

MASARYK UNIVERSITY
FACULTY OF INFORMATICS



Visual testing something catchy

DIPLOMA THESIS

Juraj Húska

Brno, 2015

Declaration

Hereby I declare, that this paper is my original authorial work, which I have worked out by my own. All sources, references and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Juraj Húska

Advisor: Mgr. Marek Grác, Ph.D.

Acknowledgement

Some people helped me a lot and some not at all. Nevertheless, I would like to thank all.

Abstract

This thesis is very important!

Keywords

key word1, and so on

Table of contents

1	Introduction	3
2	Visual testing of software	4
2.1	<i>Visual testing in release testing process</i>	4
2.2	<i>Need for automation</i>	5
2.3	<i>Requirements for automation</i>	5
2.3.1	Low cost of test suite maintenance	6
2.3.2	Low percentage of false negative or false positive results	7
2.3.3	Reasonable time to execute a test suite	7
2.3.4	Concise yet useful test suite output	7
2.3.5	Support for Continuous Integration systems	8
3	Analysis of existing solutions	9
3.1	<i>Mogo</i>	9
3.1.1	Mogo drawbacks	10
3.2	<i>BBC Wraith</i>	10
3.2.1	PhantomJS	11
3.2.2	CasperJS	11
3.2.3	BBC Wraith drawbacks	12
3.3	<i>Facebook Huxley</i>	13
3.4	<i>Rusheyeye</i>	14
3.4.1	Rusheyeye drawbacks	15
3.5	<i>Conclusion of analysis</i>	15
4	New approach	18
4.1	<i>Hypothesis</i>	18
4.2	<i>Process</i>	18
4.3	<i>Analysis of useful tool output</i>	18
5	Implemented tool	19
5.1	<i>Client part</i>	19
5.1.1	Arquillian	19
5.1.2	Arquillian Graphene	19
5.1.3	Rusheyeye	19
5.1.4	Graphene visual testing	19
5.2	<i>Server part</i>	19
5.2.1	Web application to view results	19
5.2.2	Storage of patterns	19
6	Deployment of tool and process	20
6.1	<i>Deployment on production application</i>	20
6.2	<i>Deployment on development application</i>	20

6.3	<i>Usage with CI</i>	20
6.4	<i>Cloud ready</i>	20
6.5	<i>Results</i>	20
7	Conclusion	21
	Bibliography	21

1 Introduction

There is a big demand for this thesis. Need and cost of manual testing, space for improvement.

2 Visual testing of software

Testing of software in general is any activity aimed at evaluating an attribute or capability of a program and determining that it meets its required results [1]. It can be done either manually by actual using of an application, or automatically by executing testing scripts.

If the application under test has also a graphical user interface (GUI), then one has to verify whether it is not broken. Visual testing of an application is an effort to find out its non-functional errors, which expose themselves by changing a graphical state of the application under test.

Typical example can be a web application, which GUI is programmed usually with combination of HyperText Markup Language (HTML) and Cascading Style Sheets (CSS). HTML is often used to define a content of the web application (such as page contains table, pictures, etc.), while CSS defines a structure and appearance of the web application (such as color of the font, absolute positioning of web page elements, and so on).

The resulting web application is a set of rules (CSS and HTML) applied to a static content (e.g. pictures, videos, text). The combination of rules is crucial, and a minor change can completely change the visual state of the web application. Such changes are very difficult, sometimes even not possible to find out by functional tests of the application. It is because functional tests verify a desired functionality of the web application, and do not consider web page characteristics such as red color of heading, space between two paragraphs, and similar.

That is why a visual testing has to take a place. Again, it is done either manually, when a tester by working with an application, is going through all of its use cases, and verifies, that the application has not broken visually. Or automatically, by executing scripts which assert a visual state of an application.

In this thesis we are going to focus on the visual testing of web applications only. As we mentioned above, the way how web page looks like is mainly determined by CSS script. There are two ways of automated testing used:

1. asserting the CSS script
2. or comparing screen captures (also known as screenshots) of new and older versions of the application.

2.1 Visual testing in release testing process

Nowadays software is often released for a general availability in repetitive cycles, which are defined according to a particular software development process. Such as Waterfall [2], or Scrum [3].

Testing of software has an immense role in this release process. While automated tests are often executed continuously, as they are quicker to run than manual tests, which are carried

out at a specific stage of the release process.

For example in RichFaces¹ Quality Engineering team² visual testing was done manually, before releasing the particular version of RichFaces library to a community. In practice it involves building all example applications with new RichFaces libraries, and to go through its use cases with a particular set of web browsers.

To be more specific, consider please a web page with a chart elements showing a sector composition of gross domestic product in the USA (as figure 2.1 demonstrates). To verify its visual state is not broken, would involve e.g.:

1. Checking the size, overflowing and transparency of all elements in charts.
2. Checking colors, margins between bars.
3. Putting a mouse cursor over a specific places in the chart, and verifying whether a popup with more detailed info is rendered in a correct place.
4. Repeat this for all major browsers³, and with all supported application containers⁴.

2.2 Need for automation

The chapter 2.1 tried to outline how tedious and error prone might manual visual testing be. From our experience in the RichFaces QE team, any activity which needs to be repeated, and does not challenge tester's intellect enough, become a mundane activity. The more one repeats the mundane activity, the more likely an mistake is introduced: one forgets to try some use cases of an application, overlooks some minor errors, etc.

Automated visual testing addresses this shortcomings, as it would unburden human resources from mundane activities such as manual testing, and would allow spending their time on intellectually more demanding problems. However, it introduces another kind of challenges, and needs to be implemented wisely. Following are minimal requirements for a successful deployment of an automated visual testing.

2.3 Requirements for automation

An overall cost of the automation has to be taken into consideration. It is necessary to take into account higher initial cost of automation, and consequences it brings: such as increased time to process relatively huge results of testing, cost of test suite maintenance.

1. RichFaces is a component based library for Java Server Faces, owned and developed by Red Hat
2. Quality Engineering team is among the other things responsible for assuring a quality of a product
3. Major browsers in the time of writing of this thesis are according to the [4]: Google Chrome, Mozilla Firefox, Internet Explorer, Safari, Opera
4. Application containers are special programs dedicated to provide a runtime environment for complex enterprise web applications, e.g. JBoss AS, Wildfly, Apache Tomcat

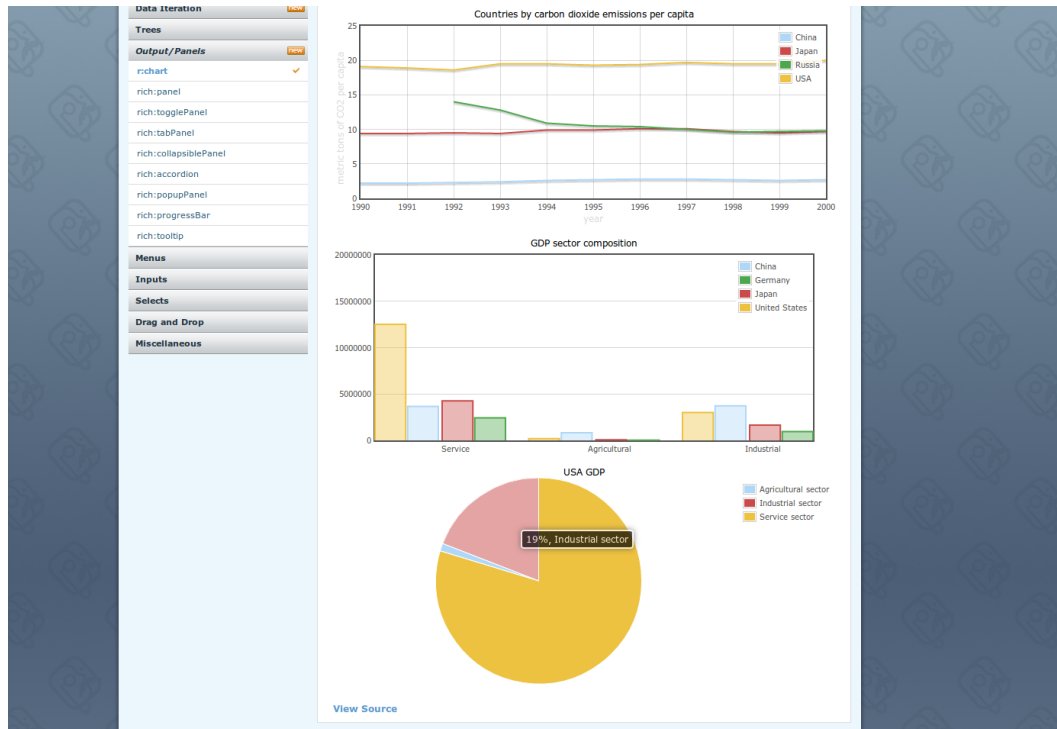


Figure 2.1: RichFaces chart component shown in Showcase application

Therefore, to foster an effectiveness in quality assurance teams, while keeping the cost of automation reasonable low, automated visual testing would require:

1. A low cost of a test suite maintenance;
2. a low percentage of false negative or false positive tests results;
3. a reasonable time to execute the test suite;
4. a concise yet useful test suite output;
5. a support for Continuous Integration systems⁵.

2.3.1 Low cost of test suite maintenance

A test suite needs to reflect a development of an application under test. Therefore, with each change in the application, it is usual that the test suite has to be changed as well. Making a

5. Continuous Integration (CI) system is software to facilitate a practice of CI, which in short is about merging all developer copies with a shared mainline several times a day [5].

change in the test suite can often introduce a bug, and cause false negative or false positive tests results.

To keep this cost as low as possible, the test suite script has to be readable and meaningful, so when the change is about to be introduced, it is clear where and how it should be done.

A test framework in which the test suite is developed should provide high enough abstraction. That would enable better re-usability for various parts of the test suite, while lowering the overall cost of maintenance.

Specifically for visual testing, when done by comparing screen captures, it is very important how well a repository of screen captures is maintainable. Secondly, how reference (those correct ones, other screen captures will be compared with) screen captures are made.

2.3.2 Low percentage of false negative or false positive results

False negative test results incorrectly indicate a bug in an application under test, while it is a bug in the test suite itself. They are unwanted phenomenon as they take time to process and assess correctly.

False positive tests results hide actual bugs in an application. They provide an incorrect feedback by showing the tests as passing, even when there is a bug in the application.

Specifically for visual testing, when it is done by comparison of screen captures, it is very easily to be broken by small changes on a page. For example if the page outputs a current date, then it would break with every different date. There has to exist techniques, which would prevent these situations.

2.3.3 Reasonable time to execute a test suite

Reasonable time is quite subjective matter, but in general, it depends on how many times e.g. per day one needs to run whole test suite. Nowadays trend is a Continuous Integration, when a developer or a tester commits changes of an application several times per day to a shared source code mainline. Each such commit should trigger the test suite, which verifies whether the change did not introduced an error to the application.

According to creators of Continuous Integration practice, the whole point of CI is to provide a rapid feedback. A reasonable time for them is 10 minutes. If the build takes more time, every minute less is a huge improvement (considering a developer/tester runs test suite several times a day).

2.3.4 Concise yet useful test suite output

One of drawbacks of automated testing is its ability to produce huge amount of logs, test results etc. The output therefore needs to be as concise as possible, while still providing an useful information. A tester needs to be able to quickly recognize where the issue might be.

The best situation would be if the tester does not need to run the test again in order to spot the issue. The output should give him a clear message where the issue is.

For visual testing specifically, this can be achieved by providing a tester with screen captures together with difference of old and new version.

2.3.5 Support for Continuous Integration systems

This is quite easily to be achieved, but still, a developer of a tool for visual testing should have this in mind beforehand. Today's CI systems support variety of build systems, for various platforms, and languages. For example Jenkins supports build systems like Maven or Gradle, but it can run also shell scripts.

3 Analysis of existing solutions

As we introduced in 2.3, there are many aspects which need to be taken into consideration when automating visual testing. Following analysis is going to compare existing solutions to automated testing with those requirements in mind, while introducing different approaches to visual testing.

The representative list of tools for comparison was made also according to an ability to be used in enterprise context. In an enterprise company, there is a stress on stability and reliability of employed solutions. It is quite vague requirement, and it is usually hard to find out which tools are a good fit for enterprise companies, however, some indicators, which we used as well, might be helpful:

- Is a project actively developed? When was the last release of the project, or how old is the last commit to a source code mainline?
- How many opened issues the project has? When was the last activity with those issues ?
- What is the quality of the source code? Is it well structured? Does it employ best development practices?
- Does the project have a user forum? How active are the users?
- Is a big enterprise company behind the solution, or otherwise sponsoring it ?
- What are the references if the project is commercialized ?

For each tool in following sections we are going to show an example usage and its main drawbacks together with some basic description.

3.1 Mogo

Mogo [6] approach to visual testing can be in short described like:

1. One provides set of URLs of an application under test to a cloud based system.
2. Application URLs are loaded in various browsers, detection of broken links is done.
3. Screenshots are made and are compared with older screenshots from the same URL to avoid CSS regressions.

There is no programming script required, therefore it can be used by less skilled human resources. It can be configured in shorter time, and thus is less expensive.

3.1.1 Mogo drawbacks

Drawbacks of this approach are evident when testing dynamic pages, which content is changed easily. Applications which provide rich interaction options to an end user, and which state changes by interacting with various web components (calendar widget, table with sorting mechanism etc.), require more subtle way of configuring what actions need to be done before the testing itself. Mogo is suitable for testing static web applications, not modern AJAX enabled applications full of CSS transitions.

Above mentioned drawbacks might lead to a bigger number of false negative test results when used with real data (any change, e.g. showing actual date may break testing), or to a bigger number of false positive tests results when such a tool is used to test mocked data ¹.

3.2 BBC Wraith

Wraith is a screenshot comparison tool, created by developers at BBC News [7]. Their approach to visual testing can be described like:

1. Take screenshots of two versions of web application by scripting either PhantomJS 3.2.1, or SlimerJS² by another JavaScript framework called CasperJS 3.2.2 [20].
2. One version is the one currently developed (which run on localhost³), and the other one is a live site.
3. Once screenshots of web page from these two different environments are captured a command line tool `imagemagic` is executed to compare screenshots.
4. Any difference is marked with blue color in a created picture, which is the result of comparing two pictures (It can be seen at Figure 3.1).
5. All pictures can be seen in a gallery, which is a generated HTML site (It can be seen at Figure 3.2).

To instruct BBC Wraith tool to take screenshots from the web application, one has to firstly script PhantomJS or SlimerJS to interact with the page, and secondly, creates a configuration file, which will tell the PhantomJS instance which URLs need to be loaded and tested. PhantomJS script is one source of distrust to this tool, and therefore is introduced furthermore.

1. Mocked data is made up data for purpose of testing, so it is consistent and does not change over time
2. SlimerJS is very similar to PhantomJS 3.2.1, it just runs on top of Gecko engine, which e.g. Mozilla Firefox runs on top of. [10]
3. In computer networking, `localhost` means this computer. [11]

3.2.1 PhantomJS

PhantomJS [8] is stack of web technologies based on headless⁴ WebKit⁵ engine, which can be interacted with by using of its JavaScript API.

For the sake of simplicity we can say that it is a browser which does not make any graphical output, which makes testing with such a engine a bit faster and less computer resources demanding.

One can instruct PhantomJS to take a screenshot of a web page with following script:

```
var page = require('webpage').create();
page.open('http://google.com/', function(status) {
  if(status === 'success') {
    window.setTimeout(function() {
      console.log('Taking screenshot');
      page.render('google.png');
      phantom.exit();
    }, 3000);
  } else {
    console.log('Error with page ');
    phantom.exit();
  }
});
```

When executing such a script it will effectively load `http://google.com/` web page, waits 3000 milliseconds, and after that, creates a screenshot to the file `google.png`.

In most environments it will be sufficient to wait those 3000 milliseconds in order to have the `www.google.com` fully loaded. However, in some resource limited environments, such as virtual machines⁶, it does not have to be enough. It will result in massive number of false negative tests. There is a need for more deterministic way of finding out whether the page was loaded fully in given time, and taking of the screenshots itself can take place.

PhantomJS API is wrapped by CasperJS, which is furthermore described below.

3.2.2 CasperJS

CasperJS is navigation scripting and testing utility written in JavaScript for the PhantomJS and SlimerJS headless browsers. It eases the process of defining a full navigation scenario and provides useful high-level functions for doing common tasks [20].

4. Headless software do not require graphical environment (such as X Windows system) for its execution.

5. WebKit is a layout engine software component for rendering web pages in web browsers, such as Apple's Safari or previously a Google Chrome [9]

6. Virtual machines are created to act as real computer systems, run on top of a real hardware

Following code snippet shows a simple navigation on Google search web page. It will load *http://google.com* in a browser session, will type into the query input string *MUNI*, and will submit it.

```
casper.start('http://google.com/', function() {
  // search for 'MUNI' from google form
  this.fill('form[action="/search"]', { q: 'MUNI' }, true);
});

casper.run(function() {
  this.exit();
});
```

The problem with this script, which we identified, is its low-level abstraction of the browser interactions. It makes tests less readable, and thus more error prone when a change is needed to be introduced.

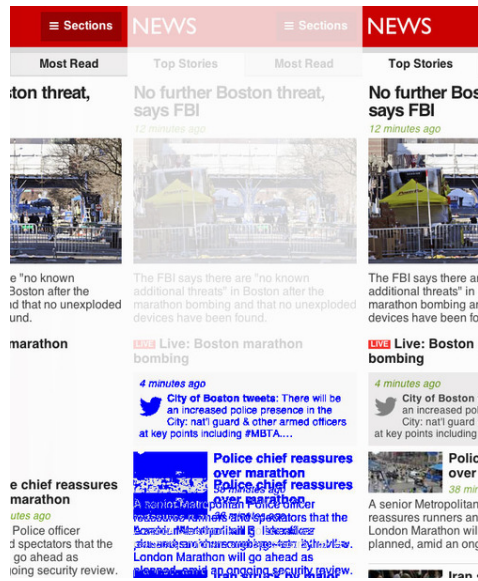


Figure 3.1: BBC Wraith picture showing difference in comparison of two web page versions [12]

3.2.3 BBC Wraith drawbacks

Two of the drawback were described in the previous sections, 3.2.1 and 3.2.2.

List of screenshots for shots

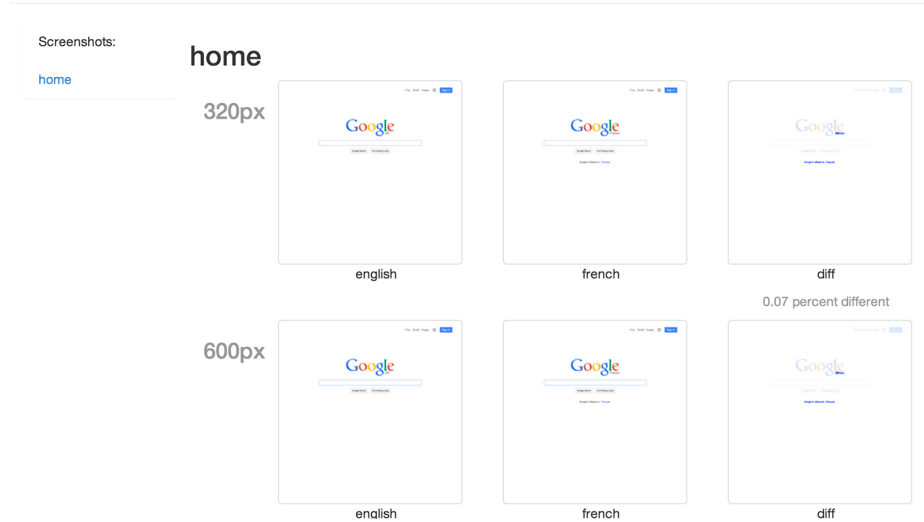


Figure 3.2: BBC Wraith gallery example [13]

Another problem which might occur when testing with BBC Wraith, is cross browser compatibility. As it supports only PhantomJS, and therefore, one can not assure that the page will be looking the same in all major browsers. The incompatibility comes from the fact, that browsers interpret CSS rules differently, and because they have different JavaScript engines. Thus for example web page might look differently in Google Chrome and Microsoft Internet Explorer, and PhantomJS will not catch this issues.

3.3 Facebook Huxley

Another visual testing tool [15], supported by a big company Facebook, Inc. [14], uses similar approach in terms of comparing taken screenshots. The process of taking them, and the process of reviewing results is different though.

1. One creates a script which would instruct Huxley tool, what web pages screenshots should be taken from. Such a script might look like:

```
[test-page1]
url=http://localhost:8000/page1.html

[test-page2]
url=http://localhost:8000/page2.html
```

```
[test-page3]
url=http://localhost:8000/page3.html
```

2. One runs Huxley in the Record mode. That is the mode when Huxley loads the pages automatically in a fresh browser session, and a tester by hitting Enter keyboard button instructs Huxley to take a screenshot. Screenshots are stored in a folder with test name (one given in the square brackets in the example above), together with a JSON⁷ file describing mainly how long should Huxley wait, when doing visual comparison, to have a tested web page fully loaded. Time is measured during this record mode.
3. To start visual testing, one has to run Huxley in the Playback mode. Huxley will start fresh browser session, and will playback loading of the pages, with waiting for the pages to be fully loaded.
4. When there is a change in an application, Huxley will log a warning, and takes a new screenshot automatically. In continuous integration environments, one can instruct Huxley to finish with error, in case screenshots are different. In that case, one can run Huxley again with an option to take new screenshots (if the change is desired, and is not an error).

Main drawback of Facebook Huxley we can see, is similar to BBC Wraith, and that is its non deterministic approach to waiting for a fully loaded web page. It is again a fixed amount of time, which can be different from environment to environment. The time to wait can be configured, however, it is still quite error prone, as for first visual testing run e.g 4 seconds can be would be enough, and for another run would not.

Secondly it lacks a proper way of viewing results of comparisons, leaving with only one option to check the command line output, together with manual opening of the screenshots. This would degrade cooperation among various QA team members, and it is harder to deploy in a software as a service cloud solutions⁸, where such a cooperation might take a place.

3.4 Rusheye

Rusheye [18] is a command line tool, which is focused on automated bulk or on-the-fly inspecting of browser screenshots and comparison with predefined image suite. It enables automated visual testing when used together with Selenium 1 project [19].

7. JSON stands for JavaScript Object Notation, a standard format that uses human readable format to transmit data objects [16]

8. Software as a service is on demand software, centrally hosted, accessed typically by using a thin client via web browser [17].

The process has subtle differences in comparison with previous solutions. It consists from these steps:

1. Screenshots are generated, for example by Selenium 1 tool, while functional tests of web application are executed.
2. First screenshots are claimed to be correct (are controlled manually), they are called patterns.
3. After a change in web application under test, another run of Selenium 1 tests generates new screenshots. They are called samples.
4. Patterns and samples are compared, its visual difference is created, and result is saved in an XML file.
5. The results can be viewed in a desktop application, Rushey manager [22].

Rushey has one very important feature, which another tools lack. It is a possibility to add masks on particular parts of the screenshots. Those parts are ignored when two screenshots are compared. It is a huge improvement to protect from false negative tests, as some all the time changing parts (such as actual date, etc.) can be masked from comparison, and thus their change will not break testing.

3.4.1 Rushey drawbacks

Core of the Rushey is only able to compare screenshots generated by some other tool. Integration with Selenium 1 is advised, however, functional tests written in Selenium 1 suffer from the same problems [21] as scripts written for BBC Wraith 3.2.2. And that is bad readability caused by their low level coupling with HTML and lack of higher abstraction.

Another problem we can see is only desktop version of the tool for viewing results (Rushey Manager). Cooperation on some test suite among QE team members and developers would be more difficult. As they would need to solve persistence of patterns, samples and descriptor of the test suite.

3.5 Conclusion of analysis

All previously listed tools have some nice features, which we would like to be inspired with. However, all of the solutions lack something, what we suppose to be an inevitable part of an effective automated visual testing.

Figure 3.3 summarizes requirements we have for a visual testing tool, and the fact whether the tool satisfies the particular requirement.

Tests readability is a problem we discussed with a particular tool description. It is a level of abstraction over underlying HTML, in which the application is written. It is quite subjective matter, however, there are some clues by which it can be made more objective. For example the way how tests are coupled with the HTML or CSS code. Because the more they are, the less they are readable [21]. A scale we used for evaluation supposes insufficient as lowest readability, which in long term run of the test suite might cause lot of issues.

By tests robustness we suppose a level of stability running of the tests with particular tool has. It means how likely there are false positive and false negative tests, whether are caused by not fully loaded page, or by dynamic data rendered on the page. If the robustness is low, you can find a red mark in a particular cell. Green one otherwise.

Cross browser compatibility issue deals with ability to test web application across all major browsers 3.

By cloud ready features we suppose whether tool has web based system for reviewing results, and thus enables cooperation across QA team members and developers of the particular software.

Continuous Integration friendliness in this context means the fact whether tool is suitable for usage in such systems. It actually means whether output of the tool is clear enough, how much work a tester would be required to do manually to deploy such a tool in a test suite. Whether testers would need to review just logs to find visual testing failure, or whether it will be somehow visible, e.g. whole test would fail.

















	Test script readability	Tests robustness	Cross browser compatibility	Cloud ready	CI friendliness
Mogo	<i>not applicable</i>				
BBC Wraith	<i>insufficient</i>				
Facebook Huxley	<i>insufficient</i>				
Rusheye + Selenium 1	<i>insufficient</i>				

Figure 3.3: Existing solutions features comparison

As Figure 3.3 shows, none of the tools met our requirements fully. Therefore, we decided to proceed with developing of a new tool, which would address all issues, and which would

integrate existing parts of the solutions when it is feasible as well as reasonable. Creation of a new process which would enable effective usage of such tool by QA team members is inevitable. Following chapters describe this new tool and the new process.

4 New approach

4.1 Hypothesis

Simply: reuse of functional tests of the application for visual testing

4.2 Process

How one would use my tool and where in testing stack such visual testing has its place, written in business process notation

4.3 Analysis of useful tool output

Requirements for useful output of such a tool based on questionnaire for RichFaces team, or maybe I will ask all JBoss employees

5 Implemented tool

An answer to the new process, requirements: CI viable, reusing what can be reused, extensible, cloud ready, multiple users

5.1 Client part

5.1.1 Arquillian

Integration testing, starting containers, event based machine

5.1.2 Arquillian Graphene

Functional testing of Web UI, screenshooter

5.1.3 Rusheyeye

Screenshots comparison, rewritten to Arquillian core

5.1.4 Graphene visual testing

An adaptor between Rusheyeye and Arquillian Graphene

5.2 Server part

5.2.1 Web application to view results

Its architecture, reasoning for chosen solutions, screenshots of app, key functionality

5.2.2 Storage of patterns

Description of solution, reasoning

6 Deployment of tool and process

6.1 Deployment on production application

Deployment on stable app

6.2 Deployment on development application

Deployment sooner on application which is in Alpha phase, my hypothesis is that it will not be worth to deploy it on such a app, due to too many changes

6.3 Usage with CI

Jenkins job and its cooperation with the tool, more particullary tool ability to handle multiple jobs, apps, versions, etc.

6.4 Cloud ready

The app can be easily deployed on Openshift

6.5 Results

The percentage of improvement of QA effectiveness

7 Conclusion

What I developed, What I improved, What can be better, Possible ways of extensions: Open-shift cartridge

Bibliography