeability" id="0">
    12.37&nbsp;&pm&nbsp;1.14</ai>f
</td>
<td class="align-char">
  <ai type="H2O Permeability" id="1">
    6.67&nbsp;&pm&nbsp;0.23</ai>f
</td>
<td class="align-char">
  <ai type="H2O Permeability" id="2">
    7.88&nbsp;&pm&nbsp;2.39</ai>f
</td>
</tr>
<tr>
<ri type="H2O Permeability_relation" id="0">
  <ai type="Packaging" id="0"> </ai>
  <ai type="H2O Permeability" id="0"> </ai>
  <aII type="Temperature"> </aII>
  <aII type="Relative_Humidity"> </aII>
</ri>
<ri type="H2O Permeability_relation" id="1">
  <ai type="Packaging" id="0"> </ai>
  <ai type="H2O Permeability" id="1"> </ai>
  <aII type="Temperature"> </aII>
  <aII type="Relative_Humidity"> </aII>
</ri>
<ri type="H2O Permeability_relation" id="2">
  <ai type="Packaging" id="0"> </ai>
  <ai type="H2O Permeability" id="2"> </ai>
  <aII type="Temperature"> </aII>
  <aII type="Relative_Humidity"> </aII>
</ri>
<td class="align-left">
  <ai type="Packaging">
    PECSEinv

```

```

        </ai>
    </td>
    <td class="align-char">
        <ai type="H2O Permeability" id="0">
            9.14&nbsp;&nbsp;&nbsp;±&nbsp;&nbsp;&nbsp;1.09</ai>f
    </td>
    <td class="align-char">
        <ai type="H2O Permeability" id="1">
            6.41&nbsp;&nbsp;&nbsp;±&nbsp;&nbsp;&nbsp;3.28</ai>f
    </td>
    <td class="align-char">
        <ai type="H2O Permeability" id="2">
            2.85&nbsp;&nbsp;&nbsp;±&nbsp;&nbsp;&nbsp;0.34</ai>f
    </td>
</tr>
<tr>
    <ri type="H2O Permeability_relation" id="0">
        <ai type="Packaging" id="0"> </ai>
        <ai type="H2O Permeability" id="0"> </ai>
        <aII type="Temperature"> </aII>
        <aII type="Relative_Humidity"> </aII>
    </ri>
    <ri type="H2O Permeability_relation" id="1">
        <ai type="Packaging" id="0"> </ai>
        <ai type="H2O Permeability" id="1"> </ai>
        <aII type="Temperature"> </aII>
        <aII type="Relative_Humidity"> </aII>
    </ri>
    <ri type="H2O Permeability_relation" id="2">
        <ai type="Packaging" id="0"> </ai>
        <ai type="H2O Permeability" id="2"> </ai>
        <aII type="Temperature"> </aII>
        <aII type="Relative_Humidity"> </aII>
    </ri>
    <td class="align-left">
        <ai type="packaging" id="0">
            CSA
        </ai>
    </td>
    <td class="align-char">
        <ai type="H2O Permeability" id="0">
            4161.31&nbsp;&nbsp;&nbsp;±&nbsp;&nbsp;&nbsp;656.17</ai>a,b
    </td>
    <td class="align-char">
        <ai type="H2O Permeability" id="1">
            2199.80&nbsp;&nbsp;&nbsp;±&nbsp;&nbsp;&nbsp;1048.33</ai>d,e
    </td>
    <td class="align-char">
        <ai type="H2O Permeability" id="2">
            25.71&nbsp;&nbsp;&nbsp;±&nbsp;&nbsp;&nbsp;2.20</ai>f
    </td>
</tr>
<tr>
    <ri type="H2O Permeability_relation" id="0">
        <ai type="CSE plasticized films" id="0"> </ai>
        <ai type="H2O Permeability" id="0"> </ai>
        <aII type="Temperature"> </aII>

```

```

        <aii type="Relative_Humidity"> </aii>
    </ri>
    <ri type="H2O Permeability_relation" id="1">
        <ai type="CSE plasticized films" id="0"> </ai>
        <ai type="H2O Permeability" id="1"> </ai>
        <aii type="Temperature"> </aii>
        <aii type="Relative_Humidity"> </aii>
    </ri>
    <ri type="H2O Permeability_relation" id="2">
        <ai type="CSE plasticized films" id="0"> </ai>
        <ai type="H2O Permeability" id="2"> </ai>
        <aii type="Temperature"> </aii>
        <aii type="Relative_Humidity"> </aii>
    </ri>
    <td class="align-left">
        <ai type="CSE plasticized films" id="0">
            CSE
        </ai>
    </td>
    <td class="align-char">
        <ai type="H2O Permeability" id="0">
            4100.77&nbsp;±&nbsp;588.88</ai>a,b
    </td>
    <td class="align-char">
        <ai type="H2O Permeability" id="1">
            2884.37&nbsp;±&nbsp;346.43</ai>b,c,d,e
    </td>
    <td class="align-char">
        <ai type="H2O Permeability" id="2">
            38.71&nbsp;±&nbsp;2.61</ai>f
    </td>
</tr>
<tr>
    <ri type="H2O Permeability_relation" id="0">
        <ai type="Packaging" id="0"> </ai>
        <ai type="H2O Permeability" id="0"> </ai>
        <aii type="Temperature"> </aii>
        <aii type="Relative_Humidity"> </aii>
    </ri>
    <ri type="H2O Permeability_relation" id="1">
        <ai type="Packaging" id="0"> </ai>
        <ai type="H2O Permeability" id="1"> </ai>
        <aii type="Temperature"> </aii>
        <aii type="Relative_Humidity"> </aii>
    </ri>
    <ri type="H2O Permeability_relation" id="2">
        <ai type="Packaging" id="0"> </ai>
        <ai type="H2O Permeability" id="2"> </ai>
        <aii type="Temperature"> </aii>
        <aii type="Relative_Humidity"> </aii>
    </ri>
    <td class="align-left">
        <ai type="Packaging" id="0">
            CSAGLY
        </ai>
    </td>
    <td class="align-char">
        <ai type="H2O Permeability" id="0">
            5410.08&nbsp;±&nbsp;1543.67</ai>a

```



```

</td>
<td class="align-char">
<ai type="H2O Permeability" id="1">
1905.39&nbsp;&nbsp;&nbsp;±&nbsp;&nbsp;&nbsp;149.64</ai>e

</td>
<td class="align-char">
<ai type="H2O Permeability" id="2">
26.14&nbsp;&nbsp;&nbsp;±&nbsp;&nbsp;&nbsp;1.24</ai>f

</td>
</tr>
<tr>
<ri type="H2O Permeability_relation" id="0">
<ai type="CSE plasticized films" id="0"> </ai>
<ai type="H2O Permeability" id="0"> </ai>
<aII type="Temperature"> </aII>
<aII type="Relative_Humidity"> </aII>
</ri>
<ri type="H2O Permeability_relation" id="1">
<ai type="CSE plasticized films" id="0"> </ai>
<ai type="H2O Permeability" id="1"> </ai>
<aII type="Temperature"> </aII>
<aII type="Relative_Humidity"> </aII>
</ri>
<ri type="H2O Permeability_relation" id="2">
<ai type="CSE plasticized films" id="0"> </ai>
<ai type="H2O Permeability" id="2"> </ai>
<aII type="Temperature"> </aII>
<aII type="Relative_Humidity"> </aII>
</ri>
<td class="align-left">
<ai type="CSE plasticized films" id="0">
CSEGLY
</ai>
</td>
<td class="align-char">
<ai type="H2O Permeability" id="0">
3481.46&nbsp;&nbsp;&nbsp;±&nbsp;&nbsp;&nbsp;343.88</ai>b,c,d

</td>
<td class="align-char">
<ai type="H2O Permeability" id="1">
2635.38&nbsp;&nbsp;&nbsp;±&nbsp;&nbsp;&nbsp;414.28</ai>c,d,e

</td>
<td class="align-char">
<ai type="H2O Permeability" id="2">
105.17&nbsp;&nbsp;&nbsp;±&nbsp;&nbsp;&nbsp;6.57</ai>f

</td>
</tr>
</tbody>
</table>
</div>
<p class="legend">
<p id="spar0035">
PE, polyethylene; CS
coated PE, chitosan (CSE) coated polyethylene; PESCEinv, coating exposed to dry compartment; CSA, chitosan film prepared
with aqueous acid solvent; CSE, chitosan film prepared with hydroalcoholic acid solvent; CSAGLY and CSEGLY, glycerol
plasticized samples.

</p>
<p id="spar0040">

```

indicate significant differences between formulations (

</p>
</div>

Different letters (a–f)

p
< 0.05).

</p>