



# Nerdball

## Ein automatischer IRC Image Collector

Aleksander »watz«Paravac

Nerd2Nerd  
watz@nerd2nerd.org  
<http://www.nerd2nerd.org>



# Übersicht

- 1 Wieso und warum?
  - Was ist Nerdball?
  - HOWTO
- 2 Umsetzung
  - The Choice is yours
  - Client-Server Prinzip
- 3 Probleme
  - A bunch of Problems



# Übersicht

## 1 Wieso und warum?

- Was ist Nerdball?
- HOWTO

## 2 Umsetzung

- The Choice is yours
- Client-Server Prinzip

## 3 Probleme

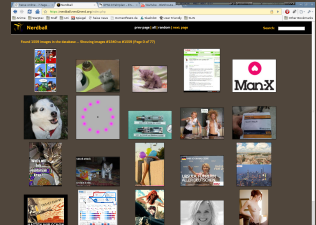
- A bunch of Problems



# Was ist Nerdball?

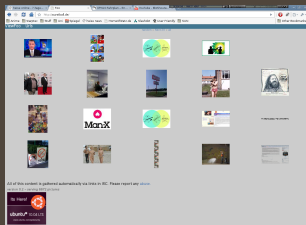
## Nerdball

- Jemand postet einen Link zu einem Bild im IRC
- Nerdball zieht das Bild
- Nerdball baut ein Thumbnail
- und sammelt die Bilder auf einer Webseite





# Was ist Nerdball?

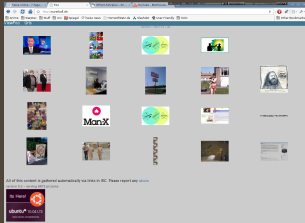


## Warum?

- 1 Weil es den Wurstball gibt
- 2 Weil wir sowas auch haben wollten
- 3 urs hat nen Preis ausgelobt
- 4 ...
- 5 Profit
- 6 Wurstball fehlen Features



# Was ist Nerdball?

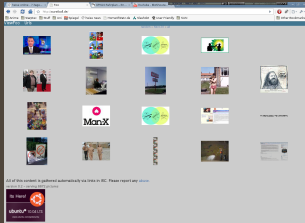


## Warum?

- 1 Weil es den Wurstball gibt
- 2 Weil wir sowas auch haben wollten
- 3 urs hat nen Preis ausgelobt
- 4 ...
- 5 Profit
- 6 Wurstball fehlen Features



# Was ist Nerdball?

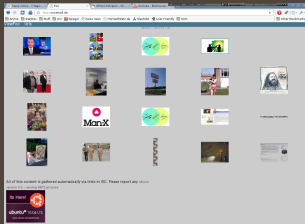


## Warum?

- 1 Weil es den Wurstball gibt
- 2 Weil wir sowas auch haben wollten
- 3 urs hat nen Preis ausgelobt
- 4 ...
- 5 Profit
- 6 Wurstball fehlen Features



# Was ist Nerdball?



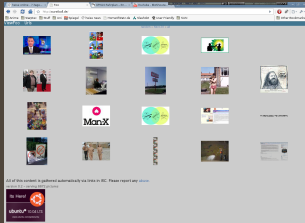
## Warum?

- 1 Weil es den Wurstball gibt
- 2 Weil wir sowas auch haben wollten
- 3 urs hat nen Preis ausgelobt
- 4 ...
- 5 Profit
- 6 Wurstball fehlen Features





# Was ist Nerdball?

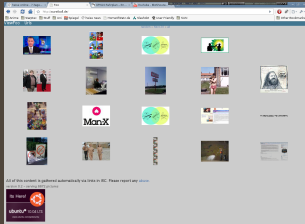


## Warum?

- 1 Weil es den Wurstball gibt
- 2 Weil wir sowas auch haben wollten
- 3 urs hat nen Preis ausgelobt
- 4 ...
- 5 Profit
- 6 Wurstball fehlen Features



# Was ist Nerdball?



## Warum?

- 1 Weil es den Wurstball gibt
- 2 Weil wir sowas auch haben wollten
- 3 urs hat nen Preis ausgelobt
- 4 ...
- 5 Profit
- 6 Wurstball fehlen Features



# Was braucht man dazu?

## Ingredients

### Die üblichen Verdächtigen

- Einen IRC Bot oder IRC Logs aus den man URLs matchen kann
- Eine Scriptsprache seiner Wahl
- Eine Grafikbibliothek um Bilder zu bearbeiten
- Einen Webserver seiner Wahl
- Eine Datenbank seiner Wahl für das Tagging



# Übersicht

## 1 Wieso und warum?

- Was ist Nerdball?
- HOWTO

## 2 Umsetzung

- The Choice is yours
- Client-Server Prinzip

## 3 Probleme

- A bunch of Problems

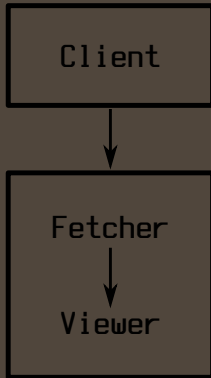


## Was wurde verwendet

- irssi
- PHP
  - PEAR/MDB2
  - curl
- imagemagick
- MySQL
- Apache



# Client-Server Prinzip



## Client

- sammelt URLs
- Schickt URLs an den Server

## Server

- fetched URLs
- baut Thumbnails
- generiert Webseite



# Probleme

## 1 Wieso und warum?

- Was ist Nerdball?
- HOWTO

## 2 Umsetzung

- The Choice is yours
- Client-Server Prinzip

## 3 Probleme

- A bunch of Problems



# Worauf man achten sollte

Manchmal steckt der Teufel im Detail ...

## Stoplersteine

- Kaputte HTTP-Header
- Restriction Nazis
- Doppelt gemoppelt hält besser
- Imagemagick





# Kaputte HTTP-Header

## Informationen über das Bild

Bevor man ein Bild herunterlädt, wäre es gut einiges zu wissen

- Mime-Typ
- Filesize

## Idee!

Fragen wir doch über den HTTP-Header entsprechende Daten ab.



# Kaputte HTTP-Header

## Wie es sein soll

```
socat - TCP4:watz.nerd2nerd.org:80,crnl  
HEAD /bier.jpg HTTP/1.1  
Host: watz.nerd2nerd.org  
  
HTTP/1.1 200 OK  
Date: Sat, 12 Jun 2010 12:26:42 GMT  
Server: Apache  
Last-Modified: Wed, 09 Jun 2010 16:32:21 GMT  
ETag: "39003b-c13a3-4889b7334ef40"  
Accept-Ranges: bytes  
Content-Length: 791459  
Content-Type: image/jpeg
```



# Kaputte HTTP-Header

## Kaputt

```
socat - TCP4:media.fukung.net:80,crln
HEAD /images/15761/6cbb1a49466dc8bb5ea459c02087aa3b.jpg HTTP/1.1
Host: media.fukung.net

HTTP/1.1 200 OK
X-Powered-By: PHP/5.2.6-1+lenny8
Last-Modified: Sat, 12 Jun 2010 08:27:09 -0400
Accept-Ranges: bytes
Pragma: public
Cache-Control: maxage=1209600
Expires: Sat, 26 Jun 2010 12:27:09 GMT
Content-Type: image/gif
Date: Sat, 12 Jun 2010 12:27:09 GMT
Server: lighttpd/1.4.19
```



# Was denn nun?

## Kaputt

```
wget http://media.fukung.net/images/15761/
6cbb1a49466dc8bb5ea459c02087aa3b.jpg
Resolving media.fukung.net... 208.53.170.43
Connecting to media.fukung.net|208.53.170.43|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [image/gif]
Saving to: "6cbb1a49466dc8bb5ea459c02087aa3b.jpg"
[...]

file 6cbb1a49466dc8bb5ea459c02087aa3b.jpg
6cbb1a49466dc8bb5ea459c02087aa3b.jpg: JPEG image data,
JFIF standard 1.01
```

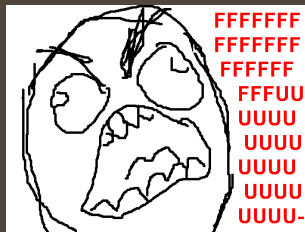


# LOL, wut?

## text/html

```
socat - TCP4:apina.biz:80,crlf
HEAD /30812.gif HTTP/1.1
Host: apina.biz

HTTP/1.1 200 OK
[...]
Content-Type: text/html; charset=UTF-8
```





# LOL, wut?

## text/html

```
GET /30812.gif HTTP/1.1
[...]
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Frameset//EN"
    "http://www.w3.org/TR/xhtml1/DTD/xhtml1-frameset.dtd">
<html xmlns="http://www.w3.org/1999/xhtml"
    xml:lang="en" lang="en">
<head>
<title>http://apina.biz/30812.gif</title>
</head>
<body style="text-align: center; margin: 0 auto; padding: 0;">
    <a href="/">APINA</a> - <a href="http://apina.biz/30812.gif">
http://apina.biz/30812.gif</a><br /><br />
    <a href="http://apina.biz/30812.gif">
    </a>
</body>
</html>
```



# Du kommst hier nicht rein

## Cookies, Referer und User Agents

Ein paar Seiten lassen Bilder nicht direkt per wget/curl ziehen. Sie achten auf

- Session Cookies!
- Kommt man vom richtigen Referer?
- Ist das ein Skript oder ein Browser, der da nachfragt?

Gut, dass z.B. CURL das abkann!



# Du kommst hier nicht rein

## Ein paar nette CURL Optionen

```
$ch = curl_init($url);
$curl_param = array(
    CURLOPT_RETURNTRANSFER => true,
    CURLOPT_TIMEOUT => 120,
    CURLOPT_USERAGENT => 'Mozilla/5.0 (Windows; U; Windows NT 6.1;
        en-US; rv:1.9.1.2) Gecko/20090729 Firefox/3.5.2 GTB5',
    CURLOPT_REFERER => dirname($url),
    CURLOPT_FOLLOWLOCATION => true,
    CURLOPT_MAXREDIRS => 2
);
curl_setopt_array($ch, $curl_param);
$result = curl_exec($ch);
```





# Doppelte Bilder und Reposts

## Das Problem ist vor dem Monitor

Wie vermeidet man doppelte Bilder durch Reposts?

- War die URL schon mal da?
- Hashes des Bildes

## Gute Idee, aber?

- “Wasserzeichen” im Bild
- Unterschiedliche Bildgrößen
- “Alzheimer”



# It's Magick!

## convert it your way

Leider hat “convert” auch ein paar kleine Probleme:

- Probleme beim resize von GIFs
- Endet ein GIF auf “.jpg”, so baut er “.jpg” Thumbnails für jedes Frame im GIF.

## Daher:

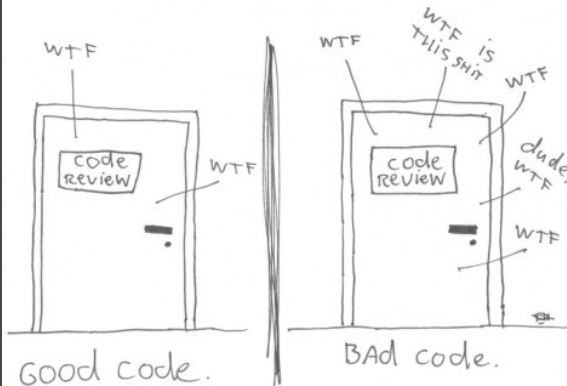
Use `-identify`!

```
convert -identify <input> -adaptive-resize <geometry> <output>
```



# Fragen?

The ONLY VALID MEASUREMENT  
OF CODE QUALITY: WTFs/MINUTE



(c) 2008 Focus Shift