

HTML DIFF PROJECT:

(the original idea)¹

PROJECT DESCRIPTION:

The goal of the "HTML diff" project is to provide a Java library that compares XML files. The minimum inputs are 2 XML files, and the minimum output file is a Java representation of the basic operations that turned the first data structure into the second. This output can be converted to an XML-based representation or, given the proper formatting guidelines (in a yet undecided data type), into an html file with a certain syntax to denote changes.

In addition, distance metrics can be computed that allow the user to quantify by how much 2 files differ. This enables them to distinguish minor from major changes, and tells them what the (chronological) relationships are between 3 files. For computing the changes there can be several algorithms:

- A simple LCS (Longest Common Subsequence) implementation that will work on plain text or simple HTML files. This algorithm is the algorithm used in the Unix "diff" command, in CVS, patch, etc. It's computationally the best algorithm but it will destroy complex XML hierarchies. There are many open source implementations available of this algorithm. For instance the Wikipedia and Drupal software uses LCS, as does the current text-based implementation in Daisy.
- A more complicated graph algorithm that uses a greedy method to compute a difference but not the optimal one. this can be optimal for XML files with little reproduction.
- A complex algorithm with perfect accurateness.

For each of these possibilities there are papers that describe algorithms. The graph based ones could be used as a first pass run on the general structure, followed by the LCS one as a second pass to annotate small changes in markup.

There should also be an option that chooses the optimal algorithm for a given XML file at runtime (based on length, complexity, leafs, ...)

While this is a project description for a very general XML comparator, in reality there might be the need to integrate knowledge of the XML format used in DiasyCMS into the code, to make it more stable and accurate. (vide infra for the column deletion example)

WHAT'S IN IT FOR DAISY:

The current implementation in Daisy is very basic and not intuitive for the average user. When for example a dot is added to the end of a sentence, the entire line of html is removed and added with the extra dot. It would be better to indicate in the margin that a change took place at that line, and then show the added dot. For removed pieces of code there can be an inline icon of sorts that can expand to show the removed data. Changes could be played like a movie (see <http://weblog.infoworld.com/udell/gems/umlaut.html>). More versions could be displayed in one document with different colors for different version/author changes, etc.

In general the fact that html tags are shown on the compare page makes it not very user friendly. When a column is removed from a table you will see alot of removed lines of code in the entire table structure, when it actually took just one basic action. (see <http://cocoondev.org/daisyscratchpad/393-daisy/version/1/diff?otherVersion=2>). This library would make it easy to represent the changes in a way that html-ignorant people will understand what actually happened.

An example of how things can go very very badly wrong can be found here: <http://cocoondev.org/daisyscratchpad/291-daisy/392-daisy/version/5/diff?otherVersion=1>

In time the library could be used to compare more than just the scratchpad html. The possibilities are endless really. The integration with daisy itself however is beyond the scope of this project and needs brainstorming with the entire community.

DELIVERABLES:

- A literature study with an outline of all the possible algorithms + Testing of existing (proprietary) code. + Explaining the algorithms on the community scratchpad
- A study of the XML files used by DaisyCMS, which I would need to prioritize on. (Do we want to compare ordered or unordered XML files?)
- A LCS implementation (could be copied from the current Daisy codebase?)
- A HTML representation/change syntax that can be used for DaisyCMS
- One or more graph-based implementations
- A library with all the options from the project description
- Documentation
- Progress updates posted to the community scratchpad (at least weekly)

TIME LINE:

April 11th - June 30th: Given the fact that in Belgium the exam period ends somewhere in the end of June this will be my least productive period of the summer: Getting to know Daisy, and the literature study will be all I can cope with.

July 1st - July 16th: To make up for the slow month of June, I will work day and night to finish the LCS Implementation and the HTML representation.

July 16th - July 26th: A well deserved vacation in Greece.

July 27th - August 15th: One or more graph algorithms.

August 15th - August 31st: Finish the package, complete any unfinished documentation before the final deadline.

September 1st - September 25th: Integrate the package into DaisyCMS and finish the unfinished.

BIO:

I'm a Belgian student living in Vosselaar. In June I hope to graduate magna cum laude in "bachelor of science in engineering: option computer science & electrical engineering" at the KULeuven and then move on to a master in computer science. I'm a very experienced Java programmer and I like working in teams. I'm active in the open source community as member of the Ubuntu-be loco team and support point. I'm a very quick learner with great analytical skills. A more exhaustive resumé can be found at <http://users.telenet.be/guyvdb/cv-en.pdf>.

I'd like to work in the offices in Ghent and do this as an internship for my university. The advantage for Google and Daisy is that my commitment will be guaranteed as this project would also be part of my studies.

Fields

Name	Value
Category	Google SoC 2007 proposal

1. /daisy-wiki/291-daisy/382-daisy.html
2. <http://weblog.infoworld.com/udell/gems/umlaut.html>
3. <http://cocoondev.org/daisyscratchpad/393-daisy/version/1/diff?otherVersion=2>

4. <http://cocoondev.org/daisyscratchpad/291-daisy/392-daisy/version/5/diff?otherVersion=1>
5. <http://users.telenet.be/guyvdb/cv-en.pdf>