1.1 PANACEA Travelling Object 1

1.1.1 Introduction

This document describes Travelling Object 1 (TO1), i.e. the first of the corpus encoding formats endorsed by the PANACEA project. The Language Resources and Technologies community has not reached a consensus in defining an encoding for (un)annotated corpora. On the other hand, in the context of PANACEA there was a need to harmonize the output of the large variety of tools to be integrated in the PANACEA factory. To address this issue, we tried to avoid reinventing the wheel, at least in some aspects of the definition of this format. Thus, TO1 is a derivative of the XCES Corpus Encoding Standard[1]. It is also based on, among other sources:

- PANACEA partners' descriptions of tools and encodings they already use[2]
- Gr. Thurmair's "Proposal for corpus representation in PANACEA"[3]
- informal communications concerning similar efforts in the Accurat project

This format was used in delivering the monolingual and bilingual corpora collected and annotated by components of the PANACEA factory during the first 2 years of the project.

In the following sections of this document, we describe how TO1 accomodates storage and annotation of monolingual and parallel corpus files. Section 1.1.2 describes the TO1 that PANACEA monolingual and bilingual corpus acquisition tools should generate. Section 1.1.3 discusses how the TO1 can be augmented with linguistic annotations generated by NLP processors. In Section 1.1.4, we sketch how TO1 can be used for aligning parallel documents and document parts. Finally, Section 1.1.5 deals with encoding of revision and distribution metadata.

1.1.2 Output of corpus acquisition tools

In this section, we describe the TO1 that PANACEA monolingual and bilingual corpus acquisition tools should generate. In this and the following sections, we will use two web pages[4] (and their derivatives) as examples in two PANACEA languages, English and Spanish. The two web pages focus on the same international news, i.e. the EU aid for Haiti after the 2010 earthquake.

English (http://ec.europa.eu/news/external_relations/100218_en.htm)

```
1 <!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"
   "http://www.w3.org/TR/html4/loose.dtd">
2 <html lang="en">
3     <head>
4         <META http-equiv="Content-Type" content="text/html; charset=UTF-
   8">
5         <meta name="Title" content="Haiti on our minds">
6         <meta name="Creator" content="">
7         <meta http-equiv="Content-Language" content="en">
8         <meta name="Type" content="57">
9         <meta name="Classification" content="26000">
```

[1] http://www.xces.org
[2] http://projectmanagement.panacea-lr.eu:9950/projects/panacea-project/conversations/17
[3] http://projectmanagement.panacea-lr.eu:9950/projects/panacea-project/conversations/12
[4] Both of these web page extracts include the title, some sentences, and some boilerplate text from the actual web pages on the European Commission news site. Sentences are modified versions of the original, for exemplification purposes.

```
10          <meta name="Keywords"
    content="EU,Europe,European,commission,Haiti,earthquake,homeless,aid,ass
    istance,humanitarian,response,relief,shelter,hurricane,rainy,season,fund
    ing,support,ECHO,donors,conference,rebuilding,reconstruction,gendarme,po
    lice,military">
11          <meta name="Description" content="Shelter seen as the top
    priority as EU ups aid to Haiti">
12          <meta name="Date" content="18/02/2010">
13          <title>Haiti on our minds</title>
14       </head>
15       <body>
16       <h1>Haiti on our minds</h1>
17       <p>Commission calls for €90m more in aid for the quake-stricken
    country. This amount will be drawn from EU emergency funds. </p>
18         <div><a href="notice.html">Legal notice</a>| <a
    href="#top">Top</a></div>
19       </body>
20 </html>
```

Spanish (http://ec.europa.eu/news/external_relations/100218_es.htm)

```
1  <!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"
    "http://www.w3.org/TR/html4/loose.dtd">
2  <html lang="es">
3       <head>
4           <META http-equiv="Content-Type" content="text/html; charset=UTF-
    8">
5           <meta name="Reference" content="EUROPA/">
6           <meta name="Title" content="La UE enviará más ayuda a Haití">
7           <meta name="Creator" content="">
8           <meta http-equiv="Content-Language" content="es">
9           <meta name="Type" content="57">
10          <meta name="Classification" content="26000">
11  <meta name="Keywords"
    content="UE,Europa,Europea,Comisión,Haití,terremoto,personas sin
    hogar,ayuda,asistencia,humanitaria,respuesta,auxilio,refugio,huracán,de
    lluvias,estación,financiación,apoyo,ECHO,donantes,conferencia,reconstruc
    ción,gendarme,policía,militar">
12          <meta name="Description" content="Los refugios, máxima prioridad
    de las ayudas de la UE a Haití">
13          <meta name="Date" content="18/02/2010">
14          <title>La UE enviará más ayuda a Haití</title>
15        </head>
16       <body>
17       <h1>La UE enviará más ayuda a Haití</h1>
18       <p>La Comisión pide otros 90 millones de euros de los fondos de
    emergencia europeos.</p>
19         <div><a href="aviso.html">Aviso jurídico</a> | <a
    href="#top">Comienzo</a></div>
20       </body>
21    </html>
```

We assume that a bilingual crawler has fetched both pages and stored them[5] in a local repository. Alternatively, a monolingual crawler targetting Spanish documents may have fetched the ES page only. In both cases, for each locally stored HTML page, a cleaner module has marked the structural

[5] See section 1.1.4 for encoding of information concerning alignment of parallel documents.

elements of certain paragraphs and has marked some paragraphs (navigation links, advertisements, disclaimers, etc.) as boilerplate.

All this information is stored in an XML file rooted by a *cesDoc* element that contains a header with automatically extracted metadata and a link to the html file. Each PANACEA *cesDoc* file can be validated against a modified version of the XCES standard schema available from http://panacea-lr.eu/en/info-for-professionals/documents. The next listing is an example of such a file corresponding to the ES document on the Haiti earthquake.

```
<?xml version="1.0"?>
<cesDoc id="news_20100514_haiti_es" version="0.4"  xmlns="http://www.xces.org/schema/2003">
  <cesHeader version="0.4">

    <fileDesc>
      <titleStmt>
         <title>La UE enviará más ayuda a Haití</title>
          <respStmt>
            <resp>
            <type>Crawling</type>
            <name>Panacea partner</name>
            </resp>
          </respStmt>
      </titleStmt>
      <sourceDesc>
       <biblStruct>
        <monogr>
        <author>EU web author if available</author>
        <imprint>
        <publisher>EU</publisher>
        <pubDate>2010-02-20</pubDate>
        <eAddress type="web">http://ec.europa.eu/news/external_relations/100218_es.htm</eAddress>
        </imprint>
        </monogr>
       </biblStruct>
      </sourceDesc>
    </fileDesc>

    <profileDesc>
      <langUsage>
          <language iso639="es"/>
      </langUsage>
      <textClass>
          <keywords>
            <keyTerm>Comisión</keyTerm>
            <keyTerm>Haití</keyTerm>
           <keyTerm>terremoto</keyTerm>
           <keyTerm>. . .</keyTerm>
         </keywords>
         <domain>International News</domain><!--  or (automotive, environment,  legal )-->
```

```
41          <subdomain>Optional information on subdomain</subdomain>
42          <subject>Optional information on the subject</subject>
43       </textClass>
44       <annotations>
45         <annotation ann.loc="news_20100514_haiti_es.html" type="htmlsource"/>
46       </annotations>
47     </profileDesc>
48   </cesHeader>
49 </cesDoc>
50
```

All metadata for this file is contained inside a *cesHeader* element[6]. Extensive documentation for this and all other elements in the XCES standard can be obtained from the XCES site. Here we can briefly discuss some crucial subelements of the header.

- The <*fileDesc*> element can be used for information about the title of the document and any annotations added. The <*sourceDesc*> subelement can be used for information on the original author and publication date, the publisher of the document, and the URL it was downloaded from. One or more <*respStmt*> subelements can be used to describe operations and people/groups responsible for these operations on this particular document.

- The <*profileDesc*> element groups information describing the language(s) of the document (<*langUsage*>) and the nature or topic of a text (<*domain*>, <*subdomain*>, <*subject*>, <*keywords*>).

- The <*annotations*> subelement of the <*profileDesc*> can be used for storing links to other documents relevant to this basic version. In the example above, a link to the original html document is shown.

The *cesDoc* file will also contain the paragraph-segmented texual content of the HTML pages as in the following listing:

```
1  <?xml version="1.0"?>
2  <cesDoc id="news_20100514_haiti_es" version="0.4"
    xmlns="http://www.xces.org/schema/2003">
3  <cesHeader>
4  <!- . . . ->
5  <!-- We add another respStmt for the cleaning. Everything else as
    in the header above  -->
6     <respStmt>
7     <resp>
8     <type>Boilerplate removal, text extraction, paragraph
    detection, etc.</type>
9     <name>Panacea partner</name>
10     </resp>
11    </respStmt>
12  </cesHeader>
13 <!- . . . ->
14
15 <!-- We add a text and a body element for storing the clean text.
    These are necessary elements so that this file can be validated
    against XCES schemas. -->
16  <text>
17   <body>
18 <p id="p1" type="title">
19 La UE enviará más ayuda a Haití
20 </p>
21 <p id="p2">
22 La Comisión pide otros 90 millones de euros de los fondos de
    emergencia europeos.
23 </p>
24 </body>
25 </text>
```

[6] It should be noted that similar headers document manually and automatically annotated files in large corpora like the American National Corpus.

```
26
27 </cesDoc>
```

A tool can either keep the paragraph elements <*p*> from the HTML source, or perform paragraph detection using a specific tool. To each paragraph element we add an obligatory *id* attribute whose values follow the convention *p1,p2,...,pN*.

In the context of PANACEA corpus acquistition process, if a paragraph is not in the language targetted by the crawler, an optional *crawlinfo* attribute with value *ooi-lang* (meaning "out-of-interest because of a language different from the main document language") is added. For an example, see the *p63* paragraph in the listing below.

```
1 <p id="p61" topic="delta;marsh">The waters of the Danube, which
   flow into the Black Sea, form the largest and best preserved of
   Europe's deltas. The Danube delta hosts over 300 species of birds
   as well as 45 freshwater fish species in its numerous lakes and
   marshes.</p>
2 <p id="p62" crawlinfo="ooi-length">Delta du Danube</p>
3 <p id="p63" crawlinfo="ooi-lang">Les eaux du Danube se jettent
   dans la mer Noire en formant le plus vaste et le mieux préservé
   des deltas européens. Ses innombrables lacs et marais abritent
   plus de 300 espèces d'oiseaux ainsi que 45 espèces de poissons
   d'eau douce.</p>
```

Moreover, the value *ooi-length* can be added to the *crawlinfo* attribute for very short paragraphs so that they are marked as "out of interest because they contain few words". For an example, see *p41* and *p43* paragraphs in the listing below and *p62* in the listing above.

```
1 <p id="p40" type="listitem" topic="forest;nature reserve">National
   Trust membership gives you access to green space and helps fund
   conservation. The trust manages 250,000 hectares of land,
   including forest, woods, nature reserves, farmland and moorland,
   as well as 707 miles of coastline in England, Wales and Northern
   Ireland.</p>
2 <p id="p41" crawlinfo="ooi-length">Plantlife</p>
3 <p id="p42">Plantlife works to protect wild plants and their
   habitats. Activities include rescuing wild plants from the brink
   of extinction, and ensuring that common plants don't become rare
   in the wild. It actively campaigns on a number of issues
   affecting wild plants and fungi. The Plantlife website has a
   wealth of downloadable information about wild plants and plant
   conservation. Find out how you can support the organisation here
   .</p>
4 <p id="p43" crawlinfo="ooi-length">Buglife - The Invertebrate
   Conservation Trust</p>
```

A *boileplate* value for the *crawlinfo* attribute is used for paragraphs that have been classified as boilerplate.

```
1 <p id="p1" crawlinfo="boilerplate">Home</p>
2 <p id="p2" crawlinfo="boilerplate">Partners</p>
3 <p id="p3" crawlinfo="boilerplate">Main Menu</p>
4 <p id="p4" crawlinfo="boilerplate">Home</p>
5 <p id="p5" crawlinfo="boilerplate">Background</p>
6 <p id="p6" crawlinfo="boilerplate">The Theme for 2011</p>
7 <p id="p7" crawlinfo="boilerplate">How can you participate?</p>
8 <p id="p8" crawlinfo="boilerplate">Register your Activity</p>
```

```
9  <p id="p9" crawlinfo="boilerplate">WMBD Around the World</p>
10 <p id="p10" crawlinfo="boilerplate">WMBD Community</p>
11 <p id="p11" crawlinfo="boilerplate">Press / Materials</p>
12 <p id="p12" crawlinfo="boilerplate">Related Links</p>
13 <p id="p13" crawlinfo="boilerplate">Partners</p>
14 <p id="p14" crawlinfo="boilerplate">Translate this Site:</p>
15 <p id="p15" crawlinfo="boilerplate">Partners &amp; Sponsors</p>
16 <p id="p16" crawlinfo="ooi-length">WMBD Partners:</p>
17 <p id="p17" topic="sustainable development">United Nations
   Environment Programme (UNEP) is the voice for the environment in
   the United Nations system. It is an advocate, educator, catalyst
   and facilitator, promoting the wise use of the planet's natural
   assets for sustainable development.</p>
```

PANACEA focused crawlers are used to acquire web documents relevant to a specific topic defined in a topic definition file containing lists of relevant terms. Thus, an optional attribute for <*p*> elements is *topic*, which contains all terms from the topic definition detected in a specific paragraph. For examples, see *p61, p40* and *p17* above.

Finally, <*p*> elements may include an optional *type* attribute. The value of this attribute is indicative of the paragraph's function in the text. Allowed values for the *type* attribute are *title, heading*, and *listitem* (see *p40* in the listing above).

1.1.3 Output of NLP tools

In this section, we describe how TO1 can be augmented with linguistic annotations generated by NLP processors. We discuss here simple, "token-level" annotations, while we skip syntactic or any other relational annotations: these are addressed by the TO2 format GrAF (GrAF is explained at http://panacea-lr.eu/en/info-for-professionals/documents) endorsed by PANACEA.

At this processing stage, we assume that a series of NLP tools have processed the *cesDoc* file described in the previous section.

At the beginning, each paragraph has been further segmented into sentences and each sentence has been tokenized. The updated *cesDoc* file generated by the NLP tool, contains sentence elements <*s*> which have an obligatory *id* element and an optional *lang* attribute, if the sentence language is different from the main language of the document. Another optional attribute for <*s*> elements is *topic*, to be used in cases where the topic has been detected as different from the one described in the <*domain*>, <*subdomain*>, <*subject*> elements of the header. The tokens for each sentence are <*t*> elements inside the sentence they belong to. The <*t*>< elements minimally consist of an *id* attribute (*t1,t2, …,tN*), and a *word* attribute.

```
1  <?xml version="1.0"?>
2  <cesDoc id="news_20100514_haiti_es" version="0.4"
    xmlns="http://www.xces.org/schema/2003">
3  <cesHeader>
4  <!- . . . ->
5  <!-- We add another respStmt for sentence splitting and the
    tokenization. Everything else as in the header above  -->
6     <respStmt>
7     <resp>
8     <type>Sentence splitting and tokenization.</type>
9     <name>Panacea partner</name>
10     </resp>
11    </respStmt>
```

```
12 </cesHeader>
13 <!- . . . ->
14
15 <!-- We add nested sentences and tokens. -->
16 <text>
17  <body>
18    <p id="p1" >
19      <s id="s1">
20       <t id="t1_1" word="La"/>
21       <t id="t1_2" word="UE"/>
22       <t id="t1_3" word="enviará"/>
23       <t id="t1_4" word="mas"/>
24       <t id="t1_5" word="ayuda"/>
25       <t id="t1_6" word="a"/>
26       <t id="t1_7" word="Haiti"/>
27     </s>
28   </p>
29    <p id="p2" >
30     <s id="s2">
31     <t id="t2_1" word="La"/>
32 ...
33  </s>
34  </p>
35 </body>
36 </text>
37 </cesDoc>
```

An optional *casing* attribute for <*s*> elements can be used to denote specific features for a sentence. For example, in the following listing, the value *uppercase* shows that all tokens in this sentence are uppercased. The following values are allowd for the casing attribute: *{sentence, uppercase, titlecase, lowercase}*. The *sentence* value (meaning "regular sentence case") is the default and need not appear explicitly.

```
1 <s id="s7" casing="uppercase">
2 <t id="t135" word="COORDINATION"/>
3 <t id="t136" word="OF"/>
4 <t id="t137" word="SOCIAL"/>
5 <t id="t138" word="SECURITY"/>
6 <t id="t139" word="SYSTEMS"/>
7 </s>
```

At a next processing stage, we assume that a POS tagger has augmented the the *cesDoc* file with morphosyntactic and lemma information. This information is encoded in the *tag* and *lemma* attributes of <*t*> elements.

```
1 <?xml version="1.0"?>
2 <cesDoc id="news_20100514_haiti_es" version="0.4"
   xmlns="http://www.xces.org/schema/2003">
3 <cesHeader>
4 <!- . . . ->
5 <!-- We add another respStmt for tagging and lemmatization.
   Everything else as in the header above  -->
6     <respStmt>
7     <resp>
8     <type>Tagging and lemmatization.</type>
9     <name>Panacea partner</name>
10     </resp>
```

```
11    </respStmt>
12  </cesHeader>
13 <!- . . . ->
14
15 <!-- We add tags and lemmas attributes. -->
16  <text>
17   <body>
1       <p id="p1" >
2        <s id="s1">
3         <t id="t1_1" tag="AFS" lemma="el" word="La"/>
4         <t id="t1_2" tag="N4666" lemma="UE" word="UE"/>
5         <t id="t1_3" tag="VDU3S-" lemma="enviar" word="enviará"/>
6         <t id="t1_4" tag="D" lemma="mas" word="mas"/>
7         <t id="t1_5" tag="N5-FS" lemma="ayuda" word="ayuda"/>
8         <t id="t1_6" tag="P" lemma="a" word="a"/>
9         <t id="t1_7" tag="N4666" lemma="Haiti" word="Haiti"/>
10      </s>
11    </p>
12     <p id="p2" >
13      <s id="s2">
14      <t id="t2_1" tag="AFS" lemma="el" word="La"/>
15 ...
16   </s>
17   </p>
18
18 </body>
19 </text>
20 </cesDoc>
```

1.1.4 Output of alignment tools

In this section, we describe the TO1 for storing alignments between *cesDoc* files in two or more languages. The structure of this alignment file is based on the schema for alignment[7] described in the latest (1.0.4) release of XCES. We briefly discuss below the elements most relevant to PANACEA acquisition and processing of bilingual data.

In the first alignment example, we show a *cesAlign* file pointing to one EN and one ES *cesDoc* files as these have been stored by the corpus acquisition tools (see Section 1.1.2 for input sources).

```
1  <?xml version="1.0"?>
2  <cesAlign version="1.0"  xmlns="http://www.xces.org/schema/2003">
3    <cesHeader version="1.0">
4      <profileDesc>
5        <translations>
6          <translation lang="en" trans.loc="http://www.panacea-lr.eu
    /.../news_20100514_haiti_en.xml"
7            wsd="UTF-8" n="1"/>
8          <translation lang="es" trans.loc="http://www.panacea-lr.eu
    /.../news_20100514_haiti_es.xml"
9            wsd="UTF-8" n="2"/>
10       </translations>
11     </profileDesc>
12   </cesHeader>
13 </cesAlign>
```

[7] http://www.xces.org/schema/#align

As specified in XCES, this document contains a <*cesHeader*> element, followed by a <*linkList*> element. The <*cesHeader*> element may contain metadata information for the alignment process. It also contains links to the path where the aligned documents are stored. The *wsd* attribute is used for the description of the encoding. The *n* attribute allows <*align*> elements (see below) to specify which <*translation*> they refer to.

In the next example, we show a cesAlign file pointing to the EN-ES *cesDoc* files, after these have been augmented with linguistic information from NLP processors (see Section 1.1.3).

```
1  <?xml version="1.0"?>
2  <cesAlign version="1.0"  xmlns="http://www.xces.org/schema/2003">
3    <cesHeader version="1.0">
4      <profileDesc>
5        <translations>
6          <translation trans.loc=" http://www.panacea-lr.eu
   /.../news_20100514_haiti_en.tag.xml"
7            wsd="UTF-8" n="1"/>
8          <translation trans.loc=" http://www.panacea-lr.eu
   /.../news_20100514_haiti_es.tag.xml"
9            wsd="UTF-8" n="2"/>
10       </translations>
11     </profileDesc>
12   </cesHeader>
13   <linkList>
14     <linkGrp domains="p1 p1" targType="s">
15       <link>
16         <align xlink:href="#s1"/>
17         <align xlink:href="#s1"/>
18       </link>
19     </linkGrp>
20     <linkGrp domains="p2 p2" targType="s">
21       <link>
22         <align xlink:href="#xpointer(id('s3')/range-
   to(id('s4')))"/>
23         <align xlink:href="#s2"/>
24       </link>
25     </linkGrp>
26   </linkList>
27 </cesAlign>
```

This time, a <*linkList*> element follows the <*cesHeader*>. This element contains one or more <*linkGrp*> elements. Groups of links apply to data within a particular text division, like paragraphs, etc. In the example above, we indicate this by creating <*linkGrp*> elements for each paragraph in the *cesDoc* files and storing the *id*s of paragraphs in the *domains* attribute of each <*linkGrp*>. The *targType* attribute of each <*linkGrp*> is used to indicate the type of links to be stored inside this element. In the case of the first two <*linkGrp*> elements, the links refer to sentences.

The <*link*> elements in the <*linkGrp*> elements contain the actual links. According to the XCES documentation "the order of the <*align*> elements within a <*link*> element is significant. Unless otherwise specified the order is assumed to match the ordering of <*translation*> elements in the header. If a different ordering is required the attribute n in the <*translation*> element and the attribute n in the <*align*> element can be used to explicitly link an <*align*> element with a specific translation."

The XLink locators in the <*align*> elements identify the aligned elements from the annotated files. As again specified in the XCES documentation "many-to-one alignments and many-to-many alignments can be represented by providing a range for the XPointer expression." This is the case in the alignment between sentences 2 and 3 in the EN, and sentence 2 in the ES document (cf. the HTML input in section 1.1.2**¡Error! No se encuentra el origen de la referencia.** above). Notice that similar N-to-N alignments could also be produced for tokens, i.e. <t> elements in annotated *cesDoc* files.

1.1.5 Revision and distribution metadata

Revisions to the *cesDoc* files descibed above can be documented via multiple <*change*> elements in a <*revisionDesc*> child element of the <*cesHeader*>.

```
1  <?xml version="1.0"?>
2  <cesDoc id="news_20100514_haiti_es" version="0.4"
    xmlns="http://www.xces.org/schema/2003">
3    <cesHeader version="0.4">
4      <fileDesc>
5      <!-- As above -->
6      </fileDesc>
7      <profileDesc>
8      <!-- As above -->
9      </profileDesc>
10   <revisionDesc>
11   <!--Summarizes the revision history for a file.  -->
12     <change>
13       <changeDate>2012-05-20</changeDate>
14       <respName>Panacea Partner</respName>
15       <item>Corrected an error in the POS tags.</item>
16     </change>
17   </revisionDesc>
18   </cesHeader>
19 </cesDoc>
20
```

Once the corpora are ready for distribution, information on rights and availability can be documented via the <*publicationStmt*> child element of the <*cesHeader*>.

```
1  <?xml version="1.0"?>
2  <cesDoc id="news_20100514_haiti_es" version="0.4"
    xmlns="http://www.xces.org/schema/2003">
3    <cesHeader version="0.4">
4      <fileDesc>
5      <!-- . . .  -->
6      <publicationStmt>
7        <distributor>Panacea project</distributor>
8        <eAddress>http://www.panacea-lr.eu</eAddress>
9        <availability>Free for research purposes or
    ...</availability>
10       <pubDate>2013-05-20</pubDate>
11     </publicationStmt>
12     <!-- . . .  -->
13     </fileDesc>
14     <!-- Rest as above -->
15   </cesHeader>
16 </cesDoc>
```