ENSEMBL UPLOAD FORMAT
VERSION 2

Summary

The main purpose of updating the format is to enable gene-based and protein-based DAS annotations. As a matter of fact the new format will allow for the upload of features based on any database identifier. It can be a reference sequence, Ensembl Peptide, UniProt Accession Name etc.

Another goal is to enable the use of all tags allowed by the DAS specification. The first version of the upload format only deals with the genomic location based features and only allows for a limited use of DAS features, namely the feature grouping. But more and more developers want to put a note and link back to their websites. DAS specification already allows for it so why not use it.

Another DAS aspect that is getting wider acceptance is the stylesheet support, which is also taken into account.

Description of the format

The format of the file is based on GFF version 3 by Lincoln Stein (http://flybase.bio.indiana.edu/annot/gff3.html), but with some variations.

First of all the file will have several sections separated by section headers. At the moment only three sections are recognized:

[annotations]

[groups]

[stylesheet]

If there are no section headers in the file it will be assumed that file contains only annotations.

Annotations section:

The format of annotations stays pretty much the same as in version 1 of the upload format, except for the first column ('group') has been removed to allow for features that belong to more than one group. Thus the format consists of 9 required columns and 1 optional, separated by tabs (NOT spaces). The following unescaped characters are allowed within fields: [a-zA-Z0-9. :^*$@!+_?-]. All other characters must be escaped using the URL escaping conventions. Unescaped quotation marks, backslashes and other ad-hoc escaping conventions that have been added to the GFF format are explicitly forbidden. The `=` `,` `;` and `%` characters have reserved, and must be escaped when used in other contexts. Note that unescaped spaces are allowed within fields. Parsers must split on tabs, not spaces.
Undefined fields are replaced with the "." character, as described in the original GFF spec.

Below is the full list of fields:

Column No	Field name	Field Description
1	feature	feature id
2	type	feature type id
3	subtype	method
4	segment	reference sequence, or database identifier
5	start	feature range
6	stop	feature range
7	strand	feature orientation
8	phase	frame
9	score	score
10	attributes	optional attributes

Apart from missing first column the only difference to the first version is that all optional attributes go in the field 11, and in version 1 optional fields 11 and 12 could hold the feature similarity alignment region. It does not look like anyone used them anyway, but if someone will want to use it there is target attribute that can be place in `attributes` column.

Last optional column is a list of feature attributes in the format tag=value. Multiple tag=value pairs are separated by semicolons. URL escaping rules are used for tags or values containing the following characters: `,` `=` `;`
Below is the full list of the recognized attributes, i.e. what else DAS specification allows for but what is not required.

Attribute	Description
feature-label	Human readable feature label
type-label	Human readable type label
method-label	Human readable method label
category	Feature category, which is an attribute of the TYPE tag in DAS spec
note	Arbitrary human-readable note in plain text format
link	A link to a web page that provides more information about this feature, e.g. Read more
target	The target sequence in a sequence similarity match. The format id:start,stop The **id** provides the reference ID for the target sequence, and the **start** and **stop** indicate the segment that matched across the target sequence

Group section

To accommodate the fact that sometimes people want to group features and provide info about the group rather than a single feature there can be [groups] section. This will consist of only two columns as none of group's attributes and tags are required by DAS specifaction, thus it will be only Group ID and Group Attributes. Below is the list of group attributes:

Attribute	Description
label	Human readable group label
type	group type
note	Arbitrary human-readable note in plain text format
link	A link to a web page that provides more information about this group, e.g. Read more
target	The target sequence in a sequence similarity match. The format id:start,stop The **id** provides the reference ID for the target sequence, and the **start** and **stop** indicate the segment that matched across the target sequence

Stylesheet section

This is purely to make data look pretty. The section contains plain XML as per DAS specification

http://www.biodas.org/documents/spec.html#stylesheet

Other syntax

Comments are preceded by the # symbol. Meta-data and directives are preceded by ##. The following directives are recognized:
##euf-version 2

The EUF version, always 2 in this spec. This must be the topmost line of the file.
##coordinate-system Ensembl Peptide ID

Describes the `segment` field of annotations section. This is what would a user select in Dasconfview when attaching the source in Ensembl. At the moment valid options are

- Ensembl Location
- Ensembl Gene ID
- Ensembl Peptide ID
- Ensembl Transcript ID
- Entrez Gene ID
- HUGO ID
- IPI Accession
- IPI ID
- Uniprot/Swiss-Prot Name
- Uniprot/Swiss-Prot Acc

Sample File

```
##euf-version 2
##coordinate-system Ensembl Location

[annotations]
AL137655.1.1 homology wutblastn 13 31787660 31787740 +
. 373.0000 group=Similarity1
AK024248.1.1 homology wutblastn 13 31787660 31787740
+ . 373.0000 group=Similarity1
BC006361.1.1 homology wutblastn 13 31787660 31787740
+ . 384.0000 group=Similarity1,Similarity2
AL137733.1.1 homology wutblastn 13 31787660 31787740
+ . 384.0000 group=Similarity1
Hs.326048.1 homology wutblastn 13 31787660 31787740 +
. 373.0000 group=Similarity2
AL137655.1.2 homology wutblastn 13 31788406 31788934
+ . 373.0000 group=Similarity1
AK024248.1.2 homology wutblastn 13 31788406 31788934
+ . 373.0000 group=Similarity1
BC006361.1.2 homology wutblastn 13 31788406 31788934
+ . 384.0000 group=Similarity1,Similarity2
AL137733.1.2 homology wutblastn 13 31788406 31788934
+ . 384.0000 group=Similarity1
Hs.326048.2 homology wutblastn 13 31788406 31788934 +
. 373.0000 group=Similarity2

[groups]
Similarity1 label=Similarity Type A; note=note1,note2;
link=http://www.sanger.ac.uk/myproject/groupid=Similarity1

Similarity2 label=Similarity Type H; note=note1,note2,note3;
link=http://www.sanger.ac.uk/myproject/groupid=Similarity2

[stylesheet]
<?xml version="1.0" standalone="no"?>
<!DOCTYPE DASSTYLE SYSTEM "http://www.biodas.org/dtd/dasstyle.dtd">
<DASSTYLE>
<STYLESHEET version="1.0">
 <CATEGORY id="default">
  <TYPE id=”default”>
  <GLYPH>
   <BOX>
    <HEIGHT>4</HEIGHT>
    <FGCOLOR>black</FGCOLOR>
    <BGCOLOR>red</BGCOLOR>
   </BOX>
    </GLYPH>
  </TYPE>
  <TYPE id="homology">
    <GLYPH>
      <ANCHORED_ARROW>
            <HEIGHT>25</HEIGHT>
```

```
            <BGCOLOR>darkgreen</BGCOLOR>
            <FGCOLOR>black</FGCOLOR>
                <NO_ANCHOR>1</NO_ANCHOR>
            <BUMP>0</BUMP>
            <FONT>sanserif</FONT>
          </ANCHORED_ARROW>
        </GLYPH>
      </TYPE>
    </CATEGORY>
    <CATEGORY id=”Similarity1”>
      <TYPE id="homology">
        <GLYPH>
          <ANCHORED_ARROW>
                <HEIGHT>20</HEIGHT>
            <BGCOLOR>palegreen4</BGCOLOR>
            <FGCOLOR>black</FGCOLOR>
                <NO_ANCHOR>1</NO_ANCHOR>
            <BUMP>0</BUMP>
            <FONT>sanserif</FONT>
          </ANCHORED_ARROW>
        </GLYPH>
      </TYPE>
      </CATEGORY>
    <CATEGORY id=”Similarity2”>
      <TYPE id="homology">
        <GLYPH>
          <ANCHORED_ARROW>
                <HEIGHT>20</HEIGHT>
            <BGCOLOR>darkgreen4</BGCOLOR>
            <FGCOLOR>black</FGCOLOR>
                <NO_ANCHOR>1</NO_ANCHOR>
            <BUMP>0</BUMP>
            <FONT>sanserif</FONT>
          </ANCHORED_ARROW>
        </GLYPH>
      </TYPE>
      </CATEGORY>
    <CATEGORY id=”group”>
      <TYPE id="similarity1">
        <GLYPH>
       <LINE>
         <FGCOLOR>black</FGCOLOR>
         <LINE_STYLE>hat</LINE_STYLE>
       </LINE>
        </GLYPH>
       </TYPE>
      <TYPE id="similarity2">
        <GLYPH>
       <LINE>
         <FGCOLOR>red</FGCOLOR>
         <LINE_STYLE>hat</LINE_STYLE>
       </LINE>
        </GLYPH>
       </TYPE>
     </CATEGORY>
  </STYLESHEET>
  </DASSTYLE>
```