# - Guidelines -

# Encoding of Header

# training data for Grobid

| | |
|---|---|
| Document version: | 0.02 |
| Author(s): | Patrice Lopez |
| Date: | 08-09-2013 |

## 1. GENERATED TRAINING FILES

For producing pre-annotated training files for Grobid based on the existing models, see the instructions for running the software in batch on the Wiki page: https://github.com/kermitt2/grobid/wiki/Grobid-batch-quick-start

After running Grobid on a set of PDF files using methods for creating training data, each article comes with at least:
- the PDF used to generate the training data, for instance *toto.pdf*,
- a pre-annotated file for the header segments: *toto.header.tei.xml*,
- the list of tokens considered for the header and the associated features: *toto.header* - this file can be ignored to a large extend, but is necessary for training Grobid.

In addition, depending on the metadata available in the article's header, the following training files are also produced:
- a pre-annotated file for the detailed affiliation and address recognition: *toto.affiliation.tei.xml*,
- a pre-annotated file for the detailed authors recognition: *toto.authors.tei.xml*,
- a pre-annotated file for the detailed reference segment analysis: *toto.header-references.xml*,
- a pre-annotated file for the detailed date analysis: *toto.date.xml*.

*The pre-annotated affiliation-address file is ignored in the current correction process.*

All the other files must be reviewed and corrected manually before being added to the training data. Taking into account additional training data requires Grobid to re-create its models.

The most important principle when correcting the pre-annotated training data is to **keep the stream of text untouched**. Only the tags can be moved, the text itself shall not be modified or corrected. The stream of text present in the training file after extraction of the content of the PDF, is similar to the stream of text Grobid will have to processed once the models created. It is thus important to have Grobid trained on these realistic input.

In the standard Grobid installation, examples of existing annotations can be found under *grobid-trainer/resources/dataset/header/corpus*.

*In the current correction process, no xml file shall be added:*
*- if an XML file is incorrectly produced, e.g. toto.header-references.xml is produced while a chunk of text was incorrectly identified as reference in the header file, then the XML file must be removed,*
*- if an XML file is missing, e.g. a chunk of text was a reference but was pre-annotated automatically by Grobid as a note, then the additional XML file toto.header-references.xml shall not be created.*
*XML files are therefore either modified or deleted, but never created.*

## 2. ENCODING OF HEADERS

### 2.1 General tagset

For header training files (training files **.header.tei.xml*), we follow the TEI and we use the following tags:

**<titlePart>** for title (included in a <docTitle> element).

**<docAuthor>** for the complete sequence of authors. The content of this tag will be subsequently parsed by the header "name" model.

**<affiliation>** for the complete affiliation field. The content of this tag together with the content of the address tag will be subsequently parsed by the "affiliation-address" model.

*Note that the further pre-annotated affiliation-address file is ignored in the current correction process.*

**<address>** for the complete address field. The content of this tag together with the content of the affiliation tag will be subsequently parsed by the "affiliation-address" model.

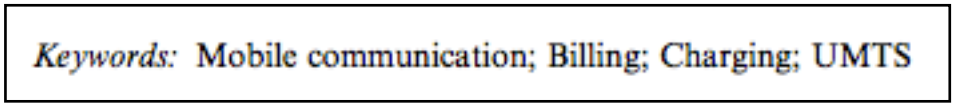
** ** for the abstract block (including some marker phrases such as “Abstract”, “Résumé de l’article: ”, etc. see the example bellow).

** ** for the start of the introduction section (i.e. just after the header ends).
(note that you may see in the older part of the corpus, alternatives such as  or even <introduction>, which are supported but not standard)

**<keyword>** for the keyword field (including possibly the marker phrases “Keywords”, or “Index Termes”, etc.). This covers also the controlled subject-headers. If type of keywords/ subject-headers is indicated in the header, an additional attribute can give some more information about the type of keywords (PACS in Physics, or ACM classes for ACM articles, etc.).

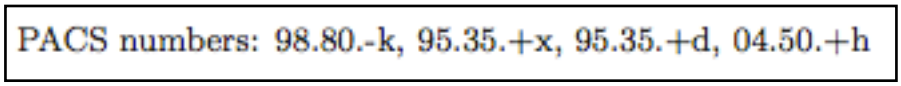
Example:

a) default keywords without any indication




```
<keyword>Keywords: Mobile communication; Billing; Charging; UMTS</keyword>
```

b) PACS keywords with subject-header schema indication




```
<keyword type="PACS">PACS numbers: 98.80.-k, 95.35.+x, 95.35.+d, 04.50.+h</
keyword>
```

**<email>** for encoding the email address (including possibly the marker phrases “Email”, etc.).

**<phone>** for encoding the contact phone number (including possibly the marker phrases “phone”, “tel.”, etc.)

**<ptr>** for web url corresponding to the processed document (e.g. where the document is available online). Other web address might appear in the header (for instance in a copyright section), but they won't be specifically encoded as web url.

**<date type="publication">** is the publication date, which is similar to the default case <date>.
For non publication date, the date can be further specified with an attribute, this is valid for the submission date <date type="submission">. However, Grobid currently does not exploit this alternative date type (e.g. <date type="submission"> will be equivalent to <note type=”submission”>, also see explanations for the tag <note type=”submission”>). The content of this field will be further parsed by the "date" model.

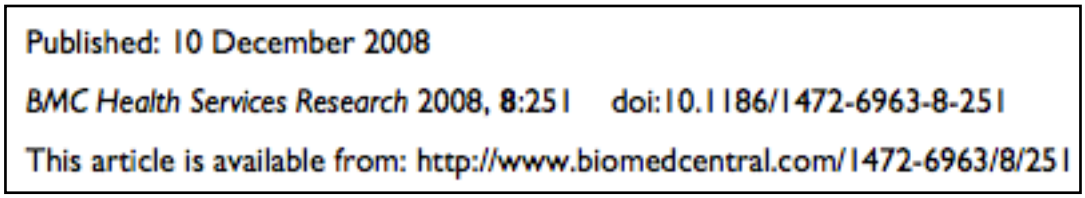
**<reference>** this is to annotate the reference information about the current article present in its header. The content of this field will be further parsed by the "citation" model. Any sub-level information corresponding to a reference must be included in the <reference> element and not further tagged, for example:


```
<reference>Chablais et al. BMC Developmental Biology 2011, 11:21</reference>

<ptr>http://www.biomedcentral.com/1471-213X/11/21</ptr>
```

The only exception is an article level identifier such as a DOI, see bellow.

**<idno>** for the article-specific identifier, in particular DOI. The type of identifier should be indicated as attribte, e.g. <idno type=”DOI”>. The identifier at article level must always be identified and encoded as such, even if it belongs to a larger field.




```
<date type="publication">Published: 10 December 2008</date>

<reference>BMC Health Services Research 2008, 8:251</reference>

<idno type="DOI">doi:10.1186/1472-6963-8-251</idno>

<ptr>This article is available from: http://www.biomedcentral.com/
1472-6963/8/251</ptr>
```

**<note type="other">** covers by default any other textual material. It is equivalent to the default case <note>, without any type attributes. The attribute *type* is used for the several specific cases listed below.

**<note type="copyright">** for the copyright info - if copyright-char © is not recognized correctly by OCR-software, use the Unicode character code **©** .

**<note type="submission">** for the information about the submission of the document. If a date is present in the submission field, it is possible to encode it as <date type="submission">, however Grobid currently does not distinguish this sort of date and process this date as it would be tagged with <note type="submission">.

**<note type="dedication">** for information about the dedication of the publication.

**<note type="page">** for the page number

**<note type="english_title">**, in the case that the main title is not in English, the English title might appear as a secondary title. It is encoded as a specific note with the indicated attribute.

Note that information such as web address, email, etc appearing in all these note fields must **not** be encoded as such, but are part of the note. For instance, in the following case, the dates are not the publication dates, so they are encoded as note:


```
<note type="submission">Received: 5 June 2008</note>

<note type="other">Accepted: 10 December 2008</note>
```

**<lb/>** for indicating a line break in the document layout

**<pb/>** for indicating a page breaks in the document layout

## 2.2 General example

```
<?xml version="1.0" ?>
<tei>
      <teiHeader>
            <fileDesc xml:id="55002195"/>
      </teiHeader>
      <text xml:lang="en">
            <front>

      <docTitle>
            <titlePart>IdeE, an IgG-endopeptidase of Streptococcus equi ssp.
equi<lb/></titlePart>
      </docTitle>

      <byline>
            <docAuthor>Jonas Lanner<lb/> ard &amp;amp; Bengt Guss<lb/></
docAuthor>
      </byline>

      <byline>
            <affiliation>Department of Microbiology, Swedish University of
Agricultural Sciences,</affiliation>
      </byline>

      <address>Box 7025, SE 750 07 Uppsala, Sweden<lb/>

      <note>Correspondence: Jonas Lanner<lb/></note>

      <phone>Tel.: 14 618 673 204; fax: 14 618 673 392;<lb/></phone>

      <email>e-mail: jonas.lannergard@mikrob.slu.se<lb/></email>

      <note type="submission">Received 25 April 2006; revised 4 July
2006;<lb/> accepted 8 July 2006.<lb/></note>

      <date>First published online 2 August 2006.<lb/></date>

      <idno>DOI:10.1111/j.1574-6968.2006.00404</idno>

      <keyword>Keywords: Streptococcus equi ; Streptococcus pyogenes ;<lb/>
IdeS; Mac; immunoglobulin; virulence.<lb/></keyword>

      <div type="abstract">Abstract<lb/> Streptococcus equi ssp. equi is the
causative agent of strangles, a highly contagious<lb/> and serious disease in
the upper respiratory tract of horses. The present study<lb/> describes the
characterization of IdeE, a homolog of the secreted IgG-specific<lb/>
protease IdeS/Mac of Streptococcus pyogenes. The activity of IdeE is compared
with<lb/> the activity of IdeZ, the corresponding enzyme of the closely
related S. equi ssp.<lb/> zooepidemicus.<lb/></div>

            </front>
      </text>
</tei>
```

### 2.3 Additional special cases

**Mixture with preceding article**: Some content from a previous article might appear in the extracted header. In the modern layout, for ease of electronic distribution, articles starts at the beginning of page. However, for older publications, an article can start just after the previous one or at the beginning of the right column. In these cases, the extracted header might be a mixture of both articles impossible to annotate. As the problem relies on the header segmentation process, it is not related to the header analysis, thus the case should be entirely ignored and deleted.

**Cover page**: Some articles can be distributed with a cover page. The cover page can be:

- A summary of bibliographical information for the current bibliographical item. This case should lead to useful results and can be further corrected.
- The front page of the journal or serial volume where the bibliographical item is published. In this case, the extracted information from the recognised header will be the ones of the journal/serials, thus not the ones of the article. This supposes an additional custom pre-process for Grobid, not yet implemented. For the moment, this exceptional case should be entirely ignored and deleted.

**Reference string present in the header, but not corresponding to the current article**: In very rare cases, the reference string present in the header might not be the ones of the current article. As there is no straightforward possibility to check such correspondence, and since the reference string is indeed a reference string, such chunk should be encoded anyway as <reference> to avoid decreasing the accuracy of the machine learning model.

## 3. ENCODING OF AFFILIATION AND ADDRESS

*Note that the further pre-annotated affiliation-address file is ignored in the current correction process.*

For the affiliation+address annotations (training files **.affiliation.tei.xml*), see the document **guidelines-affiliation_address.pdf**

Affiliation markers are annotated with the tag **<marker>**

Additional text/characters that do not belong to one of the standard elements (<orgName>, <addrLine>, <settlement>, <region>, <postCode>, <country>, <postBox>, <marker>) can be left untagged out of these elements, within the <author> block.

## 4. ENCODING OF DATES

For dates (training files **.date.tei.xml*), we do not follow the TEI but a basic XML format based on <day>, <month> and <year> elements. Additional text/characters that do not

belong to one of these specific elements (punctuations, etc.) can be left untagged under the <date> elements.

Example:

```
<?xml version="1.0" encoding="UTF-8"?>
<dates>
   <date>Received <month>August</month> <day>17</day>,
   <year>2005</year></date>
</dates>
```

## 5. ENCODING OF AUTHOR NAMES

For author names (training files **.authors.tei.xml*), we use relatively standard tags (<forename>, <middlename>, <surname>, <roleName>, <suffix>). affiliation markers are very common to indicate to which institution belongs a give author and is encoded with the tag <marker>.

Additional text/characters that do not belong to one of these elements (punctuations, syntactic sugar, etc.) has to be be left untagged under the <author> elements.

Example:

```
<?xml version="1.0" encoding="UTF-8"?>
<tei xmlns="http://www.tei-c.org/ns/1.0"
     xmlns:xlink="http://www.w3.org/1999/xlink"
     xmlns:mml="http://www.w3.org/1998/Math/MathML">
      <teiHeader>
            <fileDesc>
                  <sourceDesc>
                        <biblStruct>
                              <analytic>

                                    <author>
                                          <persName>
                                                <forename>Yongqun</forename>
                                                <surname>He</surname>,
                                                <marker>1</marker>
                                          </persName>
                                          <persName>
                                                <forename>Rino</forename>
                                                <surname>Rappuoli</surname>,
                                                <marker>2</marker>
                                          </persName>
                                          <persName>
                                                <forename>Anne</forename>
                                                <middlename>S</middlename>.
                                                <surname>De Groot</surname>,
                                                <marker>3, 4</marker> and
                                          </persName>
                                          <persName>
                                                <forename>Robert</forename>
                                                <middlename>T</middlename>.
                                                <surname>Chen</surname>
                                                <marker>5</marker>
                                          </persName>
                                    </author>

                              </analytic>
                        </biblStruct>
                  </sourceDesc>
            </fileDesc>
      </teiHeader>
</tei>
```

## 6. ENCODING OF REFERENCES

**For references (bibliographical citation typically present in the bibliographical section at the end of an article), we use the following tags:**

**<author>** for the complete sequence of authors

**<title level="a">** for article title and chapter title. Here "a" stands for analytics (a part of a monograph).

**<title level="j">** for journal title.

**<title level="m">** for non journal bibliographical item holding the cited article. Note if a book is cited, the title of the book is annotated with <title level="m">. If a thesis is cited, the title of the thesis is annotated with <title level="m">, and the type of thesis as <note>. Here "m" stands for monograph.

**<date>** the date sequence (including parenthesis, etc.)

**<biblScope type="pp">** the full range of pages

**<biblScope type="vol">** the block for volume (e.g. <volume> vol. 7,</volume>)

**<biblScope type="issue">** the block for the issue, also known as number, (e.g. no. <issue>3</issue>,)

**<orgName>** the institution for thesis or technical reports

**<publisher>** the name of the publisher

**<pubPlace>** publication place, or location of the "publishing" institution

**<editor>** for all the sequence of editors

**<ptr>** for web url

**<idno>** for the document-specific identifier, in particular DOI

**<note>** for any indications related to the reference and not covered by one of the previous tags. In the case of technical report, the indication of the document kind is encoded with the following attribute value <note type="report">

Additional text/characters that do not belong to one of these elements (punctuations, syntactic sugar, etc.) has to be be left untagged under the <bibl> elements. This is the case for instance for the tag <date>, the caracters such as parenthesis have to be put outside this element (see the example bellow).

Example:

```
<?xml version="1.0" encoding="UTF-8"?>
<tei xmlns="http://www.tei-c.org/ns/1.0"
     xmlns:xlink="http://www.w3.org/1999/xlink"
     xmlns:mml="http://www.w3.org/1998/Math/MathML">

   <listBibl>
         <bibl>
               <title level="j">Biostatistics</title> (<date>2008</date>),
<biblScope type="vol">9</biblScope>, <biblScope type="issue">2</
biblScope>, pp. <biblScope type="pp">234–248</biblScope>
         </bibl>
   </listBibl>

</tei>
```