

# Final Data Format for LDC Multiple Translation Arabic Corpora

This document applies to all the multiple translation Arabic Corpora LDC has published.

## 1. Directory Structure:

The corpus content is organized into "source" and "translation" directories. Within "translation" there is a separate subdirectory for each translation service or system.

The source directory and each of the translation subdirectories contains same number of files, one news story per file. Corresponding file names are identical across all directories.

## 2. Source Data Format:

Within each source file, the content is formatted in SGML as follows:

```
<DOC docid="...">
<hl>
<seg id=1> [Arabic text in UTF8 character encoding] </seg>
</hl>
<p>
<seg id=2> [Arabic text in UTF8 character encoding] </seg>
</p>
...
</DOC>
```

Notes:

- the docid string enclosed in double quotes matches the file name
- the <seg> tags are always assigned sequential numeric ID's starting at 1 for the first <seg> of each file, are always placed on the same line with their contents, and are always separated from the contents by a space.

## 3. Translation Data Format:

The content of the translation files is identical to the source files except:

- the initial <DOC> tag contains an additional attribute:

```
<DOC docid="..." sysid="...">
```

where the sysid string enclosed in double quotes matches the name of the directory containing the file.

- the contents of the <seg> tags are plain ASCII English text, although most of the automatic MT systems included some strings of untranslated UTF8 Arabic character data in their output, and these are retained as-is.