pdfextract

Pdf-extract is an open source set of tools and libraries for identifying and extracting semantically significant regions of a scholarly journal article (or conference proceeding) PDF.

In English, please...

The pdf-extract tools allow you to identify and extract the individual references from a scholarly journal article. References extracted using pdf-extract can, in turn, be resolved to the appropriate CrossRef DOI using CrossRef's citation resolution tools, Simple Text Query and ~~the experimental~~ CrossRef Metadata Search.

Limitations

The pdf-extract tools will only work with full text journal article PDFs. It will not work with PDFs which contain scanned bitmap images of pages. In practice, this means the pdf-extract tools are unlikely to work with older journal articles that were produced before the advent of computer typesetting.

Why have we done this?

We have built pdf-extract as part of an overall effort to make it easier for small and medium-sized publishers to meet CrossRef's linking requirements and to participate in CrossRef's Cited-by service.

When publishers join CrossRef and start assigning DOIs to their content, they also make a commitment to link references in their content to the relevant sources using DOIs. For larger publishers with skilled production departments, this requirement to link their references is relatively easy to meet. For smaller publishers, it is much more difficult. Those who do meet the obligations, often find themselves having to manually copy and resolve references for each article that they publish. Some members don't even have the resources to do this. This inability to meet CrossRef's terms & conditions for linking effects **all** CrossRef members, including our larger ones, because it means that fewer references are being followed online and because Cited-by information is incomplete.

Over the next few months we also plan on extending PDF extract to identify other semantically meaningful sections of scholarly articles including abstracts, methods sections, figures tables, captions, etc.

The pdf-extract tools are currently only designed for use by the technically savvy. To get them to work, you will need to know how to install and use software on a server running linux.

The pdf-extract tool will eventually be incorporated into a user-friendly set of web tools that will allow our members to automatically deposit article references into the CrossRef system by uploading PDFs using a simple form. We expect these more user-friendly tools to be available by Q1 2013.

Until then, we have created an experimental web form called "Extracto" that at least allows you to play with the pdf-extract tool without having to download and install the libraries.

Note that **Extracto is running on very feeble server on a very slow internet connection** and the only guarantee that we can make about it is that **it will repeatedly fall over and annoy you**. If those weasel words don't put you off, you can have a play with it here.

But your best bet is really to download and run the code locally. In order to do that, follow the instructions on github.

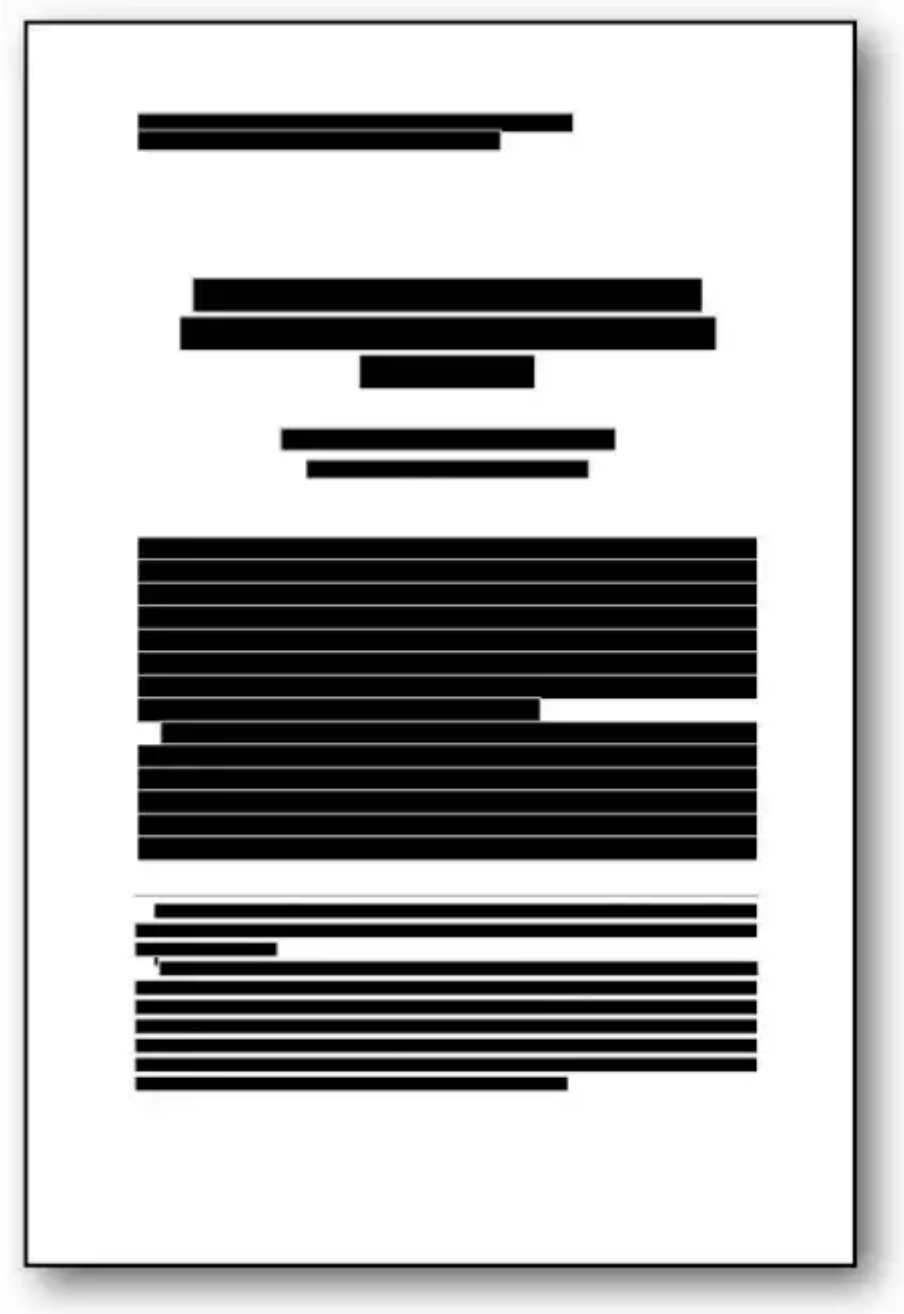
How does it work?

You can see a brief presentation we did at the CrossRef Annual meeting where we discuss, amongst other things, the pdf-extract tool.

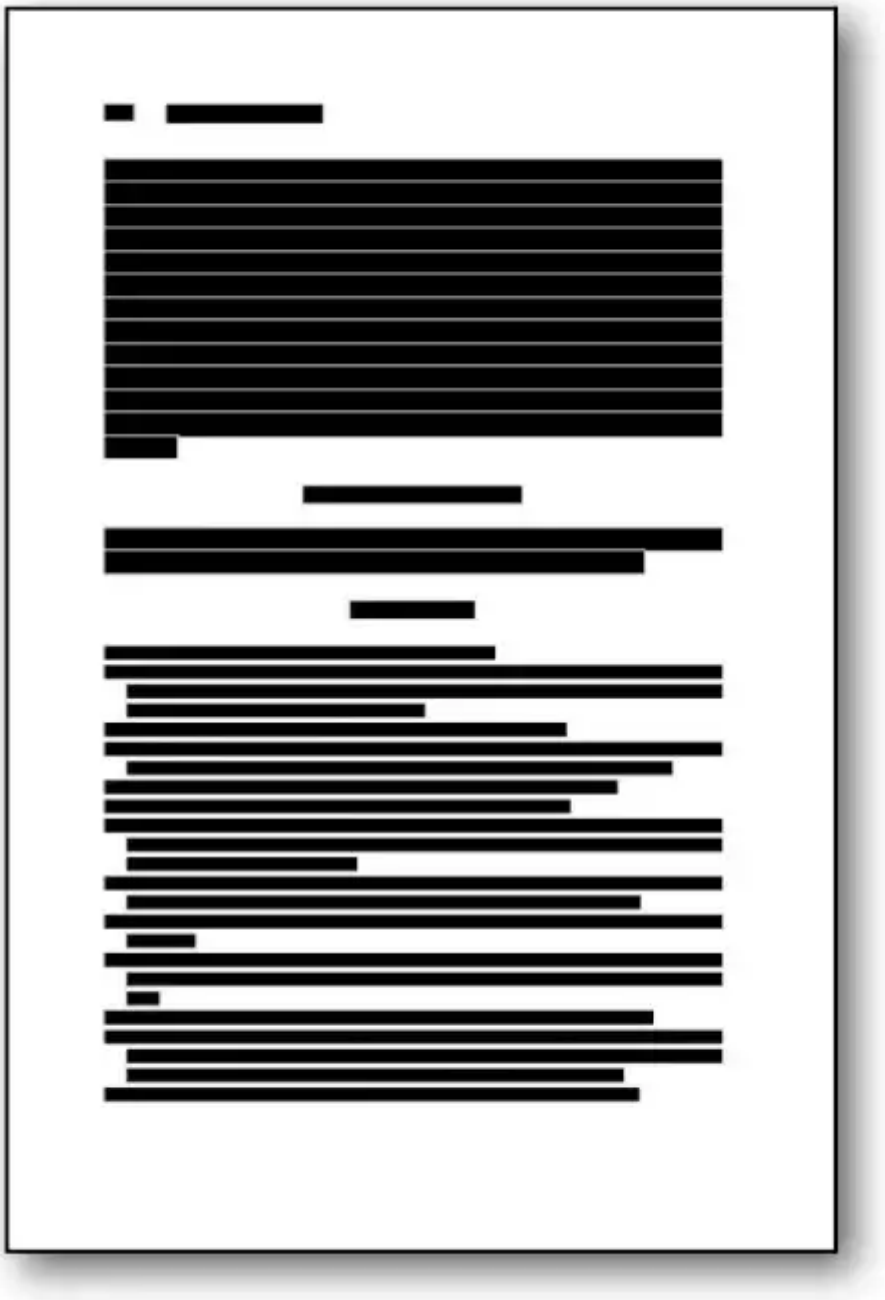
Otherwise, read on...

Most tools that attempt to extract text from a PDF have the nasty habit of throwing away formatting information. Unfortunately, this formatting information generally provides significant semantic clues to the contents of each region of a document.

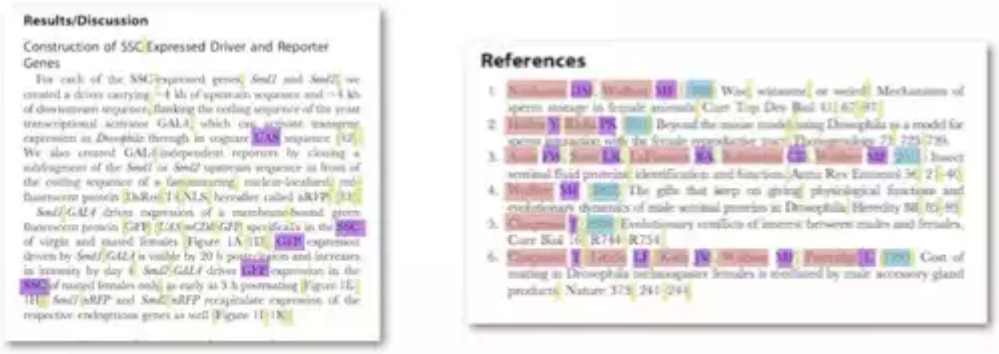
For example, if you look at the following redacted image, chances are you can immediately tell that this is an image of a scholarly article. Similarly, you can easily identify significant portions of the article, including the article's title, the authors, the author affiliations and footnotes. What is important here- it that you can do all of this without reading or understanding a single word of the article. Instead, you do this by identifying the significant "shapes" within the article page.



Similarly, in the following redacted image, it is easy to identify the references section, each individual reference, and even the acknowledgements section- all without being able to read a single word of the document.

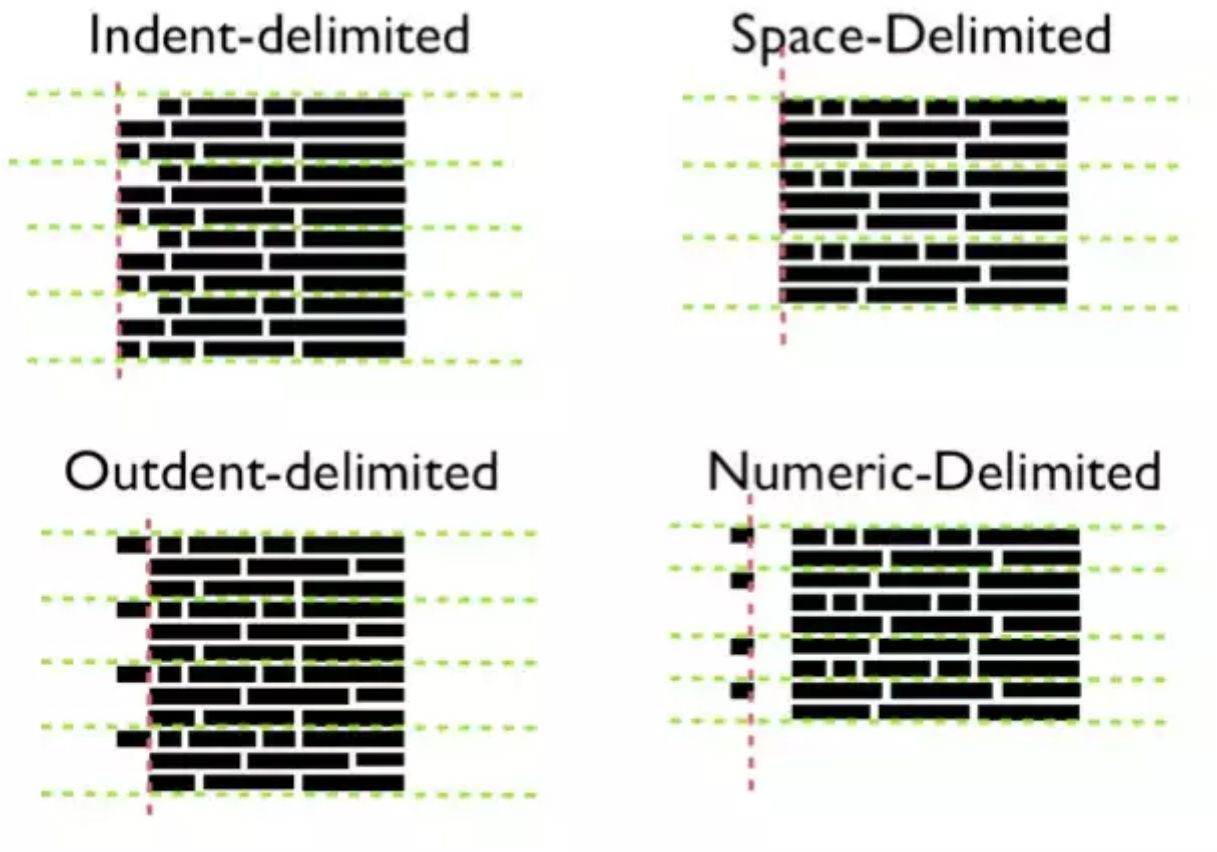


The pdf-extract tool uses a similar “visual” technique to identify semantically important areas of a PDF. After identifying semantically significant regions of text, it uses a set of heuristics to analyse certain “traits” in each region which help the tool understand what that region is doing. For example, the reference section of a PDF tends to have a significantly higher ration of proper names, initials, years and punctuation. This can be illustrated by comparing a normal paragraph within an article and the references section of the same article.



Using this combination of visual cues and content traits, the pdf-extract tool is able to detect semantically significant regions of the PDF without having to know the precise formatting conventions of any particular publisher or title.

One a region like the “references section” is detected, the pdf-extract tool can again use visual cues to identify individual references. Basically, citation styles tend to break down into the following visual categories.



Pdf-extract can detect which category a particular PDF is using simply by analyzing the margin and spacing use within the references region.

Once individual references are identified within the PDF, then we can use any of CrossRef resolution tools, such as our Simple Text Query systemor CrossRef Metadata Searchto try to resolve the reference to a CrossRef DOI.

How can you help?

We have tested the pdf-extract tools extensively over sample sets of PDFs provided to us by our members. The tool works well, but it can also be tweaked significantly as we apply it to more test cases and understand new variations in publisher formatting conventions.

If you are a developer with the requisite skills, we encourage you to contribute patches and fixes to the open source pdf-extract project.

If you are in production and encounter specific classes of PDFs that pdf-extract does not handle well, we encourage you to send us samples of said PDFs, as well as any potentially pertinent production information (e.g. tools used to produce PDFs, etc.) to:

labs@crossref.org