

HTML diff (compare) component

This project is a research-level project for an ambitious student.

Background

The main content type used for textual content in Daisy is a subset of HTML which is also well-formed XML (technically, it's not XHTML since it's not in the XHTML namespace).

Much like a source code version control system, in Daisy a version history of documents is kept. Being able to compare the content of these versions is very useful. Currently the diffs shown in Daisy are based on the HTML source text (just try Page Action -> Versions in the menu on top of this page to see an example). It would be very nice to have graphical diffs instead, thus simply a HTML document with changes, additions and deletions marked (a bit like the change tracking in word processors).

Project notes

- the project result should be a Java library which takes two HTML (well-formed XML) files as input and delivers one marked-up HTML file as result.
- this project is independent from Daisy, the student doesn't need to learn anything about Daisy. If the project succeeds successfully, Daisy developers will be happy to integrate it.
- the diff should work with the XML structure of the documents, not some simpler tokenization which might not interpret the document structure correctly.
- a good start point would probably be to investigate current products to see how they approach the problem.

Result

[DaisyDiff](#)¹ hosts and maintains the HTML diffing library developed for this project.

Fields

Name	Value
Category	Google SoC 2007 proposal

1. <http://code.google.com/p/daisydiff/>