Bücher machen mit Python

Eine praktische Einführung

Georg Hennemann, DZUG Rheinland
ghennemann@onlinehome.de

Juni 2006

.

mein FOSSiler Werdegang

- Jahrgang 1966 (Fußball-WM England-Deutschland 4:2-Sieg nach Verlängerung)
- 20 Jahre totale Computerabstinenz (gl. K.)
- Coherent Unix, SunOs, GNU Software seit 1986
- Linux seit 1990, Debian seit 1996
- Python/Zope seit 2000
- PDA Zaurus seit 2003
- Gnome & GPE 2004
- Plucker 2006

Themen die mich bewegen

- Wann kommt IPv6?
- Mobiles Internet
- Bücher und Software in der Public Domain
- freie eBook-Formate

Themen die mich bewegen

- Wann kommt IPv6?
- Mobiles Internet
- Bücher und Software in der Public Domain
- freie eBook-Formate
- mein erster Marathonlauf im Herbst ;)

Plucker

Pluck /Pluck/, v. t. [imp. & p. p. {Plucked};
p. pr. & vb. n. {Plucking}.]
[AS. pluccian; akin to LG. & D. plukken, G.
pfl["u]cken, Icel. plokka, plukka, Dan. plukke, Sw. plocka.?27.]

1. To pull; to draw.
[1913 Webster]

Mind The Gap: ’pluck’ <> ’plug’

Mind The Gap: ’pluck’ <> ’plug’

Plug ist ein Stecker oder Stöpsel

to pull the plug - den Stecker/Stöpsel ziehen

eBook-Formate

- Binäre Formate: Microsoft .lit, RocketBook .rb, Plucker .pdb etc.
- Mischformate: .pdf (Seitenbeschreibungssprache)
- Text Formate: FictionBook (xml), TAI (Lite), usw.

Binäre Formate

Vorteile

- Dokument kann maschinell erstellt werden (aus mehreren HTML- und Bilddateien)
- automatischer Ablauf (Build) eines eBooks, ePapers
- gute Kompression (DB-Format)

Nachteile

- Metadaten nicht im Dokument
- spezieller eBook-Reader erforderlich

Mischformate

Vorteil

- sieht überall gleich bescheiden aus (pdf)
- Seitenumbrüche bei Versionen für PDA

Nachteil:

- hohe Ressourcenverbrauch, langsam

Text-Formate (XML)

Vorteile

- Metadaten im Dokument
- well-formed, daher einfach gebauter Parser
- gutes Austauschformat (DRM kompatibel)
- Präsentation unabhängig von Inhalt
- notfalls auch ohne eBook-Reader lesbar

Nachteile

- Rendering, Erstellen des Inhaltsverzeichnisses im Reader
- Autor hat die gesamte Last der Dokumenterstellung
- Bilder müssen in das xml-Dokument eingefügt werden (base64 Encoding)

Was ist Plucker

- Datensauger (Download von Webinhalten)
- Parsen und Transformieren in das Plucker Format
- Syncronisation mit Palm/PDA
- off-line HTML Viewer, eBook Reader
- "Plucker PalmOS Document" eBook Format
- offen und frei (Copylefted)

Was Plucker kann

- optimiert für kleine Anzeigen
- History Funktion
- Scrollen mit Button oder Pen
- Anzeige bereits besuchter Links
- Darstellung von Bilder
- Named Anchors, zB <a href="#here">
- gute Kompression
- Suchfunktionen im Text und allen Texten der DB
- Bookmarks

Was Plucker nicht kann

- verschiedenen HTML-Tags <font>, <sup>, etc.
- Javascript, DHTML, Java, CSS
- Frames

Was ist noch zu erwarten ?

Desktop Integration in eine freie Desktop Umgebung wie Gnome/KDE

- Drag&Drop einer URL in das PGA-Gerät zum späteren Offline Lesen

Plucker‘s Konkurrenz

- AvantGO ist ein proprietärer offline-Reader für PDA’s
- komprimierte und verschlüsselte HTML-Seiten werden auf dem PDA geparst und dargestellt
- viele Channels auch für Plucker verwertbar :)

In 3 Schritten zum Buch

- WebInhalte herunterziehen und ins Plucker Format übertragen (Desktop)

In 3 Schritten zum Buch

- WebInhalte herunterziehen und ins Plucker Format übertragen (Desktop)
- Pluckerdateien vom PC auf Handheld übertragen

In 3 Schritten zum Buch

- WebInhalte herunterziehen und ins Plucker Format übertragen (Desktop)
- Pluckerdateien vom PC auf Handheld übertragen
- Pluckerdateien lesen (PDA)

Fertig eingerichtet laufen die Schritte 1 und 2 vollends automatisch ;)

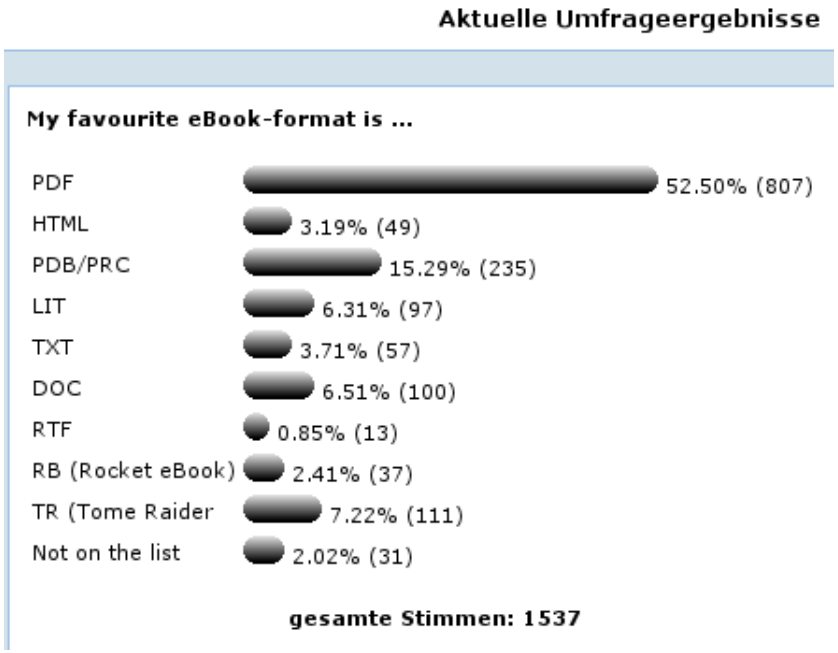
System-Anforderungen für Plucker

Plucker ist in Python programmiert

- Python
- PIL (Python Image Library)
- ZLib (Kompression Library)

Plucker läuft überall wo Python läuft

Beliebte eBook Formate



Sitescooper

- Sitescooper (perl, GNU GP License)
- aktuelle Version 3.1.2 (2001)
- kennt mehrere eBook Formate (Plucker, iSilo, Palm DOC)
- großes Repository an Newsquellen, aber kaum gepflegt
- Mailingliste: http://sitescooper.org/mailman/listinfo/sitescooper-talk
- Aktuelle Scoops http://scoops.sitescooper.org/

Perl und Python in Harmonie

Sitescooper (Perl) als Frontend für Plucker (Python)

- sitescooper hat keine eigene Funktion für das Plucker-Format
- mit der Option -plucker bzw. -mplucker wird das Python-Script plucker-build aufgerufen das die Umwandlung in das Pluckerformat erledigt

- Das ist eine gelungene Arbeitsteilung.

Plucker unterstützt keine RSSFeeds :(

Plucker unterstützt keine RSSFeeds :(

dafür aber Sitescooper :)

- rss-to-site.pl Perl-Script konvertiert URL eines RSS file in ein Sitescooper .site file
- rss-to-site.pl http://url.rss > whatever.site

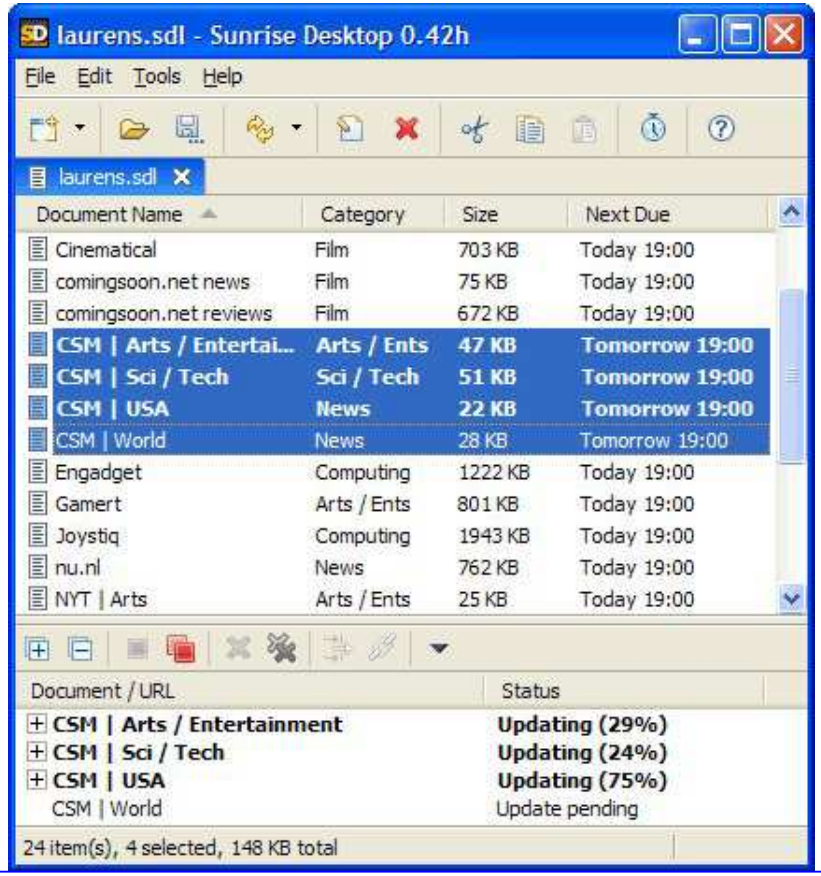
- Beispiel: http://www.spiegel.de/schlagzeilen/rss/0,5291,,00.xml

sitescooper -refresh -install /home/gh/scoops/ -mplucker -fixlinks -maxcolors 4

-

Plucker für Windows Clients

- Sunrise Projekt (java, BSD Lizenz, früher JPluck)
- Tool mit Java GUI, Version 0.42j, Febr 2006
- Homepage http://sourceforge.net/projects/sunrisexp/
- Installer für Windows XP Clients (funktioniert prima!)
- Javascript Scripting Engine (Rhino) für Dokument Konversion
- gute Wahl für Windows User ohne Shell-Kenntnisse


laurens.sdl - Sunrise Desktop 0.42h
File Edit Tools Help
laurens.sdl
Document Name Category Size Next Due
Cinematical Film 703 KB Today 19:00
comingsoon.net news Film 75 KB Today 19:00
comingsoon.net reviews Film 672 KB Today 19:00
CSM | Arts / Entertai... Arts / Ents 47 KB Tomorrow 19:00
CSM | Sci / Tech Sci / Tech 51 KB Tomorrow 19:00
CSM | USA News 22 KB Tomorrow 19:00
CSM | World News 28 KB Tomorrow 19:00
Engadget Computing 1222 KB Today 19:00
Gamert Arts / Ents 801 KB Today 19:00
Joystiq Computing 1943 KB Today 19:00
nu.nl News 762 KB Today 19:00
NYT | Arts Arts / Ents 25 KB Today 19:00
Document / URL Status
CSM | Arts / Entertainment Updating (29%)
CSM | Sci / Tech Updating (24%)
CSM | USA Updating (75%)
CSM | World Update pending
24 item(s), 4 selected, 148 KB total

Plucker Reader für Windows

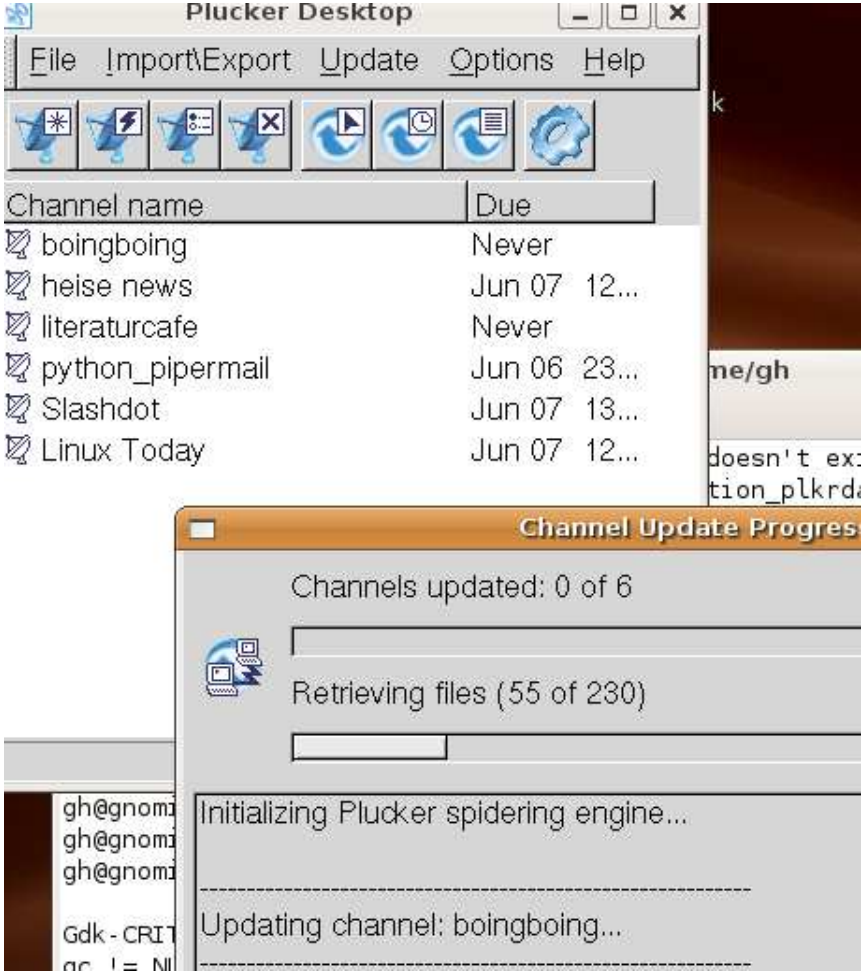
- Pocket PC: Vademecum ("geh mit mir")
- http://vade-mecum.sourceforge.net/

- Windows XP: Palm_OS_Simulator
- (habs nicht ausprobiert)

Plucker Desktop Tool

- grafische Benutzeroberfläche (wxWindows-GUI)
- Versionen für Linux, Windows, etc.
- Konfigurationsmenü für die Spidering Engine
- Setup-Wizard für Channels
- Statusanzeige bei Channel Updates
- Konfigurationsdatei ~/.pluckerc

Plucker Desktop beim Nachrichten Auffrischen



Plucker auf der Kommandozeile

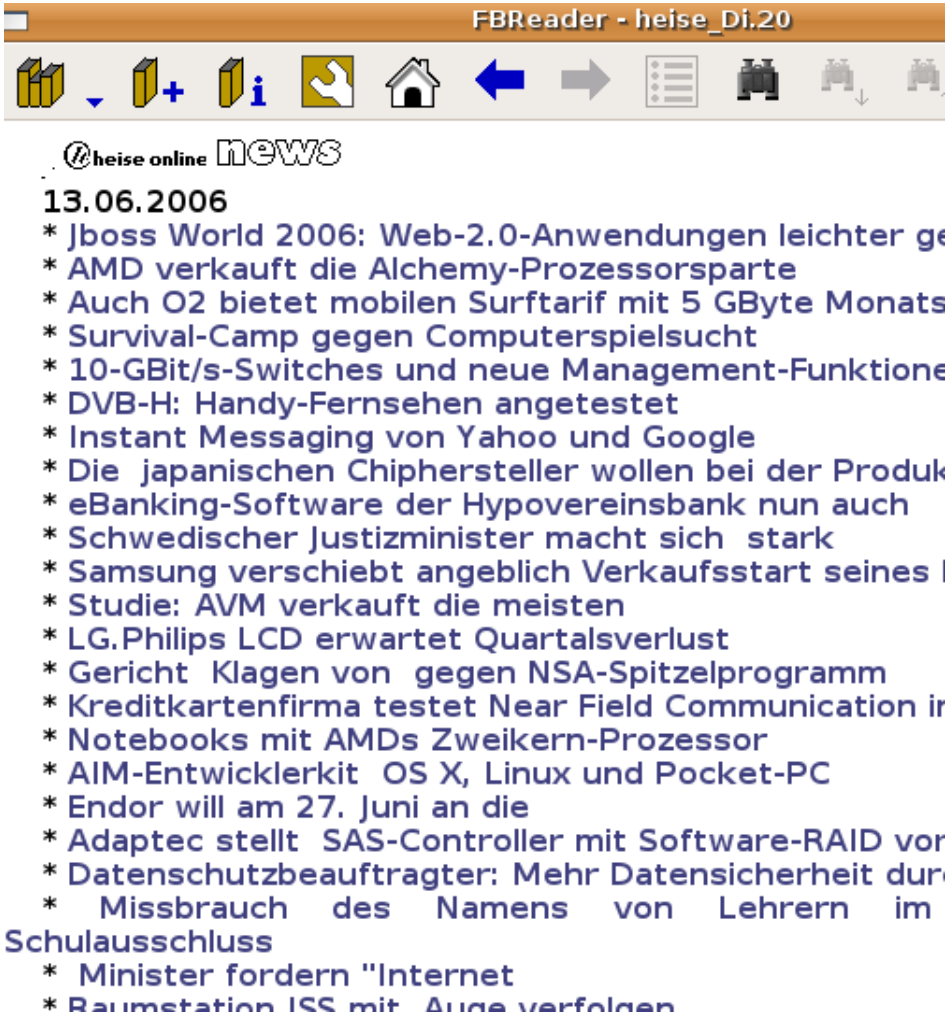
- plucker-build: Python-Skript mit vielen Optionen
- PLUCKERHOME: ~/.plucker/ (default)
- Start von plucker-build manuell oder über Crontab

Beispiel:

plucker-build -f heise_$(date +%a-%I)am -p ~/pnews/
http://www.heise.de/pda/newsticker/

erzeugt eine Datei ’heise_Mo-10am.pdb’ im Verzeichnis ~/pnews/

Und so siehts dann im FBReader aus



FBReader - offen, schnell und vielseitig

FBReader ist ein e-book Reader für Linux PDAs und Desktop Computer. FBReader arbeitet auf dem Sharp Zaurus, Siemens Simpad with Opensimpad ROM, Nokia 770 Internet Tablet und Linux Desktop Computern. FBReader unterstützt verschiedene e-book Formate: plucker, palmdoc, zTXT, HTML, fb2, TCR (psion text), OEB, RTF and plain text.

http://freshmeat.net/projects/fbreader

eBooks Lesen auf dem Linux Desktop

- pdf-Dokumente mit Acrobat-Reader
- alle anderen Formate mit dem FictionBook-Reader
- exclusiver Linux Reader, Windows User haben leider das Nachsehen ;)
- Debian-Packete von der FBReader-Homepage
- http://only.mawhrin.net/fbreader/desktop/
- Libs expat, bzip2, enca (Extremely Naive Charset Analyser)
- Alternativ Sourcen übersetzen und installieren

FBReader Formate

- fb2 e-book format (style attributes are not supported yet).
- Html format (tables are not supported).
- plucker format (embedded images are supported, tables are not supported)
- Palmdoc (aportis doc)
- zTxt (Weasel format)
- TCR (psion text) format
- RTF format (stylesheets and tables are not supported)
- OEB format (css and tables are not supported)
- Plain text format

FBReader Goodies I

- direkt Lesen aus tar, zip, gzip und bzip2 Archiven
- Mehrere Bücher in einem Archiv
- Encodings: utf-8, us-ascii, windows-1251, windows-1252, koi8-r, ibm866, iso-8859-*, Big5, GBK
- automatische Erkennung des Encodings (naja, klappt wohl nicht immer)
- Automatisch erstelltes Inhaltsverzeichnis
- Unterstützung von eingebetteten Bildern
- Fußnoten/Hyperlinks
- Positions Indikator

FBReader Goodies II

- Behält die letzte Seite und die letzte Leseposition für alle geöffneten Bücher zwischen Sitzungen.
- Liste der zuletzt geöffneten Bücher
- Automatische Hyphenation (gleicher Algorithmus wie in TeX/LaTex)
- Hyphenation Patterns für Tschechisch, Englisch, Esperanto, Französisch, Deutsch und Russisch
- Textsuche im Text und in Bibliothek
- Full-screen Modus
- Screen Rotation in 90, 180 and 270 Grad

FBReader geplante Features

- Dictionary Integration
- Automatisches Scrollen
- Lesezeichen
- HTML-Tabellen
- andere e-book Formate
- text-to-speech Plugin Flite (Festival Lite)

FBReader Sourcen kompilieren

- Libs: libgtk2.0-0, libexpat1, libenca0, libbz2-1.0,libbz2-dev,
- fbreader-sources-0.7.4c.tgz auspacken
- Targetplatform (desktop, zaurus) und GUI (gtk, QT) anpassen
- make, make install

Plucker Links

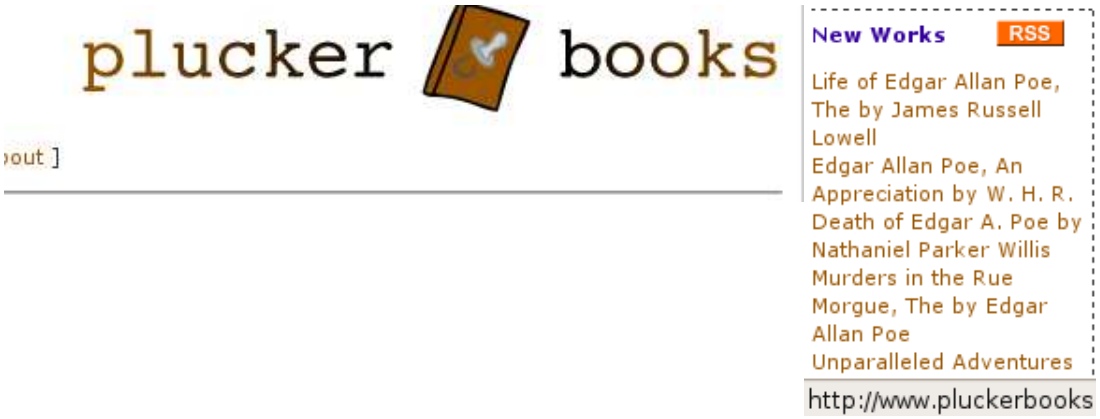
Palmtop-Portal

http://www.palmtop-portal.de/
deutschsprachige Site mit HyperLink-Datenbank zu diversen Themenbereichen

Plucker Books

http://www.pluckerbooks.com/

nette Site mit Public Domain Büchern für den "besten" eBook-Reader



Palm Docs

http://www.xecu.net/bcollins/PalmDocs.htm

Auswahl an klassichen und Science-Fiction Texten

Memoware

http://www.memoware.com/

große Auswahl an freien und käuflichen eBooks, Online-Shop

Fiction Book

http://fictionbook.ru/en/

- grosse Auswahl von Englischen & Russischen Büchern im FictionBook Format
- sortiert nach Autor und Titel (Thesaurus)
- darunter auch einige hier nicht frei zugängliche Titel

Allgemeine eBook und MobileReader Links

http://www.mobileread.com/

Copyright und Public Domain

- Rechtslage von Land zu Land unterschiedlich
- was in Russland erlaubt ist ist hier noch lange nicht erlaubt

Einstiegsseite

- http://onlinebooks.library.upenn.edu/okbooks.html

Neue Bücher in der ’Public Domain’

Onlinebooks
- http://onlinebooks.library.upenn.edu/new.html
- Liste mit Neuerscheinungen von Büchern, Aufsätzen usw. in der Public Domain

Gutenberg Projekt
- http://www.gutenberg.org/
- 18.000 freie Bücher im Online Katalog

Public Domain Books gefunden mit Google
- http://zuhause.org/dp/gfound1.html

Übungen mit plucker-build

plucker-build -M 5 --stayonhost -f itlexikon -p /home/gh/plucker
file:///home/gh/Desktop/eBooks/Lexikon/Lexikon/anfang.htm

plucker-build -M 2 --stayondomain --staybelow="http://de.wikipedia.org/wiki" -f
ipv6lexikon -p /home/gh/plucker http://de.wikipedia.org/wiki/IPv6

plucker-build -f golem$(date +%u) -p /home/gh/plucker
http://www.golem.de/pda/pdahome.html

Übungen mit sitescooper

rss-to-site http://www.spiegel.de/schlagzeilen/rss/0,5291,,00.xml > de_spiegel_rss.site
cp de_spiegel_rss.site /usr/share/sitescooper/site_samples/regional_germany/
sitescooper -refresh -install /home/gh/scoops/ -mplucker -fixlinks -maxcolors 4