PANACEA TRAVELLING OBJECT version 2 (TO2) documentation

This document describes Travelling Object 2 (TO2), i.e. the final corpus encoding format endorsed by the PANACEA project.

TO2 is a stand-off annotation format based on the LAF data model[1], it is serialised in the GrAF format, and it integrates the TO1 XCES.

TO2 Structure

Being a stand-off annotation format, TO2 is composed by several documents/files:

- One or more *primary data documents*;
- One or more documents referencing the primary data that provide the *base segmentation* for other annotations;
- Additional *annotation documents* that contain the actual linguistic information. Annotation documents reference the base segmentation document or other annotation documents;
- *header documents* which are associated with each primary data and annotation document (or a set of primary data or annotation documents treated as a logical unit).

Source data

In the typical PANACEA setting, Basic XCes.xml would be the source data, as it is the output of the Corpus Acquisition components. It is thus the typical initial data format of PANACEA workflows that need data crawled from the web. Basic XCES is the TO1 format (please read its documentation for more details).

Primary data

Primary data is the basic electronic representation of the source data. In the case of the current PANACEA setting it is text.

The original file: Is a text file containing only the relevant paragraphs and keep details about the paragraph number of the original xces and parapraph type. It contains all text and tags inside the body of the original input data. By default the GrAF converter filters out all crawlerinfo="boilerplate", but there is a parameter (--keep boilerplate yes/no) that allows to overright the default. In the latter case the original.txt file will contain all paragraphs (including tags and attributes) contained in the body or the source files.

See an example of original file here:
http://ws02.iula.upf.edu/panacea/examples/graf/example01/1-graf_original.txt

Mark-up in the original data (e.g., HTML or XML tags) is treated as a part of the data stream by referring annotations.

The raw data file: Is a text file containing only the text of the original document. This is the main input to further processing steps. See an example of plain file here:
http://ws02.iula.upf.edu/panacea/examples/graf/example01/1-graf_plain.txt

Primary data file are the input-only data to processing modules; their integrity should be preserved in order to preserve the integrity of references to locations within the document or documents.

Corrections and modifications to the primary data should be treated as annotations and stored in separate (annotation) documents.

[1] www.cs.vassar.edu/~ide/papers/LAF.pdf

Raw data filed do not contain mark-up of any kind.

References to primary data

In a stand-off annotation, primary text needs to be referenceable in a clear way. In LAF/GrAF "direct reference to locations in primary data is achieved by virtual nodes addressed by a location index (an anchor) [...] These nodes are located between each base unit of the primary data representation". In PANACEA, the default base unit is one character; locations are thus identified by character offsets.

[From LAF:] For example, consider the text "My dog has fleas":

|**M**|**y**| |**d**|**o**|**g**| |**h**|**a**|**s**| |**f**|**l**|**e**|**a**|**s**|

The anchors for each word are:

My : start=0, end=2
dog : start=3, end=6
has : start=7, end=10
fleas : start=11, end =16

By means of the anchors, each annotation layer may define different regions. The regions correspond to the basic segmentation of the primary data, i.e. different minimal annotation units: e.g. sentences for sentence splitters, word tokens for tokenisers, etc..

In TO2, at present, there are 2 segmentation files, and therefore two levels of segmentation: sentences and tokens.

See an example of the *_sent.xml here, for the sentence segmentation:
http://ws02.iula.upf.edu/panacea/examples/graf/example01/1-graf_sent.xml

See an example of TO2 *_seg.xml document here for tokenisation:
http://ws02.iula.upf.edu/panacea/examples/graf/example01/1-graf_seg.xml.

Primary data document header

Following LAF/GrAF, each primary data document is associated with an standalone XML header file containing information describing it contents. Headers are mutuated from CES (the Corpus Encoding Standard) and thus follow its specifications.
The primary data document header provides all the relevant information about the primary data.

In TO2 the primary document header (*_header.xml) has the following tags:

Tags	**Explanation**
<primaryData>	provides the filename/location and/or type of the primary data. The location should be provided with the attribute loc="" containing the full path of the primary data file, or its name. The attribute medium="" is used to indicate the file type: e.g. XML.
<annotations>	is used to list all the files containing the annotations that refer to the primary data.
<annotation>	is used to refer to each annotation file related to the given primary data file. As for primaryData, the attribute loc="" should be used to indicate the full path of the primary data file, or its name, and the

	attribute type="" to refer to the type of annotation.
<fileDesc>	Contains all the information that describe the primary data
<titleStmt>	Contains statements that contain information about the title of the primary document
<title>	Is the title of the primary data
<respStmt>	Within PANACEA, it contains the type of automatic component(s) that created the data. When there are several modules that contributed to the creation /annotation of the primary data, more than one <respStmt> will be present.
<resp>	
<type>	Indicate the type of component that generated the primary data or one of the annotation files: e.g Crawling and normalization, POS tagging etc.
<name>	Is used to provide the name of the person or institution owning/responsible for the component.
<extent>	Indicates the size of the resource. The size can be given either in tokens or in bytes, or both using the attributes: wordCount="" and byteCount=""
<sourceDesc>	Contains all the information about the source of the primary data (e.g. website, corpus, collection, etc.). It includes tags for specifying the title of the source, the publisher - for which one can specify the type (**<publisher** type="org"/>), the edition and publication date if available **<edition/>**, <pubDate/>. **<pubPlace>** Indicates the location of the original source file or data. In the case of web pages it gives the URL of the source file.
<profileDesc>	
<textClass	**catRef="">** can be used to refer to the category of the source text
<domain>	can be used to indicate the domain: e.g. Environment, Law,...
<subdomain>	can be used to specify the subdomain to which the content of the document belongs.
<subject>	can be used to specify the subject(s) of the text, <audience> for the intended or targeted audience, and <medium/>...

See an example of stand-alone file header here, which also illustrates the full hierarchy: http://ws02.iula.upf.edu/panacea/examples/graf/example01/1-graf_header.xml

The fundamental information that the primary data document header must provide is: the PID, URI or filename for the primary data document and all associated annotation documents.

Annotation documents

Annotation documents contain information describing the primary data. They constitute the actual annotations of the data contained in the primary data files.

In the PANACEA TO2 annotation files will correspond to the annotations done by the various different text processing components (taggers, lemmatisers, morphological analisers, parsers,…)

In the TO2 stand-off format, the annotation information may be directly associated with the primary data, or it may be associated with another annotation layer(s), i.e. it can be *layered* with one annotation layer depending on another., as in the syntactic dependency layers (which depend on, and refer to, the pos tag layer).

The information unit to which the annotation applies depends on the components that produce it: e.g for a sentence splitter the minimal unit, i.e the region to which annotations apply, is the sentence; for a tokeniser it is a word token, and so on (remember that the minimal units, the regions are defined in the segmentation files).

Each annotation layer, produced by any tool, should be represented in a separate annotation document.

At present the relevant annotation layers for PANACEA are:

- sentence splitting (see the *_sent.xml segmentation example file above)
- tokenization (see the *_seg.xml segmentation file above)
- lemmatization and/or part-of-speech and morphological tagging annotations
- syntactic dependency annotation

A full example of a pos and morphological annotation file can be found here: http://ws02.iula.upf.edu/panacea/examples/graf/example01/1-graf_pos.xml

Here below you find a commented excerpt of the file

```
<?xml version="1.0" encoding="UTF-8"?>

<graph xmlns="http://www.xces.org/ns/GrAF/0.99/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.xces.org/ns/GrAF/0.99/
http://www.xces.org/ns/GrAF/0.99/graf-0.99.xsd">

  <header>

        <tagsDecl>

        <tagUsage gi="tok" occurs="594"/>  this is used to list all or some tags used in the
annotation file with their frequency.

        </tagsDecl>

        <dependencies>

        <dependsOn loc="seg"/>

        </dependencies>

        <annotationSets>
```

<annotationSet name="xces" type="http://www.xces.org/schema/2003"/> *this is used to refer to the annotation schemas used. It can also be used to refer to schemas or names of the specific tagsets used in the layer.*

</annotationSets>

</header>

<node xml:id="freeling-n1"> *a node is defined referring to a segment in the segmentation file. Here it corresponds to* seg-r1 *which in turn corresponds to the first token in the plain text file.*

<link targets="seg-r1"/>

</node>

 this is to define a label for the node. The label also containd the feature structure describing the content of the node. In this case the token at hand.

<fs>

<f name="word" value="Agua"/> *the word token*

<f name="lemma" value="agua"/> *the lemma*

<f name="postag" value="NCFS000"/> *the postag in the Freeling tagset*

<f name="probability" value=""/> *the probability of the pos tag. This may be used by statistical postaggers.*

</fs>

NB: the features in the feature structure highly depends on the tool used to produce the annotation. PANACEA recommends to use the feature names proposed for the same kind of notions.

Also, notice that here lemmas, and postags are given in the same layer both because the tool (e.g. freeling) produces this analysis in one step and because this is the typical information used by other subsequent NLP modules.

Layout file

Since in PANACEA the first automatic acquisition tools in a typical workflow would be crawlers and therefore typical source files would be web pages, TO2 allows to keep track of the original HTML formatting and the crawlinfo in a layout file.

See an example of the file here: http://ws02.iula.upf.edu/panacea/examples/graf/example01/1-graf_layout.xml