OCRFeeder

Converting printed documents into digital formats

Joaquim Rocha
jrocha@igalia.com

Berlin, May 2011
What is it?

Document Analysis and Optical Character Recognition for GNOME
Why?

Paper has a number of problems

No applications for GNU/Linux to do a fair job
Paper problems: Security
Paper problems: Preservation
Paper problems: Data processing
Paper problems: Ecology
No fair conversion apps for GNU/Linux

apart from OCR engines, but...
OCR != Document Conversion

(it only deals with chars)
(does not consider the layout)
(does not distinguish contents)
What's needed is Document Analysis and Recognition (conversion of documents to an electronic format) (first projects in the 80s)
Where are we at?

* Some closed solutions
* Only for proprietary systems
* Various prices
* still... arguable results
How

Images Input → Image Pre-processing → Layout Analysis

OCR Engines → Optical Character Recognition

Document Generation

Document Analysis and Recognition
So many layouts...
Layouts vary with the type of document

What works on detecting one, won't work on others
OCRFeeder focuses on contents, not on layouts!
Key concept:

If a document image can be divided in windows of 1 (content) or 0 (not content), then it is possible to group all the 1s and outline the contents.
Recognition:

System-wide OCR engines are used

Engines are configured from the GUI or XML files
Tesseract engine

Name: Tesseract
Image format: TIF
Failure string: 
Engine path: /usr/bin/tesseract
Engine arguments: $IMAGE $FILE; cat $FILE.txt; rm $FILE

Cancel  OK
Most known free OCR engines are detected and configured automatically:

* Tesseract
* GOOCR
* OCRAD
* Cuneiform
Exportation formats:

ODT
HTML
Plain text
User interaction:

Users can edit everything and review the algorithm's results.

So, UI can work in attended and unattended ways.

CLI only works in an unattended mode.
Demo time!
Other features:

* PDF importation
* Unpaper preprocessor
* Font style edition
* Image deskewing
* OCR results cleaning
* Project saving/loading
A11y:

* OCRFeeder is a very useful tool for visually impaired users
* Last year, the main target of its development was to improve a11y
Future:

* Integrate Ocropus as an alternative analysis backend
* More exportation formats: HOCR, PDF, etc.
* Make OCR engines' management easier
Webpage:  
http://live.gnome.org/OCRFeeder

git:  
http://git.gnome.org/ocrfeeder

Bugzilla:  
http://bugzilla.gnome.org  
product: OCRFeeder
Manual in German:

http://wiki.ubuntuusers.de/OCRFeeder
Thank you!