

# OCR mit cuneiform



Jürgen Weigert  
openSUSE user  
2010-06-01



Novell.



# OCR - Optical Character Recognition

---

Was gibt es für Linux?

- GOCR / KOOCR
- Tesseract
- Cuneiform
- ...



# Was gibt es da zu Hacken?

---

- Ausprobieren, dokumentieren was geht
- Trainieren, Erkennungsraten erhöhen
- GUI? - oft Fehlanzeige → Klaas fragen
- Scripte für Formatkonvertierungen
  - Dateiformat hocr ← offener Standard in XML
- ...



# Scannen für cuneiform

- Graustufen TIFF, oder schwarz/weiss
  - Nicht “mTIFF” mit unseren OfficeScannern!
- 200 oder 300 dpi
  - Mehr ist besser ...
- Seitenorientierung muss stimmen
  - Wenn nötig mit gimp drehen
- Fotos in Graustufen umwandeln,
  - Auf gleichmäßige Ausleuchtung achten



# Software Installation

---

## cuneiform-0.8.0

Aus `home:jnweiger:branches:home:Lazy_Kent / openSUSE_11.2, openSUSE_11.3`

- In Version 0.9.0 ist die Ausgabe von hocr defekt: [bugs.launchpad.net/cuneiform-linux/+bugs/548801](http://bugs.launchpad.net/cuneiform-linux/+bugs/548801)
- In Version 0.8.0 führen PNG Bilder zu SEGFAULT

**exactimage-0.8.0** aus `home:jnweiger`

Weitere “soft dependencies”

→ `scanimage, gimp, xv, gwenview, okular, ...`

Vorführung



# Projektidee “wo-ist-oben?”

---

- Ein Script um scanimage
  - Rotieren um  $0^\circ$ ,  $90^\circ$ ,  $180^\circ$  und  $270^\circ$
  - In allen vier Richtungen OCR laufen lassen
  - Erkennungsrate jeweils messen
  - Das mit der höchsten Rate ist richtig herum!



# Projektidee “Sandwich-PDF”

- Office-Scanner schickt PDF per e-mail
- PDF auspacken → nach TIFF konvertieren
- Projektidee “wo ist oben” anwenden

```
$ cuneiform -l ger -f hocr -o X.hocr X.tif  
$ hocr2pdf -i X.tif -o X.pdf -s < X.hocr
```





# Projektidee “Sandwich-PDF”

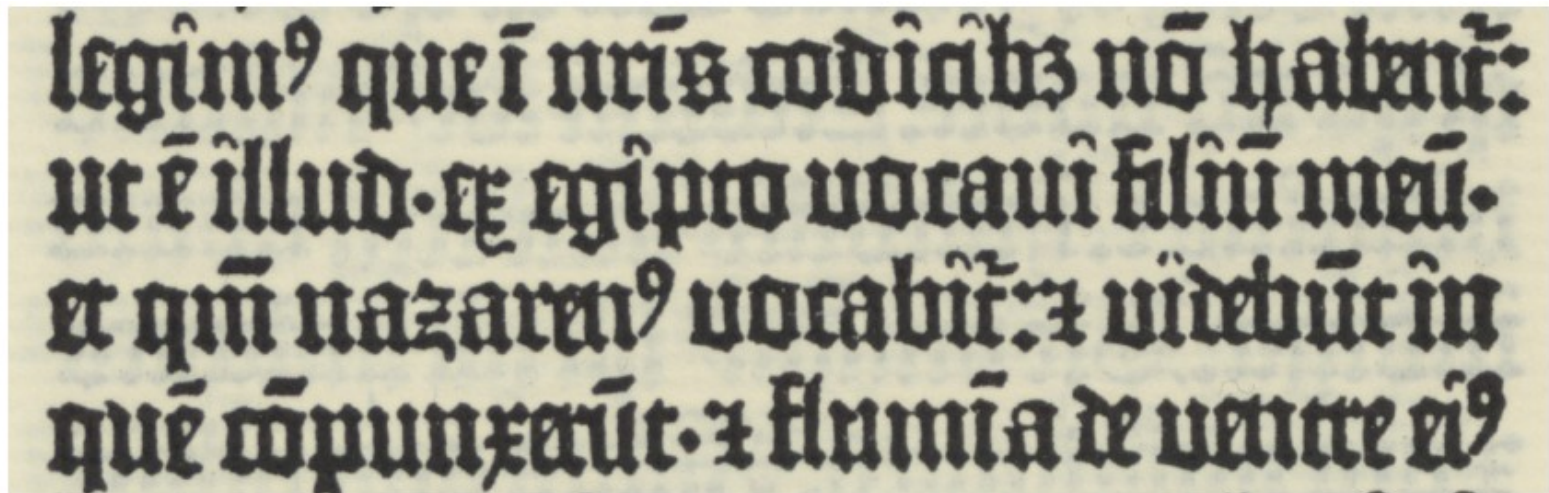
- Office-Scanner schickt PDF per e-mail
- PDF auspacken → nach TIFF konvertieren
- Projektidee “wo ist oben” anwenden

```
$ cuneiform -l ger -f hocr -o X.hocr X.tif  
$ hocr2pdf -i X.tif -o X.pdf -s < X.hocr
```

- Automatisieren mit einem e-mail Dienst  
Office-Scanner schickt an <username>@ocr

# Projektidee “Geheimschrift Fraktur”

- Einen Font für gotische Schriften trainieren
- Besser lesbar nach OCR?

A sample of Gothic Fraktur script, showing a passage of Latin text in a highly stylized, blackletter font. The text is arranged in four lines on a light-colored background.

legim⁹ que ī nr̄is codicib⁹ nō habent:  
ut ē illud. ex egipto vocavi filiū meū.  
et quū nazaren⁹ vocabit⁹: ⁊ uidebūt in  
quē cōpuxerūt. ⁊ flumīa de uentre ei⁹

# Die Typen der 42zeiligen Bibel

## a. Gutenbergtypen

À	B	C	D	E	F	G	H	I		J	K	L	M	N	O	P	Q	R	S		
À	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S			
						T	U	V	X	Y	Z										
						T	U	V	X	Y	Z										
a	a	ā	ā	ā	ā	a'	a'	b	b	b	ba	ba	bā	bā	be	be	bē				
bo	bo	bo	bo	c	c	c̄	c̄	c̄	c̄	c̄	c̄	c̄	c̄	d	d	d	d	d̄	d̄		
da	da	dā	dā	de	de	dē	dē	dē	dē	do	do	e	e	ē	ē	ē	ē				
e	e	e'	e'	f	f	ff	ff	g	g	ḡ	ḡ	ḡ	ḡ	ḡ	h	h	h				
ha	ha	hā	hā	he	he	ho	ho	i	i	i	i	i	i	i	i	i	i				
j	k	l	l	l'	l'	m	m	m̄	m̄	m̄	n	n	n̄	n̄	n̄	n̄	n̄				
o	o	ō	ō	p	p	p̄	p̄	p̄	p̄	p̄	p̄	p̄	p̄	p̄	p̄	p̄	p̄				
pp	pp̄	pp̄	pp̄	pp̄	pp̄	pp̄	q	q	q̄	q̄	q̄	q̄	q̄	q̄	q̄	q̄	q̄				
pp̄	pp̄	pp̄	r	r	r̄	r̄	r̄	r̄	r̄	r̄	r̄	r̄	r̄	r̄	r̄	r̄	r̄				
st	st	s	s	s	t	t	t̄	t̄	t̄	t̄	t̄	t̄	t̄	t̄	u	u	u				
ū	ū	ū	u	u	ū	u	u	w	x	e	y	y	z	o	z	z	z				
.	.	:	:	:																	

Viel Spass!

Novell®

## General Disclaimer

This document is not to be construed as a promise by any participating company to develop, deliver, or market a product. Novell, Inc., makes no representations or warranties with respect to the contents of this document, and specifically disclaims any express or implied warranties of merchantability or fitness for any particular purpose. Further, Novell, Inc., reserves the right to revise this document and to make changes to its content, at any time, without obligation to notify any person or entity of such revisions or changes. All Novell marks referenced in this presentation are trademarks or registered trademarks of Novell, Inc. in the United States and other countries. All third-party trademarks are the property of their respective owners.

This work is licensed under the Creative Commons Attribution-Noncommercial-Share Alike 2.5 License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/2.5/>.

For other licenses contact author.



**Novell.**