

# Patent OCR Net Action

<http://swpat.ffii.org/group/epatext/index.en.html>

Workgroup

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.

## Contents

<b>1</b>	<b>introduction</b>	<b>1</b>
<b>2</b>	<b>Annotated Links</b>	<b>2</b>
<b>3</b>	<b>Questions, Things To Do, How you can Help</b>	<b>3</b>

## 1 introduction

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 Annotated Links

- **European Software Patents: Comprehensive Documentation**<sup>1</sup>

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.

- **European Software Patents: Assorted Examples**<sup>2</sup>

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

- **Help us Win the Plenary Vote!**<sup>3</sup>

This directory contains some more scripts related to this action.

---

<sup>1</sup><http://swpat.ffii.org/patents/txt/index.en.html>

<sup>2</sup><http://swpat.ffii.org/patents/samples/index.en.html>

<sup>3</sup><http://swpat.ffii.org/group/internal/bin/index.en.html>

- **Patentability Legislation Benchmarking Test Suite**<sup>4</sup>

In order to test a law proposal, we try it out on a set of sample innovations. Each innovation is described in terms of prior art, a technical contribution (invention) and a small set of claims. Assuming that the descriptions are correct, we then test our proposed legislation on them. The focus is on clarity and adequacy: does the proposed rule lead to a predictable verdict? Which of the claims, if any, will be accepted? Is this result what we want? We try out different law proposals for the same test series and see which scores best. Software professionals believe that you should “first fix the bugs, then release the code”. Test suites are a common way of achieving this. Pursuant to Art 27 TRIPS, legislation belongs to a “field of technology” called “social engineering”, doesn’t it? Technology or not, it is time to approach legislation with the same methodological rigor that is applicable wherever bad design decisions can significantly affect people’s lives.

- **Help us Win the Plenary Vote!**<sup>5</sup>

The ingredients needed for persuading the European Parliament to vote against the Software Patent Directive and how you can help make it happen – a letter to FFII/Eurolinux supporters

### **3 Questions, Things To Do, How you can Help**

- **How you can help us end the software patent nightmare**<sup>6</sup>
- **maintain the CDs and the list of participants, send the CDs out**

---

<sup>4</sup><http://swpat.ffii.org/analysis/testsuite/index.en.html>

<sup>5</sup><http://swpat.ffii.org/group/epatext/swnerfatri/index.en.html>

<sup>6</sup><http://swpat.ffii.org/group/todo/index.en.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.