CL Metaheaders
Open Standard
Revision 1.0

John Vidler j.vidler@lancaster.ac.uk
Stephen Wattam s.wattam@lancaster.ac.uk

Compiled November 1, 2015

This standard is currently under heavy development, and is incomplete in a large number of places, please ensure you use the latest edition by examining the compiled date above.

Comments and suggestions to be directed to
j.vidler@lancaster.ac.uk and/or s.wattam@lancaster.ac.uk

Contents

1 Data Format **3**

2 Recognised Fields **4**

3 Examples **5**
3.1 Valid ARFF 5
3.2 Valid JSON 5
3.3 Valid TEI XML 5

1 Data Format

2 Recognised Fields

Draft version '1.0' of the specification only mandates two fields, and has two optional fields, which are described as follows:

`version` *Optional, Float*
: The specification version that this metaheader complies with. When missing, the latest specification draft is to be assumed - at this time, the only valid version is '1.0'.

`encoding` *Required, String*
: Character encoding of current text, as defined by the IANA list of preferred text encoding names[1].

`mime` *Required, String*
: The MIME type used to describe what this file is, examples include standard ones such as `text/xml`, and `application/json` as described in[2], alongside extended ones such as `text/arff` and `text/tei` which define specific, known file types.

  It is expected that this collection will expand over time with new file standards, but in the absence of an file type-specific MIME type, the next nearest standard one should be used. In the case of TEI files, if the specific `text/tei` MIME did not exist, `text/xml` could be used.

`group` *Optional, String or Object*
: Group fields can be used to track which files belong to collections, for example where files from a corpus are processed in fragments, or must be re-assembled after processing.

The preliminary set of features identified for inclusion have been selected to allow the identification of key features of the subsequent texts, and allow them to be correctly loaded by software, and are:

- common to all machine-readable text representations;
- necessary-yet-uninteresting features of the dataset;
- generally useful across many tool types.

3 Examples

3.1 Valid ARFF

This example has been truncated, as it is substantially long, and the metaheader is near the top of the file.

```
@RELATION relation
@ATTRIBUTE class {"1st","2nd","3rd","crew"}
@ATTRIBUTE age {"adult","child"}
@ATTRIBUTE sex {"male","female"}
@ATTRIBUTE survived {"yes","no"}

% meta { "version":"1.0", "encoding":"UTF-8", "mime":"text/arff" }

@DATA
1st,adult,male,yes
1st,adult,male,yes
1st,adult,male,yes
1st,adult,male,yes
1st,adult,male,yes
1st,adult,male,yes
```

3.2 Valid JSON

```
{
   "__meta__":{
      "version":1.0,
      "mime":"application/json",
      "encoding":"UTF-8"
   }
}
```

3.3 Valid TEI XML

```
<!-- meta {
   "version":"1.0",
   "encoding":"UTF-8",
   "mime":"text/xml-tei",
   "extra":{
      "author":"John Vidler",
      "tool":"Nulltech handicraft"
   }
} -->
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title><!-- title of the resource --></title>
         </titleStmt>
```

```
      <publicationStmt>
         <p><!-- Information about distribution of the resource --></p>
      </publicationStmt>
      <sourceDesc>
         <p><!-- Information about source from which the resource derives --></p>
      </sourceDesc>
   </fileDesc>
</teiHeader>
<text xml:lang="en">
   <!-- ... -->
</text>
</TEI>
```

References

[1] Ned Freed Martin Drst. Iana, character sets. http://www.iana.org/assignments/character-sets/character-sets.xhtml, 12 2013.

[2] IANA. Media types. http://www.iana.org/assignments/media-types/media-types.xhtml, 10 2015.