Language Specification in Arboreal

Arboreal's language architecture (cf. schematic) requires that the language of text be somehow specified. Here are the rules Arboreal uses:

1. Language may be specified in the document metadata. Arboreal determines the language by using the XPath query under the **<locator>** tag in the **<metadata>** definition in the docspec file. For Archimedes texts, the language is specified in the **<lang>** section under **<info>.** E.g.:

   ```
   <info>
   ...
       <lang>it</lang>
   ...
   </info>
   ```

2. Any element (tag) may have a **lang** attribute. The value set here applies to the entire subtree for which the element is the root (unless the setting is overridden by a **lang** attribute of some descendant node or nodes). This language setting overrides the language (if any) that is specified in the document metadata. Note that the language may be set for the entire document simply by supplying a **lang** attribute for the root element. E.g.:

   ```
   <root lang="la">
   ...
   ```

3. The text under certain elements is considered as a single unit, called an **amalgamation**. Nodes to which this behavior applies are called **container** nodes. The nodes considered containers are enumerated under **<containers>** in the docspec file. (Also: any node that is the root of a subtree containing only text nodes is automatically considered a container node.) In the Archimedes DTD, **<s>** is a container. The amalgamation belonging to a container may consist of text in only a *single* language. In the case of multilingual documents, however, it will sometimes be necessary for a container (e.g., a sentence) to contain text in more than one language. To allow for this possibility, elements may defined as **subcontainers** in the docspec file. Text that belongs to a subcontainer is treated as the amalgamation of the subcontainer, not of the (parent) container. In the Archimedes doctype, **<foreign>** is defined as a subcontainer. Thus we can have something like:

   ```
   <s id="Academica2.18.3" lang="la">Cum enim ita negaret quidquam esse
   quod comprehendi posset (id enim volumus esse <foreign
   lang="el">a)kata/lhpton</foreign>), si illud esset, sicut Zeno
   definiret, tale visum (iam enim hoc pro <foreign
   lang="el">fantasi/a|</foreign> verbum satis hesterno sermone trivimus),
   visum igitur impressum effictumque ex eo unde esset quale esse non
   posset ex eo unde non esset...</s>
   ```

4. If no language is specified anywhere in the document, the document is considered to be in the default language. This default may be set in the preferences dialog.

The code used for the language is always the two- or three-letter code specified in ISO 639. These codes are *not* case-sensitive. The codes for languages we're currently using are:

`ar` Arabic
`de` German
`en` English

- `el` Greek
- `fr` French
- `it` Italian
- `la` Latin
- `zh` Chinese

For an sample document that illustrates language embedding, see http://archimedes.fas.harvard.edu/text/en/testbed.xml.

---

webmaster@archimedes.fas.harvard.edu
Last modified: Wed May 19 15:23:29 EDT 2004