Dear Robert, dear all,

in following I´ll try to show up the technical background of our/your project, the Patterm.org project. We aim to make translation and content generation starting from given documents, more than 100 millions on every technical area, since as the USA supreme court once said “every thing made by man under the sun is patentable”. The approach consists of four/five steps:

Step 1: Search
Step 2: Download
Step 3: Convert
Step 4: Align
Step 5: Structure
Step 6: Integrate

Some of this steps can be skipped depending on the format of the original document, since patent applications/patent specifications can be generated from an image (scanned PDF), or directly from an XML file converted than to PDF. The European Patent Office has developed a simple, but efficient filing procedure based on XML-standards to make patents more accessible to the public, especially to SMEs, therefore some patent PDFs can be converted back to XML depending on the tools available. In the first use case we start from the assumption that the PDF has been created from an image. This strategy will than illustrate the "complicated way", since the conversion from PDF to RTF (for aligning) puts up some problems...So let´s get started:

1 Search

Searching for documents to pattern can be quite funny, depending on your personal interests. First of all you have to know where to search. The data mine is ESPACENET, a huge database put up by the international patent office in order to provide an interface where everybody can get insights into the system. To my experience, the best working database in Europe is to be found under this address:
http://gb.espacenet.com/search97cgi/s97_cgi.exe?Action=FormGen&Template=gb/en/advanced.hts

This is the server of the British Patent Office and usually works quiet stable and fast. There are different search fields:

1. Keywords in title: searches for words in natural language featured in the application´s title

2. Keywords in abstract: quite obvious, right? This is the best field to make natural language search

3. Publication number: WO=World (intellectual property) Organisation appears when the applicant seeks for worldwide protection.

4. Application number: Patent offices (DE=Germany, UK, US, ES, PL, AT, RU) that have already published the application, being in those countries where the patent will be enforced first

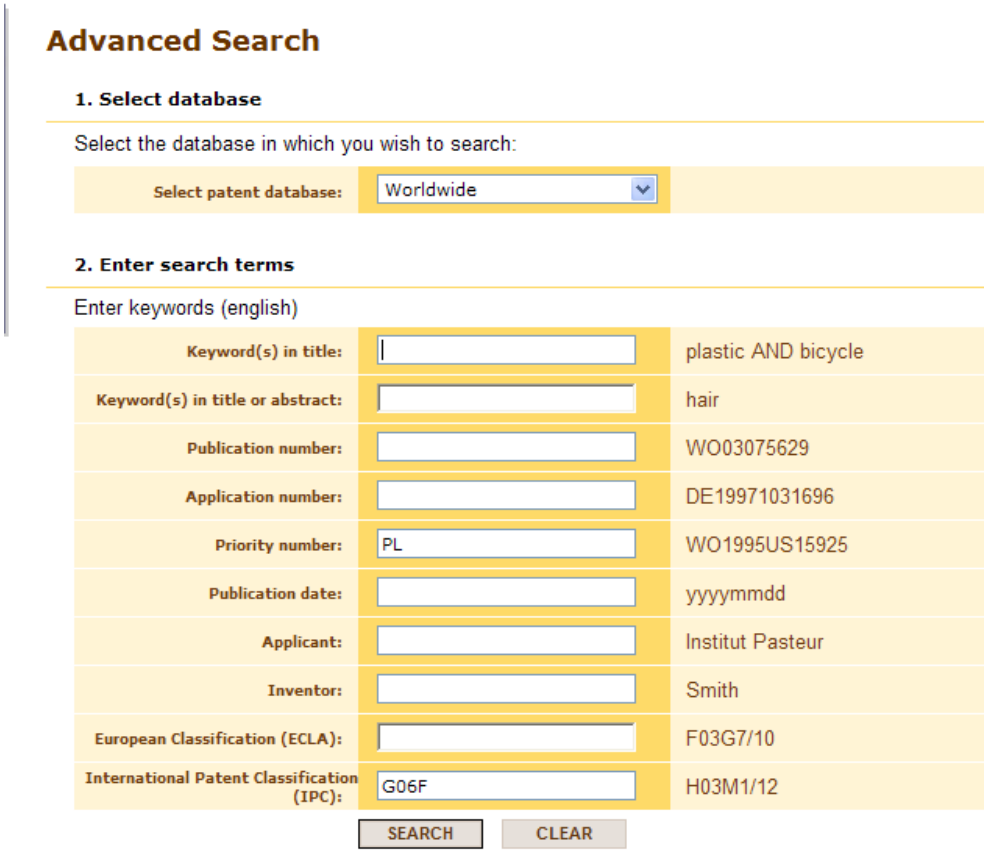
5. Priority number: this is the main field, since it proves WHO filed for the FIRST and WHEN its application

6. Date: quite obvious…

7. Applicant: the applicant and the inventor can be the same entity/person depending on the relationship between inventor and the entity who has financed the application process. In Austria for example and as long as it has not been stated in the working contract, every invention made by an employee using the company´s instruments belongs to the company.

8. Inventor: WHO is pushing the limits of scientific applied investigation or who is just been misled by an patent attorney to spend money on patenting bullshit.
9. The two classifications below will be mentioned later since they serve burocratic tasks, i.e. finding application based on more specific categories. More than 20 000...

So you seem to be into computer science. Therefore we will work with computer related patent applications from Polska:



The IPC category G06F relates to computer software methods, programming and implementations of open source software that people are trying to get patented, as well as every system controlled or driven by a computer.

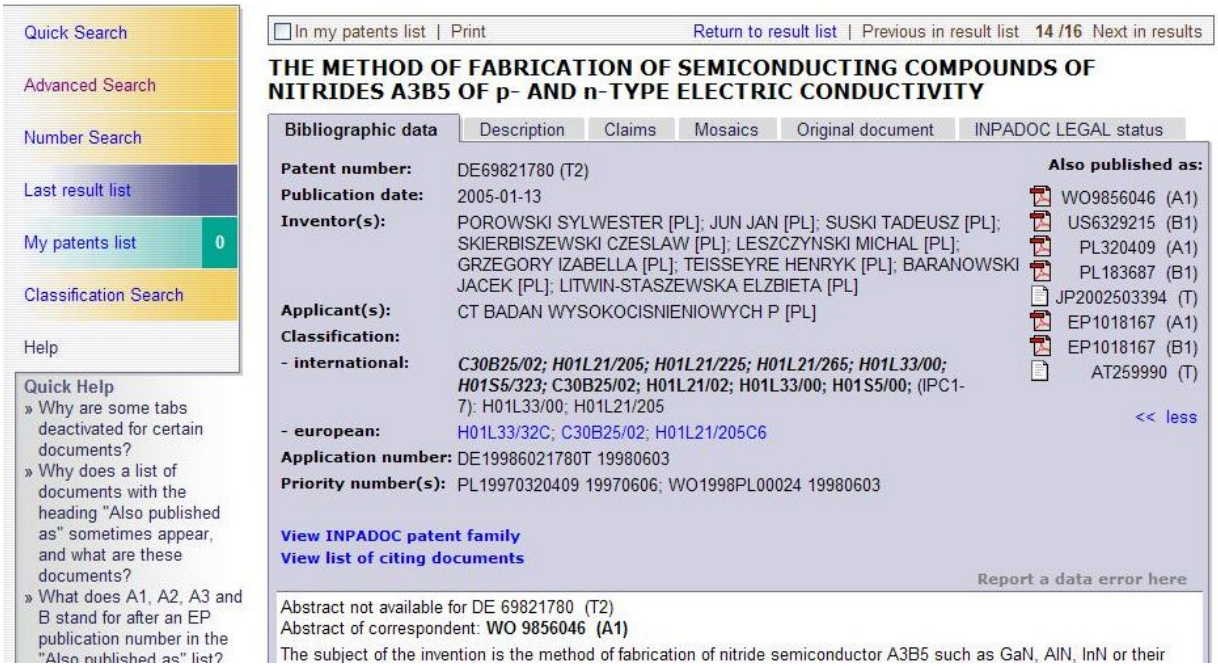
In the USA patenting software is quite easy. In Europe it has been made possible since 2000 after years of prohibition, since the European policy makers have tried to protect the European companies/inventors from the USA and Japan patents, in order to give the European developers/"inventors" time to put up patentable technology, mostly on open source basis.

So after typing this data, you get following image displayed:

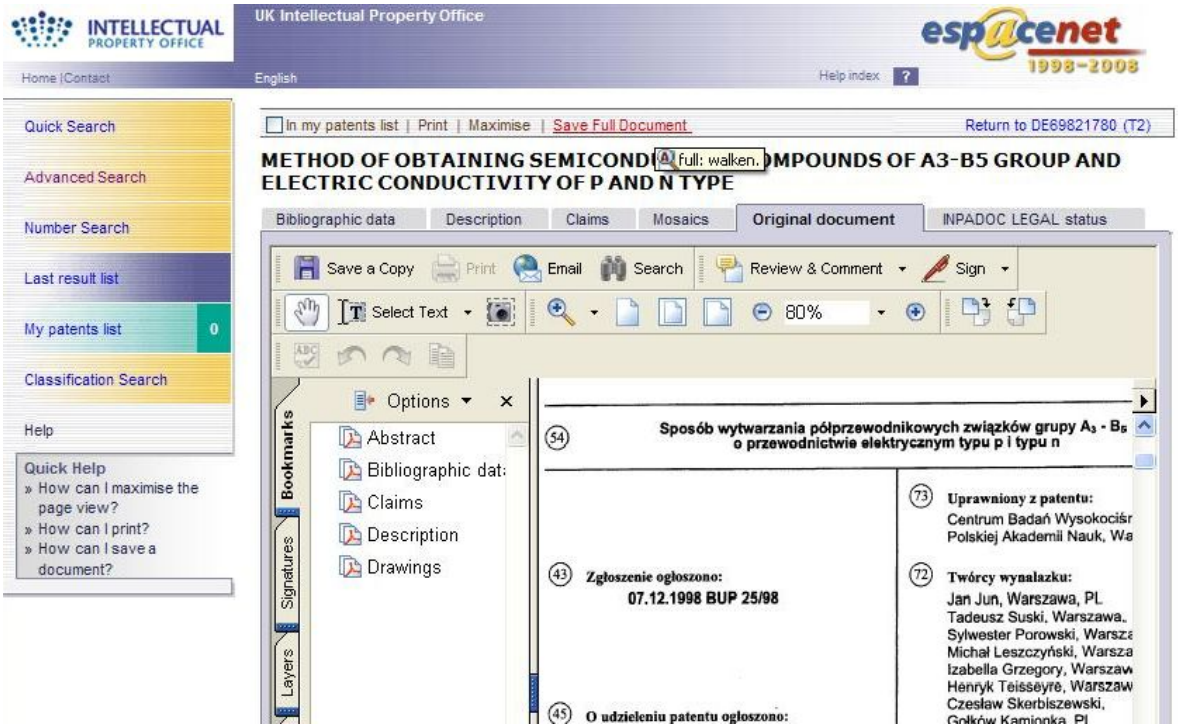
The register shows the title, the name of the inventor, his country, when the application has been published by a foreigner office, the category and other data…So according to this, in this category there are 1 287 patent applications where the inventor(s) are from Polska. The database can only display 500 results at once, so you will have to reduce the search scope using years or other parameters to get all 1 287 applications progressively displayed.

So let´s watch behind this. In order to find application filed originally in polish (is this the way you call your language?), you will have to search using PL as priority parameter; to find application or already granted patents from a polish patentee that are valid in Spain type in the publication field ES.
You get more or less 220 matches. But this two parameters cover all technical fields and we are interested in computer or information technology. So type in G or G0# in order to display theme related results. I have done so and found quiet a freaky application (Nr. 7 detection plate for an identification system). This match provides lots of translations of the original filed polish application, but sadly the polish patent office hast not published the application content in your language, therefore all other countries companies and their attorneys can claim, use, license the content of the document before the inventor even know what is happening with his invention outside Polska...

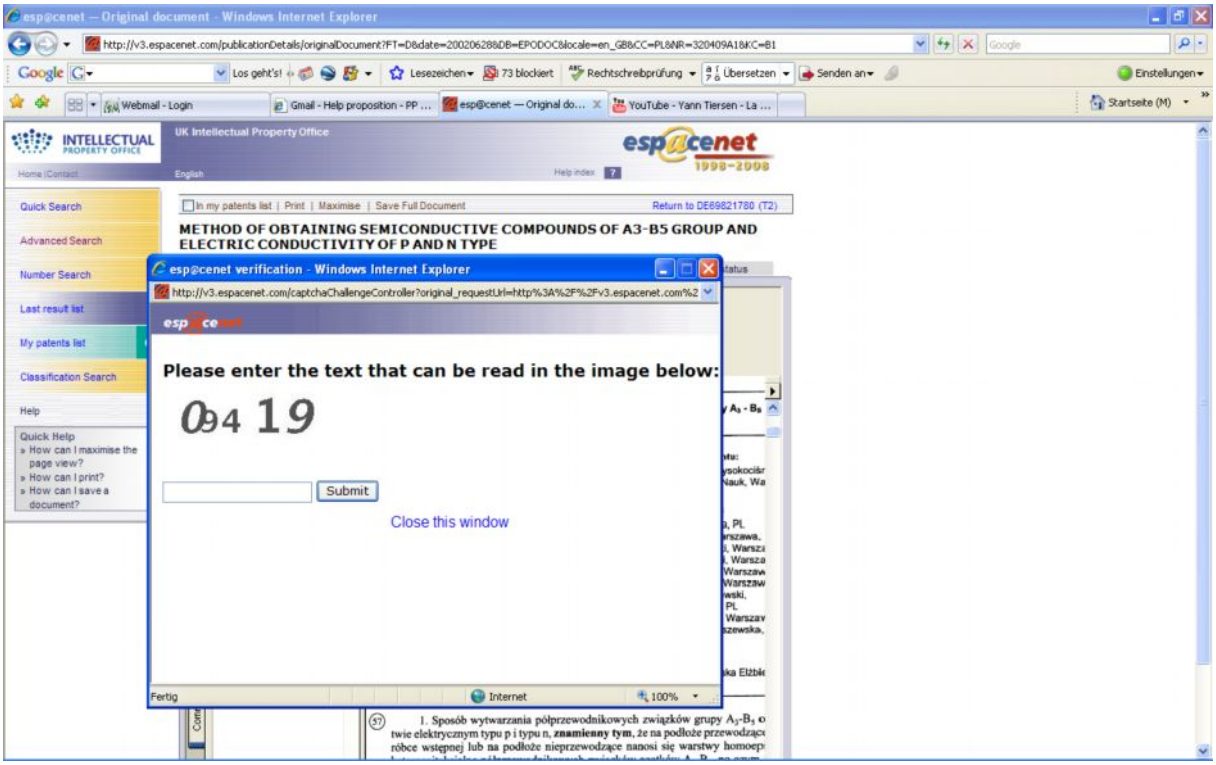
So lets take an application with more scope of our interest:



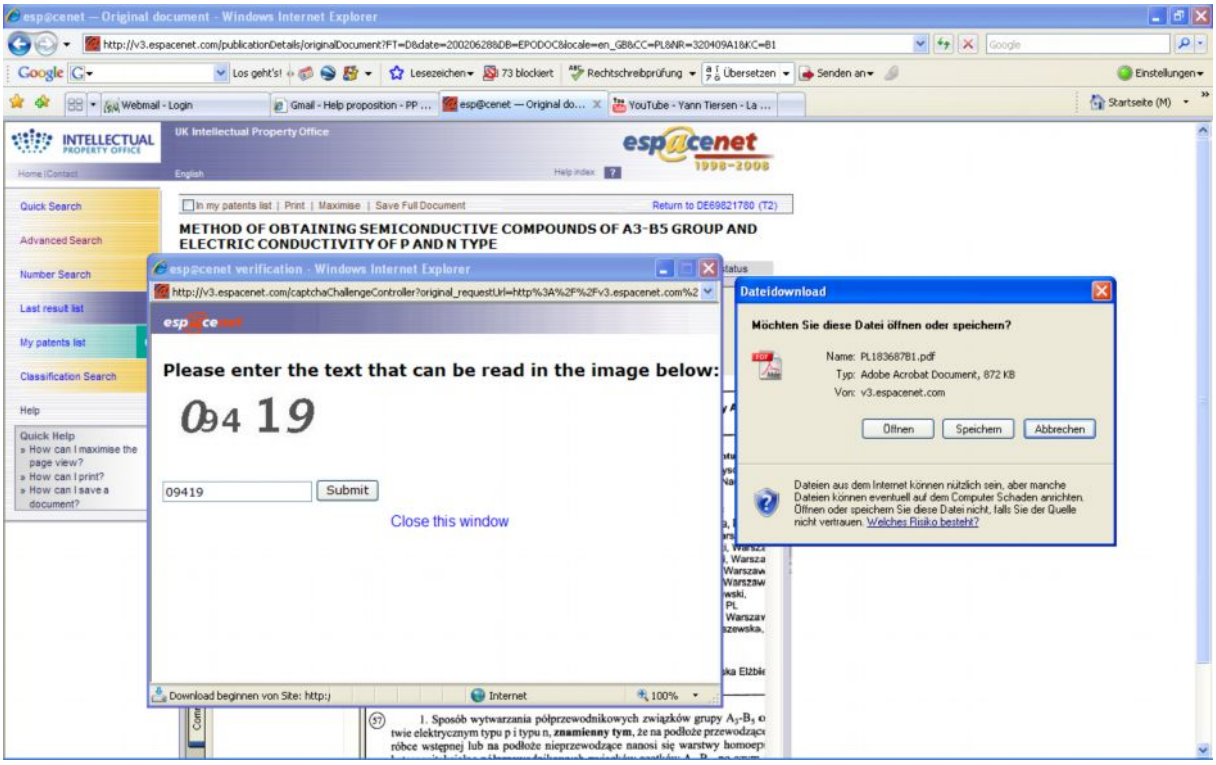
In this case you find an application from Polska, that has been translated into US-English, and partially into German, since the European Patent Office publishes the claims of the documents, i.e. the part of the document which is legal enforceable in German, English and French. The original document in polish is also available and if you click on the PDF link, yo´ll be directed to a embedded PDF, you can download, although the download is protected by a digital signature:



After clicking on the link you are asked to type in the code:



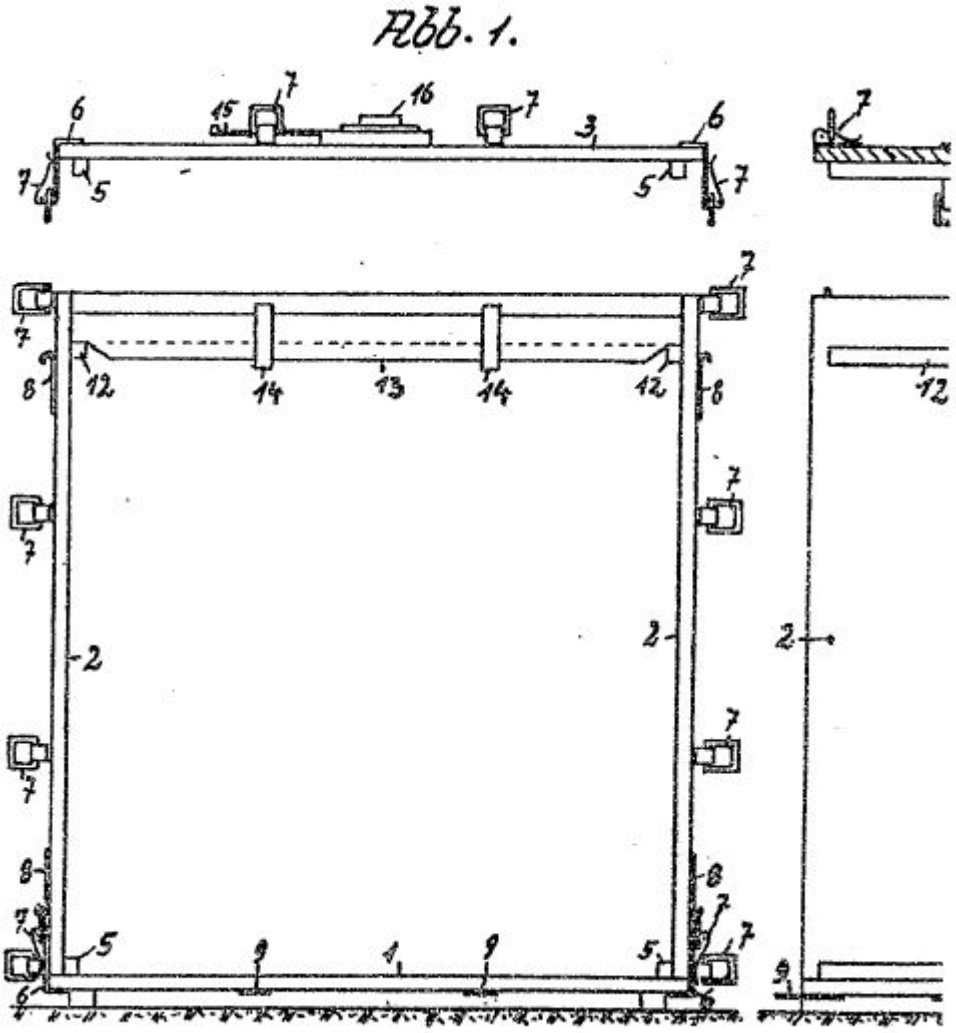
After typing the code you are asked to download the file and/or to open it...



Store the document. This is probably an image pdf. We need it in RTF or in a similar format in order to align it using a quite simple implementation of the alignment algoritmuses...
I´m quite tired after working in the shop...

So if you don´t mind we´ll keep on talking tomorrow. I hope after reading this you are able to search documents in ESPACENET in order to get insights on the way patents are been filed, in which fields and by whom in your country.

Brainstorming before shutting down my brain for today: tomorrow converting and aligning PDFs to RTFs, XML:TM standard, DTDs XSLs for patents, political dimensions, London Agreement (maybe you can check the stand of implementation of this agreement in your homeland).



(Drawings from German Patent 735749 (1943): detachable gas chamber for disinfecting and desparasitating any objects)