MADCAT Data Format Spec
Version 4.0h

Created by Linguistic Data Consortium and NIST

1 Summary

In discussing a data format for MADCAT, the major consideration is the logical storage of many layers of annotation that each applies to different sub-sections of a document. Some of these sub-sections are defined semantically (e.g. Semantic Unit, etc), while others are defined physically (e.g. page, line, etc). The semantic sub-sections are tied directly to the text transcript, while the physical sub-sections are tied to the original document image (tiff). The challenge is to create one extensible data format that will store all of the following, so that cross-references between the text data and the image data are perfectly clear.

2 Elements

Certain elements are associated with the image itself, while others are associated with the content.

- **Image**
  - page
  - zone
  - token-image
  - polygon
  - point

- **Content**
  - section (type = chapter / paragraph / etc.)
  - segment (the semantic unit e.g. a sentence)
  - token (the smallest semantic unit e.g. a word)
  - transcription
  - translation

3 XML Structure

A proposed XML structure is described in the following DTD:

```
<!ELEMENT madcat (doc)>

<!ELEMENT doc (writer, image, content?)>

<!ELEMENT writer EMPTY>

<!ELEMENT image (page+)>
<!ELEMENT page (zone+)>
<!ELEMENT zone (polygon, token-image*)>
<!ELEMENT token-image (polygon)>

<!ELEMENT polygon (point, point, point+)>
<!ELEMENT point EMPTY>

<!ELEMENT content (section+)>
<!ELEMENT section (segment+)>
<!ELEMENT segment (token+, transcription?, translation?)>
<!ELEMENT token (source?)>
<!ELEMENT source (#PCDATA)>
<!ELEMENT transcription (#PCDATA)>
<!ELEMENT translation (#PCDATA)>

<!ATTLIST madcat version CDATA #REQUIRED>

<!ATTLIST doc id ID #REQUIRED>
<!ATTLIST doc src CDATA #REQUIRED>
<!ATTLIST doc nbpages CDATA #REQUIRED>
<!ATTLIST doc type CDATA #REQUIRED>

<!ATTLIST writer id ID #REQUIRED>

<!ATTLIST page id ID #REQUIRED>
<!ATTLIST page dpi NMTOKEN #IMPLIED>
<!ATTLIST page colordepth NMTOKEN #IMPLIED>
<!ATTLIST page width NMTOKEN #REQUIRED>
<!ATTLIST page height NMTOKEN #REQUIRED>

<!ATTLIST zone id ID #REQUIRED>
<!ATTLIST zone type CDATA #REQUIRED>

<!ATTLIST token-image id ID #REQUIRED>

<!ATTLIST point x NMTOKEN #REQUIRED>
<!ATTLIST point y NMTOKEN #REQUIRED>

<!ATTLIST section id ID #REQUIRED>
<!ATTLIST section type CDATA #REQUIRED>

<!ATTLIST segment id ID #REQUIRED>

<!ATTLIST token id ID #REQUIRED>
<!ATTLIST token ref_id IDREF #REQUIRED>
<!ATTLIST token status CDATA #IMPLIED>
```

3.1 Attributes and IDs

The <madcat> tag marks the start and end of an autonomous XML document. The tag must include the attribute:

- version: identifies the evaluation year and revision of the XML specs

The <doc> tag indicates the start and end of a document and must include the following attributes:

- id: unique id string or number
- src: indicates the filename
- nbpages: number of pages in the entire document
- type: indicates the genre (e.g., letter, form, …)

The <writer> tag contains the information of the document writer and must include the following attribute:

- id: unique id string or number

3.1.1 Image annotation

The information between the <image> tags references the tiff and stores all Ground Truth annotation generated by GEDI. Each tag includes coordinates describing zone location, orientation as well as a unique id.

The <page> tag attributes will be:

- id
- dpi
- colordepth
- width
- height

The <zone> tag is flexibly defined based on the granularity of the Ground Truth annotation, and includes the following attributes:

- id
- type: indicates the unit being zoned (e.g. type=”line”)

The <token-image> tag marks a particular type of zone, which GEDI would output as <zone type=″token″>. It indicates the location and coordinates of each token in the tiff.

The DTD provides one way of storing geometric information of `zone` and `token-image`: `polygon`. Both the `zone` element and the `token-image` element can take `polygon` as a sub-element. The polygon element takes three or more points as its sub-elements. The point element takes coordinates of the page with respect to the page’s upper left corner. Here is an example of a token-image represented as a polygon.

```
<token-image id="t0000031">
  <polygon>
    <point x="100" y="10"/>
    <point x="100" y="100"/>
```

```
    <point x="300" y="10"/>
    <point x="300" y="100"/>
  </polygon>
</token-image>
```

For the first year of the MADCAT program, only quadrilaterals will be used, but the DTD accommodates any polygons that may be used in future phases of the MADCAT program.

3.1.2 Content annotation

The information between the `<content>` tags includes the text transcript and – when applicable – the translation. Each tag includes pointers to the structural `<image>` element referenced by the content using the unique id.

A `<section>` is group of segments or semantic units, and includes the following attributes:

- id
- type

A `<segment>` is a sentence or a semantically defined sentence-like unit. It is comprised of any number of tokens, and may be part of a line, a full line, or many lines. Preserving these units is necessary for consistency and compatibility between Ground Truth annotation and transcription, translation, and evaluation. A `<segment>` stores the references to the token-images. These references can optionally store the actual source text.

Within each content segment, the whole transcription for the segment is optionally stored between the `<transcription>` tags.

```
<segment id="s00036">
  <token id="s00036-1" ref_id="t000122">
    <source>This</source>
  </token>
  <token id="s00036-2" ref_id="t000123">
    <source>is</source>
  </token>
  <token id="s00036-3" ref_id="t000124">
    <source>Arabic</source>
  </token>
  <transcription>This is Arabic</transcription>
</segment>
```

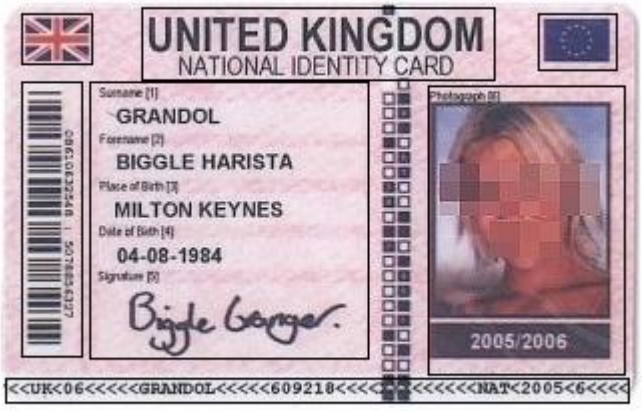
The natural reading order should be encoded in the final part (-1, -2, -3, …) of the token ids as shown above.

When available, a non-tokenized translation segment is also stored within the `<segment>`, between the `<translation> </translation>` tags.

The `status` attribute of the `token` element stores the status of the token such as, “typo”, “missing”, and “extra”.

4 Annotated Examples

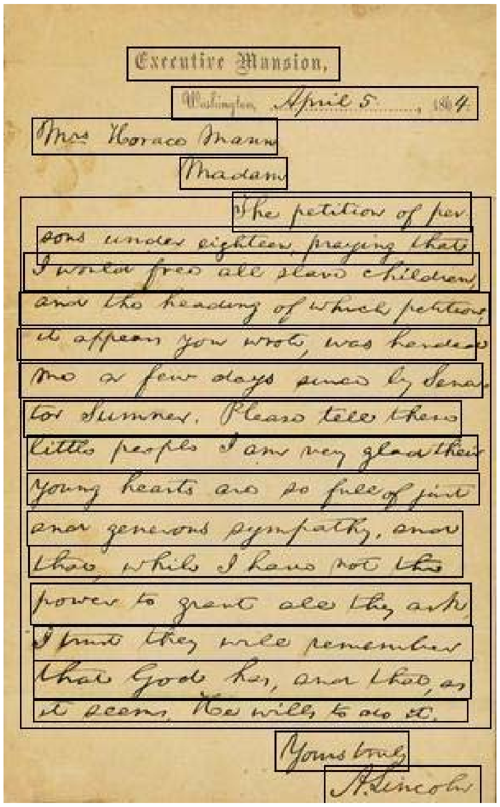
4.1 A photo ID




```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE madcat SYSTEM "madcat.v1.0.5.dtd">
<madcat version="2008.1">
  <doc id="d0001" src="uk-id.tif" nbpages="1" type="photo id">
    <writer id="w038"/>
    <image>
      <page id="p0004" dpi="600" colordepth="2" width="3980" height="2690">
        <zone id="z00094" type="logo">
          <polygon>
            <point x="500" y="400"/>
            <point x="500" y="600"/>
            <point x="860" y="400"/>
            <point x="860" y="600"/>
          </polygon>
        </zone>
        <zone id="z00095" type="line">
          <polygon>
            <point x="1140" y="400"/>
            <point x="1140" y="600"/>
            <point x="2850" y="400"/>
            <point x="2850" y="600"/>
          </polygon>
          <token-image id="t0000192">
            <polygon>
              <point x="1140" y="400"/>
              <point x="1140" y="600"/>
              <point x="1840" y="400"/>
              <point x="1840" y="600"/>
            </polygon>
          </token-image>
          <token-image id="t0000193">
            <polygon>
              <point x="1900" y="400"/>
              <point x="1900" y="600"/>
              <point x="2850" y="400"/>
              <point x="2850" y="600"/>
            </polygon>
          </token-image>
        </zone>
        <zone id="z00096" type="code">
          <polygon>
            <point x="520" y="740"/>
            <point x="520" y="2100"/>
            <point x="815" y="740"/>
            <point x="815" y="2100"/>
          </polygon>
        </zone>
      </page>
```

```
    </image>
    <content>
     <section id="sec0072" type="title">
       <segment id="s0001">
         <token id="s0001-1" ref_id="t0000192">
           <source>UNITED</source>
         </token>
         <token id="s0001-2" ref_id="t0000193">
           <source>KINGDON</source>
         </token>
        <transcription>UNITED KINGDOM</transcription>
         <translation>The translation</translation>
       </segment>
     </section>
    </content>
  </doc>
</madcat>
```

4.2 A letter



Executive Mansion,

Washington, April 5, 1864.

Mrs Horace Mann

Madam

The petition of persons under eighteen, praying that I would free all slave children, and the heading of which petition it appears you wrote, was handed me a few days since by Senator Sumner. Please tell these little people I am very glad their young hearts are so full of just and generous sympathy, and that, while I have not the power to grant all they ask, I trust they will remember that God has, and that, as it seems, He wills to do it.

Yours truly

A. Lincoln

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE madcat SYSTEM "madcat.v1.0.5.dtd">
<madcat version="2008.1">
  <doc id="d003" src="lincoln-letter.tif" nbpages="1" type="letter">
    <writer id="w005"/>
    <image>
      <page id="p0004" dpi="600" colordepth="256" width="2460" height="3990">
        <zone id="z00095" type="line">
          <polygon>
            <point x="630" y="220"/>
            <point x="630" y="350"/>
            <point x="1670" y="220"/>
            <point x="1670" y="350"/>
          </polygon>
          <token-image id="t0000031">
            <polygon>
              <point x="630" y="220"/>
```

```
              <point x="630" y="350"/>
              <point x="1100" y="220"/>
              <point x="1100" y="350"/>
            </polygon>
          </token-image>
          <token-image id="t0000032">
            <polygon>
              <point x="1170" y="220"/>
              <point x="1170" y="350"/>
              <point x="1670" y="220"/>
              <point x="1670" y="350"/>
            </polygon>
          </token-image>
        </zone>
      </page>
    </image>
    <content>
      <section id="sec1205" type="title">
        <segment id="s0007">
         <token id="s0007-1" ref_id="t0000031">
           <source>Executive</source>
         </token>
         <token id="s0007-2" ref_id="t0000032" status="typo">
           <source>Mantion</source>
         </token>
         <transcription>Executive Mantion</transcription>
         <translation>The translation</translation>
       </segment>
      </section>
    </content>
  </doc>
</madcat>
```