

Unplagged Developers Manual

Building the Plagiarism Detection Cockpit

Term paper for the master project I

Mentoring Teacher: Prof. Dr. Debora Weber-Wulff

Department Economics II

HTW Berlin – University of Applied Sciences

Elsa Mahari (s0534556) <Elsa.Mahari@gmx.de>

Dominik Horb (s0534217) <horb@htw-berlin.de>

Tien Nguyen (s0512510) <s0512510@htw-berlin.de>

Benjamin Oertel (s0522720) <contact@benjaminoertel.com>

Heiko Stammel (s0534218) <heiko.stammel@googlemail.com>

Contents

Introduction	VI
Chapter Overview	VII
Conventions	VIII
1. A Plagiarism Primer	1
1.1. Plagiarism definition and history	1
1.2. Basic Classification of Plagiarisms	2
1.2.1. Definition of Citation	2
1.3. How to detect plagiarism	5
1.3.1. Commercial software systems	7
1.4. Vroni Plag	25
1.4.1. Plagiarism detection steps	25
1.4.2. Technical support	28
1.4.3. Pros & Contras	28
2. Project Workflow and Requirements	30
2.1. The Workflow	31
2.1.1. Product Owner — “The Debbie Meetings”	32
2.1.2. Team Meetings	33
2.2. Target Group	35
2.3. User roles	35
2.4. Basic functionalities	36
2.4.1. Register	36
2.4.2. Login	37
2.4.3. Files overview	38
2.4.4. Parsed files overview	41
2.4.5. Edit profile	43

2.4.6. Googlesearch	43
2.5. Document Parser	45
2.6. Use Cases	45
2.6.1. Register	45
2.6.2. Login	46
2.6.3. File Upload	47
2.6.4. Parse	48
2.6.5. Googlesearch	48
3. Developing Unplugged	49
3.1. Development Environment	50
3.1.1. Git	50
3.1.2. Local Deployment	55
3.1.3. Netbeans	57
3.1.4. Additional Software	61
3.1.5. Continuous Integration	65
3.2. User Interface / User Experience	67
3.2.1. Responsive Layout using CSS3 Media Queries	69
3.2.2. Javascript and fallbacks	71
3.3. Frameworks	73
3.3.1. Zend Framework	73
3.3.2. Doctrine	75
4. Summary and Outlook	77
A. Meetings	79
B. Logged Time As Of March 11, 2012	85
C. Mockups	100
C.1. Hand-Drawn	100
C.2. Digitalized	105

List of Figures

1.1.	4-stage plagiarism detection process	6
1.2.	Overview of Commercial Detection Systems	7
1.3.	Plagaware Website	8
1.4.	Plagaware overview	10
1.5.	Plagaware side-by-side view	11
1.6.	Turnitin Website	12
1.7.	Turnitin - A lot of little matches can't be found, if the sensibility has not been raised.	13
1.8.	Turnitin - A lot of spam-sites are reported. Not all sufficient to use at work. This is one of the harmless examples.	14
1.9.	Ephorus Website	15
1.10.	Ephorus report - gives a great overview of the results	16
1.11.	Ephorus Problem with umlauts	17
1.12.	Plagscan Website	18
1.13.	Plagscan report is clear and tidy.	19
1.14.	Plagscan reports are not self-explanatory.	20
1.15.	Urkund Website	21
1.16.	Urkund List view	23
1.17.	Urkund Report	24
1.18.	Source: http://de.vroniplag.wikia.com/wiki/Lm , 19/03/2012, 08:53 . . .	27
1.19.	Source: http://de.vroniplag.wikia.com/wiki/Lm , 19/03/2012, 08:54 . . .	28
1.20.	colors	29
2.1.	Redmine Roadmap	31
2.2.	Scrum Meeting	32
2.3.	User Stories	33
2.4.	register1	36

2.5. register1	37
2.6. "Einloggen" Link	37
2.7. login form	38
2.8. after login view	38
2.9. overview of files	39
2.10. target function symbol	39
2.11. not marked file	40
2.12. marked target file	40
2.13. parse symbol	40
2.14. Download symbol	40
2.15. Delete symbol	40
2.16. file uploader	41
2.17. detect plagiarism symbol	42
2.18. delete symbol	42
2.19. e.g. simtext running	42
2.20. Example of a Simtex report	43
2.21. Profile edit form	44
2.22. Googlesearch contextmenu	44
2.23. use case register	46
2.24. use case login	47
3.1. Netbeans – New project – Name and Location	60
3.2. Netbeans – New project – Run Configuration	60
3.3. Parsing Files with Tesseract	62
3.4. Default Simtext output	63
3.5. Simtext output in diff format	64
3.6. Deployment workflow	65
3.7. Mockup – New case – hand-drawn	68
3.8. Mockup – New case – digitalized	69
3.9. Initial Screen PSD	70
A.1. meeting protocol no. 6	79
A.2. meeting protocol no. 8	80
A.3. meeting protocol no. 10	80
A.4. meeting protocol no. 11	81
A.5. meeting protocol no. 12	81

A.6. meeting protocol no. 13	82
A.7. meeting protocol no. 15	82
A.8. meeting protocol no. 17	83
A.9. meeting protocol no. 18	83
A.10.meeting protocol no. 19	84
C.1. Mockup – Compare results – digitalized	100
C.2. Mockup – Media list – digitalized	101
C.3. Mockup – New case – digitalized	102
C.4. Mockup – New fragment – digitalized	103
C.5. Mockup – New project – digitalized	104
C.6. Mockup – New case – digitalized	105
C.7. Mockup – List fragments – digitalized	106
C.8. Mockup – New fragment – digitalized	107
C.9. Mockup – Show fragment for approval – digitalized	108
C.10.Mockup – New report – digitalized	109

Introduction¹

After Minister Guttenberg had to resign, because of the plagiarisms found in his doctoral thesis, the big media coverage and interest in plagiarism in Germany has very much subsided (?). However, the initial idea for the creation of the “Unplagged” project, whose development approach will be described here, can be found in this very case of plagiarism. Related to it were the formation of the [GuttenPlag](#) and its descendant [VroniPlag](#). Both are Wiki-based communities that are collaboratively discovering and collecting plagiarism in their respective cases and are kind of the role models for the way the Unplagged system is developed.

The project idea and context were provided by our professor Dr. Debora Weber-Wulff and the two-term master project, every media informatics student at the [HTW-Berlin](#) has to take. Professor Weber-Wulff is a well known expert in Germany on the topic of plagiarism. As she has also done research in this field for over ten years and is actively involved in the VroniPlag community under her synonym “WiseWoman”(?), she came up with the idea to build a dedicated system — a “Plagiarism Detection Cockpit”(?) — that is modeled after the experiences that were made with the workflow used in VroniPlag and GuttenPlag.

So, to put it in a catchy marketing phrase, here is what Unplagged aims to become:

**Unplagged is a simple, web-based, collaborative system to help
discover, collect and document plagiarism in scientific papers.**

To make things a bit more conceivable, we also often refer to it as a mixture of a very specialised text editor, with a focus on comparing texts and marking passages and a

¹Every plagiarism found in this paper is intended and left for the reader as an exercise ;-)

modern project management tool like [Redmine](#) or [JIRA](#), to manage the collaborative aspects of the system. The big distinction we make to other plagiarism software on the market is, that the approach is not to autodetect plagiarism, but focused on aiding the workflow of the users while searching for plagiarized fragments inside a scientific paper, a homework assignment or any other kind of probable textual plagiarism.

This present document will be the handbook that gets you started if you are interested in helping us with the development of this open source project, which is licensed under the [GNU GPLv3](#).

Chapter Overview

One of the biggest problems we faced at the start was, that none of the team members had written a longer scientific text than a bachelors thesis and therefore the experience we got with actual scientific writing was very limited and very specific to the field of computer science. We understand the ethical problems, that come with the betrayal of good scientific practice of plagiarists, but we simply can not relate easily to the amount of work that has to be put into a PhD., or be as passionate about plagiarism as Prof. Weber-Wulff always is, because we never experienced it ourselves.

That is why we had a lot of catching up to do on the most important history behind VroniPlag, the different types of plagiarism, different citation styles and the research Prof. Weber-Wulff and others had already done on systems that try to help finding plagiarism. Chapter 1, [A Plagiarism Primer](#), will give a brief overview of the most important topics to get you up to speed with the domain of the software, if you are not already familiar with it.

Chapter 2, [Project Workflow and Requirements](#), will be the place, where the development process is described and a collection and description of the parts of the system, that already exist or that we identified as necessary parts of Unplagged will be given.

If you know all those things already and simply want to get started working and coding, you should probably jump to 3, [Developing Unplagged](#). This chapter will give the

technical insights into the system, the basic installation steps and all necessary tools for you as a developer.

And if this still isn't enough reading material for you, in the appendices you can find the meeting minutes, some reports on the logged time in our project management tool and a collection of the mockups of the system.

Conventions

To markup important words in the text, the following typographical conventions are used:

Italic

First used technical terms

Constant Width

Programm code, file names, paths

Bold Constant Width

Variables that have to be changed by the user

1. A Plagiarism Primer

Here we will introduce you to the topics we encountered and researched in the beginning of the Unplugged project.

1.1. Plagiarism definition and history

You will probably have learned as a child, that stealing material goods is something that will get you arrested sooner or later. Somehow many people make a huge difference when it comes to intellectual property, maybe due to the fact that stealing ideas or texts is much easier to do and much harder to catch.

Nonetheless, this is something that would fall under copyright laws and is also a crime, which could get prosecuted.

Here is what ? thinks about plagiarism:

“ Many people think of plagiarism as copying another’s work, or borrowing someone else’s original ideas. But terms like ‘copying’ and ‘borrowing’ can disguise the seriousness of the offense: [...] In other words, plagiarism is an act of fraud. It involves both stealing someone else’s work and lying about it afterward. ”

1.2. Basic Classification of Plagiarisms

In this part, we will try to cover common classifications of plagiarism, but at first we want to figure out the purpose of the classification.

There are many reasons explaining why people plagiarize. Generally those include, that they didn't have the time, energy or the ability to do the work by themselves or that they try to steal other's work on purpose with the hope that others will not recognize it. This kind of plagiarism is done intentionally and named *deliberate plagiarism(?)*.

Another kind is *accidental plagiarism*. This occurs when texts from some sources are copied or rephrased, but no reference is given. The reason is that the writer didn't know that it is a plagiarism because of "carelessness or lack of skill"(?) while writing.

Anyway, it is important to know, that working with the source carelessly may cause plagiarism. Classification of plagiarism is also a good way to help distinguish typical types of plagiarism, so that people are aware and able to reference the sources carefully and to avoid plagiarizing.

Classification of plagiarism is also good for professors or the plagiarism detection community such as VroniPlag, because it gives them a common terminology, which enables them to communicate faster and more efficiently. It also helps them to realize what category and how many percent the text is plagiarized if there is any suspicion, so that statistical data can be created.

But what is a citation, and what kinds of citation are there? Understanding of citations and the way to cite is an important thing in order to avoid and detect plagiarism.

1.2.1. Definition of Citation

"A citation is a credit or reference to another document or source"
(?)

According to this definition, citation is your reference to a source of information, which was generated by someone and should be indicated properly when this source is used in your work.

Based on ?, there are 3 forms of citation:

Direct quotation:

That is an exact word-to-word copy from one source. In this case, in order to cite you have to mark the text with quotation marks and indicate a parenthetical referencing, which includes the book's name and the page number where you found the source.

Paraphrase:

That is your own explanation of someones idea. Many people think it is not a plagiarism because the text is written with your own words. But it is still a plagiarism because basically that is not your idea or your opinion, but the author's himself. In this case of citation, some keywords are still kept in the work. Therefore the parenthetical reference must be given.

Summary:

That is the same as paraphrase, but this kind of citation is likely a summary of the text. In this case you have to give the parenthetical referencing as well.

Now we can see that citation has a lot of forms. You are free to use the source but you have to make sure to cite sources properly or you are applying a plagiarism.

After understanding what citation is and its form, now we can try to classify the plagiarism. The criteria which we choose for classification are based on the VroniPlag's plagiarism categories, because Unplagged is mostly based on the workflow of the VroniPlag community.

By VroniPlag, it does not matter what citation standard system is used in the work, but it is important to know how the citation is documented and if the reference is given properly.

VroniPlag's classification of plagiarism

According to VroniPlag there are the following plagiarism categories. The categories are originally written in German, and the translation or the similar category in English is written right after the german expression.

Komplettplagiat/Copy & Paste:

The name of this category already indicates how the text is plagiarized. The plagiariser just copies from the source exactly word-by-word and does not leave any proper reference to the source intentionally. Some is too lazy so that they just copy existing mistakes, formats etc... without checking, which later can become a proof of plagiarism.

The category is differentiated by the way of conducting a citation. Source is not cited: the original text is completely copied but the source reference is not given intentionally. Source is cited but not completely: reference is given but not correctly.

Verschleierung/Paraphrasing:

The texts from different sources are rephrased and mixed together. The plagiariser tries to hide his stealing by changing some word orders, replacing words with synonyms ... The source is not given with the hope, that the text is considerably generated by the plagiariser himself and therefore the plagiarism will not be detected.

Übersetzungsplagiat/translations:

This kind of plagiarism occurs when the sources in foreign languages are translated. The plagiariser pretend that this translated text is his own work. Sources are oft not given properly. This is a well-liked way in researching because there are many translation tools which can be found easily in Internet such as Babelfish, Google Translate ... and it is not easy to detect the plagiarism with existing plagiarism detection tools.

Alibi-Fußnote/The forgotten Footnote:

in this category, the source is cited but the real location of the source is hidden.

Bauernopfer und. Verschärftes Bauernopfer:

The source reference is embedded in footnote part but it redirects to another text part of the source which has no relation with the plagiarized text.

Other categories

Halbsatzflickerei/The Labor of Laziness:

The plagiariser takes sentences, or just parts of sentence from different sources, tries to reword them and mixes all so that they look fit together. Source reference is not or not correctly given.

Shake & Paste:

The sentence or paragraph from one source is mixed with one from another source. This plagiarism can be detected by changes of writing styles. Source reference is also not given properly.

1.3. How to detect plagiarism

Plagiarism detection means a lot of effort and hard work, because a lot of documents, books and papers must be found, scanned and compared line for line. Often it is really complicated to find available sources in libraries or scientific database systems and plenty of pages must be copied by hand. During the last ten years a lot of software companies came into market to automatize this process and developed algorithms to find plagiarisms automatically. These systems are called "Plagiarism detection systems (PDS) ". There are two different approaches for plagiarism detection systems to identify plagiarism in text documents:

1. Corpus based analysis

means to "compare suspicious documents against a set of potential original documents" PAN'07 to find similar text passages.

2. Intrinsic analysis

”identifies potentially plagiarized passages by analyzing the suspicious document with respect to changes in writing style”. PAN’07

These computer assisted detection systems alone are not appropriate to find all plagiarisms without human judgement.

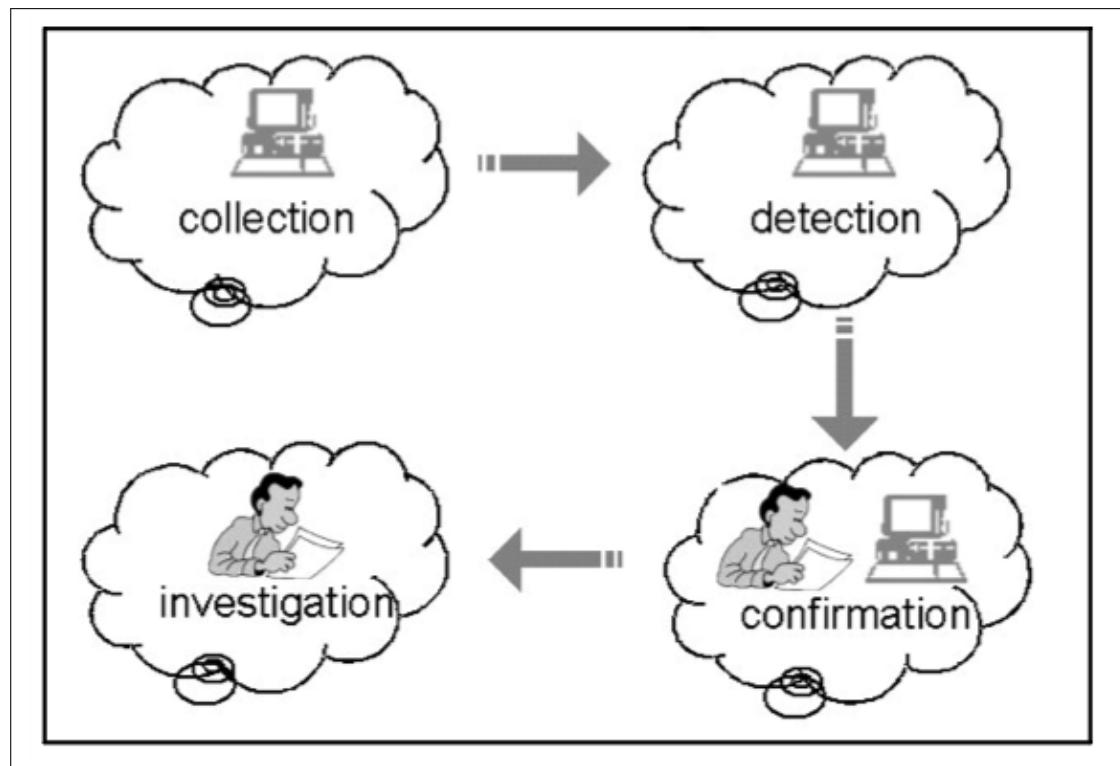


Figure 1.1.: 4-stage plagiarism detection process

We found out that the most practical way is to combine both approaches - the first step is to use the computer-based detection to find similarities between the suspicious documents and the original papers. The second step is to examine the results, validate them and continue the search in a deeper level. (see figure 4 stage)

(Culwin, Fintan; Lancaster, Thomas (2001), ”Plagiarism issues for higher education”, Vine 31 (2): 36-41, doi:10.1108/03055720010804005)

Name	No free access	Free version available	Free demo available	Website
Plagaware		X		www.plagaware.de
Turnitin			X	www.turnitin.com
Ephorus			X	www.ephorus.com
Plagscan	X			www.plagscan.com
Urkund			X	www.urkund.com

Figure 1.2.: Overview of Commercial Detection Systems

1.3.1. Commercial software systems

Several computer companies developed commercial software systems to facilitate the detection of plagiarism. They offer different terms of pricing and online as a service or offline programs. The software system compare digital content from the internet or different types of databases. They seek for similarities, report suspicious parts and try to answer the question if the present text is a plagiarism or not.

This chapter covers parts of the results of the big "Plagiarism Detection System Test 2010". The following benchmarking was done by Prof. Debora Weber-Wulff at the University of Applied Sciences HTW Berlin and her plagiarism team in 2010. The entire benchmarking includes 26 of the 47 available systems on the market and gives an overview of the strengths and weaknesses of these systems in finding plagiarism.

We used the top 5 "partially useful" software systems in this benchmarking to overview the market and find the best usable features for our own system to simplify the daily work of plagiarism finders.

PlagAware

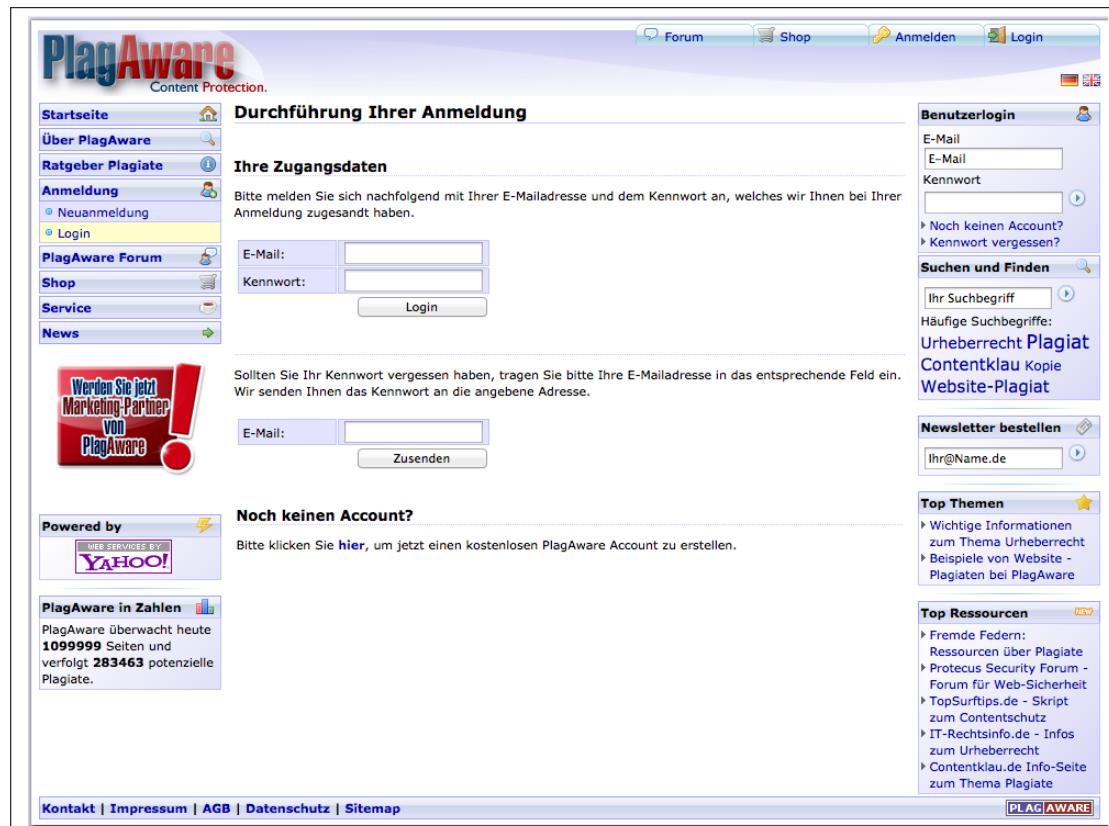


Figure 1.3.: Plagaware Website

Plagaware Business Promotion

"PlagAware is an online-service, which offers services around the topics Searching, finding, analysing and tracing of plagiarisms. The central element of PlagAware is a search engine, which is specialised in detecting identical contents of given texts. Contrary to the plagiarism scanning with classical search engines, the places of finding are not directly transferred to the user, but analysed on the rate and the type of analogy, before a message to the user is written. By this the differing result reports of PlagAware allow to recognise very fast the percentage and the distribution of the copied text contents, thus permitting an efficient and secure rating of a possible plagiarism."

“PlagAware Business Promotion”(?) [PlagAware](#)

PlagAware is software company from Ulm in Germany. The website is online for nearly 5 years and is the top-ranked system in the HTW ”Plagiarism Detection System Test” 2010, but in fact it still detect only 61,11% of the plagiarism cases. Although PlagAware ”produces excellent documentation of the plagiarism found, highlighting the commonalities in a side-by-side presentation. However, its usefulness at university is limited, as each file must be uploaded individually - no ZIP file or student-submission is possible. The system was not designed to be used in a university setting, but rather to find plagiarisms of online texts, which is important for sites trying to optimize their search machine ranking, as plagiarism will contribute to downranking.” [Plagiat Website](#)

Figure Overview shows an overview of all fragments and the results of the plagiarism detection. In figure Side-by-side shows a nice side-by-side fragment view, where all found plagiarisms are shown with different colors. Anomalies in the text are highlighted and the barcode-view is available.

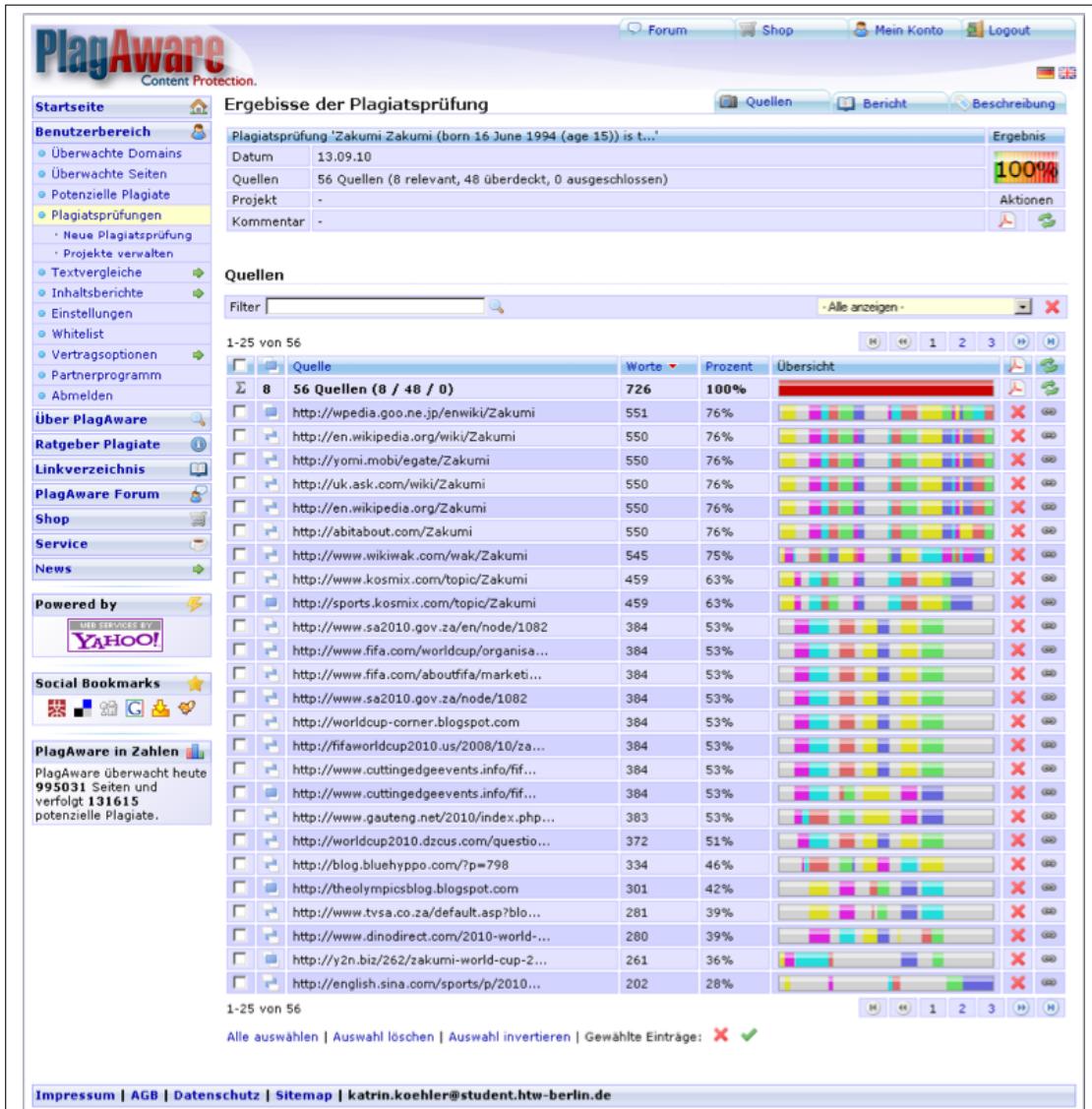


Figure 1.4.: Plagaware overview

PlagAware Costs

PlagAware has four different payment-models:

1. Free

30 scans/month for free. Every additional scan costs 3,0 ct.

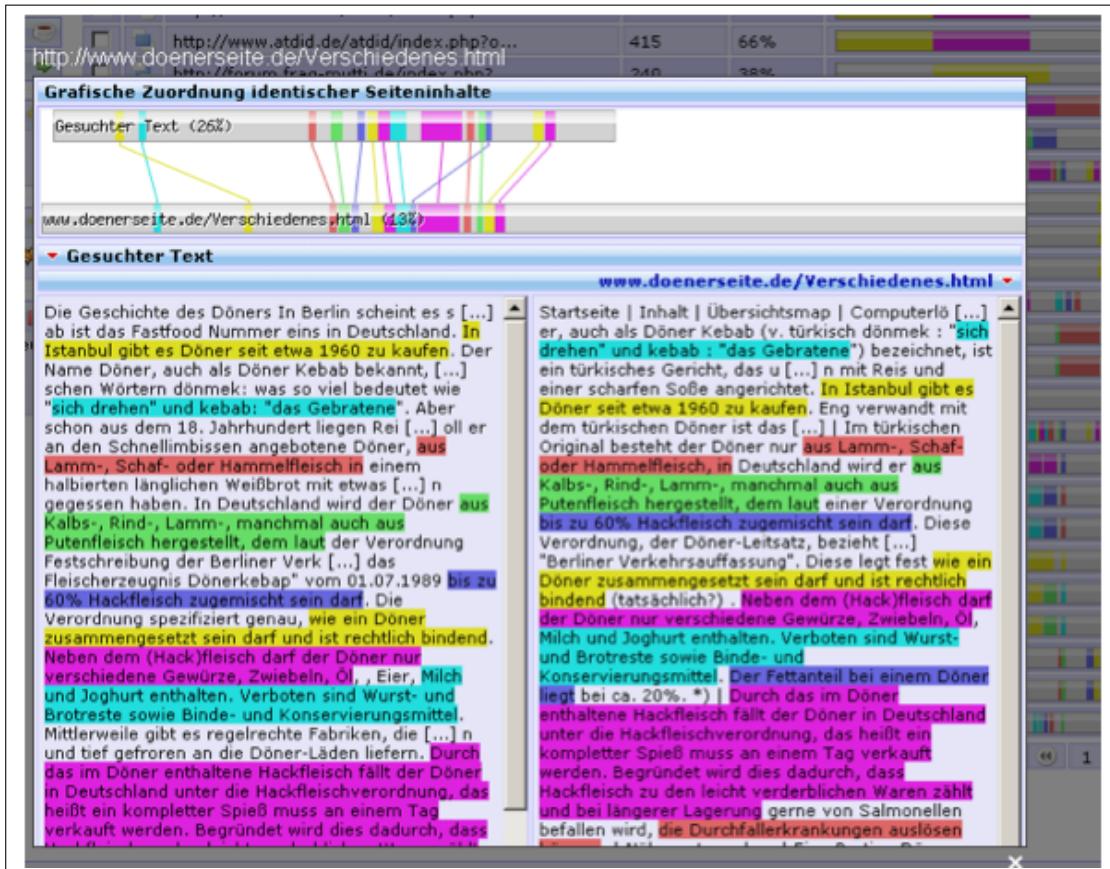


Figure 1.5.: Plagaware side-by-side view

2. Light

EUR 2,99/month. 150 scans/month included. Every additional scan costs 2,0 ct.

Mininum term of 6 month.

3. Standard

EUR 7,49/month. 500 scans/month included. Every additional scan costs 1,5 ct.

Mininum term of 6 month.

4. Premium

EUR 14,99/month. 1500 scans/month included. Every additional scan costs 1ct.

Mininum term of 6 month.

Turnitin

The screenshot shows the Turnitin homepage. At the top, there is a navigation bar with links for 'Startseite', 'Produkte', 'Kunden', 'Partner', 'Support und Schulung', 'Über uns', and a search bar. On the right side of the header are fields for 'E-mail' and 'Passwort' with buttons for 'ANMELDEN', 'Account erstellen', and 'Passwort zurücksetzen'. Below the header, there is a language selector set to 'Deutsch' and a 'suchen...' search input field.

The main content area features a large banner with the text 'Die umfassende Lösung zur Evaluierung von Texten' (The comprehensive solution for evaluating texts). It lists three benefits: '✓ Originalität sicherstellen', '✓ Arbeiten benoten', and '✓ Peer-Reviews erleichtern'. Below this, there are two red buttons: 'ERFAHREN SIE MEHR' and 'FORDERN SIE EIN ANGEBOT AN'. To the right of the banner, there is another section titled 'Plagiatsprävention mit Turnitin' (Plagiarism prevention with Turnitin) for 'Pädagogische Hochschule Zürich' on March 28, 2015, from 12.15 - 17.00 Uhr, with a red 'ANMELDEN' button.

Below the banner, there is a testimonial in German: "Ich habe Stunden damit zugebracht, mithilfe von Google nach ungewöhnlichen Formulierungen zu suchen, wenn ich vermutete habe, dass ein Student eine Arbeit nicht selbst verfasst hat. Jetzt suche ich einfach schnell mit Turnitin!" followed by '- LEHRKRÄFT'.

On the right side, there is a section titled 'Turnitin – Resultate auf einen Blick' (Turnitin – Results at a glance) with a list of statistics:

- über 200 Millionen archivierte Arbeiten
- über 90.000 Magazine und Bücher
- über 1 Million aktive Lehrkräfte
- über 17 Milliarden gecrawlte Webseiten
- 10.000 Bildungseinrichtungen
- über 20 Millionen lizenzierte Studenten
- 126 Länder

Below this, there is a note: 'ENTDECKEN SIE, WAS EINE HALBE MILLION LEHRKRÄFTE SCHON WISSEN' (Discover what half a million teachers already know) and a link to a PDF report: 'Eine kürzlich durchgeführte statistische Analyse von Millionen von im letzten Jahrzehnt bei Turnitin eingereichten Dokumenten beweist, dass sich durch den regelmäßigen Einsatz von Turnitin bei Studenten die Übernahme von Inhalten anderer Verfasser im Laufe der Zeit reduziert.' followed by a 'PDF herunterladen' link.

At the bottom of the page, there are links for 'Startseite', 'Datenschutz', 'Nutzungsrichtlinien', 'Kontakt', and copyright information: 'Copyright © 1998 - 2012 iParadigms, LLC. Alle Rechte vorbehalten.'

Figure 1.6.: Turnitin Website

Business Promotion

"Our award-winning solution discourages plagiarism and facilitates rich, meaningful feedback that improves writing skills, promotes critical thinking, and streamlines grading."

"Turnitin"(?)[Turnitin](#)

Turnitin is a product by a company called iParadigms. It is a well-known US plagiarism software system and one of the most used plagiarism detection systems in the education sector. The website is online for nearly 13 years and the system is at the second position in the HTW "Plagiarism Detection System Test" 2010.

The best results can be achieved with material which is already in the database.

In the past they had a lot of problems to deal with umlauts, and having a complex setup. They improved a lot of parts especially the german translation. Still a big problem for european countries is the copyright policy of Turnitin. They still storing copies of user material in their database without a permission. [Plagiat Website](#)

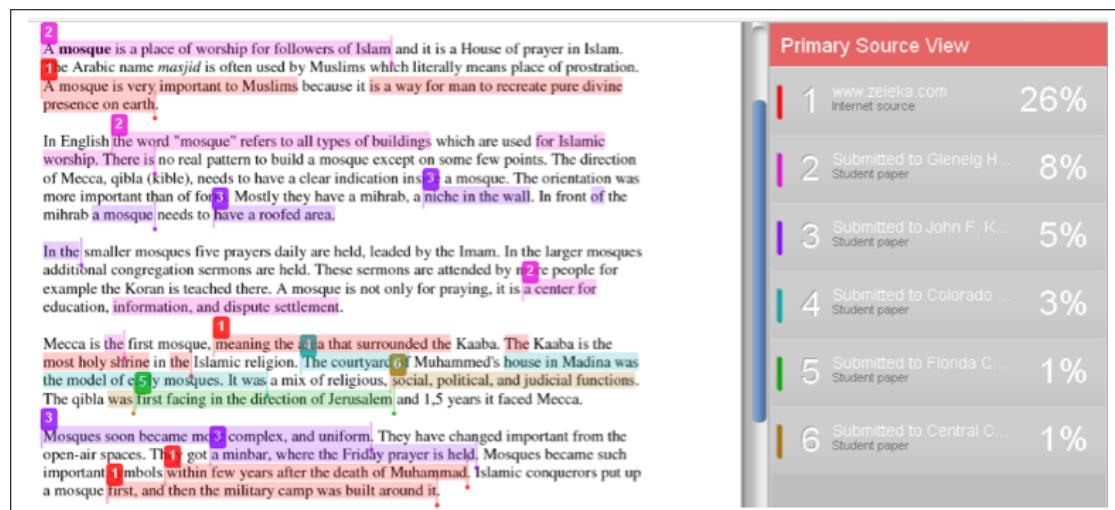


Figure 1.7.: Turnitin - A lot of little matches can't be found, if the sensibility has not been raised.

In 2008 the system was placed at the 13th position. The reason of this change is that the other systems have gotten worse. Turnitin has still problems of flagging spam sites especially when these sites are not safe for work (e.g. site with pornography content). (figure porn) On the other hand the search algorithm of Turnitin is storing sites in their database although they are still not exist. [Plagiat Website](#)

Costs

There are different license models for the education sector and the cost depends on the amount of users .



Figure 1.8.: Turnitin - A lot of spam-sites are reported. Not all sufficient to use at work.
This is one of the harmless examples.

Ephorus



The screenshot shows the homepage of the Ephorus website. At the top, there is a navigation bar with links for "Start", "Demo", "Produkte", "Referenzen", "Support", and "Kontakt". A language selector dropdown is set to "Deutsch". The main content area features a large orange header "ephorus" composed of many small, semi-transparent letters. Below the header is a black and white photograph of a spiral-bound notebook and a pen. The main text on the page reads: "Nie mehr selbst nach Plagiaten suchen? Dem Ärger ein Ende bereiten und dafür qualitativ bessere Arbeiten erhalten? Kein Problem! Mit Ephorus beugen Sie Plagiaten ohne zusätzlichen Aufwand vor! Mehr noch, mit diesem Marktführer im Bereich des Anti-Plagiatismus sind Ihnen beste Ergebnisse zu niedrigen Kosten garantiert. Mit Ephorus unterscheiden Sie die Spreu vom Weizen!" It also mentions a free demonstration by clicking a link. A "Login" button is located at the bottom left of the main content area. At the very bottom of the page, there is a footer with the text "© Ephorus 2012 | Privacy statement | Disclaimer".

Figure 1.9.: Ephorus Website

Business Promotion

"Never search for plagiarism yourself again? An end to all irritations and qualitatively better papers? No problem. With Ephorus, you can prevent plagiarism with no extra effort. Moreover, with this anti-plagiarism market leader, you will be assured of the best service and the lowest prices. With Ephorus, teaching will be fun again! Would you like to try out Ephorus?"

"Ephorus Business Promotion"(?) Ephorus

The third position in the HTW "Plagiarism Detection System Test" 2010 is Euphorus. It's a plagiarism detection system from the netherlands and the website is online for nearly 8 years. 2007 it took the first place in the test, 2008 it was only position 8. Now they redesigned and reorganized the system and old problems were solved. The usability of reports and the whole handling of the system very good. (figure: report) But their still problems with umlauts and the european copyright problematic like in Turnitin. (figure: umlauts)

Eingereichtes Dokument - Download

68% 32-zakumi.pdf - Umbenennen 02-09-2010 | 15:08

Quellenliste

68% Gesamte Übereinstimmungen

51%	<input checked="" type="radio"/> http://worldcup-corner.blogspot.com/
51%	<input type="radio"/> http://www.cbn.co.za/pressoffice/2010/fullstory/1028.htm
51%	<input type="radio"/> http://www.sa2010.gov.za/node/1082
37%	<input type="radio"/> http://www.tvsouthafrica.net/default.asp?blogname=shugesblogiwood&articleid=9369
22%	<input type="radio"/> http://www.southafrica.info/2010/zakumi.htm
16%	<input type="radio"/> http://en.wikipedia.org/wiki/Zakumi

Bericht

Zakumi
Zakumi (born 16 June 1994 (age 15)) is the Official Mascot for the 2010 FIFA World Cup. He is a cheerful and sporty leopard with green hair, presented on 22 September 2008. His name comes from "ZA", the ISO 3166-1 alpha-2 code for South Africa, and "kumi", a word that means ten in several African languages.

Eingereicht:
One thing is for sure, Zakumi will be first on the dancefloor and last off it at the biggest party in the world - the 2010 FIFA World Cup South Africa™. He wants to dance and entertain as many people as he can. He is an animator for fans, players and officials, for schoolchildren, teenagers and big kids alike!

Gefunden:
One thing is for sure, Zakumi will be first on the dancefloor and last off it at the biggest party in the world - the 2010 FIFA World Cup South Africa™. He wants to dance and entertain as many people as he can. He is an animator for fans, players and officials, for schoolchildren, teenagers and big kids alike!

Eingereicht:
Zakumi is a jolly, self-confident, adventurous, spontaneous, and actually quite shrewd little fellow. He loves to perform and always follows his instinct and intuition, yet sometimes has the tendency to exasperate a bit. You will often find him fooling about and teasing

Gefunden:
Zakumi is a jolly, self-confident, adventurous, spontaneous, and actually quite shrewd little fellow. He loves to perform and always follows his instinct and intuition, yet sometimes has the tendency to exasperate a bit. You will often find him fooling about and teasing

Figure 1.10.: Ephorus report - gives a great overview of the results

Ephorus costs

Not stated.

Buchbesprechung (Originaltitel: „The Handmaid's Tale“) In einem fiktiven Staat im Jahr 2195 in Nordamerika haben religiöse Fundamentalisten die totalitäre Republik Gilead errichtet. Nach einer atomaren Verseuchung ist ein großer Teil der weiblichen Bevölkerung unfruchtbar. Frauen werden entmündigt und in drei Gruppen eingeteilt: Ehefrauen von Führungskräften, Dienerinnen und Mägde. Letztere werden zur Fortpflanzung rekrutiert und sollen nach biblischem Vorbild für unfruchtbare Ehefrauen Kinder empfangen. Können sie ihre Aufgabe als Gebärmaschine nicht erfüllen, werden sie in entfernte Kolonien zu gefährlichen Arbeiten wie Giftmüllentsorgung abgeschoben. Desfred, Hauptfigur und Erzählerin, lebt in einer Welt, in der sie nur geduldet wird, weil sie gesund ist und gebären kann. Sonst wäre sie für ihren Widerstand gegen die herrschende Diktatur schon lange erhangt worden. Ihr Glück ist, das gebärfähige Frauen Mangelware sind: die Umwelt ist so verseucht, daß die meisten Menschen unfruchtbar sind. Darum wird sie „Magd“ bei einem altilchen Offiziersehepaar, und ihre einzige Aufgabe ist, bei den rituellen Befruchtungsakten stillzuhalten. Und bald schwanger zu werden, aber das ist nicht so einfach. Desfred wird dem Kommandanten Fred als Zweitfrau in dessen Haushalt zugewiesen, wo regelmäßig die entwürdigende Prozedur des Geschlechtsakts in Gegenwart der Ehefrau durchgeführt wird. Ihr Zimmer darf Desfred nur zu seltenen Einkäufen und zu öffentlichen Hinrichtungen verlassen. Sie hat sich dem totalitären Regime unterworfen, dem sie nur durch Zufall entfliehen kann. Desfreds Tonbandaufzeichnungen schildern in einfachen Worten den Alltag in einer entmenschlichten Gesellschaft mit vollständiger Überwachung und grausamer Unterdrückung. Unterbrochen werden die Beschreibungen von Desfreds Erinnerungen an die alte Zeit: an Mann und Kind, aber auch an die Errichtung des Überwachungsstaats. Wie es dazu kam, ist eigentlich noch interessanter: nach und nach wurden die Rechte der Frauen beschnitten. Erst funktionierten ihre Kreditkarten nicht mehr, dann durften sie nicht mehr arbeiten, usw. Die Parallelen zu Nazideutschland ist subtil, aber vorhanden: auch den Juden beschneit man nach und nach ihre Freiheiten, so langsam, daß jeder Schritt als gerade noch erträglich schien. Eindringlich entsteht das Bild der totalen Repression, das umso bedrückender auf die Leser wirkt, als es eine geringe Distanz zur Gegenwart aufweist: mit alltaglichen Requisiten wie Computer und Kreditkarte, die hier in den Dienst der Republik Gilead gestellt werden. Trotz offenkundiger Parallelen zu 1984 von George R. Orwell ist „Der Report der Magd“ aber weniger als Utopie denn als Stellungnahme zu aktuellen politischen Strukturen und Diktaturen zu verstehen.

Figure 1.11.: Ephorus Problem with umlauts

PlagScan

Business Promotion

PlagScan stands for professionalism

- All documents are treated 100% confidential
- You control whether your document is checked against others, or not
- Integration via API in your existing CMS or learning management system possible

Plagiarism check as easy as pie: PlagScan

- Annotations directly in the document, check without additional work
- No installation - complete functionality in every browser
- All popular formats can be processed

Save time with PlagScan

Figure 1.12.: Plagscan Website

- Check several documents in parallel.
- Fully automated document analysis.
- No use of your resources, all computation is carried out on our servers.

“Plagscan Business Promotion”(?) Plagscan

Plagscan is a software company from Mainz, Germany. It placed at position number 4 in the HTW "Plagiarism Detection System Test" 2010. The website is online for 3 years and in the preview check 2008 it came to the 10th position. As a user you have to buy "Plag Points" (PP). One test costs 1 PP per 100 words. The administrator sets up users and assigns them points for use. There are three different kinds of reports - a list of possible sources with links to click on, the submitted document with the suspicious areas linked to a possible source, and a docx file with the sources in comments. There's no side-by-side presentation, so it's not possible to compare the fragments. Although there are still problems, PlagScan was first place in usability, but only 8th place in overall effectiveness with only 60% of the points awarded for finding plagiarisms.

The screenshot shows the Plagscan dashboard. On the left, a sidebar includes 'Hallo Bernhard Blütentraum!', 'Institution Test HTW Berlin', 'Ihr Guthaben 1913', and links for 'Texteingabe', 'Hilfe anzeigen', 'Admin fragen', and 'Einstellungen'. A 'Plagiat Level' section shows a color scale from green (0-1%) to red (5-100%). The main area displays a table of 10 documents with columns for file name, word count, plagiarism percentage, report link, and analysis date. A note at the bottom says 'Tipp: Analysen können nach Upload automatisch gestartet werden - siehe Einstellungen.'

		Ausgewählte Dokumente	Analysieren	Löschen
<input type="checkbox"/> 18_vikinger.pdf - Die Vinland-Karte, die isla*ndische	879 Worte	Inhalt	0% Bericht>>	2010-08-03 10:36:58
<input type="checkbox"/> 17_squaredance.pdf - Square Dance Ein Square D:	531 Worte	Inhalt	7.2% Bericht>>	2010-08-03 10:36:56
<input type="checkbox"/> 16_jelenik.pdf - Elfriede Jelinkel ist am 20.10.1946 in	468 Worte	Inhalt	50.2% Bericht>>	2010-08-03 10:36:55
<input type="checkbox"/> 15_beduerfnisanstalt.pdf - Bedu*fnisanstalten, im w	721 Worte	Inhalt	30.9% Bericht>>	2010-08-03 10:36:53
<input type="checkbox"/> 14_schmeling.pdf - Max Schmeling war ein deutsche	706 Worte	Inhalt	0% Bericht>>	2010-08-03 10:36:50
<input type="checkbox"/> 13_piment.pdf - Viele von den uns heute bekannten	494 Worte	Inhalt	0% Bericht>>	2010-08-03 10:36:48
<input type="checkbox"/> 12_mikrobrauereien.pdf - Eine Mikrobrauerei, oder i	720 Worte	Inhalt	0% Bericht>>	2010-08-03 10:36:46
<input type="checkbox"/> 11_mankell.pdf - Vor dem Einsetzen des Frostes ist	670 Worte	Inhalt	57.2% Bericht>>	2010-08-03 10:36:44
<input type="checkbox"/> 10_fraktur.pdf - Die Fraktur wird oft als "alte Schrift"	1131 Worte	Inhalt	5.1% Bericht>>	2010-08-03 10:36:42
<input type="checkbox"/> 19_blogs.pdf - Der Begriff Blog geht zurück auf der	1062 Worte	Inhalt	1.2% Bericht>>	2010-08-03 10:36:40

Figure 1.13.: Plagscan report is clear and tidy.

Plagscan costs

PlagScan has four different payment-models without a contract:

1. 9 Euro

500 Plagpoint - 50.000 words - 200 Sites.

Dateiname: 21.doc Datum: 2010-08-03 12:14 Ergebnisse der Plagiarismus Analyse vom 2010-08-03 12:16:42	
127 Treffer von 43 Quellen, von denen 42 Onlinequellen sind. Plagiat Level: 	
Im Dokument anzeigen Docx-Dokument herunterladen	
Zeige besten Treffer pro Quelle - klicken Sie ihn an um alle Treffer anzuzeigen oder auf die URL für die Quelle selbst.	
29 Treffer von ein PlagScan Dokument datiert vom 2010-08-03 10:40	
<p><input type="checkbox"/> Wenige Tage nach den ersten Protesten begannen gewaltsame Ausschreitungen vorwiegend ... Tibeter in Lhasa dann auch in anderen Teilen des Landes die sich sowohl gegen chinesische Zivilisten als auch gegen die staatlichen ... und deren Einrichtungen richteten</p>	
16 Treffer von http://de.wikipedia.org/wiki/Tibetische_Uhrufen_2008	
<p><input type="checkbox"/> Wenige Tage nach den ersten Protesten begannen gewaltsame Ausschreitungen vorwiegend ... Tibeter in Lhasa dann auch in anderen Teilen des Landes die sich sowohl gegen chinesische Zivilisten als auch gegen die staatlichen ... und deren Einrichtungen richteten</p>	
14 Treffer von http://forum.gaming-universe.de/index.php?showtopic=14315&mode=threaded	
<p><input type="checkbox"/> März meldete zu diesem Vorfall die amtliche Nachrichtenagentur Xinhua unter Berufung auf Polizeiquellen beim Einsatz seien vier Demonstranten durch Schüsse verletzt worden</p>	
7 Treffer von http://www.mogelpower.de/forum/thread.php?thread_id=233840	
<p><input type="checkbox"/> Ausländische Webseiten werden in China in der Regel umgehend geblockt wenn sie regierungskritische Inhalte verbreiten</p>	
6 Treffer von http://www.computerwoche.de/nachrichtenarchiv/1858744	
<p><input type="checkbox"/> Wie der heimische Anbieter Baidu unterwerfen sich auch ausländische Unternehmen wie Google einer umstrittenen Selbstzensur in dem sie eigenhändig systemkritische Inhalte aus den Suchergebnissen herausfiltern um in China weiterhin Geschäft machen zu können</p>	
5 Treffer von http://www.adhs-anderswelt.de/index.php?topic=21220.20	

Figure 1.14.: Plagscan reports are not self-explanatory.

2. 19 Euro

1.250 Plagpoints - 125.000 words - 500 Sites.

3. 29 Euro

2.000 Plagpoints - 200.000 words - 800 Sites. .

4. 69 Euro

5.000 Plagpoints - 500.000 words - 2.000 Sites.

Urkund

The diagram illustrates the Urkund system architecture. It shows a central 'Transport' component with 'XML' and 'JSON' options. On the left, a 'Browser' box contains 'Client Objects'. On the right, a 'Server' box contains '.Net Objects'. Arrows labeled 'Serialize' point from the Browser to the Transport, and arrows labeled 'Deserialize' point from the Transport to the Server.

URKUND

HOME | ABOUT URKUND | SERVICES | SUPPORT | STUDENTS | LOGIN | REGION

DEMO

Contact information

Continental Europe

URKUND France
Tel: +33 970 447 884
Fax: +33 1 73 72 91 75
savoirplus@urkund.fr

URKUND España
Tel: +34 902 001 288
Fax: +34 902 001 289
conocer@urkund.es

URKUND Nederland
Tel: +31 (0)10 3 400 666
info@urkund.eu
www.urkund.eu

URKUND Belgium
Tel: +32 (0)2 6118715
info@urkund.be
www.urkund.be

URKUND Deutschland
Tel: +49 69 59 603 605
wissen@urkund.de

URKUND Türkiye
Etkin Proje Yönetim Dan., Tic. Ltd. Şti.
info@urkund.com.tr
Tel: +90 (0)216 325 83 89
Fax: +90 (0)216 340 19 57

URKUND Polska
Tel: +48 510 099 134
support@urkund.pl
www.urkund.pl

Central and South America

Grupo Difusión Científica
Av. Emiliano Zapata No. 285
(Eje 7A Sur)
Tel: +55 5090 2800
+55 5090 5300

Scandinavia / Others

PrioInfo (HQ)
Tel: +46 8 7385200
info@urkund.se
Box 3217
10364 Stockholm, Sweden

URKUND is the leading system for plagiarism control in the Nordic region. It is designed with user friendliness in mind and is designed to help teachers and examiners. The system helps improve the quality of education by providing pedagogical support in the writing process and to save time otherwise spent on verifying the contents of written assignments.

URKUND can be integrated into external systems, allowing users of a Learning Management System to use URKUND as part of an already familiar environment. The new web service (2011) has all the nuts and bolts to make the interaction between our partners and us both more modern and flexible. The new web service uses completely different technology and we have therefore in connection with the revamp also improved major parts of both the flow and the logic. Among other things, we move away from e-mail as a method of transmitting documents in favour of a REST-based protocol. The new integration service is not an SOAP web service. It uses a REST inspired protocol using XML or JSON over HTTP.

URKUND is already integrated with a number of common Learning Management Systems, such as Moodle, Blackboard, Fronter, SharePoint, PingPong, Vklass and others.

Integration with It's Learning is coming up during 2011.

Figure 1.15.: Urkund Website

Business Promotion

"URKUND was born from the academic world. A team of teachers developed the idea of a web based service that would help them detect and deter plagiarism and URKUND was born in the fall of 2000. The problem of plagiarism received much attention in the media and more and more began realise the scope of the problem and the need of a tool to support the pedagogical work. URKUND continued to grow and develop over the years and came to be recognised as Sweden's foremost anti plagiarism service.

Today, URKUND is present in our neighbouring countries and continental Europe as well as the USA, Asia and the Middle East.

URKUND is a natural part of the educational work of the academic world today. Both faculty and students are aware of the immediate and long term benefits of our system."

["Urkund Business Promotion"\(?\) Urkund](#)

Urkund is the last system in our comparison of partially useful systems. It ranked at the 5th position in the HTW "Plagiarism Detection System Test" 2010. The company is from sweden and started their business in 2000. In this test it ranked high in effectiveness but on the other side it's not easy to use. It has problems in the translation and after the redesign 2008 the usability was going worse. Overall "the navigation is confusing, the layout at times catastrophic with texts overlapping fields, the printed reports could be better, the error messages are cryptic, and the link descriptions are unclear." "Plagiat Team HTW"(?)

Urkund costs

Not stated.

Name des Studiengangs	Dok.	Code des Stud.	Ablaufdatum	Den Ordner ändern	
2010 0 [Japanese] Ordner ändern				<input type="button" value="Den Ordner ändern"/>	
HTW/10 43 [OSW] 14:00	2010-08-05 14:00			<input type="button" value="Den Ordner ändern"/>	
Archiv					
Datum:Zeit	Dok Nr.	Name des Dokuments	Absender-Antragsteller	Reg	A
2010-08-05 13:12	D2897850	24.pdf	<input type="checkbox"/> cassis, cornelia	<input type="button" value="Download"/>	<input type="button" value="Attachment"/>
2010-08-05 13:12	D2897851	26.pdf	<input type="checkbox"/> cassis, cornelia	<input type="button" value="Download"/>	<input type="button" value="Attachment"/>
2010-08-05 13:12	D2897852	27.pdf	<input type="checkbox"/> cassis, cornelia	<input type="button" value="Download"/>	<input type="button" value="Attachment"/>
2010-08-05 13:12	D2897853	28.pdf	<input type="checkbox"/> cassis, cornelia	<input type="button" value="Download"/>	<input type="button" value="Attachment"/>
2010-08-05 13:12	D2897854	29.pdf	<input type="checkbox"/> cassis, cornelia	<input type="button" value="Download"/>	<input type="button" value="Attachment"/>
2010-08-05 13:12	D2897855	30.pdf	<input type="checkbox"/> cassis, cornelia	<input type="button" value="Download"/>	<input type="button" value="Attachment"/>
2010-08-05 13:12	D2897856	31-pickles.pdf	<input type="checkbox"/> cassis, cornelia	<input type="button" value="Download"/>	<input type="button" value="Attachment"/>
2010-08-05 13:12	D2897857	32-zakumi.pdf	<input type="checkbox"/> cassis, cornelia	<input type="button" value="Download"/>	<input type="button" value="Attachment"/>
2010-08-05 13:12	D2897858	33-eyjafjallajoekul...pdf	<input type="checkbox"/>	<input type="button" value="Download"/>	<input type="button" value="Attachment"/>
2010-08-05 13:12	D2897859	33-eyjafjallajökul...pdf	<input type="checkbox"/>	<input type="button" value="Download"/>	<input type="button" value="Attachment"/>
2010-08-05 13:12	D2897860	34-stieglersson.pdf	<input type="checkbox"/>	<input type="button" value="Download"/>	<input type="button" value="Attachment"/>
2010-08-05 13:12	D2897861	35-xss.pdf	<input type="checkbox"/>	<input type="button" value="Download"/>	<input type="button" value="Attachment"/>
2010-08-05 13:12	D2897862	36-champagne.pdf	<input type="checkbox"/>	<input type="button" value="Download"/>	<input type="button" value="Attachment"/>
2010-08-05 13:12	D2897863	37-mosque.pdf	<input type="checkbox"/>	<input type="button" value="Download"/>	<input type="button" value="Attachment"/>

zweiter Urkundtest

Attachment: pdf.ZIP (1387433 bytes)

Figure 1.16.: Urkund List view

Resume commercial software systems

Although over the years there are more software detection systems that claim to check text reliable if it's plagiarism or not, but the quality of the systems decreases. A big problem that some of the tested systems offer "ghostwriting".

"In 2008 the systems fared slightly better on the test, but in 2010 many have lost the ability to detect plagiarisms that have been slightly edited - word orders switched, words dropped or added, or synonyms used. The best systems only reached a grade of C-, not quite reaching 70% of the possible points.

The 2010 test included not only short essays in German, but also ones in English and Japanese. Additionally, a usability metric was calculated that took into account aspects

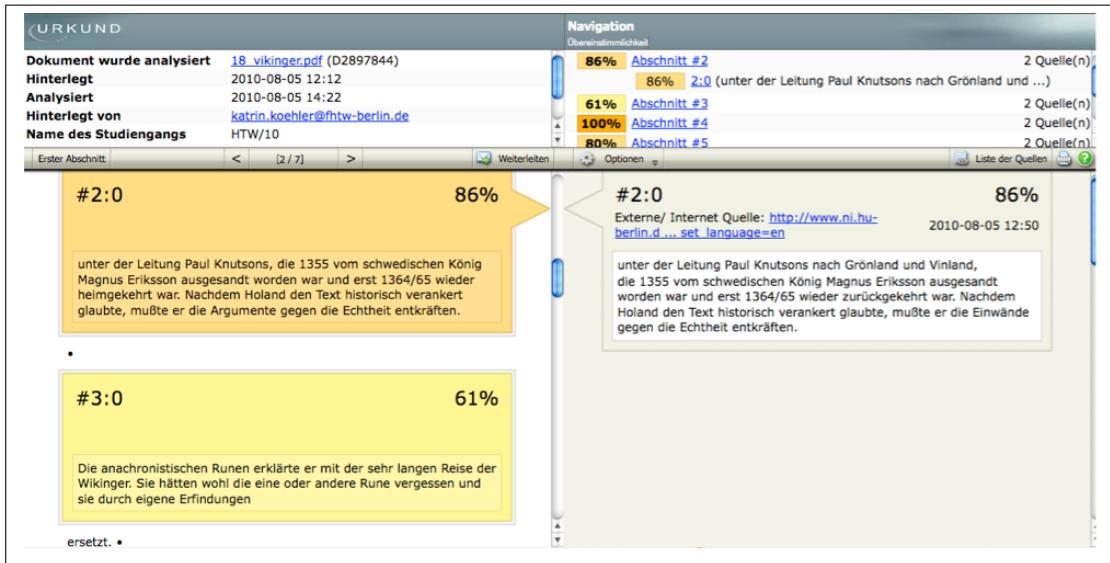


Figure 1.17.: Urkund Report

such as design, language consistency, navigation, print quality of the reports, and how well the system fits into the workflow of a university. A new professionalism metric includes giving a real street address and the name of a contact person; not advertising for paper mills or ghostwriting services; answering the phone during normal business hours and not installing malware on the computer under the guise of installing the detection software.

The systems were categorized as partially useful, barely useful, and useless for university purposes. The best systems with between 60 and 70% effectiveness are PlagAware, Turnitin, Ephorus, PlagScan and Urkund.

Our recommendation: Only use these systems when suspicions of plagiarism arise that cannot be found with 3-5 words in a search machine. The focus should be on teaching students about plagiarism and how to avoid it instead of investing time in using software. Most of the work involved later is in preparing a plagiarism case and dealing with the plagiarist, and for this good documentation is needed. Very few systems provide good documentation.” (plagiats)

1.4. Vroni Plag

VroniPlag is a wiki platform which many volunteers, who are ready to give their free time, their money or their books/resources etc. work on. They collaborate with each other in order to detect plagiarism in dissertations or habilitations.

Das VroniPlag Wiki ist ein am 28. Maerz 2011[2] auf Wikia gegrundetes Wiki, das verschiedene Hochschulschriften - hauptsaechlich Dissertationen - untersucht, die unter Plagiatsverdacht geraten sind. Die Untersuchungen fuehrten in mehreren Faellen zur Aberkennung des Doktorgrades. Der Gruender von VroniPlag ist Martin Heidingsfelder. Das ist nicht genau bekannt, da man sich fuer eine Mitarbeit nicht anmelden muss. Es gab am 18.3.2012 6335 Seiten im Wiki, 189 verschiedene Accounts haben schon mal editiert, 21 Leute sind als Admins eingetragen, und es gibt 3 Buerokraten, die Rechte vergeben koennen.

VroniPlag is named after the first case which they published. The case was the one of Veronica Saß.

Because it is a wiki page, which means open to all, every one could join in. If somebody is interested in a public case, he can feel free to edit the page without asking for an allowance. In the chat portal of this community, one could ask for more help or to take a look at the list of waiting fragments.

Before starting detection, there are some information that a collaborator might know.

1.4.1. Plagiarism detection steps

A suspicious case is called *candidate case*. This case is still not public for all. If a user has detected some suspicious part of a dissertation, he/she may first go to Chat portal to indicate his/her suspicion. He/she has also to give at least one original text source as proof for it. If it is well reasoned that there is an existing plagiarism, he/she must see if there are enough collaborators, who are ready to spend their time to work with.

The candidate case has been given an anonymous name and not public until there is proof, that there is at least 10% of the pages, in which plagiarized texts are found. After that the case will be published with the name of the author.

During the detection process, the candidate case will be divided into smaller fragments. The fragments will be checked carefully if there is a plagiarism found. If there is, a report is created which shows in which fragment the plagiarism is found. The original source is also correspondingly given.

After checking, fragment's state is changed to "to be proofed." That means, this fragment must be checked again by a second inspector. The result will be classified in corresponding category (see [VroniPlag's classification of plagiarism](#)).

After all fragments are checked, an overview of detected plagiarism will be generated. The overview includes the following parts:

Part 1:

The whole page numbers with two different colors. The page number includes also a link redirecting to corresponding fragment.

- Grey: page in which there is no plagiarism found or not checked yet.
- Blue: page with plagiarised texts found. The link goes to the plagiarized text in comparison with the source as well.

Part 2:

The second part is a generated barcode label which performs the percent of plagiarism found in one page with different colors.

Haupttext
001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020
021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040
041 042 043 044 045 046 047 048 049 050 051 052 053 054 055 056 057 058 059 060
061 062 063 064 065 066 067 068 069 070 071 072 073 074 075 076 077 078 079 080
081 082 083 084 085 086 087 088 089 090 091 092 093 094 095 096 097 098 099 100
101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200
201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220
221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240
241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260

Figure 1.18.: Source: <http://de.vroniplag.wikia.com/wiki/Lm>, 19/03/2012, 08:53

- Blue: pages which are not calculated in the dissertation such as Index, Appendix, literature list
- Grey: suspiciously plagiarized
- Black: verified that 100% of the page is plagiarized
- Brown: verified that more than 50% of the page is plagiarized
- Red: verified that more than 75% of the page is plagiarized

If there is more than 10% of the whole pages plagiarized, the case will be public on wiki. Then a report will be sent to the university, where the dissertation is finished.

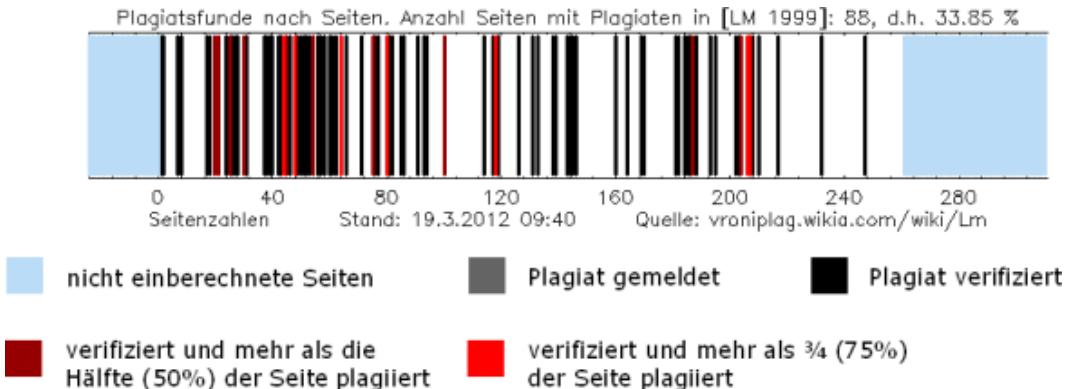


Figure 1.19.: Source: <http://de.vroniplag.wikia.com/wiki/Lm>, 19/03/2012, 08:54

1.4.2. Technical support

Most of the work by VroniPlag is done by hand. For example collaborators could use Google to search for sources, or they borrow books from the library and scan the texts. Generally there is no special software to help detect plagiarism, but collaborator could feel free to choose some of the existing software to help work faster.

1.4.3. Pros & Contras



Figure 1.20.: colors

2. Project Workflow and Requirements

First of all, we've got a confession to make: Unplagged is like a big playground of new workflows and technologies for us, as we are aiming to incorporate "best-practices" wherever possible, or at least what we currently consider to be best-practices.

We believe this approach is necessary, because of the fact, that we are essentially trying to incubate Unplagged as a real open source project and this will only work if it is well crafted and if cutting-edge workflows and technologies are used. Nearly all of the team members are also working in some kind of web related side job, so we all got enough experiences with the problems that can occur during the maintenance of badly designed web applications.

Most of the times this works pretty well, but sometimes we are still trying to figure out how to get everyone up to speed with every technology and part of the system or how to divide the responsibilities carefully.

To start this project, we opted to use *Scrum*¹ as our agile development approach. If you are familiar with this methodology, you may notice, that there could be a few problems when considering, that the team is working mostly distributed without a common office and with very different time tables for each of the members.

We struggled a bit to tweak the workflow that is required by Scrum to fit the situation we faced, but you will see in the following what we came up with.

¹A nice introduction into Scrum is "The Scrum Primer" of the Scrum Alliance: <http://www.scrumalliance.org/resources/339>

2.1. The Workflow

To make it possible to work efficiently together in this kind of environment, we chose to use [Redmine](#) as our project management tool, which you can access under:

- <http://tickets.unplagged.com>

If you register there, an administrator should grant you access to the tickets and the wiki, so that you can participate in solving the problems at hand.

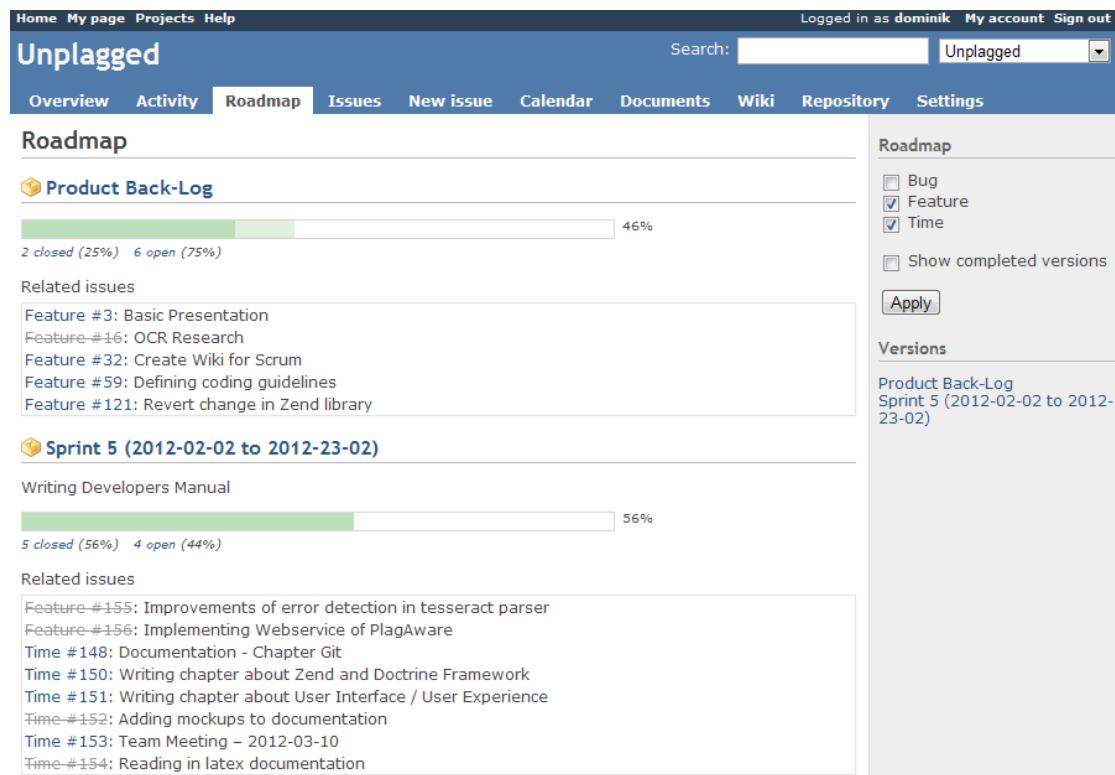


Figure 2.1.: Redmine Roadmap

What we are doing there is to map every *Sprint* and the *Product Backlog* to Redmine's notion of "Version" and every identified *User Story* to an "Issue".

You can see in figure 2.1 the view of the roadmap, with the current sprint 5 designated to create this very document at the bottom and a not very well filled product backlog at

the top.

Normally every identified user story that is not part of the current sprint should be in the product backlog (and we got plenty), but as we were still working in our small group at this point, the hassle of filling in all the tickets seemed unnecessary. This is something that will be fixed in the near future, so that you are able to see where the development is going.

Currently we are working mostly with four week long sprints, to overcome the problem that we are not working fulltime on the tickets, which is something that scrum normally assumes with its two week long sprints.

To have a nice statistical overview and more planning security for the “scrums”, it is required to log the time that was spent on an issue within redmine.

2.1.1. Product Owner — “The Debbie Meetings”



Figure 2.2.: Scrum Meeting

To figure out the user stories we mostly rely on what we internally call “Debbie Meetings”. Normally at the end of every sprint, the members of the team meet with Prof. Weber-Wulff, who we state to be our *Product Owner*. We simply sit down there and talk about

what should be implemented in the next sprint and collect it on cards as you can see in figure 2.3.

We consider this to be just a temporary way of handling this, because we hope that when we eventually have a prototype that we consider to have enough “business value” to be shown to more people, the focus will shift away from Prof. Weber-Wulff to a more broader understanding of the product owner.

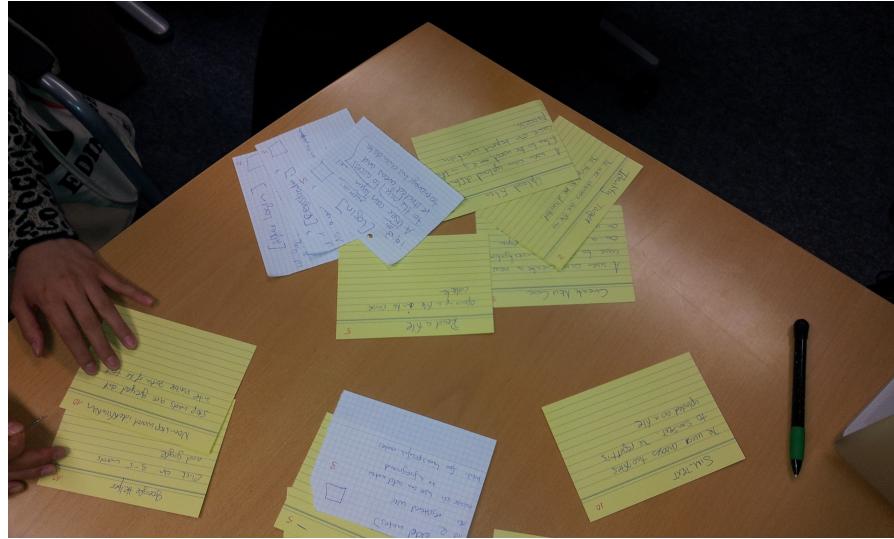


Figure 2.3.: User Stories

This means, that we want to open up our ticket system to directly collect new user stories and gather feedback for example from the VroniPlag group, which can be reached through their IRC channel #VroniPlag on:

- <http://webchat.freenode.net/>

2.1.2. Team Meetings

The Unplugged version of the *Daily Scrum* is currently a weekly meeting. Even though our meetings aren't daily, they are still respecting the other characteristics of a *Daily Scrum*. Each team member says what he did in the last week and what he is going to do the next week.

But those meetings are also the most problematic aspect of our version of Scrum, because the daily frequency of the meetings normally helps to let everybody know more about necessary changes in the logic of previously discussed ideas and makes the process really agile. Sometimes this led to situations where team members developed unnecessary complicated, in a wrong direction or not compatible with each other, which only got noticed a week later.

Our projectteam was doing the meetings every monday for at least one hour. Most of the times we met at the same place, if due to appointment collisions we couldn't manage to meet in person, we met in Skype. So we never used to skip the weekly meeting.

For each meeting that we held someone was responsible for writing the protocol about the meeting. For each protocol a new wiki-page was created in [Redmine](#). To put the protocols into this tool was really helpful for the teamwork, because everybody had access to it and could, if necessary, have a look on the decisions which were made in the past.

During the meetings at least these following points have been discussed:

- What was done last week
- What will be done for the next week
- Time estimates of user stories
- Organizational issues
- External dependency issues
- Problems encountered last week
- Further questions

All the protocols which were made during the phase one are available in the appendix ?? (page number ??).

2.2. Target Group

The main goal of the application is to support the detection of plagiarism. So the target group is composed of people who want to correct documents and be sure that this does not contain any plagiarism.

The following list is showing people who could be interested in using this application:

- Professors
- Teachers
- Academics
- Instructors
- Tutors
- ...

This people are mostly not it-specialists, so it is important to focus on it during the development of the application. This has an influence on the way they will use the application: it must be easy and naturally to use, the results must be easy to understand and reliable.

2.3. User roles

As the Unplagged system will provide a permission based user system, our goal is to make it possible, to create custom user roles from an administration area and make it possible for users to have multiple user roles in one case and also different roles for different cases. The standard roles which will be provided by the system are:

Guest A user without a valid login can only see the parts of cases that are set to be public.

Registered Registered users can get “promoted” to higher roles and contribute to publicly editable cases.

Collaborator Collaborators are registered users who were granted access to a specific case. Collaborators can access and edit these projects.

Case-Manager Case-Managers can set up new cases and manage collaborators for their cases and project versions. They may have the permission to add or remove project members.

Admin An admin owns all permissions, such as user administration or project administration. They also have the ability to block/unblock an existing case.

2.4. Basic functionalities

The following points are showing the basic functionalities of the application.

2.4.1. Register

A person who wants to use the application needs to register first. It gets the ability to do it, as soon as it opens the start-page and hits the “Jetzt anmelden” link.



Figure 2.4.: register1

Below you see the register form which has to be filled.

The screenshot shows a registration form titled "Registrierung". At the top left is a small house icon. The form is divided into sections: "Zugangsdaten" (Access Data) and "Angezeigter Benutzername" (Displayed Username). It contains three input fields: "E-Mail", "Passwort", and "Angezeigter Benutzername", each with a corresponding text input box. Below the fields is a large blue button labeled "Registrieren" (Register).

Figure 2.5.: register1

Every entry here will be saved in the database.

2.4.2. Login

The login functionality is an easy to use form, which can be successfully used from every user which is already registered. With a click of the "Einloggen" link, the login form will be visible.



Figure 2.6.: "Einloggen" Link

The received login datas that are taken from the login form will be compared with the entries in the database, are the entries valid, the user has the right to use further functions in the application.

The screenshot shows a login form titled "Einloggen". It has a header "Zugangsdaten" and two input fields: "Benutzername" and "Passwort". Below the fields is a "Einloggen" button. The background features a blue header bar with a house icon and some blurred text.

Figure 2.7.: login form

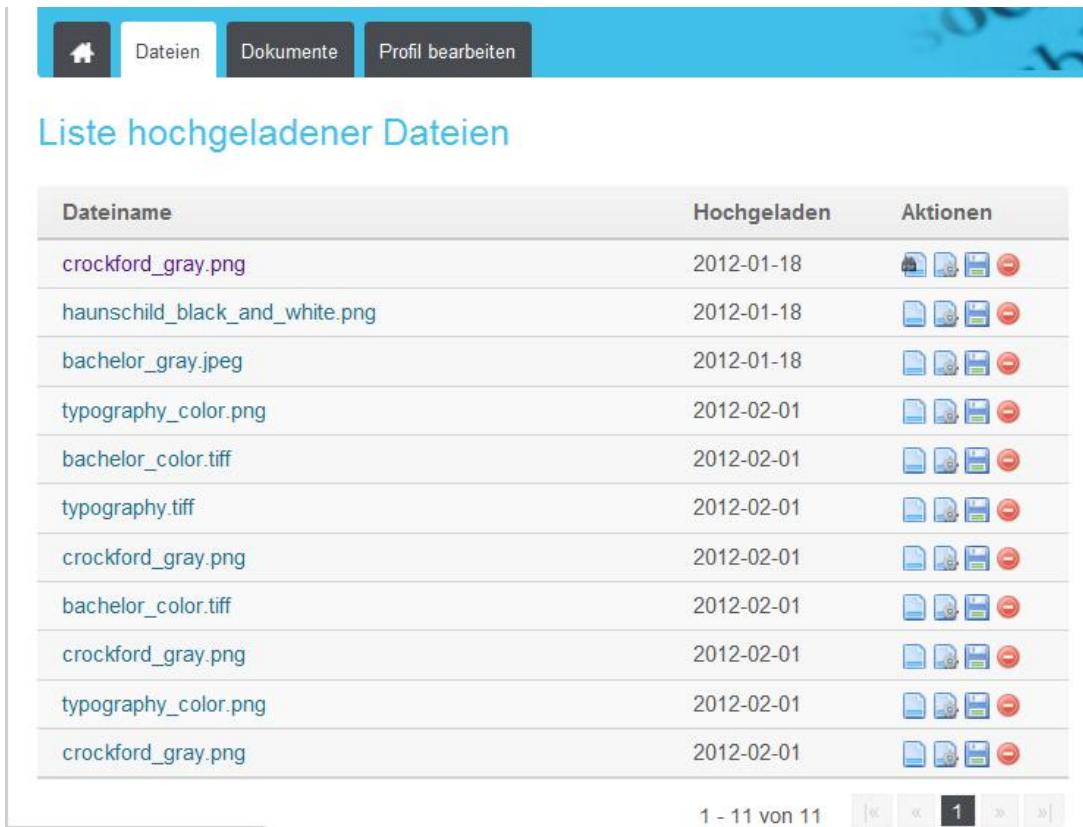
The screenshot shows the Unplagged software interface after login. At the top, there's a navigation bar with a user profile, a search bar, and menu items like "Dateien", "Dokumente", and "Profil bearbeiten". The main area has a blue header with the text "unplagged detect plagiarism". Below the header, a yellow banner says "Willkommen, yupit! Sie wurden erfolgreich eingeloggt.". A text block explains the software's purpose: "to Unplagged. This web-based software offers you a simple, collaborative approach to search and document plagiarism in scientific papers." To the right, there are sections for "Aktuelles und Kontroverses" and "Berichte und Mitteilungen", each listing three news items. A "BETA" badge is visible in the top right corner.

Figure 2.8.: after login view

2.4.3. Files overview

The tab “Dateien” is showing the list of files which were added from an user. Not only shows it an overview it also has several additionaly functions which can be processed for

a file.



The screenshot shows a user interface for managing files. At the top, there is a navigation bar with icons for home, files, documents, and profile editing. Below the navigation bar, the title "Liste hochgeladener Dateien" (List of uploaded files) is displayed. The main content is a table with three columns: "Dateiname" (File name), "Hochgeladen" (Uploaded), and "Aktionen" (Actions). The table lists eleven files, all of which were uploaded on January 18, 2012, or February 1, 2012. Each row contains a set of small icons for file operations like download, edit, and delete, followed by a red circular icon with a minus sign. At the bottom of the table, there is a page navigation bar showing "1 - 11 von 11" and a page number "1".

Dateiname	Hochgeladen	Aktionen
crockford_gray.png	2012-01-18	
haunschild_black_and_white.png	2012-01-18	
bachelor_gray.jpeg	2012-01-18	
typography_color.png	2012-02-01	
bachelor_color.tiff	2012-02-01	
typography.tiff	2012-02-01	
crockford_gray.png	2012-02-01	
bachelor_color.tiff	2012-02-01	
crockford_gray.png	2012-02-01	
typography_color.png	2012-02-01	
crockford_gray.png	2012-02-01	

Figure 2.9.: overview of files

Set target

If the user wants to mark a specific file as his target file, he can do it just with a click on the following symbol.



Figure 2.10.: target function symbol

Afterwards a binoculars icon will be shown.

Dateiname	Hochgeladen	Aktionen
bachelor_gray.jpeg	2012-01-18	   

Figure 2.11.: not marked file

Dateiname	Hochgeladen	Aktionen
bachelor_gray.jpeg	2012-01-18	   

Figure 2.12.: marked target file

Parse

The parse function parses the mentioned file and forwards the parsed file to the “Dokumente” section. Witht a click of the following symbol, the parse function will be fired.



Figure 2.13.: parse symbol

Download

With the disk-symbol the user can download the listed file.



Figure 2.14.: Download symbol

Delete

To delete a file the user only needs to click on the following symbol.



Figure 2.15.: Delete symbol

After the function was fired the path to the file in the database and the file in the related folder will be deleted.

File upload

The file uploader is an important function, it gives the user the opportunity to upload data which he wants to work on it. Also has the user the opportunity to give the file a desired name.

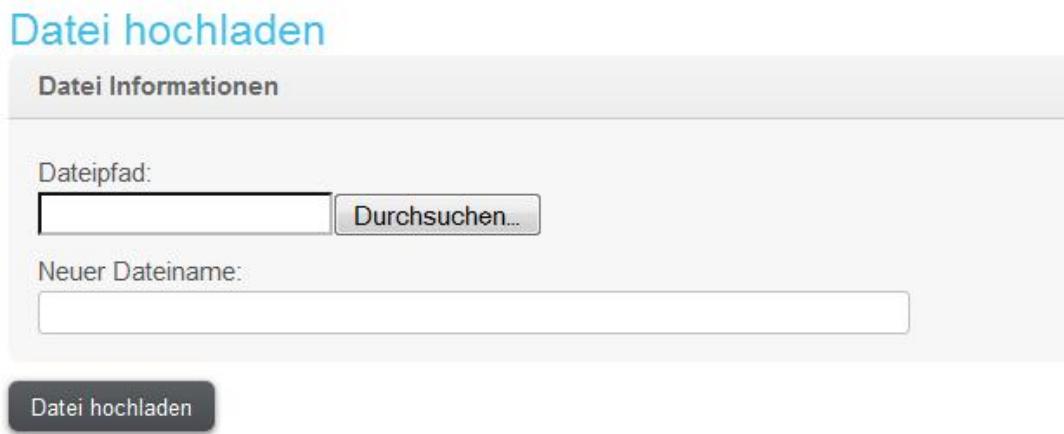


Figure 2.16.: file uploader

2.4.4. Parsed files overview

So after a file was parsed with the parsed function from the “Dateien” section, it will be listed in the “Dokumente” tab.

Integration of detection plagiarism software

The “plag-detection” functionality gives user the possibility to check a parsed file of plagiarism automatically. It uses the application interface from the company “PlagAware”.

This is possible during our coorporation with “PlagAware”. With the following symbol the “plag-detection” function can be used on the file. The detailed result report is not to see on the application. But the user has the possibility to see it on the ”PlagAware” page. The user can get several informations on the application, like the success of the scanning and the procentaged part of the plagiarism.



Figure 2.17.: detect plagiarism symbol

Delete

To delete a parsed file the user only needs to click on the following symbol.



Figure 2.18.: delete symbol

After the function was fired the path to the file in the database and the file in the related folder will be deleted.

Simtext

The simtext function is meant to compare two files with each other. So equal parts in texts can be detected.

A screenshot of a web-based application interface. At the top, there is a navigation bar with four items: a house icon, "Dateien", "Dokumente" (which is highlighted in blue), and "Profil bearbeiten". Below the navigation bar, a yellow status bar displays the message "Text analyse running...". Underneath this, a list of file paths is shown: "/var/www/preview.unplagged.com/application/storage/files/a.txt", "/var/www/preview.unplagged.com/application/storage/files/b.txt", and "/var/www/preview.unplagged.com/application/storage/reports/test_report.txt". At the bottom of the list, a message states "The report was successfully created.".

Figure 2.19.: e.g. simtext running

The simtext function also creates a report, where further detailed information can be seen.

```
test_report - Editor
Datei Bearbeiten Format Ansicht ?
File /var/www/preview.unplugged.com/application/storage/files/a.txt: 268 tokens
File /var/www/preview.unplugged.com/application/storage/files/b.txt: 397 tokens
Total: 665 tokens

/var/www/preview.unplugged.com/application/storage/files/a.txt:
line 3-8 | /var/www/preview.unplugged.com/application/storage/files/b.txt: line 3-8[72]
# $Id: READ.ME,v 2.9 2008/09/23 09:07: | # $Id: READ_ME,v 2.7 2008/09/23 09:07:

These programs test for similar (or eq These programs test for similar (or eq
files and can be used to detect common files and can be used to detect common
checkers are available for C, Java, Pa checkers are available for C, Java, Pa
natural text. natural text.

/var/www/preview.unplugged.com/application/storage/files/a.txt:
line 27-34 | /var/www/preview.unplugged.com/application/storage/files/b.txt: line 44-51[49]
Dick Grune           Dick Grune
Vrije Universiteit   Vrije Universiteit
de Boelelaan 1081    de Boelelaan 1081
1081 HV Amsterdam    1081 HV Amsterdam
the Netherlands       the Netherlands
email: dick@cs.vu.nl  email: dick@cs.vu.nl
ftp://ftp.cs.vu.nl/pub/dick  ftp://ftp.cs.vu.nl/pub/dick
http://www.cs.vu.nl/~dick http://www.cs.vu.nl/~dick

/var/www/preview.unplugged.com/application/storage/files/a.txt:
Line 1-3 | /var/www/preview.unplugged.com/application/storage/files/b.txt: line 1-3[25]
# This file is part of the software si | # This file is part of the software si
# written by Dick Grune, Vrije Univers | # written by Dick Grune, Vrije Univers
# $Id: READ.ME,v 2.9 2008/09/23 09:07: | # $Id: READ_ME,v 2.7 2008/09/23 09:07:
```

Figure 2.20.: Example of a Simtex report

In the report the equal parts on the compared files are shown.

2.4.5. Edit profile

The “Profil bearbeiten” form is a form which gives the user the opportunity to change profile infomations that he gave before. All edit and saved entries will be saved in the database.

2.4.6. Googlesearch

The “Googlesearch” function is content from the contextmenu of the application. The contextmenu only shows if the user marked text on the application. So the user can direclty search for text in google from the application. After the searchwords the user

E-Mail
test@me.com

Benutzername
yupi

Firstname

Lastname

Save

Figure 2.21.: Profile edit form

have to click on “Goolge-Suchwörter löschen”.

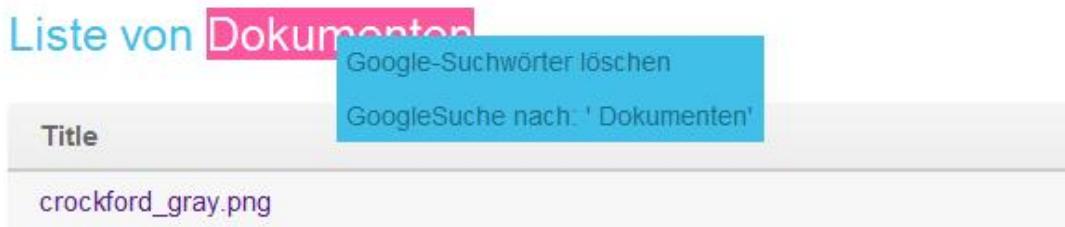


Figure 2.22.: Googlesearch contextmenu

2.5. Document Parser

The document parser is more of a background functionality that enables a user to input data into the system. Essentially it is only an interface, that can be used to implement converters for different types of data, e. g images of scanned texts, PDFs, Word documents or similar, into an internal representation of a document.

Currently this is used when a user uploads an image of scanned text and tries to parse it as described in section [Parsed files overview](#).

2.6. Use Cases

The following UML use-case diagrams are meant to visualize the interaction between the user and the system.

2.6.1. Register

The not registered actor wants to use the application. To use the application he have to first register. Which includes the follow steps:

- User opens the webpage of unplagged.
- User clicks on the "Jetzt anmelden" link.
- User puts needed information in the register form.
- User saves entries.

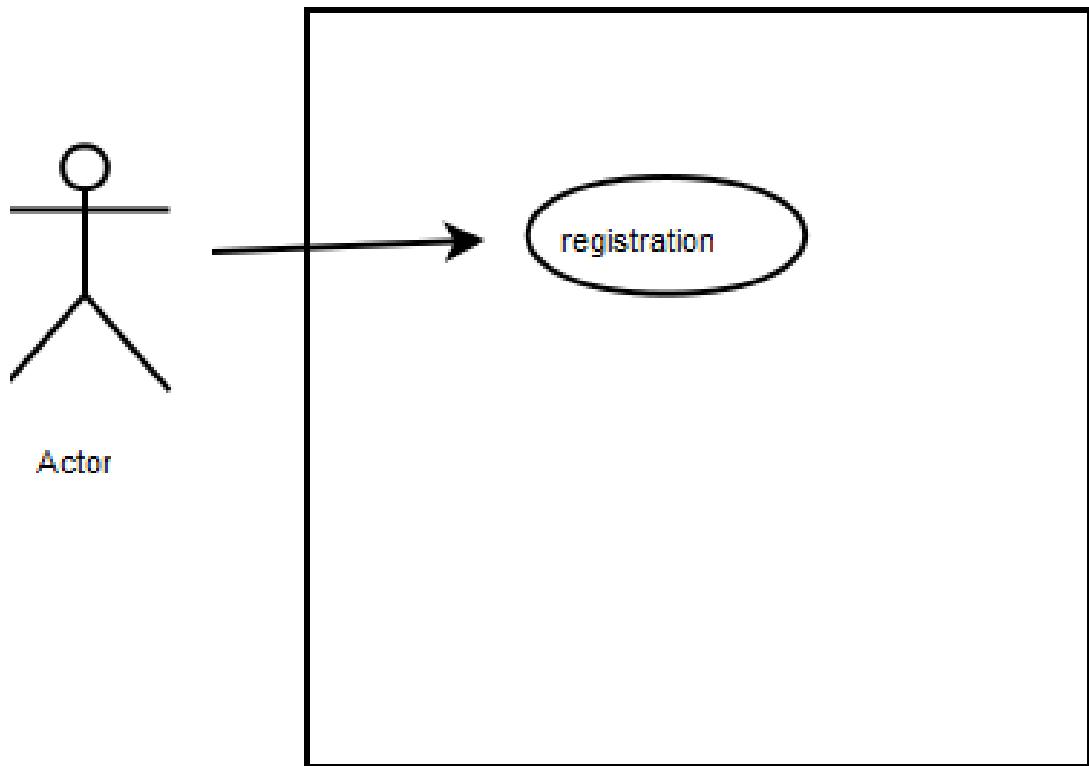


Figure 2.23.: use case register

2.6.2. Login

A registered user wants to work at the application.

Condition: user registered.

- User opens the webpage of unplagged.
- User clicks on the "Einloggen" link.
- User puts needed information in the login form.
- User sends entries.

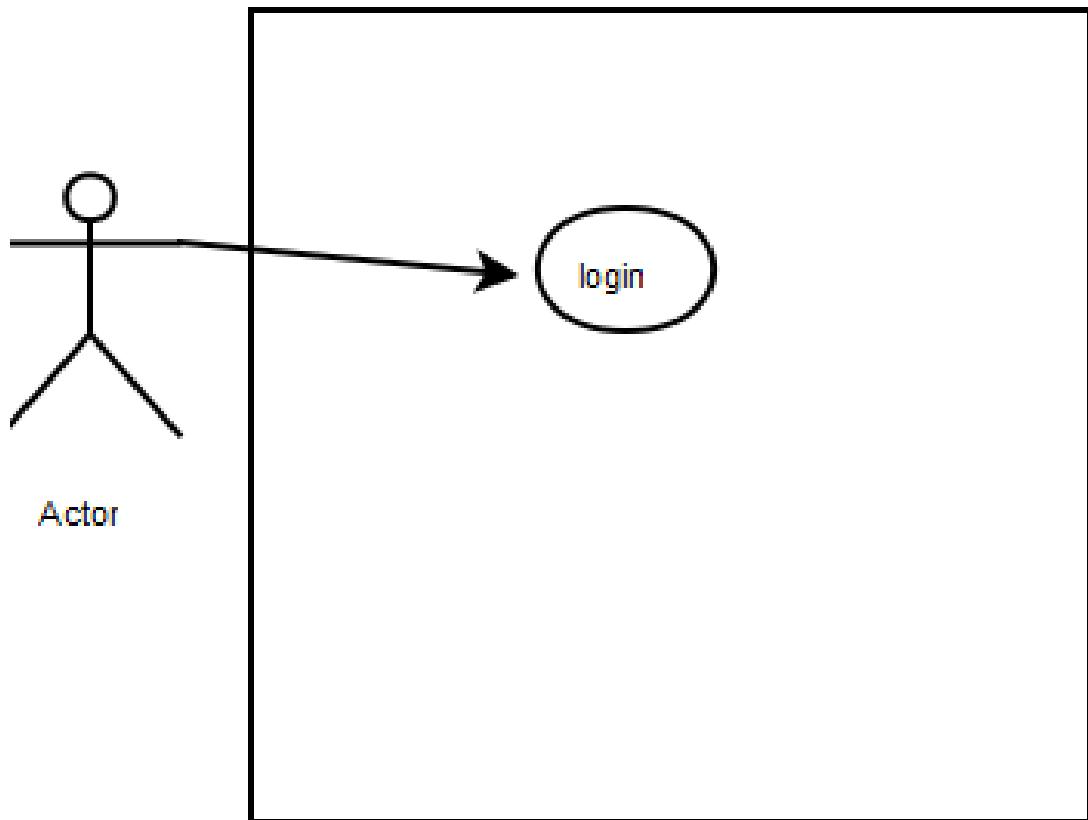


Figure 2.24.: use case login

2.6.3. File Upload

A user wants to upload a file.

Condition: user logged in and open the tab "Dateien".

- User clicks on the button "Datei hochladen".
- User browses for the file which he want to upload.
- (Optional: User gives new name for the uploaded file.)

2.6.4. Parse

A user wants to parse a file.

Condition: user already uploaded the file successfully on the section "Dateien".

- User clicks on the parse-icon button.

2.6.5. Googlesearch

A user wants to google-search a word in the application.

- User clicks on a word or text.
- User makes a right click on the marked part.
- User clicks on the context menu of the "GoogleSuche nach:" link.
- User can see results in extern browser page.

3. Developing Unplagged

Coming from the [Project Workflow and Requirements](#) here we have yet another set of requirements, before we can start with the actual description of the used technologies within the system. This time it's about what we believe will be helpful or sometimes even necessary prerequisites.

First of all, the programming languages mostly used in Unplagged are PHP and JavaScript, both of which in conjunction with a framework. Teaching programming languages is, as you probably can imagine, well beyond the scope of this document, but we will at least try to cover the most important concepts of the frameworks as they occur.

The used frameworks are [jQuery](#) for Javascript and [ZEND](#) for PHP respectively. jQuery is kind of the industry standard for unobtrusive scripting with about 50% of the Top 10.000 websites using it according to [?](#) and the Zend framework is also well established and brings a lot of features, that are useful to this project.

For most of the other topics, we will give you some (hopefully) helpful resources on the way, if it isn't covered thoroughly by us. But just to let you know, here is a list of the buzzwords, err technologies that will be mentioned:

- HTML5 and CSS3
- Continuous Integration
- Responsive Webdesign
- Progressive Enhancement

- Git, Netbeans, Redmine
- Tesseract, Imagemagick, Simtext

As said before, the system is developed, so that it should work on multiple platforms. This makes it sometimes difficult to describe certain installation processes in a way that would work for everybody. As it's often most problematic, to get some Linux software running on Windows, we will mostly concentrate on the way those things are done on this platform and give the instructions for other operating systems as an aside if necessary.

3.1. Development Environment

The following section will mostly focus on the way you can get a development version of Unplagged up and running on your system.

3.1.1. Git

The source code and files of all parts of the Unplagged project are managed through Git, with the repository hosted at [Github](#). Git is a distributed version control system, that exists since 2005 and gained more and more track in recent years. Many developers prefer it over other version control systems, because it is much easier to create different branches and merge them again or simply initialize local repositories. This made it also interesting for us to use it for Unplagged.

However, none of the team members had ever used Git before in a bigger context so it was a challenge to get a common workflow running on all the systems. But we took it, to explore all the features Git offers.

Installing Git Bash

First of all let's find out, how to install the Git console application, called Git Bash. Unfortunately all the GUIs we were evaluating didn't work consistently, so we decided to use it from the command line only. A very good instruction on how to install the Git Bash can be found on the website of the github project:

Windows: <http://help.github.com/win-set-up-git/>

Linux: <http://help.github.com/linux-set-up-git/>

Mac OS X: <http://help.github.com/mac-set-up-git/>

Getting the source code of the unplagged project

Now it is time to get the project source code on your machine. As said before, the whole unplagged project is hosted on github, so if you want to be able to contribute source code later on, you first need to create an account there:

- <https://github.com>

This isn't necessary, if you simply want to look into the source code, which can be accessed via the repository URL:

- <https://github.com/benoertel/unplagged>

If you haven't been granted write access to the above mentioned repository by a project member (which is very likely when you are reading this document for the first time), you will need to do a fork of the Unplagged project right at github, like described in:

- <http://help.github.com/fork-a-repo/>

After this, the following steps are mostly the same for everybody, with the distinction of the project URIs, which should be the one of your newly created fork.

Open up the Git Bash and switch to the directory where you want the project to be located and clone the repository as you can see in listing 3.1.

Listing 3.1: Cloning a repository

```
1 cd Sites/unplagged.local  
2 git clone  
    https://<username>@github.com/benoertel/unplagged.git
```

After this you should have a local copy of all the repository data in the specified directory.

The most important git commands

You are now ready to use Git! Here are some more instructions on the most important commands and how to properly use it. However, if the given instructions in this manual are not enough, feel free to checkout the whole Git manual on:

- <http://schacon.github.com/git/user-manual.html>

The Unplagged project consists of several branches, which are used to develop and store code independently of the other developers. Once a new feature is done, it is merged into the master branch. The master branch usually includes only fully tested and deployable source code.

As a new developer, it is important to create an own branch before doing anything else and switch to it.

Listing 3.2: Creating branches

```
1 git branch mynewfeature  
2 git checkout mynewfeature
```

Now anything in the repository can be changed. At any point changes can be versioned in the repository by using the `git commit` command. If new files were created, `git add` has to be executed as well.

Listing 3.3: Adding all new files and committing

```
1 git add .
2 git commit -m "A message that describes the changes."
```

When the feature is fully working and approved, it has to be merged back to the master branch, in order to get deployed to the staging environment. To do this, the master branch has to be checked out, updated with `git pull` and then all changes have to be merged from the new feature into the master branch. The feature branch can then be removed.

Listing 3.4: Merging branch into master

```
1 git checkout master
2 git pull
3 git merge mynewfeature
4 git branch -d mynewfeature
```

In comparison to Subversion for example, Git has one more step to really write back to the remote source repository. After a `git commit`, a `git push` has to be executed, each push can include multiple commits.

Listing 3.5: Pushing to the server

```
1 git push origin master
```

This is nearly it, the changes to the repository have been pushed to the master branch. The only thing, that probably has to be done now, is to open up a pull request on github, if you developed on your own fork of the project. This means, that you are asking the project members who have access to the “real” Unplugged github account, to integrate your changes into the actual project sources. A nice description of how this process is done can be found at github again:

- <http://help.github.com/send-pull-requests/>

Handling conflicts in merging process

It is possible, if two developers were working on the same part of a file, that a conflict is found during the merge. Such a conflict could look like this:

Listing 3.6: Merge conflict

```

1 CONFLICT (content): Merge conflict in readme.txt
2
3 To https://github.com/benoertel/unplagged.git
4 ! [rejected]          master -> master (non-fast-forward)
5 error: failed to push some refs to 'https://github.com/
       benoertel/unplagged.git'
6 To prevent you from losing history, non-fast-forward updates
   were rejected
7 Merge the remote changes (e.g. 'git pull') before pushing
   again. See the
8 'Note about fast-forwards' section of 'git push --help' for
   details.
9
10 # Unmerged paths:
11 #   (use "git add/rm <file>..." as appropriate to mark
12 #     resolution)
13 #   both modified:      readme.txt
14 #

```

To resolve the issues, open the files listed in the error message, in this case *readme.txt* and decide how the correct version should look like, by removing or changing all the “< < < < < < HEAD” and “> > > > > > b478801d68267ef479acc5ca54544634c52c545c” parts accordingly or using a dedicated merge tool, that is able to show you the changes that were made

Here is an example of how this process would work:

Listing 3.7: Conflicted file

```
1 <<<<< HEAD
2 The goal of this project is the creation of an easy-to-use,
   web-based
3 system to document and detect plagiarism in scientific papers.
4
5 hello world
6 =====
7
8 The goal of this project is the creation of an easy-to-use,
   web-based
9 system to document and detect plagiarism in scientific papers.
10
11 >>>>> b478801d68267ef479acc5ca54544634c52c545c
12 Just a change for educational purposes.
```

It could look like this after merging:

Listing 3.8: Fixed conflict after merging

```
1 The goal of this project is the creation of an easy-to-use,
   web-based
2 system to document and detect plagiarism in scientific papers.
3
4 hello world
5
6 Just a change for educational purposes.
```

3.1.2. Local Deployment

This subsection will describe how to configure a virtual host properly. A virtual host is a domain that is mapped to the local web server. It is assumed that Apache, MySQL and PHP are already running on the machine. If not, here are some tutorial to get them all running:

Windows:

<http://www.apachefriends.org/de/xampp-windows.html#1098>

Mac OS:

<http://www.djangoproject.com/blog/2011/07/24/installation-of-mysql-server-on-mac-os-x-lion/>

<http://www.quarkstar.at/index.php/2009/05/18/webserver-aktivieren-und-konfigurieren-in-mac-os-x/>

Most Linux distributions should already have this kind of server stack installed.

The main goal to make the system run, is to create a local domain and add the virtual host from listing 3.12 to the vhost config.

In Max OS X this can be done via the command line:

Listing 3.9: Mac OS X: Creating virtual host

```
1 sudo vi /private/etc/hosts
2 #add the following line:
3 "127.0.0.1 unplagged.local"
4
5 sudo vi /private/etc/apache2/extr/httpd-vhosts.conf
```

On Windows you need to open up your *hosts* file, which is mostly located in *C:\WINDOWS\system32\drivers\etc\hosts*, and add the following line on the bottom:

Listing 3.10: New host declaration

```
1 127.0.0.1 unplagged.local
```

Now you need to open your apache configuration file *C:\xampp\apache\conf\httpd.conf* and remove the hash symbol(uncomment) from the following line

Listing 3.11: httpd.conf

```
1 #Include conf/extr/httpd-vhosts.conf
```

Add the following configuration to the httpd-vhosts.conf file you just included:

Listing 3.12: Apache configuration

```
1 <VirtualHost *:80>
2   ServerName unplagged.local
3   DocumentRoot "/Users/me/Sites/unplagged.local/public"
4   SetEnv APPLICATION_ENV "development"
5
6   <Directory "/Users/benjamin/Sites/unplagged.local/public">
7     Options +Indexes +FollowSymLinks +ExecCGI
8     DirectoryIndex index.php
9     AllowOverride All
10    Order allow,deny
11    Allow from all
12  </Directory>
13 </VirtualHost>
```

You can tryout your new configuration by entering *unplagged.local* in your browser.

3.1.3. Netbeans

As a software developer who wants to develop object-oriented and use a free software, there is actually not so much software to choose from available. Either Eclipse or Netbeans, we decided to use Netbeans since it has more built-in plugins for PHP and it is faster.

This section will describe how to configure Netbeans in order to start working on the Unplagged project. It is necessary that PHPUnit is already installed on the machine, otherwise not all steps of the Netbeans setup can be done.

Setting up PHPUnit

Mac OS:

<http://kubyshkin.ru/programming/phpunit-on-mac-os-x-snow-leopard-10-6/>

<http://www.newmediacampaigns.com/page/install-pear-phpunit-xdebug-on-macosx-snow-leopard>
<http://www.phpunit.de/manual/3.6/en/code-coverage-analysis.html>

Windows:

For windows there are some more steps to. First of all it is assumed that XAMPP is already installed.

Listing 3.13: Setting up PHPUnit on Windows

```
1 # download http://pear.php.net/go-pear.phar and save it to
2 C:\xampp\php
3 # open a command prompt and go to
4 C:/xampp/php
5 # type the following commands
6 php go-pear.phar
7 pear update-channels
8 pear upgrade --alldeps
9 pear channel-discover components.ez.no
10 pear channel-discover pear.symfony-project.com
11 pear channel-discover pear.phpunit.de
12 pear install --alldeps phpunit/PHPUnit
```

After the installation process is finished, go to the preferences in Netbeans, choose the PHP tab and then the tab "Unit Testing". First try to click the search button, if the installation of PHPUnit is not found, add the following:

Mac OS:

/usr/bin/phpunit

Windows:

C:\xampp\php\phpunit.bat

Configuring Netbeans

Download the latest Netbeans release for PHP from the Netbeans webpage:

<http://netbeans.org/downloads/>

Listing 3.14: Setting up Netbeans project

- 1 Install it and go to File > New Project
- 2 Choose PHP > PHP Application with Existing Sources
- 3 Set "Sources Folder" to the directory, where the **git** repository is located.
- 4 "Project Name" can e.g. be "unplugged.com"
- 5 "PHP Version" should be "5.3"
- 6 "Default Encoding" "UTF-8"
- 7 IMPORTANT: Check the Checkbox "Put NetBeans metadata into a separate directory" and choose another directory than the **git** repository.

Now the project is created and coding can start.

Documentation

Unplagged uses **?** for the generation of a HTML page of all the source code documentation comments, because of it's superior and much more beautiful user interface in comparison to the older **?**.

Sadly the automatic generation is not yet supported by Netbeans, but as it will be soon(?), we are currently only generating this server-side as described in section [3.1.5](#).

This section will be enhanced, when the Netbeans user interface becomes available. If you are interested, you could install the software for yourself and use it over the command line.

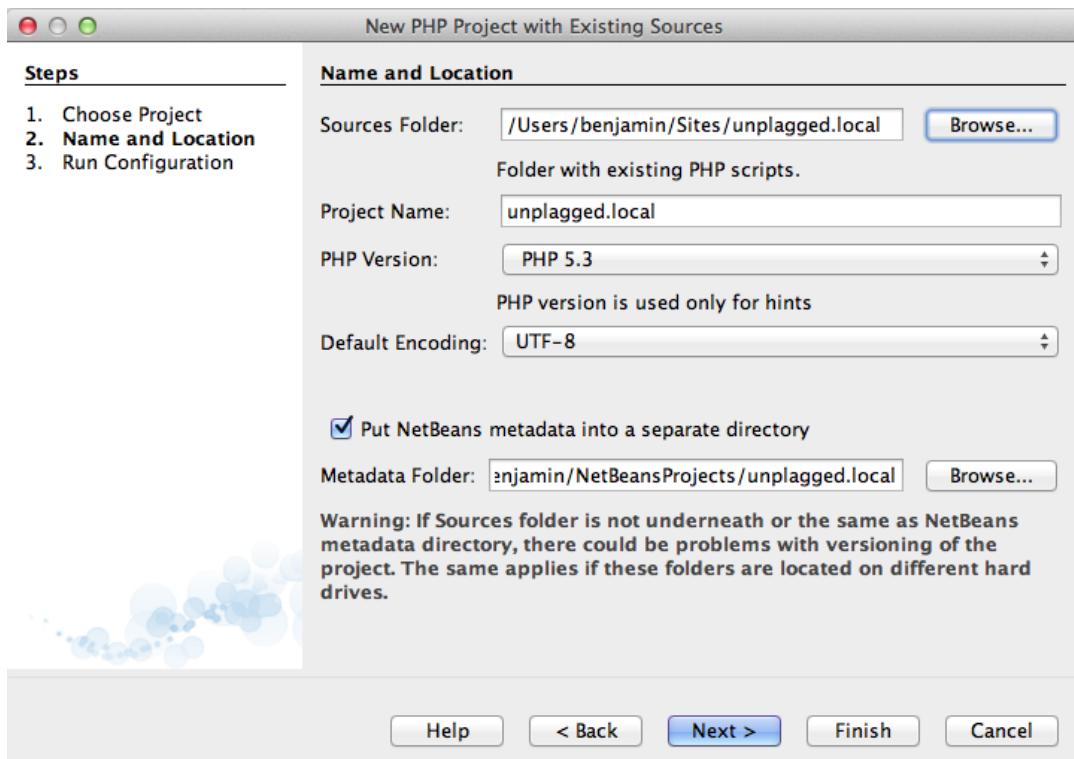


Figure 3.1.: Netbeans – New project – Name and Location

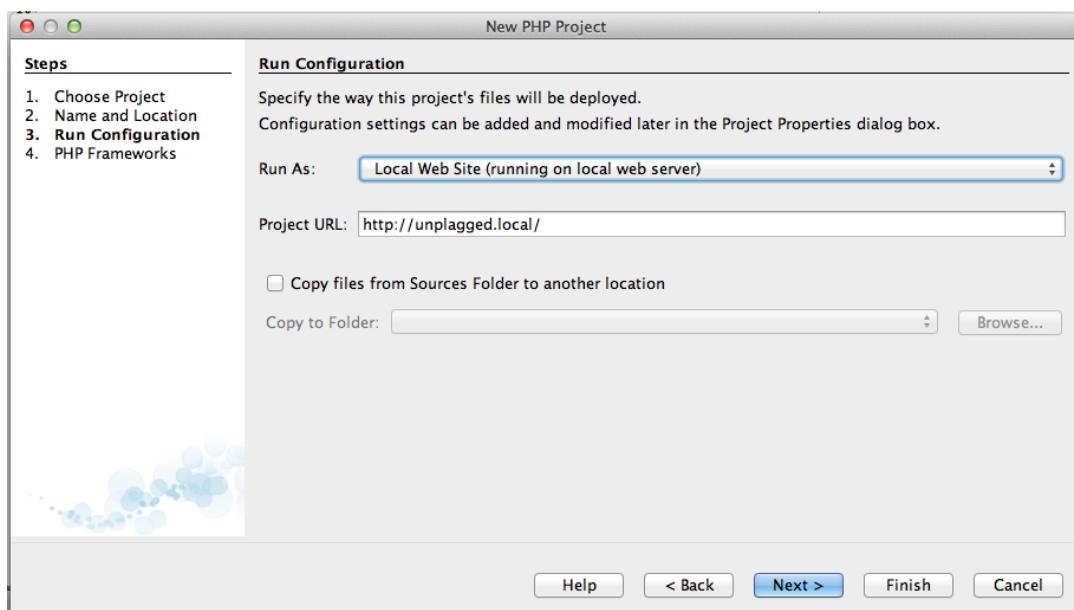


Figure 3.2.: Netbeans – New project – Run Configuration

3.1.4. Additional Software

As we currently have no installer or script that checks for installed software, you still have to install some additional dependencies to make some parts of the system work. Those are mainly command line tools that we use for the optical character recognition or text comparison.

Most of the times the software wouldn't break completely if those dependencies were not installed, but some parts would silently fail, which is of course one of the more annoying problems to debug.

Tesseract

Tesseract is an open source OCR¹ software, that is currently used to "figure out" texts from scanned images a user can upload.

The big idea is to build the system in a way that enables users to plugin their favourite OCR system, but to have one default system that produces satisfying results in Tesseract.

You can download the latest installation files from:

- <http://code.google.com/p/tesseract-ocr/downloads/list>

After installing Tesseract, the easiest way to make it work with Unplagged, is to put it on your systems execution path, so that it could be called from the command line in any directory.

If for some reason you don't want to do this, you can also change the following line in /application/configs/application.ini to the path where you installed Tesseract to:

Listing 3.15: Tesseract executable path

¹Optical character recognition

```
1 parser.tesseractPath = 'tesseract'
```

You should be able to click the “parse” icon on files now:

The screenshot shows a web-based file management interface. At the top, there is a navigation bar with icons for home, dateien, dokumente, suche, and profil bearbeiten. Below the navigation bar, the title "Liste hochgeladener Dateien" is displayed. A table lists ten files with columns for Dateiname, Hochgeladen, and Aktionen. The "Aktionen" column contains icons for download, edit, save, and delete. The file "crockford_gray.png" from February 1, 2012, has its "parse" icon highlighted with a red box.

Dateiname	Hochgeladen	Aktionen
crockford_gray.png	2012-01-18	
haunschild_black_and_white.png	2012-01-18	
bachelor_gray.jpeg	2012-01-18	
typography_color.png	2012-02-01	
bachelor_color.tiff	2012-02-01	
typography.tiff	2012-02-01	
crockford_gray.png	2012-02-01	
bachelor_color.tiff	2012-02-01	
crockford_gray.png	2012-02-01	
typography_color.png	2012-02-01	
crockford_gray.png	2012-02-01	

Figure 3.3.: Parsing Files with Tesseract

Imagemagick

Because Tesseract and probably other OCR systems that are provided via a plugin won't work with every image format a user decides to upload, we integrated Imagemagick as a tool to convert images from one format to another. To install it you can simply follow the installation instructions provided here:

- <http://imagemagick.org/script/binary-releases.php>

Similar as with Tesseract, you can change the ini directive `parser.imagemagickPath`, if you chose not to include the executable in your path.

If you installed it properly, you can try to upload an image file other than `.tiff` and parse it with Tesseract, which should work now.

Simtext

Simtext is a text comparison tool, that is able to output nice data of the differences between texts. In figure 3.4 for example, you can see the default side-by-side comparison and in figure 3.5, the output in `diff` format is shown.

```

F:\projects\unplagged\unplagged\library\SIM\bin>sim_c ./READ.ME ../READ_ME
File ./READ.ME: 268 tokens
File ../READ_ME: 397 tokens
Total: 665 tokens

./READ.ME: line 3-8          | ./READ_ME: line 3-8      [72]
# $Id: READ.ME,v 2.9 2008/09/23 09:07: | # $Id: READ_ME,v 2.7 2008/09/23 09:07:
These programs test for similar (or eq | These programs test for similar (or eq
files and can be used to detect common | files and can be used to detect common
Checkers are available for C, Java, Pa | Checkers are available for C, Java, Pa
natural text.                         | natural text.

./READ.ME: line 27-34           | ./READ_ME: line 44-51      [49]
Dick Grune                    | Dick Grune
Vrije Universiteit            | Vrije Universiteit
de Boelelaan 1081             | de Boelelaan 1081
1081 HV Amsterdam             | 1081 HV Amsterdam
the Netherlands                | the Netherlands
email: dick@cs.vu.nl          | email: dick@cs.vu.nl
ftp://ftp.cs.vu.nl/pub/dick   | ftp://ftp.cs.vu.nl/pub/dick
http://www.cs.vu.nl/~dick     | http://www.cs.vu.nl/~dick

./READ.ME: line 1-3            | ./READ_ME: line 1-3      [25]
# This file is part of the software si | # This file is part of the software si
# Written by Dick Grune, Vrije Univers | # Written by Dick Grune, Vrije Univers
# $Id: READ.ME,v 2.9 2008/09/23 09:07: | # $Id: READ_ME,v 2.7 2008/09/23 09:07:

F:\projects\unplagged\unplagged\library\SIM\bin>
  
```

Figure 3.4.: Default Simtext output

```
F:\projects\unplugged\unplugged\library\SIM\bin>sim_c -d ../READ.ME ../READ_ME  
File ..\READ.ME: 268 tokens  
File ..\READ_ME: 397 tokens  
Total: 665 tokens  
  
..\READ.ME: line 3-8  
..\READ_ME: line 3-8  
<# $Id: READ.ME,v 2.9 2008/09/23 09:07:10 dick Exp $  
<  
<These programs test for similar (or equal) stretches in one or more program  
<files and can be used to detect common code or plagiarism. See SIM.DOC.  
<Checkers are available for C, Java, Pascal, Modula-2, Lisp, Miranda and  
<natural text.  
---  
># $Id: READ_ME,v 2.7 2008/09/23 09:07:10 dick Exp $  
>  
>These programs test for similar (or equal) stretches in one or more program  
>files and can be used to detect common code or plagiarism. See sim.i.  
>Checkers are available for C, Java, Pascal, Modula-2, Lisp, Miranda and  
>natural text.  
  
..\READ.ME: line 27-34  
..\READ_ME: line 44-51  
<  
<  
<  
<  
<  
<  
<  
<  
<  
---  
>  
>  
Dick Grune  
Vrije Universiteit  
de Boelelaan 1081  
1081 HV Amsterdam  
the Netherlands  
email: dick@cs.vu.nl  
ftp://ftp.cs.vu.nl/pub/dick  
http://www.cs.vu.nl/~dick  
  
Dick Grune  
Vrije Universiteit
```

Figure 3.5.: Simtext output in diff format

Installing Simtext can sadly be a bit tricky, because only the sources and no executables(at least none that worked for us) are distributed. We already provided a Windows 64bit EXE and an executable that runs on our Ubuntu staging environment in `/library/SIM/bin`, but if you use any other environment, you will have to compile the C sources of the system for yourself. The sources can be found here:

- http://dickgrune.com/Programs/similarity_tester/

We assume, that Linux users will be familiar with the installation steps that are described in the readme file. The only thing that can be easily overlooked there is the necessity to install *Flex* first. Windows however needs some special treatment to make the compilation work:

1. Download and install “Make for Windows” from <http://gnuwin32.sourceforge.net/packages/make.htm> and “Flex for Windows” from <http://gnuwin32.sourceforge.net/packages/flex.htm>

2. Add the folder of the binaries to the path (should be something like C:\Program Files (x86) \GnuWin32\bin)
3. Rename “flex.exe” to “lex.exe” in the above mentioned directory
4. Unzip and open Simtext directory on the command line and type:

Listing 3.16: Installing and checking simtext

```
1 >make
2 >sim_c --help
3 >sim_c READ.ME READ_ME
```

And again, to make it work you can set the ini directive simtext.simtextPath to the appropriate path or simply include it in your systems path variable.

3.1.5. Continuous Integration

To always have a running version of the latest code, we use an automated workflow, that always deploys everything that has been pushed to the Github repository on the Unplugged staging server. The machine this is done with, is a simple Ubuntu web server, that is also used for hosting our collaboration tools and the webpage.

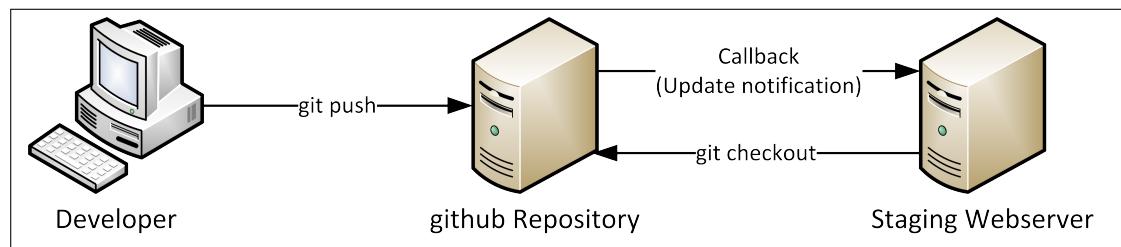


Figure 3.6.: Deployment workflow

As you can see in figure 3.6 the mechanism used for this is a callback, the *post-receive hook* of git, which github employs to let its users enter a *post-receive URL* to call a URL after someone has pushed to the repository. The URL that gets called is located on our staging server and gets answered by a Redmine plugin called “redmine_github_hook”,

that would normally only call a checkout on the server-side repository, so that the newest sources can be seen via the Redmine web frontend.

Listing 3.17: Changes to redmine.github_hook.rb

```
1 # Fetches updates from the remote repository
2 def update_repository(repository)
3   command = git_command('fetch origin', repository)
4   if exec(command)
5     command = git_command("fetch origin '+refs/heads/*:refs/
6       heads/*'", repository)
7     exec(command)
8   end
9   #custom checkout to preview area
10  system('sh /usr/local/etc/scripts/buildUnplaggedPreview.sh
11    ')
12 end
13 end
```

However, we tweaked the source code of this plugin slightly, as you can see on line 9 in listing 3.17 so that it also calls the below bash script(listing 3.18) to initiate the deployment process.

Listing 3.18: Deployment script

```
1#!/bin/bash
2
3 cd /var/git/unplagged.git/
4 GIT_WORK_TREE=/var/www/preview.unplagged.com git checkout -f
5
6 cd /var/www/preview.unplagged.com
7 #generate phpdoc
8 apigen -s application/ -s library/Unplagged/ -d docs/phpdoc --
9   title "Unplagged Documentation" --todo yes
10
11 #run database build scripts
12 cd scripts/build
13 php initdirectories.php
14 php doctrine_staging.php
```

```
15 cd /var/www  
16 chown www-data:www-data preview.unplagged.com
```

The bash script is then used to do a “clean checkout”(without hidden .git folders) of the repository and to run “Apigen”, an engine to process the PHPDoc comments inside the project. Those two things can be accessed by the already prepared vHosts on the server:

- <http://preview.unplagged.com/>
- <http://phpdoc.unplagged.com/>

If you would like to get access to the preview areas, you need to obtain the password and username from a team member.

Possible Improvements

The above described workflow is, as we believe, already on a good way, but it still has a lot of room for improvement. First of all, it would be nice to only let the deployment go through, if the unit tests ran successfully on the server and to have some sort of email notification mechanism if this wasn’t the case.

Another improvement would also be to have a separation into a staging environment with the newest commits and an actual preview environment, that can be deployed to a known stable state/commit of the system in a simple manner.

3.2. User Interface / User Experience

Here we will explain the progress and development of the user interface of Unplagged. As we tried to follow the typical project workflow, we first drew a lot of mockups in the beginning, which represented the main features of Unplagged. At first the wireframes, or

also called mockups were drawn by hand, before we digitalized them. As an example the mockups of the 'new case' page are shown below in figure 3.7. All the other mockups can be found in the appendix C.

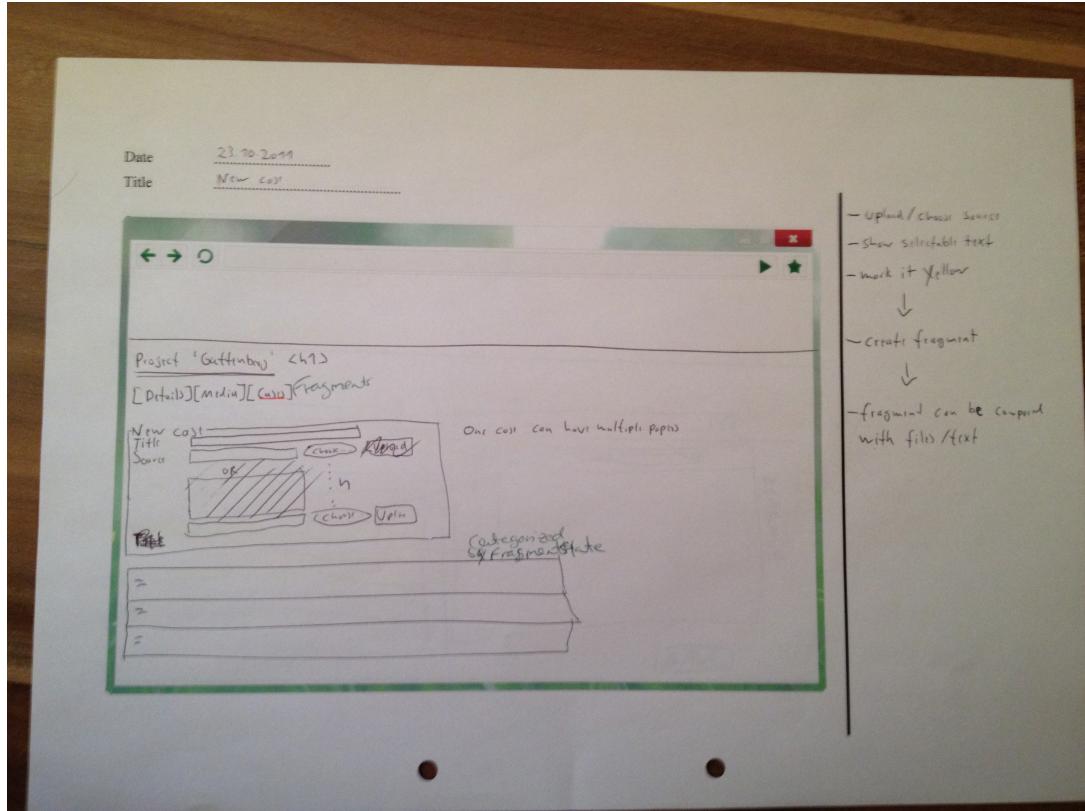


Figure 3.7.: Mockup – New case – hand-drawn

After we had a basic idea how the main interface should be structured, we created a first screen in Photoshop. Therefore we got several helpful suggestions from the website PremiumPixels:

- <http://www.premiumpixels.com/>.

The next step, before the HTML template got created, we defined the key features, our user interface should take care of:

- Responsive Layout – optimized layouts for different devices

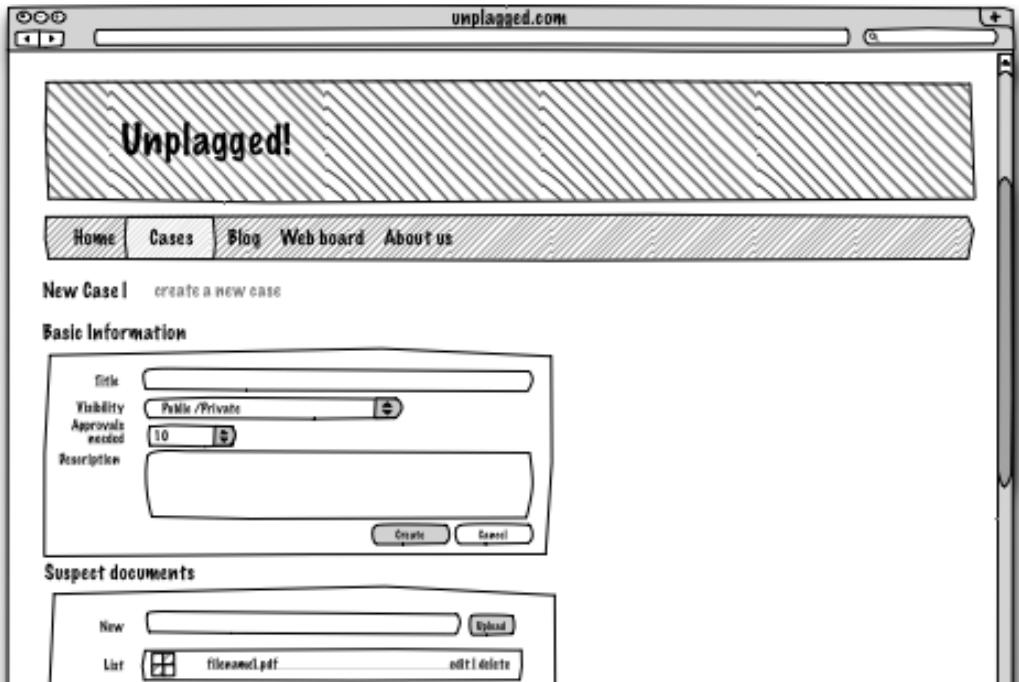


Figure 3.8.: Mockup – New case – digitalized

- Cross-Browser-Compatibility
- Light-weight and w3c-conform HTML5 Markup
- Progressive enhancement with CSS3 – CSS instead of images where possible

3.2.1. Responsive Layout using CSS3 Media Queries

Since the worldwide amount of different mobile and desktop devices is growing very fast, Ethan Marcotte coined the term *Responsive Webdesign*(?) for websites that are not optimized for any devices at all, but simply for different screen resolutions. And this is what we do.

Some functionality as uploading a file, doesn't work on iOS at all, but at least all functions that are working on mobile devices, should work. So the goal is to create a user interface that uses the same markup, but displays differently according to the device it is viewed

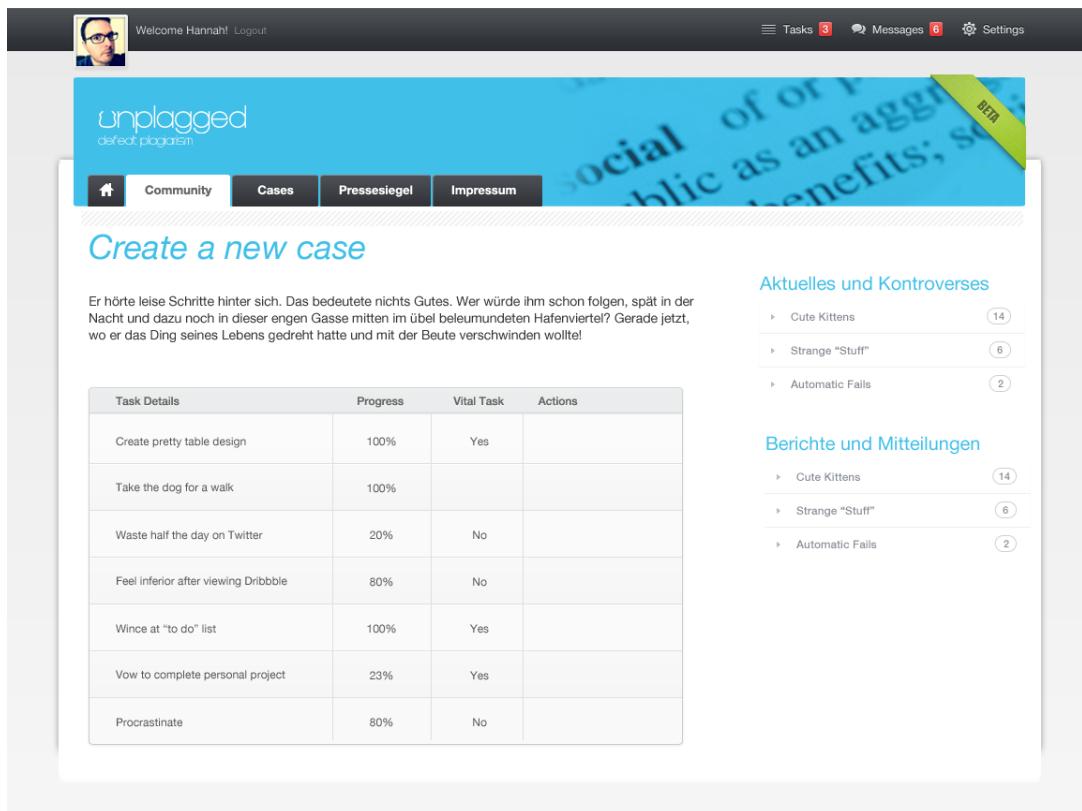


Figure 3.9.: Initial Screen PSD

on. Therefore CSS media queries are used, these are basically conditions that apply the CSS only if the condition is true. The most prominent conditions are the following:

Listing 3.19: CSS Media Queries

```
1 max-width: 600px  # max browser width 600px
2 min-width: 300px  # min browser width 300px
3 orientation:landscape # current orientation landscape mode
4 orientation:portrait # current orientation portrait mode
5 -webkit-min-device-pixel-ratio: 2 # min pixel ration (iPhone
4, Retina)
```

These media queries can be combined in any order to display an optimized page for each device. Since only the CSS changes, the HTML does not have to be touched. An example query looks like this:

Listing 3.20: CSS Media Query

```
1 @media only screen and (max-width: 600px) { }
```

Media queries are a CSS2 feature which most modern browsers, except IE9, implement. This isn't a problem though, because browsers that don't support them, just display the default css outside of the media queries, which is optimized for a width of about 1000 pixels in our case.

3.2.2. Javascript and fallbacks

Even though many websites require Javascript as mandatory for using the whole functionality range of the page, it is important to provide as much functionality as possible, when Javascript is turned off, to ensure accessibility(?, page 323). A short example will show how to implement a pagination with and without Javascript.

Usually when Javascript is disabled, the whole page will be reloaded, when the user changes to another page of the paginated content. The URL in this case will look something like this: <http://unplugged.local/document/list/page/2>. When Javascript is enabled, it

is much faster to only refresh the area of the page, that really needs to reload, in this case the table with the content of the next page. It is only possible to change the hash of an URL, the part after the hash key (#), using Javascript, the changed URL will be: <http://unplugged.local/document/list/#page/2>. An event called 'hashchange' can trigger a change of this part of the url and then reload the content through an AJAX request.

The pagination has the same HTML markup with and without Javascript, but with Javascript enabled it is much more convenient.

Listing 3.21: Javascript Pagination

```
1 $(".pagination a").live("click", function() {
2     var href = $(this).attr("href");
3     if(href) {
4         var substr = href.split('/');
5         var hash = substr.substr.length-2] + "/" + substr[
6             substr.length-1];
7         window.location.hash = hash;
8     }
9     return false;
10 });
11
12 $(window).bind('hashchange', function(){
13     var newHash = window.location.hash.substring(1);
14
15     if (newHash) {
16         var substr = newHash.split('/');
17         var hash = substr.substr.length-2] + "/" + substr[
18             substr.length-1];
19
20         var url = window.location.pathname;
21         if(url.charAt(url.length-1) != '/') {
22             url += '/';
23         }
24         url += hash;
25         $("#main-wrapper").load(url + " #main");
26     };
27 }
```

3.3. Frameworks

When we first discussed which programming language and frameworks, the Unplagged project should be built on, we figured out, that everyone was familiar with PHP. Since programming in a group requires a much better structure, than programming on your own, we needed a framework that requires a comfortable Model-View-Controller structure, we decided to use the Zend Framework. And to get rid of all the database issues as well as getting a flexibility in the used database system, we decided to use an Object-Relational-Mapping(ORM) framework, called Doctrine. What an ORM is, will be discussed later on.

3.3.1. Zend Framework

The Zend Framework is a typical PHP Framework using the Model-View-Controller pattern. Due to it's pre-defined directory structure, it is easy to get well seperated code.

The directories and their meaning:

- application — includes controller, models and view
- data — currently only includes i18n stuff (language stuff)
- docs — PHP Documentation and Developers Manual
- library — external and internal frameworks and extension to the Zend framework
- public — files that are accessible directly through the browser

- scripts — build scripts, deploying scripts
- temp — data that is overridden at any deployment
- tests — PHP Unit tests

The Zend framework offers RESTful² URLs that follow a fixed pattern: `unplagged.local/controller/`. Each controller is defined as `ControllerNameController.php` in the `application/controllers` directory and includes all possible actions. For example a file controller will look like this:

Listing 3.22: Persisting an object to the database in Doctrine

```

1 class FileController extends Zend_Controller_Action{
2     public function init(){
3     }
4
5     public function indexAction(){
6     }
7
8     public function uploadAction(){
9     }
10
11    <<<<< HEAD
12     public function listAction(){
13 =====
14     public function listAction(){
15 >>>>> 7a2ccf18eebad5d280d777f496be384bdf2a24f8
16     }
17
18     public function downloadAction(){
19     }
20 }
```

By default, if no action is defined, the `indexAction` is called. For each action the appropriate view is by default rendered in the `application/views/scripts/conntrollerName/actionName` file.

²See for example “RESTful Web Services” of Richardson and Ruby for more information

The `models` directory includes all the objects, this will be discussed in more detail in the following chapter about Doctrine.

3.3.2. Doctrine

The whole database connection management of Unplagged is implemented using the Doctrine Framework in version 2. It consists of two layers, a database abstraction layer (DBAL) and an object relational mapping framework (ORM). The DBAL uses PDO, a PHP framework for encapsulating database statements. The DBAL manages the communication with any kind of SQL database and offers an own query syntax. This has the advantage, that the database behind the framework can be changed from MySQL, to OracleSQL, PostgreSQL or SQLite at any time. The DBAL is the agent between PDO and the ORM. The ORM is the connection between PHP objects and the DBAL.

Before the use of Doctrine is described, it will be explained, how the database on a new machine can be created and how the database structure can be updated, whenever something changed in the structure. Actually this is very easy, it is required to have a local MySQL database at this point having 'root' as a username, no password and a database called 'unplagged'. It is also possible to create a new configuration in the application/configs/application.ini file, although this step will not be described in this chapter. If the database is created, the build script can be executed:

Listing 3.23: Updating database structure

```
1 php unplagged.local/scripts/build/doctrine.php
```

Now, the database is created or updated. If the database already existed, the data in it will not be removed! As described below, the big advantage of ORM is, that the programmer can stay in the PHP object context at any time. The only thing that has to be done additionally, is adding comments to the member variables of a class, that define the fields in the database:

Listing 3.24: Defining a class in Doctrine

```
1 /** @Entity */
2 class UserClass
```

```
3 {
4     /** @Column(type="integer") */
5     private $id;
6     /** @Column(length=30) */
7     private $username;
8 }
```

The whole syntax documentation of doctrine can be found here:

- <http://docs.doctrine-project.org/projects/doctrine-orm/en/latest/index.html>

To persist a new object to the database, the persist method on this object has to be called, this writes the object into the doctrine cache. It stays in the cache, until the flush method is called, which actually executes all the previous operation since the last flush. These can be deleting, updating, or editing an object.

Listing 3.25: Persisting an object to the database in Doctrine

```
1 $user = new User();
2 $user->setUsername('Max');
3 $em->persist($user);
4 $em->flush();
```

As the previous examples show, no database programming is necessary to create or persist a new object to the database, everything can be done in the PHP object context.

4. Summary and Outlook

During our first project semester, we already immersed into the field of plagiarism detection very deeply. After we read about existing websites and talked to Prof. Dr. Weber-Wulff about her experience and missing features at VroniPlag, we defined our list of requirements. In the next step we developed first user interface mockups and a basic layout, before we started the development of some requirements.

The application of a modified version of Scrum as our agile software development method was a very interesting and new experience to all of us. With the use of Redmine for keeping track of all issues, repository changes and time logs, everyone in the team was at anytime able to take a look on the current project state.

All in all the conditions for a successful development were allocated properly and we already have implemented many features of the requirements list, that can be improved in the next semester.

Though the employed development processes in this semester were not as consistent as they should have been. The in theory defined rules, how to program and how to test source code properly, were not always applied. So one of the main goals for the second semester is to improve this process. We need to focus on test driven development and to increase our velocity during the sprints, in order to get stable code more easily.

Another issue we need to work on is the staging environment we are using. Some features behave differently on Mac OS and Windows machines. Since we have only one staging environment, which updates on every commit to our git repository, it sometimes happens that a feature working on Windows crashes the staging environment. As a solution we should setup a second pre-staging environment which updates at every commit to the

repository automatically and another more stable server, which gets updated manually as soon as the pre-staing environment is testet properly.

The documentation of our work, which is represented in this developers manual was done in the end of the semester. In the next semester we are planning to extend and keep the manual up to date at every sprint, which means preferably more time for development and less time for documentation in the end.

A. Meetings

The following tables are showing the protocols of the held team meetings.

Meeting No. 06	Date: 2011-11-01
Content:	<p>Workflow requirements:</p> <ul style="list-style-type: none">• view table with page numbers, 3 digits per page number• double click words in fragment<ul style="list-style-type: none">• do google search and put found source up as a possible source• sim_text on source• turnItIn plugin in moodle -> takes document -> and sends back a report and this is stored in the database<ul style="list-style-type: none">• 2 documents are stored, can this be hooked in?• mark something as being a possible source (mark it by hand or automated)• page number, word number, line number, start number, page number word number end• candidate fragment (possible fragment, not approved yet)• merge fragments• two fragments invisible

Figure A.1.: meeting protocol no. 6

Meeting No. 08	Date: 2011-11-07
Content:	<ul style="list-style-type: none"> • Defined User Stories <p>Questions for the meeting with Mrs. Weber-Wulff</p> <ul style="list-style-type: none"> • priorising the featurelist • priorising the stories • for what exactly we need the fragmentmerging? • definition for fragments and textparts

Figure A.2.: meeting protocol no. 8

Meeting No. 10	Date: 2011-11-15
Content:	<p>User story updated</p> <p>Workflow</p> <ul style="list-style-type: none"> • User registration (1. Sprint) • User login (1. Sprint) • Upload a file/case (1. Sprint) <ul style="list-style-type: none"> - Use OCR to correct/edit file • Identify target (1. Sprint) • Add non-stop word identification • Choose a few word and check on Google to see if suspicious text is plagiarised • Find a file/source to upload • Use Simtext to compare • Read a file (by yourself) • If nothing detected --> use PDS • If plagiarism found --> create a new case (1. Sprint) • Set up permission for a new case <ul style="list-style-type: none"> - User has different roles - Inspection • Set a fragment • Generate simple report and upload • Create list of source <p>Consideration</p> <ul style="list-style-type: none"> • Merge fragments

Figure A.3.: meeting protocol no. 10

Meeting No. 11	Date: 2011-11-21
Content:	<p>Estimated Time for Tasks for Sprint 2</p> <ul style="list-style-type: none"> • create new case: 1h • create file uploader: 6.5h • create status 'set Target' to uploaded files: 3.5h • create user login: 2.5h • create user registration: 3.5h <p>Tasks assigned for Sprint 2</p> <ul style="list-style-type: none"> • Dominik: create new case • Elsa: create file uploader • Heiko: create status 'set Target' to uploaded files • Tien: create login • Benni: create registration

Figure A.4.: meeting protocol no. 11

Meeting No. 12	Date: 2011-11-28
Content:	<ul style="list-style-type: none"> - online scrum poker card - take a file in the pdf, scrape text in the pdf - sim text - fragments and marking fragments - make an xml of the pdf - open up original vs. intermediate state of a file - usually ocr does the work for you - we are interested in lines and pages <p>in january plagiarism detection software talk idea:</p> <ul style="list-style-type: none"> - a small group of documents as candidate - second step taking sim text, star plagger, .. - research: articles on plagiarism detection itself

Figure A.5.: meeting protocol no. 12

Meeting No. 13	Date: 2011-12-05
Content:	<p>Workingstatus</p> <ul style="list-style-type: none"> • Heiko: Was searching for a good design template and already found some. He showed us the design-template, and we all like it. • Tien: Finished login-part but has no Unit-tests done yet. • Dominik: Will give us a quick tutorial over skype, about how we can include PHPUnit-Test in our NetBeans environment. • Elsa: Finished basic uploader-part but has no Unit-tests done yet. <p>Appointment: The skypemeeting for the PHPUnit-Test tutorial from Dominik is on wednesday 07.12.2011 at 3 o'clock.</p>

Figure A.6.: meeting protocol no. 13

Meeting No. 15	Date: 2011-12-13
Content:	<p>I. Things that should be considered</p> <ul style="list-style-type: none"> • Vroni Plag Feeds • Search function • Invitation • Multiple targets • Folders for file system needed? • Tagging file and searching for tags • Category for fragments • Dealing with Umlaut <p>II. 2. Sprint until 17.01.2012</p> <ol style="list-style-type: none"> 1. Standard OCR 2. Display text within the system 3. Create/edit OCR-scanned text - Text dehypenmater 4. Simtext 5. Non-stop words + Google helpers <p>III. 3. Sprint</p> <ol style="list-style-type: none"> 1. Inspection 2. Merge fragments 3. Report + PDS

Figure A.7.: meeting protocol no. 15

Meeting No. 17	Date: 2012-01-18
Content:	<p>1. Current stand Following functions are already finished</p> <ul style="list-style-type: none"> • Parsing • Dehyphening • Non-stop word • Compiling Simtext • Google Search already integrated in browser <p>2. Improvement</p> <ul style="list-style-type: none"> • Simtext • Google Search for multiple words selection, for searching for fragments, notes, cases, texts etc... <p>3. Consideration</p> <ul style="list-style-type: none"> • Roll management • Workflow management

Figure A.8.: meeting protocol no. 17

Meeting No. 18	Date: 2012-01-23
Content:	<p>Focus:</p> <p>1. Simtext Improvement with comparision of more than 2 files.</p> <p>2. Workflow improvement There are following steps:</p> <ol style="list-style-type: none"> 1. Registration: user can see only public case He/she has to Sign up to have access allowance. 2. After logging in, user can see activities' stream A log table is needed to save all activities and show again if needed 3. Edit a case

Figure A.9.: meeting protocol no. 18

Meeting No. 19	Date: 2012-02-01
Content:	
	<p>Focus:</p> <ul style="list-style-type: none"> • Documentation will be written in Latex • Decided who writes what in how many time -> This will be showed in Appendix - logtime • Simtext improvement • Integration of other PDS in our system (PlagAware, Plagscan ...)

Figure A.10.: meeting protocol no. 19

B. Logged Time As Of March 11, 2012

The following tables are some example reports generated from the logged time in Redmine. To find the most recent version of these reports or to generate custom data analysis you can use the “Report” tool found in Redmine on the “Overview” page.

Table B.1.: Overview By Member and Month

Member	2011-10	2011-11	2011-12	2012-1	2012-2	2012-3	Total
Dominik Horb	27.30	10.00	55.75	51.00	28.00	16.80	188.85
Benjamin Oertel	34.00	29.00	19.00	57.50	4.50	29.00	173.00
Elsa Mahari	9.50	10.00	23.50	14.50	7.00		64.50
Tien Nguyen	1.00	24.00	13.00	12.00			50.00
Heiko Stammel	16.00	28.00	13.50	11.00			68.50
Total	91.30	101.00	124.75	146.00	39.50	45.80	548.35

Table B.2.: Overview By Member and Issue

Issue	Member	Total
none		18.30
Feature #1: Trac aufsetzen	Dominik Horb	2.00
Feature #5: Create Wiki List of Interesting Plagiarism Papers	Dominik Horb	1.00
Feature #6: Logo	Dominik Horb	1.00
Feature #8: Configure Email Notification in Trac	Benjamin Oertel Elsa Mähari	8.00 7.00 1.00
Feature #9: Github in Trac integrieren	Dominik Horb	1.50
Bug #11: Umstellung von Trac auf Redmine	Dominik Horb	1.00
Bug #12: Zeitmanagement in Trac integrieren	Dominik Horb	3.50
Bug #13: Initialize basics in wiki and issues	Dominik Horb	4.80
Bug #14: Setup Github in Redmine	Dominik Horb	1.00
	Dominik Horb	2.50
	Dominik Horb	2.50

Table B.2.: Overview By Member and Issue

Issue	Member	Total
Bug #15: Fremde Federn finden	Dominik Horb	3.00
Feature #16: OCR Research	Dominik Horb	3.50
Feature #20: Create Wiki Page for UserRoles	Dominik Horb	3.50
Time #21: Meeting 24.10.11	Tien Nguyen	1.00
	Dominik Horb	12.50
	Benjamin Oertel	2.50
	Elsa Mahari	2.50
	Tien Nguyen	2.50
	Heiko Stammel	2.50
Time #22: Meeting 17.10.11		10.00
	Dominik Horb	2.00
	Benjamin Oertel	2.00
	Elsa Mahari	2.00
	Tien Nguyen	2.00
	Heiko Stammel	2.00
Time #23: Meeting 10.10.11		10.00
	Dominik Horb	2.00
	Benjamin Oertel	2.00
	Elsa Mahari	2.00

Table B.2.: Overview By Member and Issue

Issue	Member	Total
Time #24: Debbie-Meeting 18.10.11	Tien Nguyen Heiko Stammel	2.00 2.00
	Dominik Horb	7.00
	Benjamin Oertel	1.00
	Elsa Mahari	2.00
	Tien Nguyen	1.00
	Heiko Stammel	2.00
Feature #28: Resume: Projectmeeting with HTW-Plag-Team		4.00
Feature #29: Projectmanagement Research	Heiko Stammel	4.00
Feature #30: Scrum Research	Heiko Stammel	7.00
Feature #31: Create Wiki for Project Management Steps	Heiko Stammel	7.00
Feature #32: Create Wiki for Scrum	Heiko Stammel	6.00
Bug #33: Mockup hand-drawing	Heiko Stammel	2.00
Bug #34: Mockup digital version	Benjamin Oertel	1.00
	Benjamin Oertel	6.00
	Benjamin Oertel	11.00
	Benjamin Oertel	11.00

Table B.2.: Overview By Member and Issue

Issue	Member	Total
Bug #35: Textvergleich-Algorithmen rechechieren	Tien Nguyen	1.00
Time #39: Meeting 31.10.11		12.50
	Dominik Horb	2.50
	Benjamin Oertel	2.50
	Elsa Mahari	2.50
	Tien Nguyen	2.50
	Heiko Stammel	2.50
Time #43: Meeting 2011-11-04 (Martin Heidingsfelder)		8.00
	Benjamin Oertel	2.00
	Elsa Mahari	2.00
	Tien Nguyen	2.00
	Heiko Stammel	2.00
Feature #44: Setting up initial Zend Framework + Doctrine environment		2.00
Feature #45: User Stories	Benjamin Oertel	2.00
Time #47: Meeting 07.11.11	Tien Nguyen	1.00
		10.00
	Dominik Horb	2.00
	Benjamin Oertel	2.00
	Elsa Mahari	2.00
	Tien Nguyen	2.00

Table B.2.: Overview By Member and Issue

Issue	Member	Total
Time #48: Meeting 14.11.11	Heiko Stammel	2.00
	Dominik Horb	2.00
	Benjamin Oertel	2.00
	Elsa Mahari	2.00
	Tien Nguyen	2.00
	Heiko Stammel	2.00
Time #49: Wiki Editing	Dominik Horb	2.00
Bug #50: Meeting Plagiatssteam HTW 3.11.11	Dominik Horb	6.00
	Benjamin Oertel	1.50
	Heiko Stammel	1.50
Time #53: Debbie Meeting 15.11.11	Dominik Horb	3.00
	Benjamin Oertel	1.50
	Elsa Mahari	1.50
	Tien Nguyen	1.50
	Heiko Stammel	1.50
Feature #56: User registration		6.50
Feature #57: Profile information update	Benjamin Oertel	6.50
		3.00

Table B.2.: Overview By Member and Issue

Issue	Member	Total
Time #58: Meeting 28.11.11	Benjamin Oertel Dominik Horb	3.00 2.50
	Benjamin Oertel Elsa Mahari	2.50 2.50
	Tien Nguyen Heiko Stammel	2.50 2.50
Feature #59: Defining coding guidelines	Benjamin Oertel	0.50
Feature #61: Meeting 24.11.11 - Setting up developing environment	Benjamin Oertel Dominik Horb	0.50 12.00
	Benjamin Oertel Elsa Mahari	3.00 3.00
	Tien Nguyen	3.00
Feature #62: Identify target	Benjamin Oertel	0.50
Feature #63: Read a file	Benjamin Oertel	1.00
Feature #64: Upload files	Elsa Mahari	8.00
Feature #65: User Login	Benjamin Oertel	6.50
		1.00

Table B.2.: Overview By Member and Issue

Issue	Member	Total
Feature #66: Create a new case	Tien Nguyen	5.50
	Dominik Horb	2.00
Feature #67: Setting up PHPUnit	Dominik Horb	7.50
	Benjamin Oertel	2.00
	Elsa Mahari	2.00
Time #68: Meeting 05.12.11		4.00
	Dominik Horb	2.00
	Elsa Mahari	2.00
Time #69: Reading DocBook-XML	Dominik Horb	4.00
	Benjamin Oertel	1.00
Feature #70: Writing git tutorial		1.00
Time #71: Meeting Online 07.12.11	Benjamin Oertel	1.00
	Dominik Horb	16.00
	Benjamin Oertel	4.00
	Elsa Mahari	4.00
	Tien Nguyen	4.00
Time #72: Meeting preparation PHPDoc		2.00
Feature #73: Layout design	Dominik Horb	2.00
		10.50

Table B.2.: Overview By Member and Issue

Issue	Member	Total
Bug #74: protocoll meeting minuts 05.12.2011	Benjamin Oertel	10.50
	Elsa Mahari	0.50
Time #75: Creation of Preview area on server		3.00
Time #76: Meeting 12.12.2011	Dominik Horb	3.00
	Dominik Horb	7.50
	Benjamin Oertel	1.50
	Elsa Mahari	1.50
	Tien Nguyen	1.50
	Heiko Stammel	1.50
Time #77: create Wiki page for MeetingMinutes 15.11.2011		0.50
	Tien Nguyen	0.50
Feature #78: Setting up the unplagged-blog		17.00
Feature #79: Google Helper	Heiko Stammel	17.00
Feature #80: SIM-Text	Elsa Mahari	10.00
		10.00
Feature #81: Correct/edit OCR-scanned Text	Dominik Horb	3.00
	Tien Nguyen	10.00
	Benjamin Oertel	4.00

Table B.2.: Overview By Member and Issue

Issue	Member	Total
Feature #82: De-hyphenator	Benjamin Oertel	8.00
Feature #83: Non-stop word identification	Dominik Horb	8.00
Feature #84: Display text	Heiko Stammel	2.00
Feature #85: Standard OCR	Dominik Horb	6.00
Feature #86: Change file list view columns	Benjamin Oertel	3.50
Feature #87: Decrease line-height in tables.	Dominik Horb	1.00
Time #89: create Wiki page for MeetingMinutes 13.12.2011	Benjamin Oertel	2.50
Time #90: Prepare automatic preview area	Tien Nguyen	18.00
Time #101: Preparation of Database Mocks	Dominik Horb	12.00
Feature #102: Basics for responsive layout	Benjamin Oertel	6.00
	Benjamin Oertel	0.50
	Benjamin Oertel	0.50
	Benjamin Oertel	0.50
	Tien Nguyen	0.50
	Dominik Horb	11.25
	Dominik Horb	3.50
	Dominik Horb	3.50
		9.00

Table B.2.: Overview By Member and Issue

Issue	Member	Total
Time #103: Meeting 10.01.12	Dominik Horb	9.00
		3.00
	Dominik Horb	1.50
	Elsa Mahari	1.50
Bug #104: Skype Meeting 4.01.12	Elsa Mahari	1.00
		3.00
Bug #105: Switching Menu to Zend_Navigation	Dominik Horb	3.00
		10.00
Feature #106: Add Imagick to convert to different formats	Dominik Horb	6.00
	Benjamin Oertel	4.00
Bug #107: Meeting 2012-01-09		1.50
	Tien Nguyen	4.00
Time #108: Skype Meeting 12.01.12	Dominik Horb	2.00
	Elsa Mahari	2.00
		2.50
Time #109: Cleaning up directory structure	Dominik Horb	2.50
		13.50
Feature #110: Implementing Zend_Acl for user access control	Dominik Horb	11.50
	Benjamin Oertel	2.00
Feature #111: Improving interface workflow		8.00

Table B.2.: Overview By Member and Issue

Issue	Member	Total
Time #115: Debbie Meeting 18.01.12	Benjamin Oertel	8.00
	Dominik Horb	6.00
	Elsa Mahari	3.00
Feature #116: Logger	Benjamin Oertel	3.00
	Dominik Horb	4.00
Feature #117: Create Paging mechanism	Benjamin Oertel	2.00
Bug #118: create Wiki page for MeetingMinutes 18.01.2012	Benjamin Oertel	2.00
Bug #119: Fixing several css issues	Tien Nguyen	6.00
Feature #120: Beautifying forms, tables and buttons	Benjamin Oertel	0.50
Time #122: 23.01.12 Meeting	Benjamin Oertel	5.00
	Dominik Horb	5.00
	Elsa Mahari	5.00
Time #123: 30.01.12 Meeting	Benjamin Oertel	6.00
	Dominik Horb	2.00
	Benjamin Oertel	2.00

Table B.2.: Overview By Member and Issue

Issue	Member	Total
Time #128: LaTeX preparation	Elsa Mahari	2.00
Time #142: License research and inclusion	Dominik Horb	2.00
Bug #143: Fixing tests after last refactorings	Dominik Horb	6.00
Time #144: Reading Zend Doku for Bootstrap Plugins	Dominik Horb	1.50
Time #145: Meeting 06.02.12	Dominik Horb	4.00
Time #146: Basic TOC and split up of LaTeX document	Dominik Horb	2.00
Feature #147: Preface Developers Manual	Dominik Horb	1.50
Time #148: Documentation - Chapter Git	Dominik Horb	7.00
Time #149: Meeting 04.03.12	Benjamin Oertel	4.50
Time #150: Writing chapter about Zend and Doctrine Framework	Benjamin Oertel	5.00
	Benjamin Oertel	5.00

Table B.2.: Overview By Member and Issue

Issue	Member	Total
Time #151: Writing chapter about User Interface / User Experience	Benjamin Oertel	6.00
Time #152: Adding mockups to documentation	Benjamin Oertel	2.00
Time #153: Team Meeting – 2012-03-10	Benjamin Oertel	2.00
Time #154: Reading in latex documentation	Benjamin Oertel	1.50
Feature #155: Improvements of error detection in tesseract parser	Benjamin Oertel	3.00
Feature #156: Implementing Webservice of PlagAware	Benjamin Oertel	2.00
Bug #157: Zend Acl – allowing/denying access to actions, not only controllers	Benjamin Oertel	1.00
Total		548.35

Table B.3.: Overview By Sprints

Version	2011-10	2011-11	2011-12	2012-1	2012-2	2012-3	Total
none	7.00		28.75	47.00	33.00	19.30	135.05
Sprint 1 (2011-10-01 - 2011-11-14)	81.80	67.50					149.30
Product Back-Log	2.50	0.50	3.00				6.00
Sprint 2 (2011-11-15 to 2011-12-13)		33.00	72.00				105.00
Sprint 3 (2011-12-14 to 2012-01-17)			21.00	63.00			84.00
Sprint 4 (2012-01-18 to 2012-02-01)				36.00	2.00		38.00
Sprint 5 (2012-02-02 to 2012-23-02)					4.50	26.50	31.00
Total	91.30	101.00	124.75	146.00	39.50	45.80	548.35

C. Mockups

C.1. Hand-Drawn

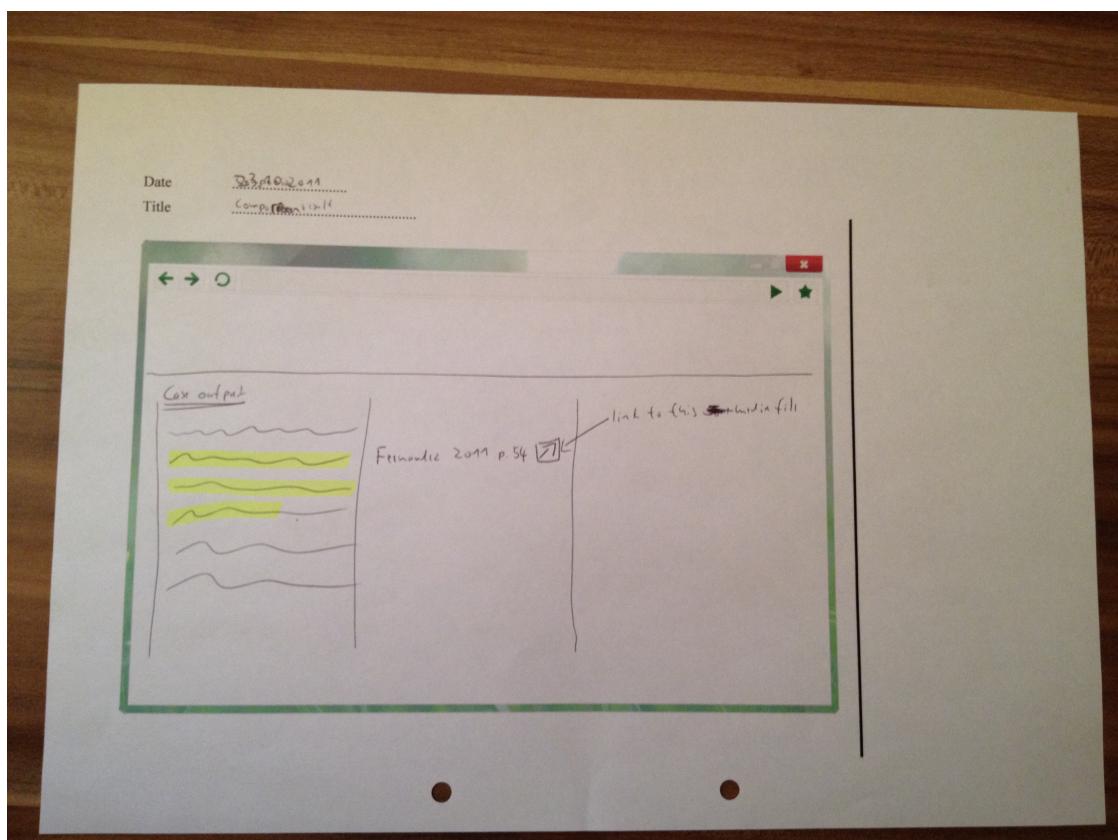


Figure C.1.: Mockup – Compare results – digitalized

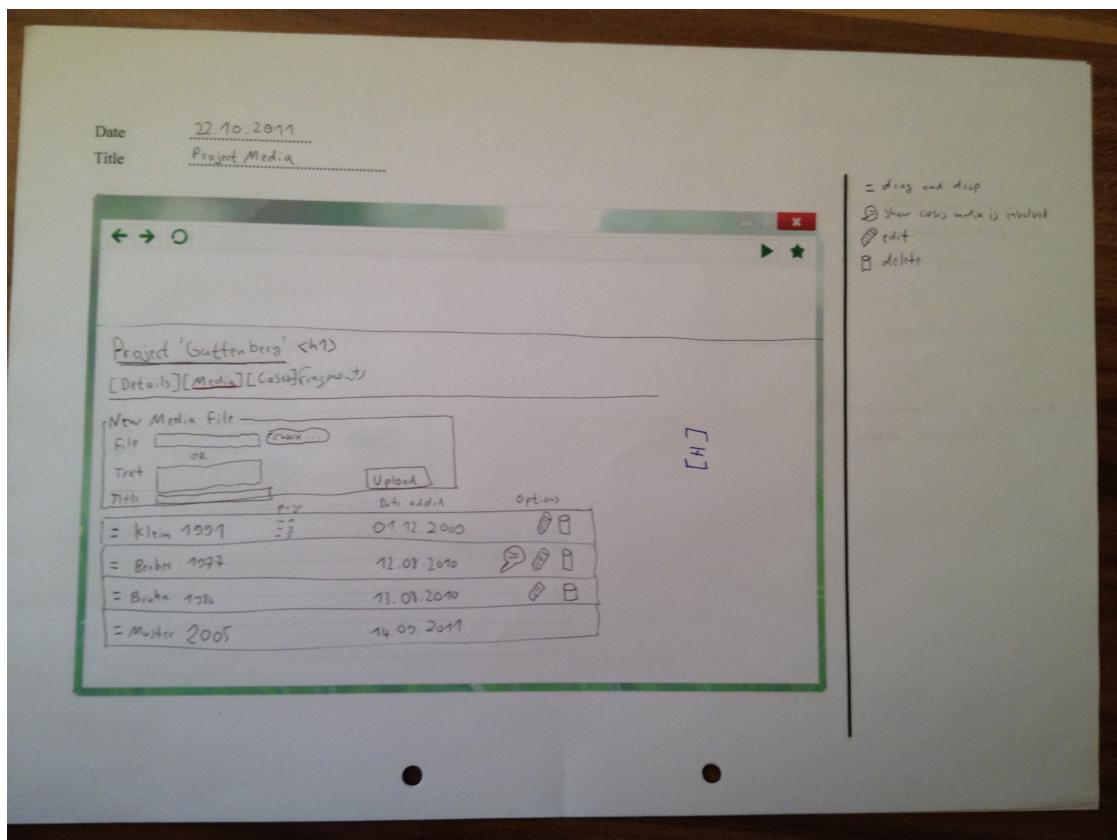


Figure C.2.: Mockup – Media list – digitalized

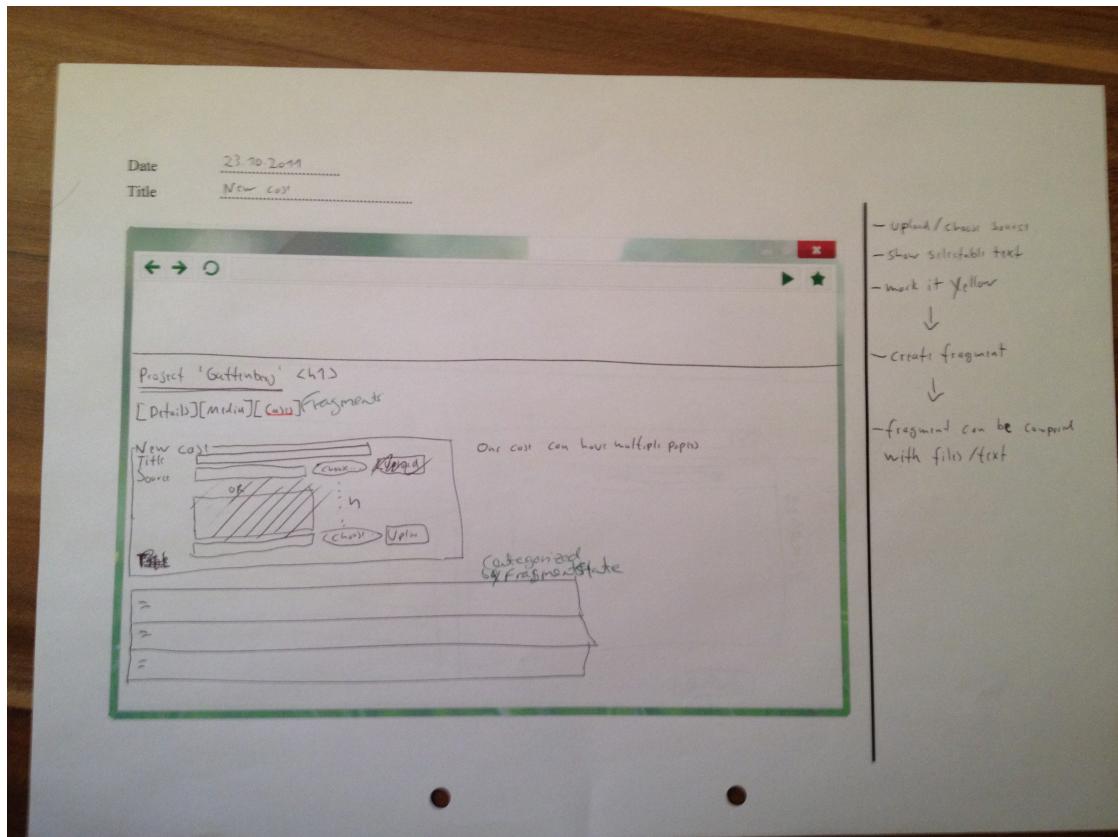


Figure C.3.: Mockup – New case – digitalized

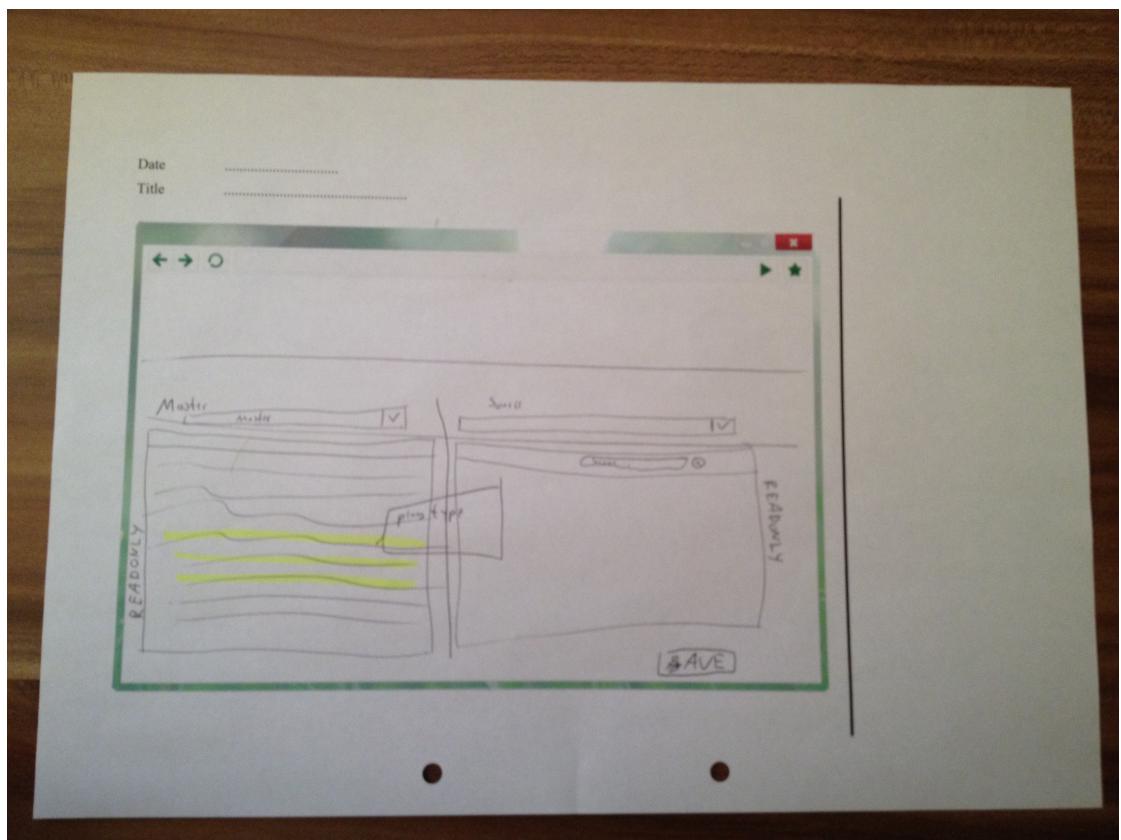


Figure C.4.: Mockup – New fragment – digitalized

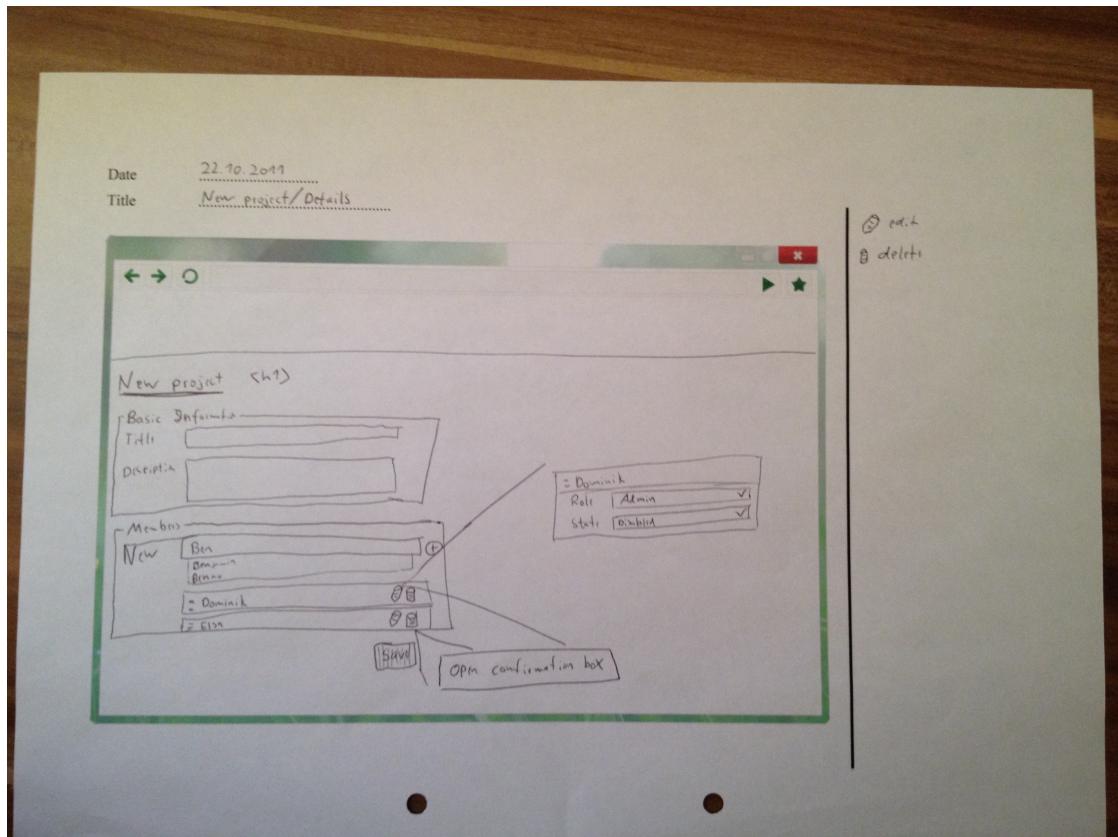


Figure C.5.: Mockup – New project – digitalized

C.2. Digitalized

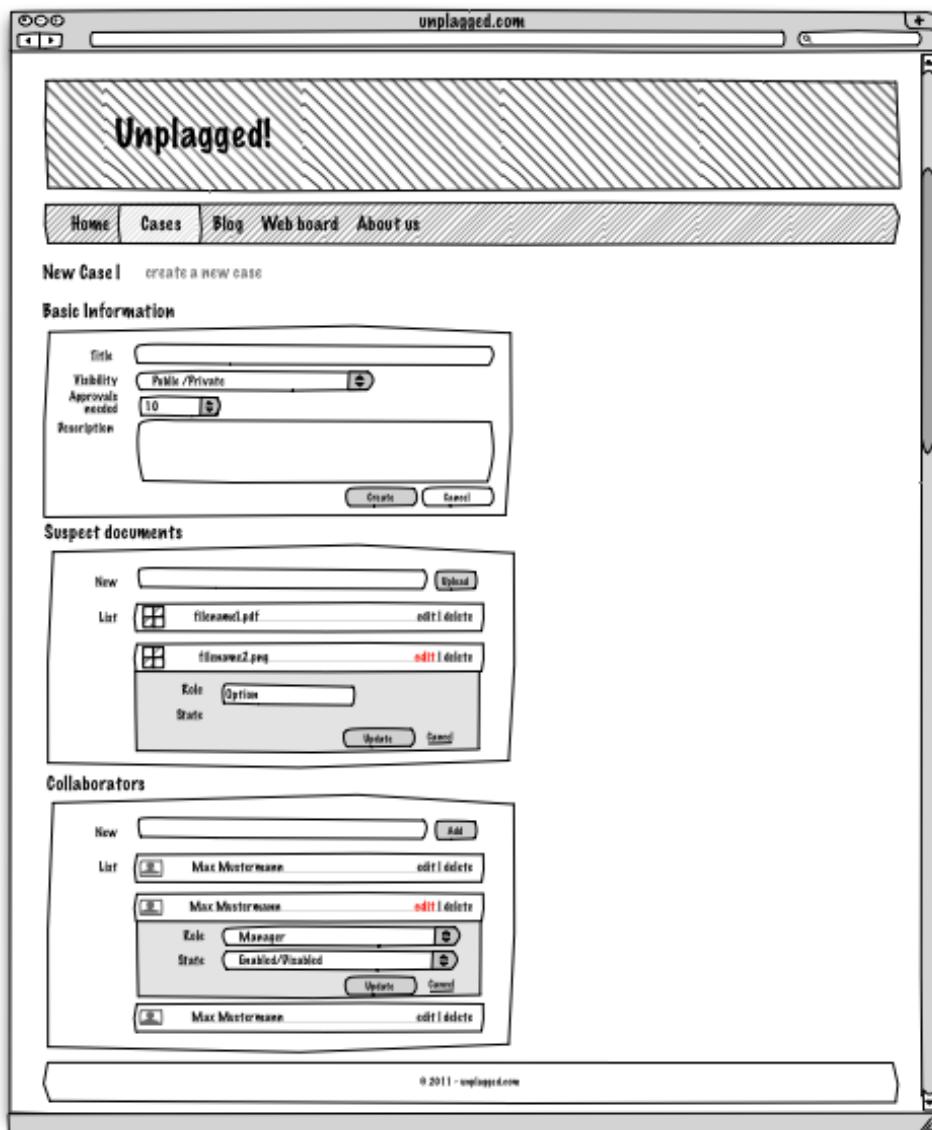


Figure C.6.: Mockup – New case – digitalized

The mockup shows a web browser window for 'unplugged.com'. The header features a hatched background with the title 'Unplugged!' in large, bold, black font. Below the header is a navigation bar with links: Home (selected), Cases, Blog, Web board, and About us.

The main content area displays the 'Fragments' page, which lists all fragments in case 'Gutenberg'. A yellow callout box points to the 'Create' button in a modal dialog, stating: 'clicking create opens a new page with suspect document and search pre-selected'.

The 'New Fragment' dialog has two input fields: 'Suspect document' set to 'Gutenberg.pdf' and 'Suspect source' set to 'Kafka'. It includes 'Create' and 'Cancel' buttons.

The 'Fragments by page' section shows a table of fragments across various pages. The table includes columns for 'Page', 'Fragments', 'Status', 'Latest activity', and 'Options'.

Page	Fragments	Status	Latest activity	Options
(+) Page 1	0 2 7 5	waiting approval	2011-08-25	
(+) Page 2	0 2 7 5	waiting approval	2011-08-25	
(+) Page 3	0 1 1 2	waiting approval	2011-08-25	
Line 19-25 Kompletthyperlink	approved	2011-08-25	edit show approve	
Line 29-44 Kompletthyperlink	waiting approval	2011-08-28	edit show approve	
Line 55-55 Kompletthyperlink	waiting approval	2011-08-25	edit show approve	
(+) Page 5	0 2 7 5	waiting approval	2011-08-25	
(+) Page 6	0 2 7 5	waiting approval	2011-08-25	

The footer contains the copyright notice '© 2011 - unplugged.com'.

Figure C.7.: Mockup – List fragments – digitalized

The mockup displays the Unplagged! website interface with the following sections:

- Header:** Shows the Unplagged! logo and navigation links: Home, Cases, Blog, Web-board, About us.
- Suspect document:** A text editor window showing a document page from "Ostenberg.pdf" (Page 1). The text discusses Josef K. being questioned about his actions after a fight. A callout bubble highlights the word "mark as plagiarism" and notes "type: lorem ipsum".
- Supposed source:** A text editor window showing a document page from "Kafka" (Page 8). The text discusses a dog named Max being questioned about his actions after a fight. A callout bubble highlights "comments possible on each line, different colors per user" and "text colors different per plagiarism type". Another callout bubble says "commenter show comment".
- Comment:** A form for creating a comment. It includes fields for "Note" (with a large text area), "Visibility" (set to "Private/Public/Group"), and buttons for "Create", "Save", "Revert", and "Cancel". A callout bubble lists comment status options: "new created but not ready for verification", "in progress", "ready", "feedback", "closed (5 approvals)", and "rejected (5 declined)".
- Footer:** Shows the copyright notice "© 2011 - unplagged.com".

Figure C.8.: Mockup – New fragment – digitalized

unplugged.com

Unplugged!

Home Cases Blog Web board About us

Show Fragment 1 fragment in case "Gutenberg"

Suspect document

File: document1.pdf
Page: 1
Line: 34-40

...Als Gregor Samza eines Morgens aus unruhigen Träumen erwachte, fand er sich in seinem Bett zu einem ungeheueren Ungeziefer verwandelt. Und es war ihnen wie eine Bestätigung ihrer neuen Träume und guten Absichten, ...
2 words below and 2 above ...

Supposed source

File: document1.pdf
Page: 1
Line: 34-40

emand musste Josef K. verleumdet haben, denn ohne dass er etwas Böses getan hätte, würde er eines Morgens verhaftet. »Wie ein Handl« sagte er, es war, als sollte die Scham ihn überleben. Als Gregor Samza eines Morgens aus unruhigen Träumen erwachte, fand er sich in seinem Bett zu einem ungeheueren Ungeziefer verwandelt. Und es war ihnen wie eine Bestätigung ihrer neuen Träume und guten Absichten, als am Ziele ihrer Fahrt die Tochter als erste sich erhob und ihren jungen Körper dehnte. »Es ist ein eigenmächtiger Apparat«, sagte der Offizier zu dem Forschungsgerüstenden und überblieb mit einem prahlsermaßen bewundernden Blick den ihm doch wohlbekannten Apparat.

(+) expand whole page (-) collapse whole page

commenter show facebook-like list with all selected users, comment and date

Ratings

appreciate : 1 rejects : 1

State: Approve

comment ratings that are not the most recent ones of a user are shown greyed out

Comment	User	Date	Rating
guten Absichten, als am Ziele ihrer Fahrt die Tochter als erste sich erhob und ihren jungen Körper dehnte. »Es ist ein eigenmächtiger Apparat«, sagte der Offizier zu dem Forschungsgerüstenden und überblieb mit einem prahlsermaßen bewundernden Blick den ihm doch wohlbekannten Apparat.	Max Mustermann	2011-11-05	no
Re: guten Absichten, als am Ziele ihrer Fahrt die Tochter als erste sich erhob und ihren jungen Körper dehnte. »Es ist ein eigenmächtiger Apparat«, sagte der Offizier zu dem Forschungsgerüstenden und überblieb mit einem prahlsermaßen bewundernden Blick den ihm doch wohlbekannten Apparat.	Emmy Watson	2011-11-06	no
Re: guten Absichten, als am Ziele ihrer Fahrt die Tochter als erste sich erhob und ihren jungen Körper dehnte. »Es ist ein eigenmächtiger Apparat«, sagte der Offizier zu dem Forschungsgerüstenden und überblieb mit einem prahlsermaßen bewundernden Blick den ihm doch wohlbekannten Apparat.	Max Mustermann	2011-11-07	no
guten Absichten, als am Ziele ihrer Fahrt die Tochter als erste sich erhob und ihren jungen Körper dehnte. »Es ist ein eigenmächtiger Apparat«, sagte der Offizier zu dem Forschungsgerüstenden und überblieb mit einem prahlsermaßen bewundernden Blick den ihm doch wohlbekannten Apparat.	Emmy Watson	2011-11-08	no
guten Absichten, als am Ziele ihrer Fahrt die Tochter als erste sich erhob und ihren jungen Körper dehnte. »Es ist ein eigenmächtiger Apparat«, sagte der Offizier zu dem Forschungsgerüstenden und überblieb mit einem prahlsermaßen bewundernden Blick den ihm doch wohlbekannten Apparat.	Max Mustermann	2011-10-27	yes

Figure C.9.: Mockup – Show fragment for approval – digitalized

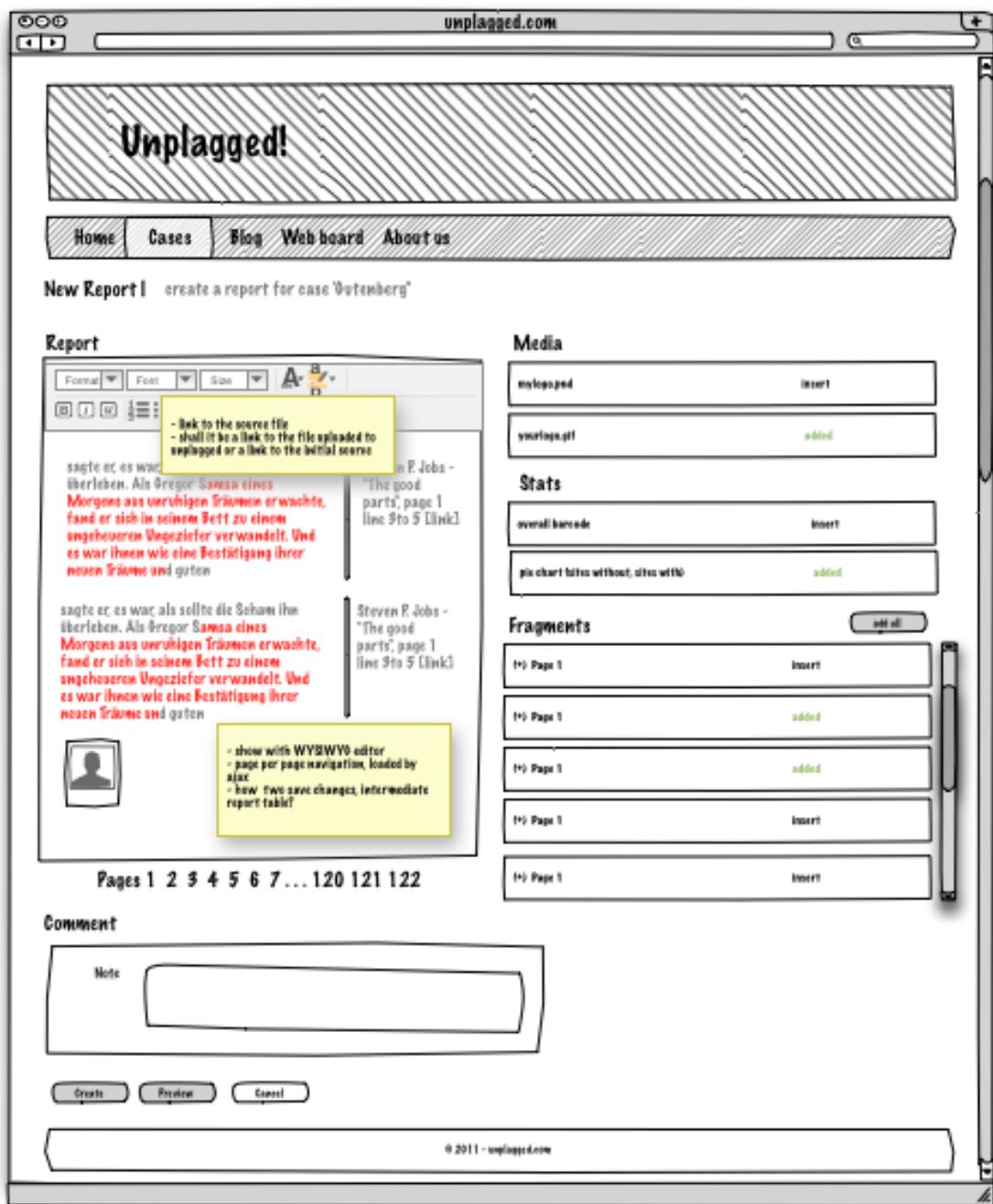


Figure C.10.: Mockup – New report – digitalized