Data tables annotation rules

Romane Guari

12/2020

Contents

Introduction

This document puts forward a semantic annotation method of data tables from scientific papers using an ontology. This method, based on adding HTML/XML tags in the scientific paper is intended to be automatic. In the HTML / XML code of a scientific article, data tables are located in a division indicated by div tags. This annotation method is exploitable to all data tables in the following format and which have an associate HTML code similar to this example:

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>
```

In order to be annotated by the method presented, a data table must have a single row of header cells and should not have rows or columns dividing themselves in the rest of the table.

As a first step, it is important to explain the ontology structure used in order to apply this method. Then, the different tags used will be presented before ending with an application on an example in the field of bio-sourced packaging using the TRANSMAT V115 ontology.

The ontology structure

The ontology structure relies on the structure of database schemas and is organised into numeric types, symbolic types and relations. For the processing of documents in the field of bio-sourced packaging, the TRANSMAT ontology available on @Web[1] is used.

Symbolic types are described by a type name, a list of synonyms for the type name and a taxonomy of possible values. For example, the symbolic concept “Food Products” has a taxonomy containing values such as “Fresh meat” or “Frog legs”.

Numeric types are described by a type name, a list of synonyms for the type name and the set of units in which the type can be expressed and eventually a numerical range. The quantity concept “Water content” has a set of units including percent with a numerical restriction [0.0,100.0].

Relations are described by the name of the relation and its signature. The signature of a relation is divided into one result type and several access types. The signature of the “Matrix properties Thickness” relation of the ontology is made up of “Thickness” as unique result concept and “Matrix” as access concept.

The structure of ontology is therefore found in the definition of the annotation method. Indeed, tags are defined in accordance with types of concepts identified in the data table.

[1] https://ico.iate.inra.fr/atWeb/

Header annotation

To annotate the data table header, two tags are used. These tags enable the annotation of a column with the identified concept. A qc tags is used for a quantity concept and a sc tag for a symbolic concept. These HTML / XML tags are placed around the content of the header cell being of the column annotated. The concept associated with a column by one of these tags does not only depend on the content of the header cell but also on all the other cells in the column.

Distinction between numerical and symbolic columns

In order to differentiate if a column is either numeric or symbolic, it is necessary to analyse the cells' content. A numeric indicator is a number and a symbolic indicator is a word. A cell is numeric if it contains a unit or a number in scientific notation. Otherwise, a comparison between the number of symbolic indicator and the number of numeric indicator is needed:

- if the number of numeric indicator is superior than the number of symbolic indicator, the cell is considered numeric
- if the number of numeric indicator is inferior than the number of symbolic indicator, the cell is considered numeric
- else if the number of numerical indicator is equal to the number of symbolic indicator, the cell is considered of unknown type.

Once all the cells have been classified, the column is itself classified knowing that:

- if the number of numeric cells is superior or equal to the number of symbolic cells then the column is numeric
- else if the number of numeric cells is inferior than the number of symbolic cells then the column is symbolic

qc tag

The qc tag has a "type" attribute to specify the quantity concept of the ontology represented in the content of all cells including the header cell of the column. The tag has also an optional "unit" attribute for the unit that can appear in the column. It has another optional attribute "exponent" associated to the exponent of the values in the column, if it exists. All these attributes are empty if there is no quantity concept of the ontology identified in the column.

sc tag

The sc tag has a "type" attribute to specify the symbolic concept of the ontology represented in the content of all cells including the column header cell. Such as the qc tag, this attribute is empty if there is no symbolic concept of the ontology identified in the column.

Example

```
<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>
```

Caption annotation

rc tag

This rc tag is used to annotate a relation concept of the ontology which has been identified in the data table. It must be placed after the tag denoting the id of the caption. This tag has a "type" attribute to specify relation concepts of the ontology represented in the content of the data table and its caption. Note that a data table can have several relations. However, if a relation concept is represented more than once in a data table, only one rc tag is needed. It is possible that a data table does not contain any relation concept belonging to the ontology. For example, when a data table has only symbolic columns annotated by the sc tag or when there is no quantity concept result of the relation's signature. Indeed, a relation concept should not be annotated with a rc tag if the result concept has not been previously annotated in the header of one of the table columns.

Example

```
<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>
```

Rows annotation

In rows of the data tables, three tags are added to annotate instances of relation concepts that have been identified. These three tags are: ri, ai and aii.

ri tag

This tag is used to annotate an instance of a relation concept in the data table. It should be nested in the <tr> tags indicating the beginning of a row in a data table and placed before the <td> tags used for the declaration of a cell. This tag has two attributes:

- a "type" attribute taking for value a concept of relation which has been previously identified in the data table in the annotation of the caption

- an "id" attribute, a positive integer, used to uniquely identify a concept of relationship in a row. This attribute is used to deal with the case when a relation concept has several instances in one row.

The ai and aii tags are nested in the ri tags, they allow to annotate instances of symbolic or numeric concepts linked to the instantiation of the relation.

ai tag

An ai tag is used to identify an instance of concept present in a cell of the data table. It has a "type" attribute associated to a term of the ontology. This term must belong to the taxonomy of the concept annotated in the column in which the cell is located. An “id” attribute is used to uniquely identify an instance of a concept in a row. This tag has to be found in two places on the same line. On the one hand, it is nested in the ri tags knowing that the annotated term by the ai tag must be a term in the taxonomy of a concept and this concept has to belong to the signature of the relation annotated in the ri tag. On the other hand, this tag is nested in the td tags where the value necessary for the instantiation of the concept is found. In the case where a cell of the data table belongs to a numeric column, the ai tag just includes the numeric value. For example, if the cell in the table contains the value "2.34 ± 0.07f", the ai tag surrounds "2.34 ± 0.07".

aii tag

An aii tag also has a “type” attribute associated with an ontology term belonging to the taxonomy of a numerical or symbolic concept. Unlike the ai tag, an aii tag only appears nested within the ri tags. This tag is used to annotate an implicit instance of a concept belonging to the signature of the relation annotated by the ri tag. This instance can be found in the legend, in the header, or in the notes of the table. It is therefore a term of a concept that belongs to the relation annotated by the ri tag but not represented in the line.

Note that among the ai or aii tags nested in a ri tag annotating a relation concept, only terms belonging to the taxonomy of a quantity or symbolic concept belonging to the signature of the relation concept must be find. Moreover, each concept has only one instantiation, in other words, each concept has a unique tag either ai or aii with a term of its taxonomy.

Example

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

Example

The example has been annotated according to the above presented approach and using based the TRANSMAT ontology. This data table is taken from «Barrier properties of chitosan coated polyethylene»[2] Mia Kurek, Mario Ščetar, Andree Voilley, Kata Galić, Frédéric Debeaufort, Journal of Membrane Science,Volumes 403–404, 2012, Pages 162-168, ISSN 0376-7388.

Table 1 . Water vapour permeability (WVP) at 25 °C of PE films coated with chitosan compared to chitosan self standing films prepared with different casting solvents and plasticizers.

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

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.

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

```
<div class="tables frame-topbot rowsep-0 colsep-0" id="tbl0005">
	<span class="captions">
		<span>
			<rc type="H2O Permeability_relation"> </rc>
			<p id="spar0030">
				<span class="label">Table 1</span>
				. Water vapour permeability (WVP) at 25 °C of PE films coated with chitosan compared to chitosan self standing films prepared with different casting solvents and plasticizers.
			</p>
		</span>
	</span>
	<div class="groups">
		<table>
			<thead class="valign-top">
				<tr>
					<th scope="col" class="rowsep-1">
						<sc type="Packaging">
							Sample
						</sc>
					</th>
					<th scope="col" class="rowsep-1 align-left">
						<qc type="H2O Permeability" unit="Gram by reciprocal meter by reciprocal second by reciprocal pascal" exponent="10^-13">
						WVP × 10
						<sup>
							−13
						</sup>
						(g/m s Pa)
						Δ RH 70%
						</qc>
					</th>
```

[2] https://doi.org/10.1016/j.memsci.2012.02.037

```
                                        <th scope="col" class="rowsep-1 align-left">
                                                <qc type="H2O Permeability" unit="Gram by reciprocal meter by reciprocal second by reciprocal pascal" exponent="10^-13">
                                                WVP&amp;nbsp;×&amp;nbsp;10
                                                <sup>
                                                        −13
                                                </sup>
                                                 (g/m&amp;nbsp;s&amp;nbsp;Pa)
                                                 Δ RH 45%
                                                 </qc>
                                        </th>
                                        <th scope="col" class="rowsep-1 align-left">
                                                <qc type="H2O Permeability" unit="Gram by reciprocal meter by reciprocal second by reciprocal pascal" exponent="10^-13">
                                                WVP&amp;nbsp;×&amp;nbsp;10
                                                <sup>
                                                        −13
                                                </sup>
                                                 (g/m&amp;nbsp;s&amp;nbsp;Pa)
                                                 Δ RH 33%
                                                 </qc>
                                        </th>
                                </tr>
                        </thead>
                        <tbody>
                                <tr>
                                        <ri type="H2O Permeability_relation" id="0">
                                                <ai type="PE film" 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 film" 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 film" 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 film" id="0">
                                                PE
                                                </ai>
                                        </td>
                                        <td class="align-char">
                                                <ai type="H2O Permeability" id="0">
                                                4.62&amp;nbsp;±&amp;nbsp;0.73</ai>f

                                        </td>
                                        <td class="align-char">
                                                <ai type="H2O Permeability" id="1">
                                                5.55&amp;nbsp;±&amp;nbsp;0.23</ai>f

                                        </td>
                                        <td class="align-char">
                                                <ai type="H2O Permeability" id="2">
                                                7.72&amp;nbsp;±&amp;nbsp;2.58</ai>f
```

```
            </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 ± 1.14</ai>f

            </td>
            <td class="align-char">
                        <ai type="H2O Permeability" id="1">
                        6.67 ± 0.23</ai>f
            </td>
            <td class="align-char">
                        <ai type="H2O Permeability" id="2">
                        7.88 ± 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 ± 1.09</ai>f

                </td>
                <td class="align-char">
                    <ai type="H2O Permeability" id="1">
                    6.41 ± 3.28</ai>f

                </td>
                <td class="align-char">
                    <ai type="H2O Permeability" id="2">
                    2.85 ± 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 ± 656.17</ai>a,b

                </td>
                <td class="align-char">
                    <ai type="H2O Permeability" id="1">
                    2199.80 ± 1048.33</ai>d,e

                </td>
                <td class="align-char">
                    <ai type="H2O Permeability" id="2">
                    25.71 ± 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 ± 588.88</ai>a,b

</td>
<td class="align-char">
<ai type="H2O Permeability" id="1">
2884.37 ± 346.43</ai>b,c,d,e

</td>
<td class="align-char">
<ai type="H2O Permeability" id="2">
38.71 ± 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 ± 1543.67</ai>a

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

                                    </td>
                                    <td class="align-char">
                                        <ai type="H2O Permeability" id="2">
                                        26.14 ± 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 ± 343.88</ai>b,c,d

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

                                    </td>
                                    <td class="align-char">
                                        <ai type="H2O Permeability" id="2">
                                        105.17 ± 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">
```

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