

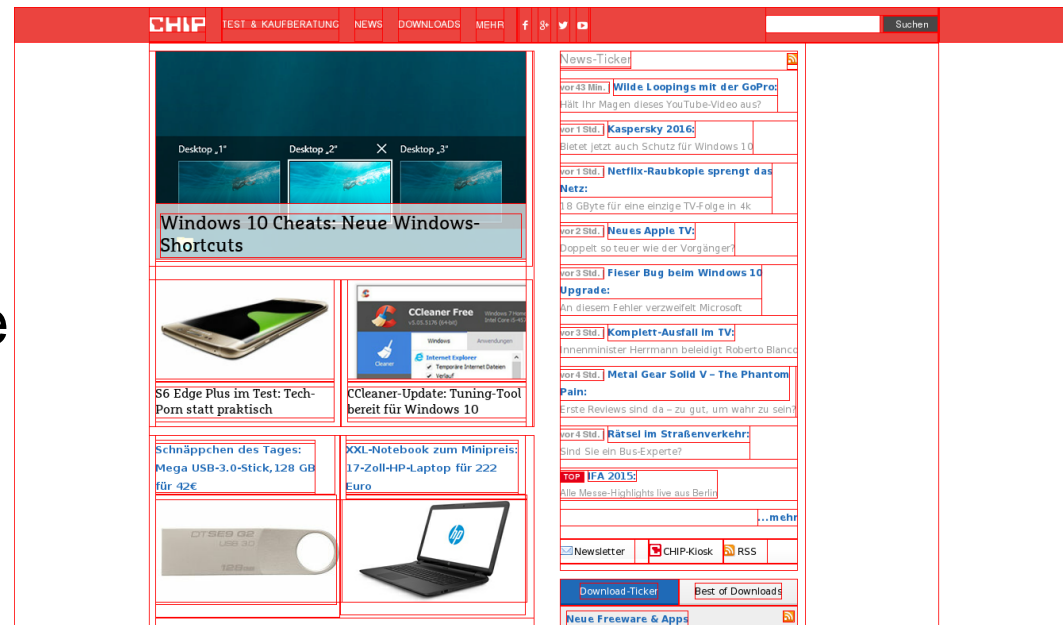
# **Visuelle Segmentierung von Webseiten mit Hilfe von Crowdsourcing**

Florian Kneist. Bauhaus-Universität Weimar.  
Bachelor Verteidigung. 29.04.2016

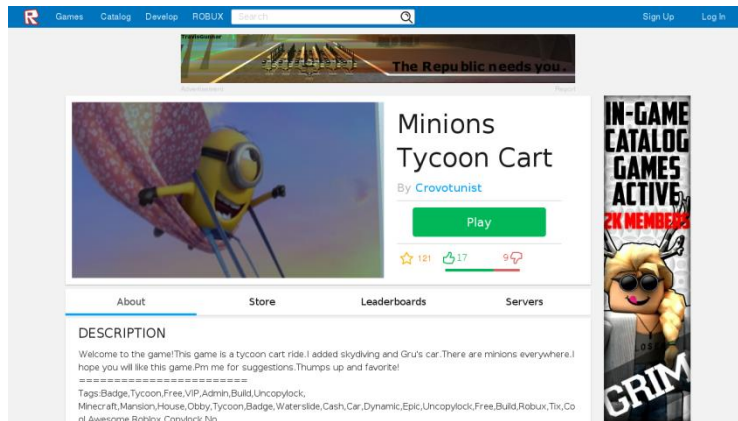
# Einleitung

# Einleitung: Webseiten Segmentierung

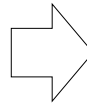
- Zusammengehörige Inhaltsblöcke
  - Navigationsbereich
  - Werbung
  - Kommentare
  - Artikeltext
  - Bilder
- Rechteckige Segmente
- Webseitengestaltung
  - Positionierung
  - Farbe
  - Leerzeilen/Freiraum
- Lösung nicht immer eindeutig



# Einleitung: Webseiten Segmentierung



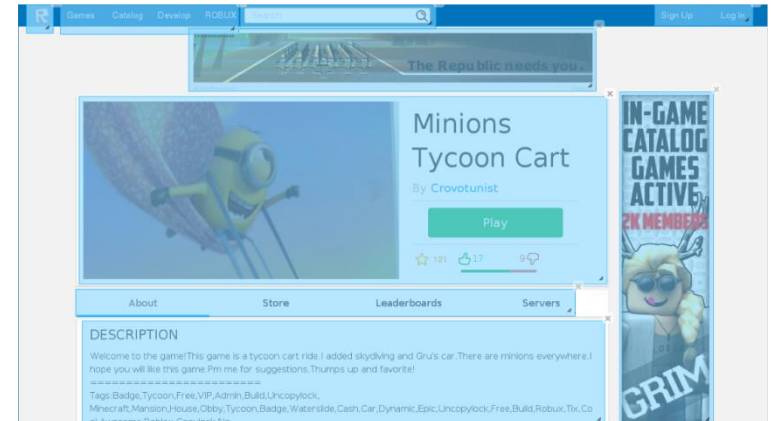
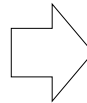
Original Webseite



# Einleitung: Webseiten Segmentierung



Original Webseite



Segmentierung

- Wozu segmentieren?
  - Vorbereitender Schritt zur Kategorisierung
  - Vorbereitender Schritt zur Content Extraction
  - Interessant für Screenreader, Barrierefreiheit

- **Andere Ansätze**
  - Analyse der HTML-Struktur und Textbausteinen
  - Automatisches Finden von visuellen Separatoren
  - Linguistische Betrachtung: Unterschiede im Schreibstil von Textblöcken
- Verfügbare Datensätze zu klein (bis 600 Seiten)
- Web hat sich weiterentwickelt
- **Größere und aktuelle Datensätze benötigt**
  - Vergleich von bestehenden Segmentierungsverfahren
  - Evaluierung von neuen Ansätzen

- Softwarekomponenten
  - Arbeitnehmer Interface
  - Arbeitgeber Interface
  - Archivierungspipeline
- Erzeugung eines Goldstandards



# Interfaces

- Virtuelle Microjobs (Bezahlung in Cent-Beträgen)
- Zahlungsbetrag frei wählbar
- Zugang für Arbeiter und Arbeitgeber
- Jede Aufgabe wird manuell erledigt → Ergebnisse können variieren
- Ablehnen von Abgaben durch Arbeitgeber möglich
- Typische Aufträge: Bilder kategorisieren, Umfragen

## Interfaces: Generelle Anforderungen

- Im Browser
- Schnelle Erledigung der jeweiligen Aufgabe
  - Skalierbarkeit

# Interfaces: Arbeiter - Übersicht

The screenshot displays the Amazon Mechanical Turk interface for a task titled "Mark Content Elements". The task is assigned to a worker named "amazonmechanical\_turk" with a qualification of 100%. The task instructions are as follows:

**Mark Content Elements**

Please draw rectangles around groups of content that belong together. A few instructions:

1. Protect areas of the website that belong together.
2. You can make multiple selections.
3. You can refine your selection.
4. You can delete rectangles.

Some bad and good examples:

The examples show two columns of screenshots. The left column, labeled "Bad", shows examples of incorrect selections (e.g., selecting individual elements instead of groups). The right column, labeled "Good", shows examples of correct selections (e.g., selecting entire sections of the page).

The task is currently in progress, with a progress bar showing 100% completion. The task is titled "Key Payment and Service Information" and includes a description of the service and a link to the "PayPal Service Essentials" page.



Anleitung

Beispiele


Arbeitsbereich

Feedback

# Interfaces: Arbeiter - Demo

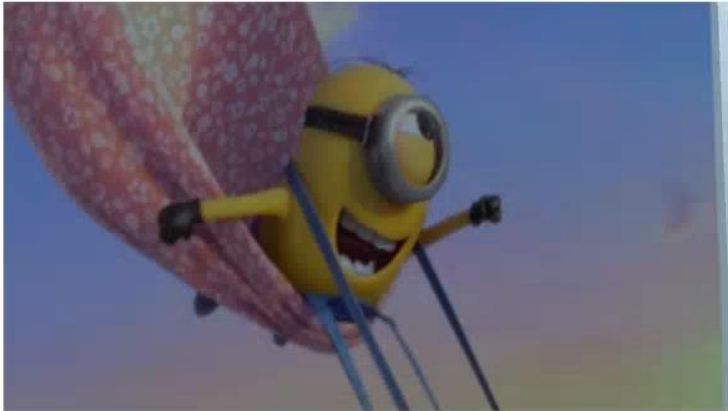
[Games](#)[Catalog](#)[Develop](#)[ROBUX](#)

[Sign Up](#)[Log In](#)



Advertisement


Report





## Minions Tycoon Cart

By [Crovotunist](#)

[Play](#)

 121

 17

 9


[About](#)[Store](#)[Leaderboards](#)[Servers](#)

### DESCRIPTION

Welcome to the game! This game is a tycoon cart ride. I added skydiving and Gru's car. There are minions everywhere. I hope you will like this game. PM me for suggestions. Thumbs up and favorite!

=====

Tags: Badge, Tycoon, Free, VIP, Admin, Build, Uncopylock, Minecraft, Mansion, House, Obby, Tycoon, Badge, Waterslide, Cash, Car, Dynamic, Epic, Uncopylock, Free, Build, Robux, Tix, Cool, Awesome, Roblox, Copylock, No



# Interfaces: Arbeiter – Auswahlmethode

## Content selektieren

Browser: nativ

[Selection - Web APIs | MDN](#)  
developer.mozilla.org › ... › Web technology for developers › Web APIs  
Jan 29, 2016 - toString() method returns the text contained in the selection, e.g. ... is subject to change in future versions of browsers as the specification changes. ... To obtain a Selection object for examination or modification, call window.



Browser: Dev. Tools

[div.s 512px × 71px | MDN](#)  
developer.mozilla.org › ... › Web technology for developers › Web APIs  
Jan 29, 2016 - toString() method returns the text contained in the selection, e.g. ... is subject to change in future versions of browsers as the specification changes. ... To obtain a Selection object for examination or modification, call window.

Console Sources Network Timeline Profiles Resources Security Audits Adblock Pl

▼<div class="sng">  
▼<div class="g">  
<!--m-->  
▼<div class="rc" data-hwid="22">  
▶<h3 class="r"></h3>

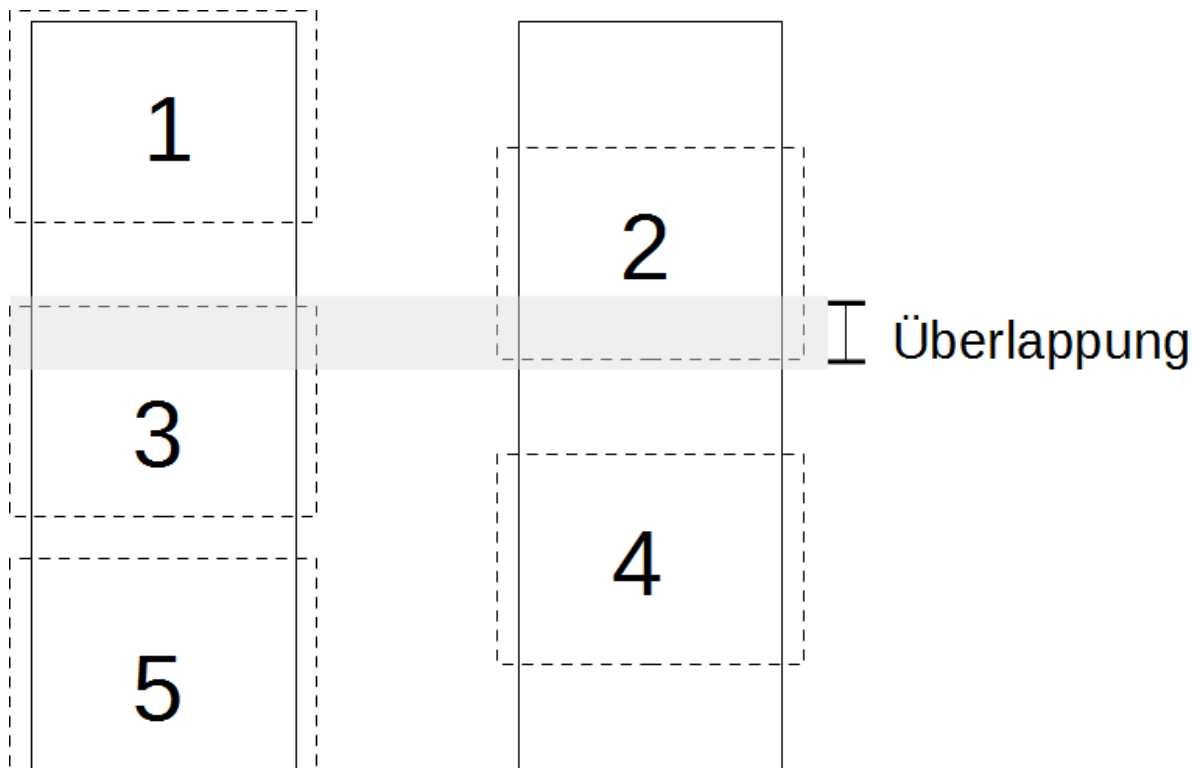


Auswahlrechteck



## Interfaces: Arbeiter – Aufbau des Tasks

- Webseiten als Screenshots
- Screenshots unterteilt
- Überlappung



## Interfaces: Arbeiter – Vorteile von Screenshotabschnitten

- Keine versehentliche Interaktion mit Webseite
- Vergleichbarer Arbeitsaufwand pro Abschnitt



- Segmentierungen: Rechtecke
  - Zeitlicher Verlauf der Segmentierung
  - Feedback der Arbeiter
  - ...
- 
- Zeitzone
  - Auflösung
  - Browser/Betriebssystem
  - ...

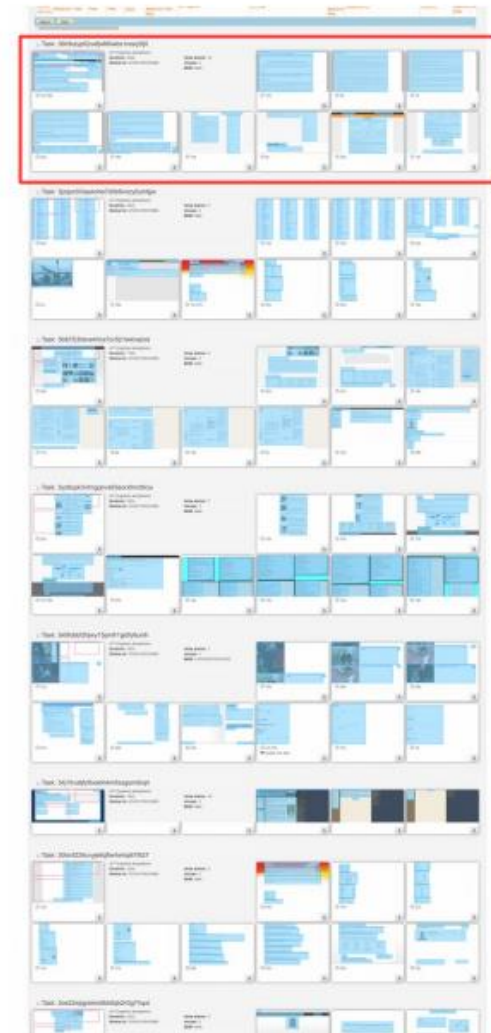
## Interfaces: Arbeiter – Task Aufbau

- 10 Screenshotabschnitte
- Davon 2 Referenzabschnitte zur Kontrolle
- 10 verschiedene Arbeiter bearbeiten gleichen Task

## Interfaces: Arbeiter - Limitierungen

- Verdeckter Inhalt
- Inhalt, der Interaktion erfordert
- Scrolleffekte

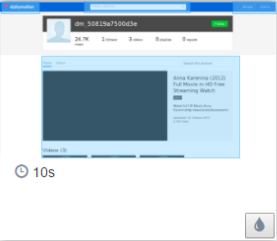
- Ergebnisse auf AMT nur in Tabellenform
  - Visualisierung der Annotationen durch Browsererweiterung
- Übersicht über Arbeiter
  - Erfolgreich absolvierte Aufgaben
  - Technische Daten über verwendete Hardware/Software
  - Vergleich mit Referenzabschnitten

[illegible]

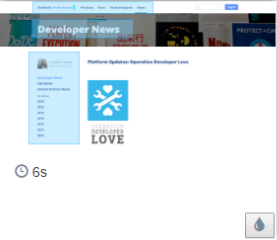
# Interfaces: Arbeitgeber - Detail

Task: 34s9dkfk73pgy0co5ugratfb7lmnyx

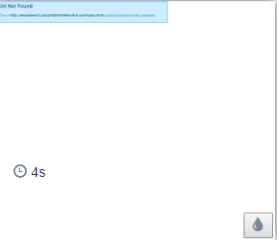
Duration: 76s  
Worker Id: A7W013PM199BS  
Submit time: Wed Apr 13 03:00:23 PDT 2016  
Status: Approved  
Approval (7d, 30d, lifetime):  
0% (0/0),  
92% (156/170),  
94% (275/293)  
Language: de  
Time Zone: Mitteleuropäische Sommerzeit  
OS: Windows  
Frame Width: 1850px



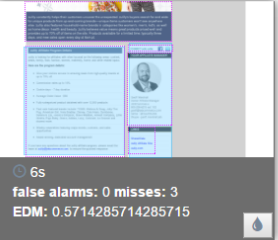
10s




6s



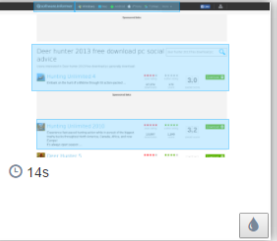
4s



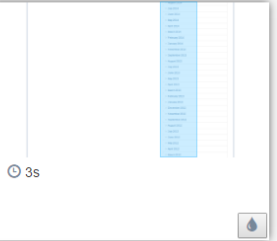
6s  
false alarms: 0 misses: 3  
EDM: 0.5714285714285715



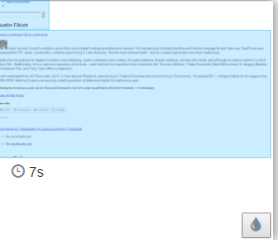
8s  
false alarms: 0 misses: 3  
EDM: 0.6666666666666666




14s



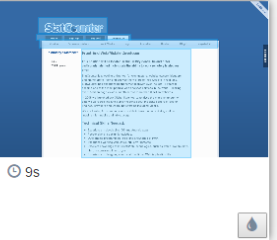
3s



7s

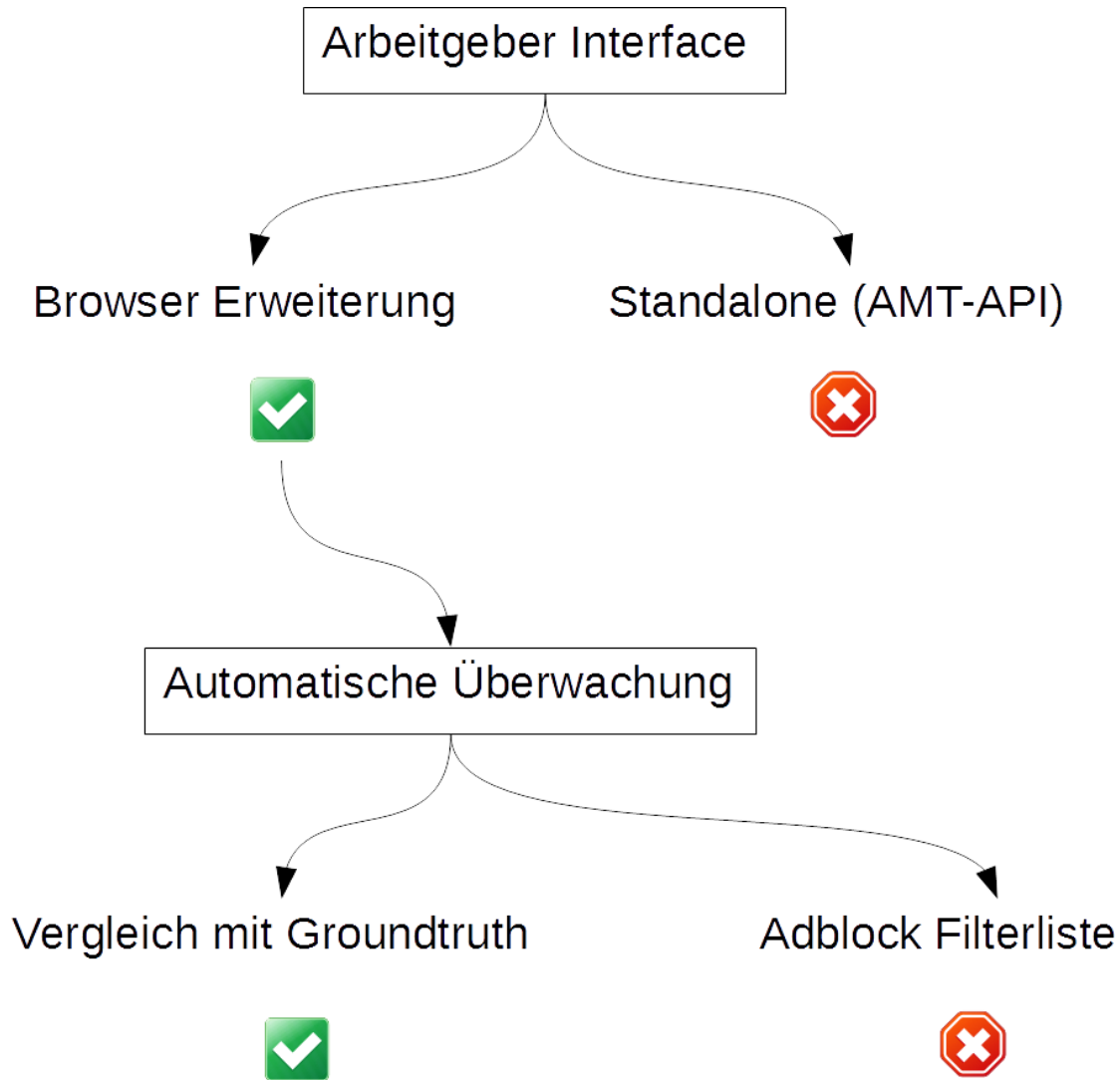


4s



9s

# Interfaces: Arbeitgeber – Entscheidungen



# Interfaces: Arbeitgeber – Vergleich mit Goldstandard

Medved on **October 3, 2015 at 9:51 am** said:  
Как скачать то? How download?

BongaCams on **October 7, 2015 at 7:42 am** said:  
You don't need to download it, just visit the website on your mobile browser, and that's it 😊

Alegei on **October 18, 2015 at 12:38 pm** said:  
Жду любовь!

sebeni\_cornel on **October 21, 2015 at 10:01 am** said:  
vreau sex

kivan on **October 21, 2015 at 10:30 pm** said:  
BongaCams on October 7, 2015 at 7:42 am said: You don't need to download it, just visit the website on your mobile browser, and that's it



# Interfaces: Arbeitgeber – Vergleich mit Goldstandard



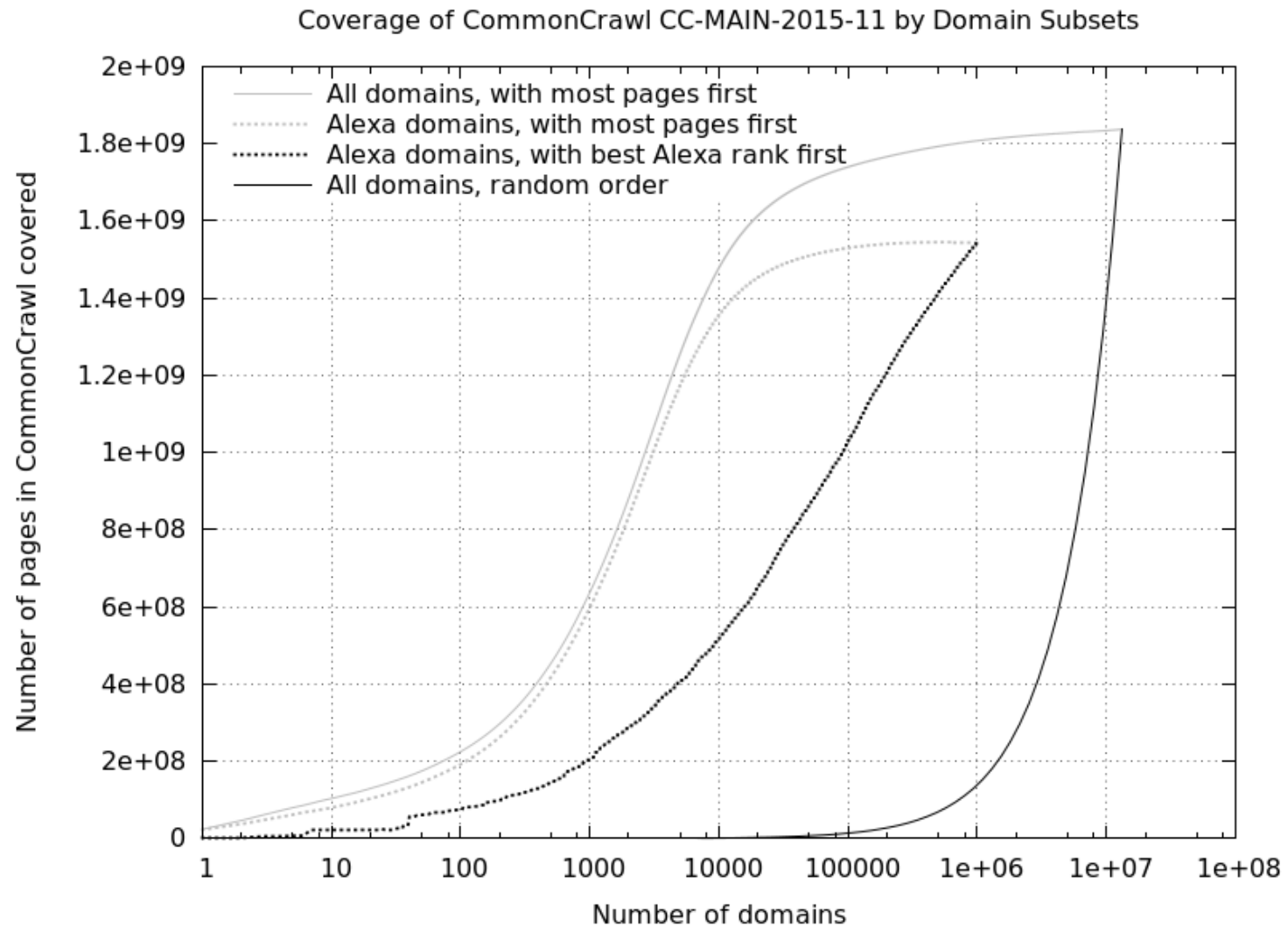
Pink: von Experten erstellt

Blau: von Arbeitern erstellt

# Korpuskonstruktion

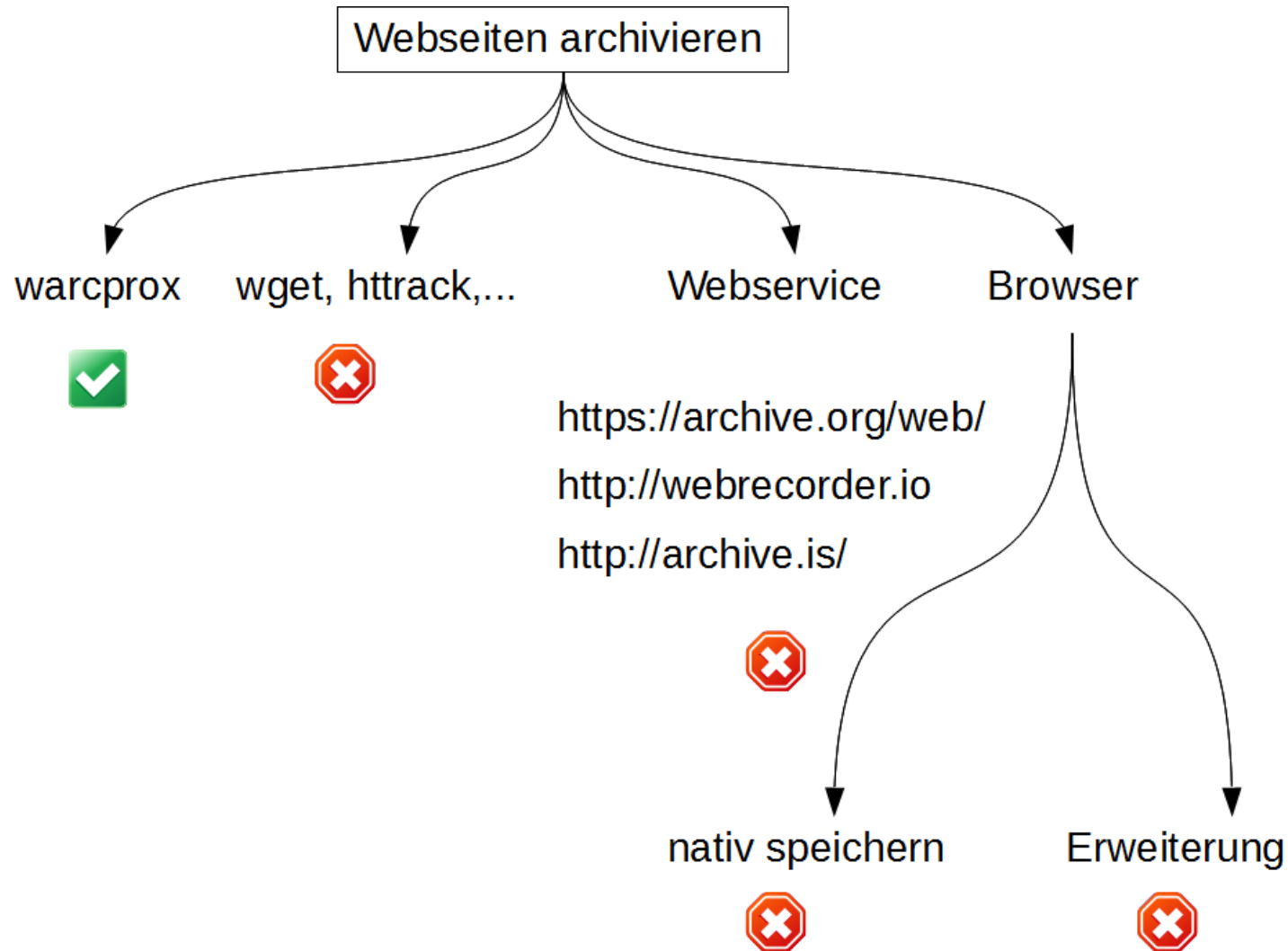
- Archivierte Webseite
- Zugehörige Segmentierungen
  - Rechtecke
  - Liste von xPaths zu den DOM-Fragmenten
- Screenshot von Live- und Archivversion der Webseite

# Korpuskonstruktion: Quelle der Webseiten

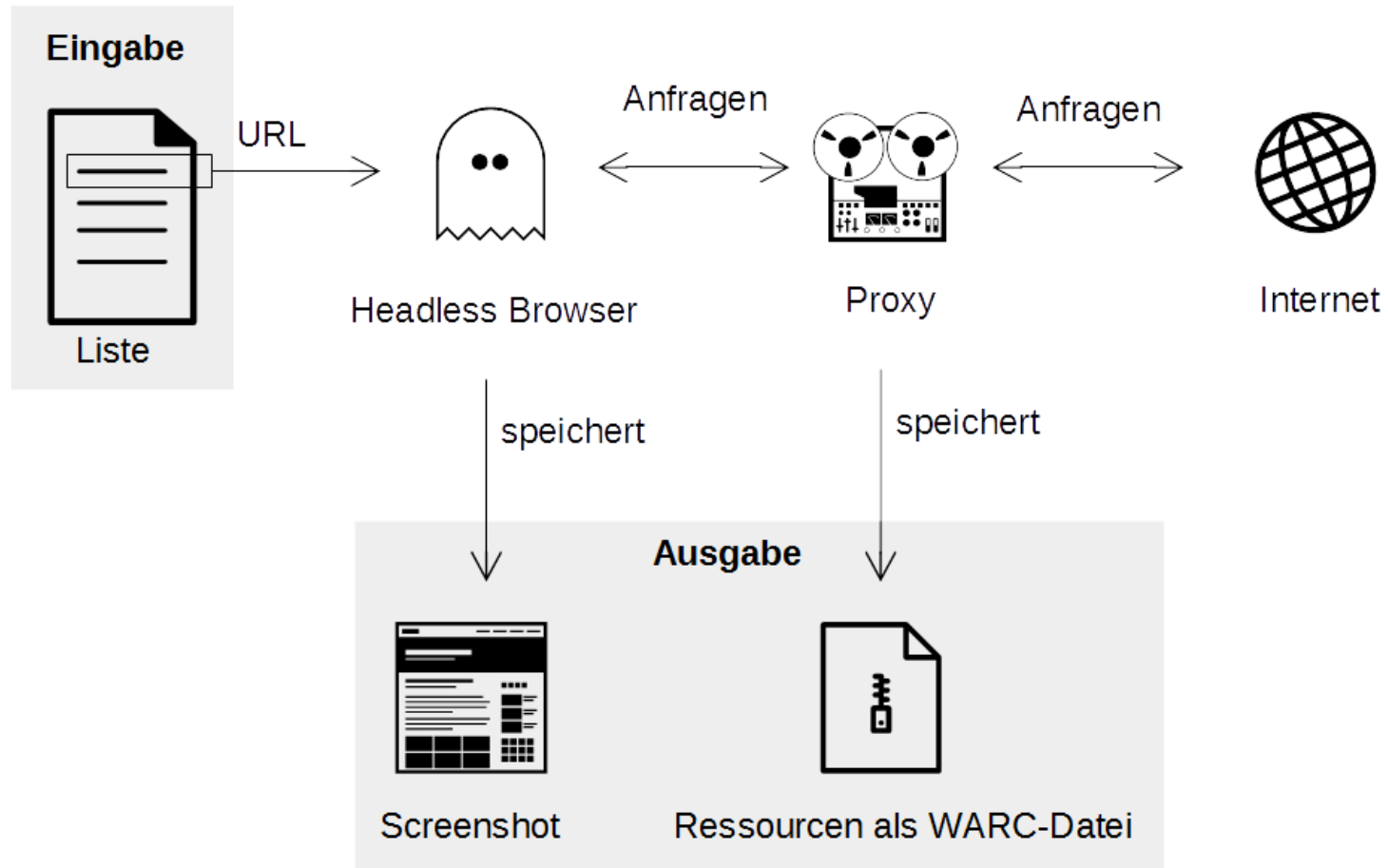


# **Langzeitarchivierung**

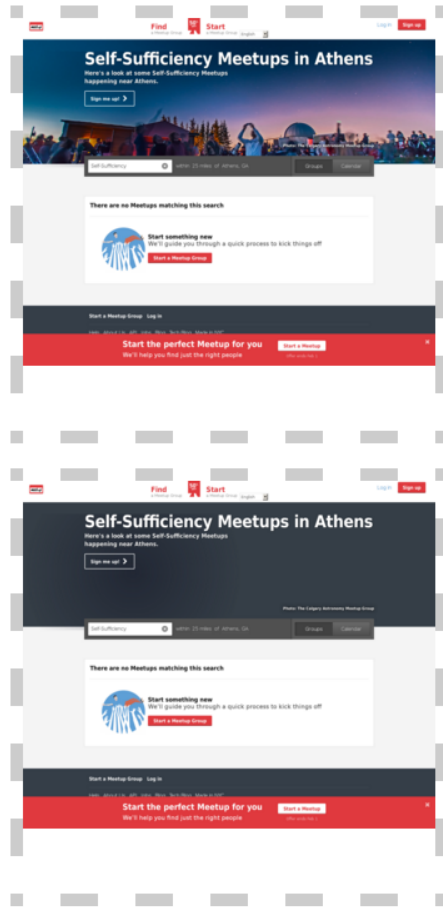
# Langzeitarchivierung: Qualität



# Langzeitarchivierung: Zusammenspiel der Komponenten



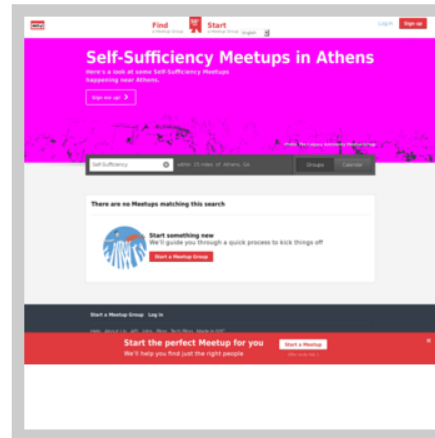
# Korpuskonstruktion: Werkzeug Resemble.js



## Compare two images?

Drop two images on the boxes to the left. The box below will show a generated 'diff' image, pink areas show mismatch. This example best works with two very similar but slightly different images. Try for yourself!

Don't have any images to compare? [Use example images](#)



[Ignore nothing](#) [Ignore colors](#) [Ignore antialiasing](#)

[Pink](#) [Yellow](#)

[Flat](#) [Movement](#) [Flat with diff intensity](#) [Movement with diff intensity](#)

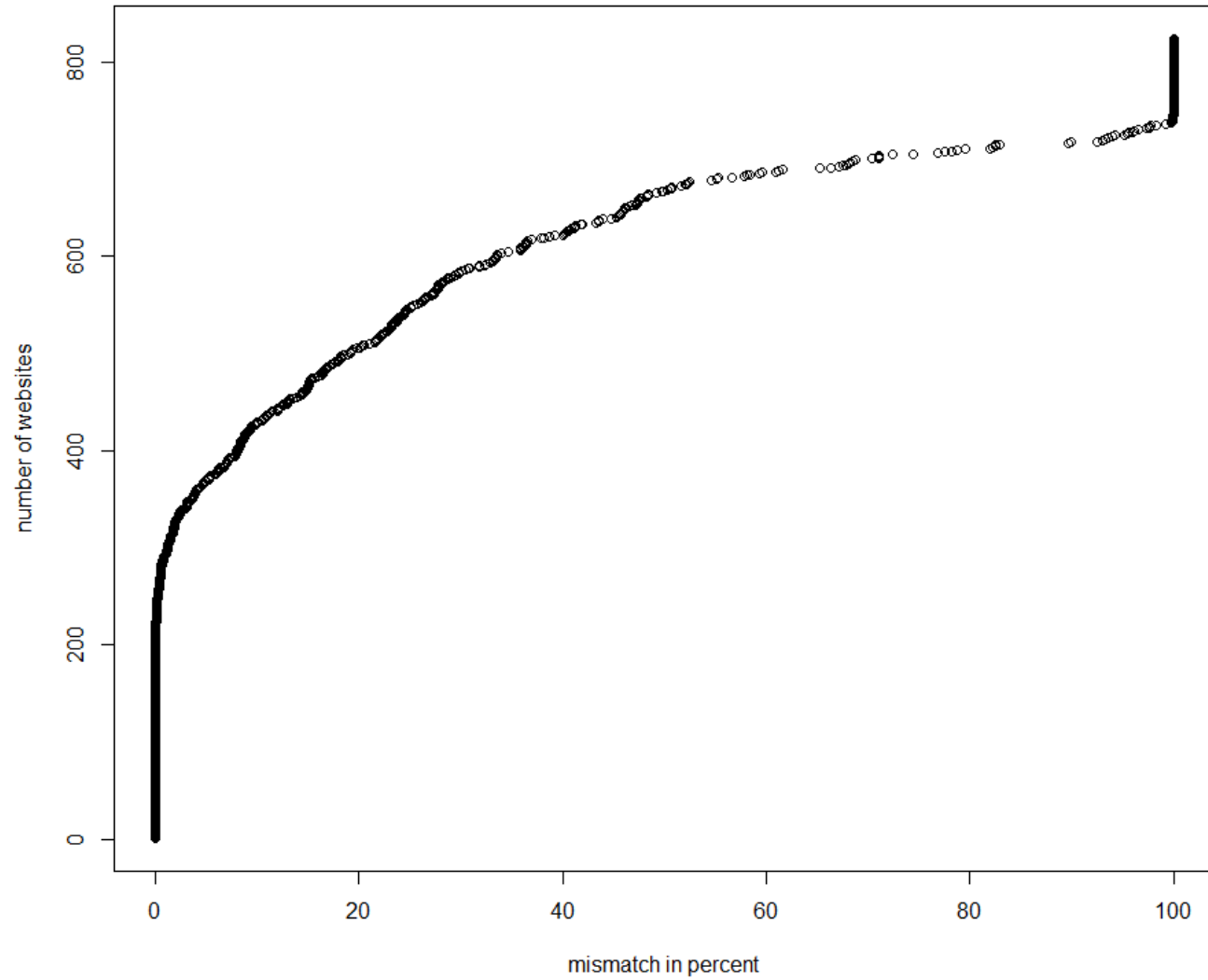
[Opaque](#) [Transparent](#)

**The second image is 32.02% different compared to the first.**

Use the buttons above to change the comparison algorithm. Perhaps you don't care about color? Annoying antialiasing causing too much noise? Resemble.js offers multiple comparison options.

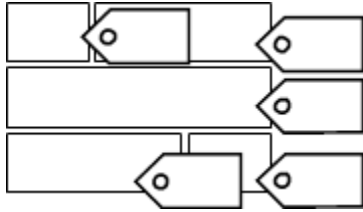


# Langzeitarchivierung: Qualität



- Vorteile:
  - Archivierung und Wiedergabe von clientseitig rendernden Webseiten
  - Größtmöglicher Anteil an Originalressourcen

- Verbesserte Überführung von Rechtecken auf DOM-Fragmente
- Erstellung eines Korpus
- Kategorisieren der Segmente



**Danke für Ihre Aufmerksamkeit!**

**Fragen?**