

Netzaktion Patenttexterkennung

<http://swpat.ffii.org/gruppe/epatext/index.de.html>

Arbeitsgruppe

swpatag@ffii.org

2003-10-04

Let's pool our computing resources so as to make the software patent documents of the European Patent Office (EPO) digitally accessible. So far the information is available only in the form of single-page graphic documents. Fortunately there is a free OCR program that can produce usable text output from these graphic files. We have already acquired most of the relevant graphical data. You can help us by locally running an OCR script on these CDs. The OCR output will then appear on our website as well as a ring of cooperating websites, depending on your configuration.

Inhaltsverzeichnis

1 Einführung	1
2 Kommentierte Verweise	2
3 Fragen, Aufgaben, Wie Sie helfen können	3

1 Einführung

We have a set of currently 15 CDs with EPO software patent graphic files. Turning all the graphics into text may take a week per CD on a 586 computer.

In the future we will have more and more CDs (up to 800) and we may wish to run the recognition software on them over and over again, as we adapt the software and/or its input data.

Therefore we need participation of some people who take charge of a few CDs each and run the script on them.

Moreover, we may also need people with good network bandwidth to help us download more graphic files for this and other projects.

In the long run, this volunteer network may be able to solve many worthwhile problems, wherever the upgrading of massive amounts of data is concerned, which, for whatever reason, were presented to the public in a crippled or otherwise unsatisfactory form.

2 Kommentierte Verweise

- **Europäische Softwarepatente: Umfassende Dokumentation¹**

Here you can already find many text versions of EPO software patents, some of which contain a section with OCR output from this action. But we are not providing the PDF graphics online, because we can't afford the fees which we have to pay when other people download our data.

- **Europäische Softwarepatente: Einige Musterexemplare²**

This collection contains some PDF files ready to be compared with the GOCR output.

- **Helft uns, das Plenum zu gewinnen!³**

This directory contains some more scripts related to this action.

¹<http://swpat.ffii.org/patente/txt/index.de.html>

²<http://swpat.ffii.org/patente/muster/index.de.html>

³<http://swpat.ffii.org/gruppe/intern/bin/index.de.html>

- **Testsuite für die Gesetzgebung über die Grenzen der Patentierbarkeit**⁴

Um eine Patentierbarkeitsrichtlinie auf Tauglichkeit zu prüfen, sollten wir sie an Beispiel-Innovationen ausprobieren. Für jedes Beispiel gibt es einen Stand der Technik, eine technische Lehre und eine Reihe von Ansprüchen. In der Annahme, dass die Beispiele zutreffend beschrieben wurden, probieren wir dann unsere neue Gesetzesregel daran aus. Unser Augenmerk liegt auf (1) Klarheit (2) Angemessenheit: führt die vorgeschlagene Regelung zu einem vorhersagbaren Urteil? Welche der Ansprüche würden erteilt? Entspricht dieses Ergebnis unseren Wünschen? Wir probieren verschiedene Gesetzesvorschläge an der gleichen Beispielserie (Testsuite) aus und vergleichen, welches am besten abschneidet. Für Programmierer ist es Ehrensache, dass man “die Fehler beseitigt, bevor man das Programm freigibt” (first fix the bugs, then release the code). Testsuiten sind ein bekanntes Mittel zur Erreichung dieses Ziels. Gemäß Art. 27 TRIPS gehört die Gesetzgebung zu einem “Gebiet der Technik” namens “Sozialtechnik” (social engineering), nicht wahr? Technizität hin oder her, es ist Zeit an die Gesetzgebung mit derjenigen methodischen Strenge heran zu gehen, die überall dort angesagt ist, wo schlechte Konstruktionsentscheidungen das Leben der Menschen stark beeinträchtigen können.

- **Helft uns, das Plenum zu gewinnen!**⁵

Zutaten, die wir brauchen, um das Europäische Parlament zu überzeugen, gegen die Softwarepatentrichtlinie zu stimmen, und wie Sie helfen können, dass es auch passiert - ein Brief an Unterstützer von FFII/Eurolinux

3 Fragen, Aufgaben, Wie Sie helfen können

- **Wie Sie uns helfen können, dem Logikpatent-Schildbürgerstreich ein Ende zu bereiten**⁶
- **maintain the CDs and the list of participants, send the CDs out**

⁴<http://swpat.ffii.org/analyse/testsuite/index.de.html>

⁵<http://swpat.ffii.org/gruppe/epatext/swnerfatri/index.de.html>

⁶<http://swpat.ffii.org/gruppe/aufgaben/index.de.html>

- **adapt gocr**

1. make it take b/w inverted graphics as input so as to eliminate the time-consuming pnminvert procedure
2. extend GOCR so that it can use existing similar texts (e.g. corresponding patent applications, for which text files already exist) to improve its recognition efficiency.
3. write configuration files that specify the structure of patent descriptions so that GOCR handles them better, e.g. recognises layout and text structures better. As needed, develop such configuration formats and/or improve gocr so as to use them and/or make them unnecessary in more and more cases.
4. write a frontend that lets gocr directly interact with PDF files and insert OCR results of PDF images into the files as specified by Adobe's PDF format.

- **create CDs with non-EP patents, e.g. DE, FR, GB, SE, JP, US etc on them**

Servers such as DepatisNet and others need to be studied.

Perhaps it is also possible to obtain some of this stuff for a reasonable price from the patent offices. (A few years ago the pricing was prohibitive: thousands of EUR/USD for the patents of one year. But this policy has probably changed. With enough perseverance, it should be possible to get everything very cheaply.)

- **write an MSWin version of the bash script if possible**

some people have win-specific ocr programs that may be worth trying out for comparison.