

CADIXE XML Annotation Editor

User manual (editor version 1.08 22/03/04, translation v0.2)

Contact: gilles.bisson@imag.fr

(poibeau@lipn.univ-paris13.fr for the English version of the manual)

Objectives of the editor

The CADIXE XML Annotation Editor has been developed for easy and interactive annotation of documents using a tag set described by an XML DTD. Below are two websites concerning XML and related stuff.

- XML: <http://www.brics.dk/~amoeller/XML/overview.html>
- DTD: http://www.w3schools.com/dtd/dtd_intro.asp

Contrary to some other editors, the end-user using CADIXE does not have to structure the text as an XML tree. New tags can be introduced one by one, in any order. Text mark-up can be seen using a style sheet that can be freely modified by the end-user.

Minimal configuration

The software is written in java. It can thus be launched on a large number of operating systems (it has been tested on Linux, MacOS and Windows). A virtual machine v1.4.1 or higher is required and the system should have at least 256 Mo RAM.

Installation

The archive containing the software must first be unzipped. A directory called "Editeur-XML" is created. This directory contains the application and related resources that are structured as follows:

- XmlJava.bat (launching script)
- SgmlJava.jar (java runnable file)
- Example.txt
- Preferences/
 - Default.DTD
 - Default.Style
 - Xmljava.conf (editor configuration file)
 - ...
- Icons/ (icons used by the system)
- Splash/ (welcome screen)
- Bugreport/ (log directory)

Starting the software

A java virtual machine (v1.4.1 or higher) must be available in the working environment. The annotation editor on Unix (or Linux) can be launched in a console (xterm), from the "Editeur-XML" directory, using the following command line:

```
java -classpath SgmlJava.jar MainFrame
```

With other operating systems (MacOS X or Windows for example), the annotation editor can also be launched with a double-click on the SgmlJava.jar icon in the file explorer. After a while the editor window appears.

The command line can include the following options to modify the default behaviour of the annotation editor (relative paths are related to the XML-EDIT-x.x directory):

```
-root path: to set the directory containing the editor and its resources
-text filename: to load a non annotated text file in the editor
-dtd filename: to load another DTD than the one indicated in Xmljava.conf
-style filename: to load another DTD than the one indicated in Xmljava.conf
```

The `-root` parameter does not have any influence on `-text`, `-dtd` and `-style`. This option is only useful if one want to launch the annotation editor from another location than its home directory (in this case, the option is mandatory for the editor to have access to its resources).

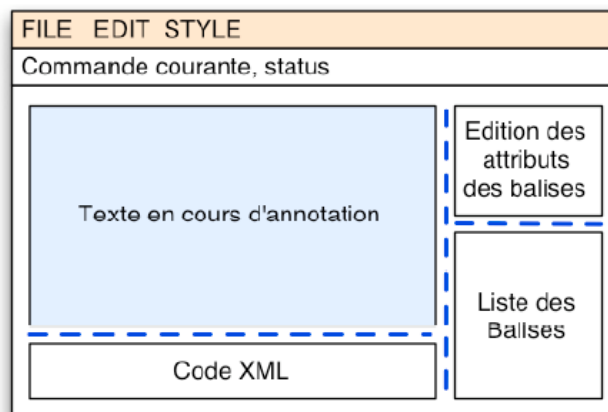
Below is an example allowing to load a text file from the command line:

```
java -classpath SgmlJava.jar MainFrame -text /Volumes/Bio/myText
```

Description of the Graphical User Interface

Besides the menu bar and the icon bar corresponding to menu items, the window is separated in four different zones that can be resized using the separations bar between them. The four zones are the following:

- TXT: the original document being annotated (upper left)
- XML: XML tags introduced in the text (lower left)
- ATT: value of the XML attributes of the current tag (upper right)
- BAL: list of available tags from the current context (lower right)



When the preference directory contains a DTD and a style sheet, the list of available tags is automatically initialised when the editor is launched. Then, the user must either

- load a new ASCII file or
- continue to annotate a partially annotated file.

Using the current version of the annotation editor, we recommend not to work on files longer than 10.000 words, for efficiency reasons. If the user wants to start with an example, the file example.txt can be loaded using the "load text document" item in the file menu.

An introduction to text annotation using the CADIXE editor

The annotation process mainly includes two different steps:

- 1) select with the mouse the chunk of text (words, paragraph) to annotate
- 2) apply a annotation tag on this chunk by clicking on the corresponding tag in the right side of the window.

The corresponding chunk of text will then appear using the display format associated with the tag (in the style sheet) and the new tag appears in the XML area in the lower part of the editor. The display format can be modified by changing the associated style sheet (from the style menu).

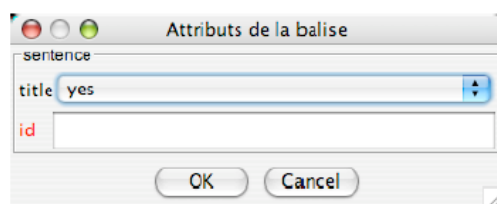
To facilitate the annotation process, pressing the ALT key automatically adjust the current selection to plain words (if a word is partially selected on the left or on the right-hand-side of the current selection, pressing the ALT key will automatically completely include it in the selection). This behaviour can be set by default in the editor by modifying an option in the preference box, so that the user does not have to press the ALT key (see edit menu).

```
We isolated a temperature-sensitive sporulation defective mutant
of the #SigA#
+ ALT: We isolated a temperature-sensitive sporulation defective
mutant of the #SigA#
```

The list of available tags is dynamically modified according to the position of the cursor in the text (only tags that are more specific than the current tag can be used, according to the DTD). Other tags are coloured in grey and can't be used. This behaviour has been defined to avoid tagging errors and guide the end-user in the annotation process.

According to the DTD, each tag can be associated with a list of attributes to specify its meaning. These attributes and their value appear in the ATT part of the editor window and can be modified by the end-user. If values are included in a named list, these values can be selected from a menu. Otherwise, an editable window is provided to enter the value of the attribute.

When a new tag is inserted, a dialog box appears to specify the values of the different attributes (if any). Clicking on the cancel button cancels the current insertion process. Values associated with a REQUIRED value appear in red (it is the case of the ID attribute below):



The current tag is always defined by the cursor position. Thus, to check a specific annotation, the end-user only needs to click on the relevant chunk of text. The tag name, the list of attributes together with their values automatically appears in the ATT area of the annotation editor.

The lower left area of the editor window displays an XML view of the annotated text. This area can be useful to check the result of the annotation (for example to check tag inclusion). Navigation between the TXT and XML area of the editor is synchronized, that is to say that clicking on one area automatically position the other area to the corresponding text chunk.






Modifying a text and its annotation

The end-user is free to insert, delete or copy-and-paste text in a document. To delete an annotation tag, the user must put the cursor on the corresponding text area and click on the "Delete current tag" button. The current tag will be deleted, together with all the subsequent tags (from the XML sub-tree). In some cases however, several different unrelated tags can be applied on the same chunk of text. In such a case, to indicate the layer to which one wants to apply a process, it is possible to specifically click on the concerned tag in the chunk. The editor has the following behaviour:

- 1st click: select the tagged chunk of text
- 2nd click: the cursor is positioned at the beginning of the chunk
- 3rd click: the cursor is positioned at the end of the chunk.

File menu

The set of commands available to load a file is contained in the FILE menu and most of them are directly available from the icon bar.

- : create a new empty document
- : load a non annotated text file
- : load a (partially annotated) XML file
- : save the current file
- : load a new DTD

When saving an annotation, the editor builds two different files: the first one (with an .xml extension) contains the annotated text and the second one contains the current style sheet (with a .style extension). XML files can be exported using the HTML or RTF









format with the same appearance but these formats can not be loaded for further use in the editor.

: save in HTML format

: save in RTF format

Edit menu

The edit menu contains the following commands (the two first commands are not active in this version of the annotation editor):

- : cut the selected text (without tags)
- : copy the selected text (without tags)
- : paste the selected text (without tags)
- : search a string or a regular expression in the text
- : next occurrence
- : delete the current tag (does not delete included tags)
- : delete the current tag and recursively all included tags (a conformation box appears for this command since all existing annotation can disappear using this command)
- : modify the editor preferences (see below)

Editor preferences

In the edit menu, the "preferences" item allows the end-user to modify the default behaviour of the editor. The preferences box contains three different tabs:

- File tab (to set default files)
 - working folder (default resource directory)
 - default DTD
 - default style
- Font tab (default fonts)
 - plain text (font used in the annotation area)
 - tag body (font used for XML tags)
 - sample text (font used in the sample XML text)
- Option tab
 - extend the selection at the border of the word
- show ending tag in XML (deprecated)
- allows to insert all the most specific tags (if the option is active, the end-user has access to all the tags, otherwise he only has access to accessible tags according to the context; this last possibility is the default behaviour and is more convenient for beginners)







- Hide the inactive tags in the list (if this option is active, inaccessible tags are not visible, otherwise they are displayed on a red background)
- Text size limit (maximum size of the text sample in the XML area)
- Tag size limit (maximum size of the XML tag with attributes in the XML area)

Toolbar panel

This option allows to modify the number of icons appearing in the tool bar. The left part of this panel gives the set of possible items and the right part shows if the item should appear or not in the tool bar). The relative position of an icon is specified by two values: a group number (0 to 9, 0 referring to the left-hand side of the bar) and its position in the group (0 to 9). A group is a set of related items.

Style menu

Each tag from the DTD is related to a display format so that each tag can be easily identified from the shape and appearance of the corresponding chunk of text. Styles in CADIXE are comparable to Word processor styles or CSS style for HTML web pages. The style menu contains the following items:

-  : load a style sheet. A style sheet must have a .style extension.
-  : save the current style sheet (with a .style extension)
-  : save the current style sheet as the new default style sheet (in preferences). When a new document is created, this style sheet will then applied.
-  : create a new style sheet
-  : duplicate a style sheet
-  : edit the current style sheet

Once the "edit the current style sheet" item has been selected, a new dialog box appears with two different areas:

- on the left, the set of tags contained in the current DTD
- on the right, a new panel with two tabs: graphic style and semantic style

Graphic style tab

This tab allows the end-user to associate a given tag with a font, a size, a shape (bold, italic, underlined), a letter and a background colour. Once the dialog box is closed, the new options are applied to the current XML file. Modifications that are made on the style sheet are automatically saved together with the annotated XML document. If one wants to save the modifications to create a new style sheet, the save command must be explicitly called.

Semantic style tab

During the annotation process, introducing structuring tags (that is to say tags that are not related to semantic information on the text) can be a tiresome and time-consuming task. To partially solve this problem, the annotation editor includes an automatic completion mode. This mode associates to some tags, a set of other tags that will be automatically inserted according to the XML tree, if they are missing.

The dialog box contains two sets of related options. The first one specifies whether an action should be performed if the first upper tag is missing according to a specific tag. If yes, the end-user has to specify what tags must be inserted and where these tags must be inserted (at the beginning and at the end of the chunk tagged text). Four different actions are possible:

- directly add the new tag around the current tag
- add the new tag around the current paragraph
- add the new tag at the beginning and at the end of the document
- add the new tag using a regular expression describing the position of the tag to insert. The insertion can be made before or after the chunk of text described by the regular expression.



The automatic insertion recursively apply (up to the top of XML tree) while no relevant tag according to the DTD is found. Inserted tags have the default shape and appearance.

Help menu

This menu only contains one item (about) to display the current version of the software and the date of its release.

Document validation

The annotation editor allows the end user to annotate documents in any desired order. The end-user can always, however, check that the current annotation is valid according to the DTD. Two options are available:

-  : partial validation (annotated parts of the documents are valid according to the DTD). This option is of interest only if the annotation has been made using another word processor. Due to its contextual behaviour, the CADIXE annotation editor guarantees that the annotation is always correct.
-  : strict validation (the whole document is valid according to the DTD).

The end-user can always save the document, whether it is valid or not according to the DTD.

Contact

This software is developed in the framework of the CADERIGE project (<http://caderige.imag.fr/>). This research project is funded by CNRS (ACI Bioinformatique). For any further information, please contact Gilles Bisson (gilles.bisson@imag.fr). If you want to report a bug, please include as far as possible

your data (document, DTD, style sheet) and a small textual description of the behaviour of the software, so that we can reproduce the bug. Thanks!

English translation (v0.2: Thierry Poibeau)