An elaboration on some of the problems with a few files from Springer

Ching-Li Chai
Department of Mathematics
University of Pennsylvania
David Rittenhouse Lab.
209 S. 33rd St.
Philadelphia, PA 19104-6395
U.S.A.
chai@math.upenn.edu

Amnon Neeman
Centre for Mathematics and its Applications
Mathematical Sciences Institute
John Dedman Building
The Australian National University
Canberra, ACT 0200
AUSTRALIA
Amnon.Neeman@anu.edu.au

Takahiro Shiota
Department of Mathematics
Kyoto University
Kitashirakawa-Oiwake-cho
606-8502 Kyoto
JAPAN
shiota@math.kyoto-u.ac.jp

November 10, 2010

Before all else: if you want to have a quick look at the kind of quality we're talking about, just have a glance at Volume 1 of the Collected Works of Vladimir Arnold, published by Springer in 2009.[1] This is the kind of product one can expect from Springer if one doesn't jump up and down and make a big fuss.[2]

In footnote 3 of *The Saga of Springer-Verlag and Volume II of Mumford's Selecta* we promised the reader some more detail of the problems with the files we were sent to proofread. Most of this note focuses on two Mumford papers, comparing different reproductions of them. To put this in context we begin with a little background, explaining how these documents came our way.

Let us remind the reader that, after our initial shock at the poor quality, we turned to Catriona Byrne and Joachim Heinze for help. This certainly made waves: the first we heard was an email from the typesetter SPi in Chennai, asking us what all the fuss was about. As it happens the person who sent it was kind enough to include, at the bottom of the message, a string of panicky internal emails from Springer, in which the employees asked one another, now that they [the three of us] have threatened to complain to the Editorial Director, what should we do?

Rachel informed us she was going to have a meeting with the production people, after which they would present us with a proposal on how to proceed. We tried to preempt the meeting by making a proposal of our own, in which we basically suggested tossing out the current product and starting all over again. Then we received a copy of an irate email, probably intended more for fellow employees at Springer than for us. It was from someone in the production unit, a man we will call Jim (not his real name), who said: stop rushing me, I'm doing what I can, this isn't my project, John (not the real name) is on vacation this week, I'm going through this and documenting all the places where the product fails to meet Springer's quality standards, you could help me by joining in the proofreading. We wrote that we had no intention of proofreading the SPi product; it was worthless trash and should be discarded.

A couple of days later Rachel called up Ching-Li. She told him that actually SPi has had a higher quality product all along, the files delivered to us were just the smaller, low quality version for proofreading purposes, would we be willing to take a look at the high quality files. You might forgive a certain skepticism on our part, but Ching-Li agreed to look at the higher quality material, as long as all of it was sent to us immediately. Two weeks later we finally received two "chapters" worth of high quality material, a total of 14 pages. The verdict was as follows:

[1] There is also an ebook version.

[2] For instance lines on pages 15–24, 259–266 and 428–432 are curved.

- The first of the two chapters happened to be the introduction Mumford wrote to the book *The Unreal Life of Oscar Zariski.* The book had been reprinted by Springer only a short time earlier, and SPi had been provided with the Springer pdf file for this introduction. Somehow they managed to corrupt it, making the file much larger (12.4 times) and poorer in quality. In the remainder of this note, we will always refer to this introduction as [90].

- The second chapter was a document also available on JSTOR, and the JSTOR version was smaller and higher quality. Our name for this 4-page document will be [62b].

We had a great deal of correspondence with Springer concerning this issue. After they finally let us see the original scans, Taka sent Springer, on 1 December 2009, a detailed analysis comparing the quality of several documents. That message is reproduced below.

In his analysis Taka only studied five electronic files, in the PDF and EPS formats. The two versions of [90] which he compared were the Springer ebook version and the high-quality SPi product, a file named `Chapter-90_PrintPDF.pdf`. The three versions of [62b] that Taka discussed were the original Springer scan, SPi's high-quality rendition (a file named `149784_1_En_2_Chapter_PrintPDF.PDF`), and the JSTOR version.

Finally: there were six graphic files attached to Taka's message, in the JPG and GIF formats. The two JPG files among the six are reproduced after the message body; all six attachments are available from `http://www.math.upenn.edu/~chai/story.html`.

The names of Springer employees have been changed. Other than that the message is identical with Taka's email of 1 December 2009, up to minimal corrections to the English.

. . . . . .

```
Dear Rachel and John,

I'm writing this message because I live and work in Japan and it is
not practical for me to visit your New York office.  If you have a
chance to visit Kyoto I'll be happy to show you the problems.

This message turned out to be rather long.  Due to its length and
my poor background in English I am not fully confident in the use
of language in this message.  In particular, in some places I might
be repeating what was previously said, I might have made obvious
```

errors, and I might sound overly offensive. I hope you excuse those possible problems with my English, and read it through and understand it. In this message I listed some fundamental problems in your work, explained how they affected the image quality, and warned you that the nature of the problems is such that it is obvious to an ebook owner that the scanned materials were not taken care of by skilled workers. This message doesn't quite reach a conclusion, and it doesn't include a clear recommendation on how to save the work, but one thing I can suggest to you is to promise us that you will switch from SPi to a better supplier, one with better knowledge of EPS, PDF and LaTeX. If you agree, then we can resume our discussion. If not, then I'd rather not waste any more of my time on this.

The smaller files you first wanted us to proofread have resolutions much too low to see possible problems, and indeed, we saw problems in the two larger files from SPi which you showed us (the ones for [62b] and [90], called

149784_1_En_2_Chapter_PrintPDF.PDF and Chapter-90_PrintPDF.pdf

respectively). So far those two are the only "high resolution" files from SPi we have received, but you should send us the rest, before inviting Ching-Li to your office to look at the hard copies.
I will NOT discuss problems which have NO effect on the image quality, but the way a problem affects image quality is often subtle and indirect. Electronic files help us diagnose problems in your and SPi's work more easily than hard copies printed on a particular printer do.

John's message on October 23 said: "the final quality of the images in the printed and e-books will be high resolution and comply with Springer's quality standards."

You've never given to us the real definition of Springer's quality

standards, and you've never described them as HIGH quality either, so it's hard to fault the latter half of this claim, but the former half is questionable: They are just not "high resolution" to my eyes:

Compare SPi's [62b] which has nominal resolution of 1200 dpi, with JSTOR's 600 dpi job (http://www.jstor.org/stable/1970376). Just in case you cannot print out JSTOR's file or cannot find a magnifying glass to see the difference I attached two close-up photos of print-outs, one from SPi and the other from JSTOR, named

SPi1.jpg and JSTOR1.jpg

respectively, of the same portion of the article (the word "consider" that appears two lines below the displayed formula in p.612, or the first page of [62b]). Care was taken to make sure that they were printed with no rescaling/resampling done by the printing program or the printer itself. If you see the outline of a letter more badly deformed in the former (SPi's work) than in the latter (JSTOR's), and the smooth and graceful way the curvature changes on each segment of the outline of a letter (lower case "s" is my favorite) better preserved on the latter (JSTOR's work) than on the former (SPi's), you are seeing the "residual evidence" of SPi's poor handling of the scanned data, after details got washed out by my printer's slight lack of acuity.

If you don't see what I was talking about ("outline of a letter more badly deformed," etc.), read some book on typography. If you see what I'm talking about and wonder what is going on, just "pixel-peep" the two pdf files which produced those print-outs.
For your convenience I attached another file named

Springer-SPi-JSTOR1.gif .

This consists of three "snapshots" of my computer screen put together, showing highly magnified views of (from top to bottom)

(A): Springer's original file "62bp612.eps",
(B): SPi's rendition of it "149784_1_En_2_Chapter_PrintPDF.PDF", and
(C): JSTOR's offering,

each showing the word "consider" that was photographed.

(Here and in what follows, we call the eps files in the directory Original_Springer/ORGANIZED/ in SPi's ftp server "original files," and the images stored in the original files "original scans" etc.)

A close look will reveal that the one in the middle, i.e., SPi's work, looks most damaged, with very uneven and irregular outlines, and the biggest "step size" of the jaggies (staircase effects) among the three. This is why you saw more distorted outlines in SPi1.jpg than in JSTOR's version.

As (A) was scanned at 800 dpi and SPi put it into a 1200 dpi image, you might say 'this can't be true; SPi's (B) can't be inferior to JSTOR's 600 dpi (C),' or 'how can upconverting 800 dpi data in (A) to 1200 dpi (B) degrade, not improve, the image quality?'
Sorry, but you have to face the reality. The degradation may partly be due to SPi's lack of knowledge on whatever tools they used, but the degradation itself cannot be avoided:
When the original scan has two values (black-and-white) rather than gray scale and the resampled image has to also have only two values, any sensible-looking operation like interpolation does not work, so any resampling process must, in effect, always either skip (to reduce the "dpi number") or repeat (to increase the "dpi number") sample values. E.g., the simplest way to use 800 dpi data to cook up some "1200 dpi data" is to use every other sample value twice:

s1, s2, s3, s4, s5, s6, ... -> s1, s2, s2, s3, s4, s4, s5, s6, s6, ...

Here s1, s2, ... stand for successive sample values along a row (or a column). I'd like to call it "1-2 pull down," since it works like the poplular "2-3 pull down" used in the "telecine" process (see

http://en.wikipedia.org/wiki/Telecine if you don't know what telecine means or involves). I should have drawn a diagram of 2-dimensional array of sample values to show the 2-dim effect of this resampling, but decided to look at only one row to simplify the presentation.

This means the sample values are treated UNEVENLY. In this example you can imagine that the repetition of sample values would make the maximum size of jaggies (staircases) in the resampled image TWICE as big as in a native 1200 dpi image (i.e., just as big as in a 600 dpi image), but the unevenness doesn't do us a favor either, so the perceived quality of the resulting image can be lower than in a 600 dpi image. Actually, SPi seems to have resampled in a more strange way (I guess they used some Photoshop routine which is only suited for resampling photographs and other gray scale data, and then saved the result as a black-and-white image, or maybe they did something more "creative" for resampling black-and-white images), so they got jaggies even bigger than in a 600 dpi image.

You'd argue that it must be enough for the final print-out to have "high enough image quality" when viewed by naked eyes at a normal viewing distance, so it's unfair to take close-up photos or to pixel-peep the pdf files on a computer screen to analyze them. Let me explain. I consider the use of a magnifying glass or close-up photos just an aid to let you find the problems. After finding the problems, you can often see the problems with your naked eyes (or with reading glasses if you normally use them to read fine print), maybe not as the same problem, but as some subtle feeling that something is "missing" or "wrong", "dry", "cold", "veiled", "not lucid", "not relaxed" or "not smooth", etc.
Now compare your best print-outs of (A) (= Springer's original) and (B) (= SPi's work) with, of course, no further scaling done by the printing program or the printer itself. You've learned how much harm was done by SPi's resampling job, so, do make sure to turn off any rescaling/resampling features on your printing program or your printer. Your printer may not be as silly as SPi, but rescaling is never better than no rescaling, period. This means, in particular,

that you should use a printer having at least 2400 dpi of resolution, since (A) has 800 dpi data and (B) has 1200 dpi data, and the least common multiple of 800 and 1200 is 2400.
Aside from the approx 97 percent reduction in size, can you still say that (B) is as good as (A) in the image quality?

If you still don't see the problems, so be it. Some people just don't see a problem which annoys others. Then I want to ask you "what about the ebook?"

For an ebook owner, finding irregular outlines of letters in SPi's product requires just a few clicks of the "magnify" button on the pdf viewer, so it is almost as easy as finding a skewed page in a book. You don't want to sell a book with obvious cosmetic problems, even if the problems are harmless and don't impede the reading. If you want to deskew the scans for a cosmetic reason, I think you should redo the scanning properly and process the scanned images properly.

For an ebook owner, it is also quite easy to compare both the file size and image quality with freely available versions of the same article from JSTOR, Numdam etc. I haven't said earlier, but the file size of (B) is too large. I can create pdf images which look the same (same, approx 97 percent of reduction, etc.), and WITHOUT showing irregular outlines when they are pixel-peeped on your computer screen (more on this later), and with less than a half of the file size of (B).

So far I have sounded more critical of SPi's work than of Springer's original scans. Comparing the image qualities per se, just as shown in my snapshot comparison file Springer-SPi-JSTOR1.gif , I think this makes sense. It is also true that SPi's way of handling [62b] and [90] shows their incompetence, as I'll discuss later. However, I'd like to also claim that Springer's choice of 800 dpi as the resolution when the original material was scanned was not a good choice. But before I start explaining this, let me list casually and randomly other problems I noticed

with the image quality of Springer's original scans and SPi's work.

- There are black spots in supposedly white areas and vice versa. There are more such spots in SPi's [62b] than in the same article available from JSTOR. Some are from the original scans (sorry but they aren't quite clean) which were not removed by SPi; some spots in the original did get removed by SPi but SPi seems to have put its own spots (look at the "period" at the end of the last displayed formula which reads "8(1-q) is greater than or equal to rho - 1", in the second page of SPi's [62b]).

- This may be related to the above, but Springer's original (A) often doesn't get the outlines of letters as clean as JSTOR's (C), and I wonder if the problem might come from the quality of the original material which was scanned.

- Many parentheses don't get a steady curvature that looks like a portion of a smooth arc. This is another place where the effect of rescaling/resampling is more obvious. SPi's work is thus generally not good at handling parentheses. Springer's original scans should be good at it if they are true originals, but just in one random place I looked at (I mean, I wasn't looking for problems, it just popped out on my eyes), I found that Springer's original scan had a much less steady curvature on a parenthesis than JSTOR's (last line of text before the footnote in [62b], p.612; see another attached file

sprjst.gif

in which the top row is Springer's 800 dpi original scan, the bottom row is JSTOR's 600 dpi offering). I then compared other parentheses on Springer's and JSTOR's data and the result wasn't so obvious as I first expected, but my theory, that Springer's "original scans" are not so close as the name suggests, to the real original data from the scanner's sensing device (CCD in a flatbed scanner, PMT or whatever in a drum scanner), still remains.

- Some letters got exceedingly wiggly outlines. For instance, look at the formula "R=R^*[K]" in the first page of [62b] (just below the only displayed formula in that page). You see near the top of the three uppercase Italics, i.e., two R's and one K, got wiggly; the effect is most obvious on the leftmost slanted edge of each letter. This kind of wiggliness often occurs approximately around the same horizontal line, making me suspect that the scanner might not have been on a very stable support, or some of its moving parts might not be in a perfect shape. How much wiggliness you see on the printout depends on the printer's nominal resolution (dpi number) and its real acuity. In many cases you don't see the most severe form of the wiggliness in the printout, but that only means your printer's real resolving power is somewhat low, and even in such cases you most probably see some form of wiggliness which are either not present or much less obvious on the JSTOR version. For your convenience I attached yet another file

Spr-SPi-JST2.gif

showing the "R" on the right-hand side of the above mentioned formula in three incarnations: (from left to right) Springer's original, SPi's work, and JSTOR's. See the wiggliness near the top of the "column" on the left edge of R, and how it disturbed the outline of the letter in SPi's rendition?

- Now I realized that I prepared another "comparison of three" file for the lower case "s". This is the same "s" as the one that appears in sprjst.gif, but magnified more and including SPi's big job to show my favorite "can you feel the smooth change of curvature" demo. Again, it's from left to right Springer's, SPi's and JSTOR's. The file name is

Spr-SPi-JST.gif

Let me recall that Mark Spencer promised us that the materials would be scanned at 2400 dpi to ensure the quality. Ching-Li told you this, so this is not the first time you hear it.

2400 dpi is an awful lot of resolution. It may be a tall order for the optical system in your scanner to have enough acuity to exploit that resolution, but even if it can't quite resolve two dots 1/1000 of an inch apart, data sampled at 2400 dpi and in gray scale (rather than in two values, i.e., black-and-white) will give you much more room to do whatever work is needed (e.g., when old original had a stain and may need real Photoshop work) and more room to recover from any errors. So it's good, I thought. I took Mark's word as a promise for the quality: IF your product is printed using a high quality laser printer, THEN it must look as good, even to an expert with a precision loupe, as a good 2400 dpi scan of the same material printed on the same printer, and IF your product in the form of electronic files was examined by a pixel-peeper, THEN it must be as good as the best scan made at whatever the resolution the final product would have chosen to use.

I'm not saying that all this is impossible if you start with "less than 2400 dpi" black-and-white data, but then to meet the quality standards stated above, you have to make correct decisions and have very precise control throughout. But (B), SPi's best effort, miserably failed. It has 1200 dpi "resolution" but looks worse than JSTOR's 600.

I said "correct decisions and ... throughout." Indeed, the first crucial mistake was your decision to make it necessary for SPi to resample. I explained above what is in a resampling process of 800 dpi data to get 1200 dpi data and how it affects image quality. To be precise, the resampling in this case is from 826 dpi to 1200 dpi, but you get essentially the same problem. (When you scanned [62b] at 800 dpi no rescaling was done; adjusting it to the text width of the book would slightly raise the pixel density to approx 826 dpi. Thus the number 826 came up in the above sentence.)

Even if you have 800 dpi scans (not reduced) and asked SPi to make 800 dpi files, it would need rescaling which means resampling of 826 dpi data to 800 dpi. Or if you have 1200 dpi scans (not reduced) and asked SPi to produce 1200 dpi files, it would need resampling of 1239 dpi data to 1200 dpi. (Appearance of "1239" is for the same reason as that of "826" above). But the resampling in those cases involves removal of one in every 30 or 31 sample values, so the effect on image quality would be less severe.

So, the biggest mistake you made was the choice of two different values, 800 dpi and 1200 dpi, as the original and output resolutions. The next biggest mistake is your choice not to reduce properly (i.e., to adjust to the book's required text width) at the time of the original scanning.

Here of course, I assume your scanner is good enough to be able to properly (I mean optically (in case of flatbed) or mechanically (in case of drum)) take care of reduction or enlargement. A less expensive flatbed scanner with no way to optically reduce or enlarge the image would have to reduce or enlarge by some electronic means, which is the same thing as resampling of the scanned data. Such a scanner should of course be avoided.

There is another reason to avoid 800 dpi. From Mark Spencer's mentioning of "2400 dpi" I assumed your CTP (computer-to-plate; am I using the right term?) process and your best laser printer have 2400 dpi as their "native" resolution, and so 800 dpi data would pose no problem for you. (Going from 800 to 2400 requires each sample value to be just repeated 3 times; NO uneven handling of sampled values is needed.) But most math institutions don't have a 800, 1600 or 2400 dpi printer, while 600 and 1200 would be most common. As in the "1-2 pull down" scheme explained above, converting 800 dpi data to 1200 dpi isn't a good thing to do to your images.

Now let me shift our attention to SPi's work. For pixel-peepers their lack of skills and knowledge, their lack of interest in the quality of the product, and whatever problems you associate to their work, are by far the most obvious problems, although I must remind you that Springer's originals are far from flawless and not meeting our expectations (c'mon, why 800 dpi? Answer me).

We have explained to you in an earlier email the main problem with [90]: the original pdf file is from Springer's ebook version of Carol Parikh's book, so it mainly consists of text objects, but SPi has somehow converted it to a 1200 dpi RASTER (pixel) image to inflate the file size by more than ten times, and to introduce a potential problem in the image quality: in the original file the text part of the article was typeset using Adobe's popular Times families of outline fonts, so the letters in those fonts are defined as vector graphics objects so they look great even when enlarged enormously (e.g., to cover a huge billboard), but this fails after SPi converted the page image to a big raster graphics object in [90]: some enlargement lets you see jaggies and staircases instead of smooth curves.

The main problem with SPi, as seen in both [62b] and [90], is that they seem to have converted the original eps or pdf files into TIFF images first, before placing them in the output pdf file. TIFF may be a good file format, but it's a raster image format. When SPi converted each original page of [90] into a TIFF image the text objects had to cease their meaning as text to become just a huge chunk of black-and-white dots (i.e., a raster image). This may still look OK if the result will only be fed into your CTP machine having fixed 2400 dpi resolution, but it is terrible for the ebook. This shows SPi's ignorance of the nature of the files they were handling and tools they used to handle them.

When you have scanned originals like in [62b], the simplest way to

reduce or enlarge them to adjust the image width or height is to LEAVE THE SCANNED DATA INTACT and just alter the "pixel density". This method has the additional benefit of being reversible, in contrast to SPi's way of brutal resampling from which you wouldn't be able to recover the original scanned data (just think about going back from (B) to (A); even a Ph.D. in computer vision couldn't do it!). In [62b] the originals were 4.75 inches wide and scanned at 800 dpi, so to obtain the 4.6 inch text width needed for Springer's monograph style you just go to 800 x 4.75/4.6 = 826.08... dpi of pixel density. This is easy on PS, EPS or PDF files; SPi should only need to alter a few lines in the original eps files (see the next paragraph), and then save them as a pdf file. See how simple it is, compared to SPi's approach in which the images reduced to fit into the prescribed text area were resampled to yield 1200 dpi TIFF images, which were then saved as a pdf file. This is much too complicated and the result is more irreversible than that of a simple "pull down" method described earlier. (Even for a "pull down" method you need some guesswork to reverse the process if the resampling was not a simple "800 to 1200" kind of thing, but some odd variety like "826 to 1200". SPi's work is just much worse.)

So how do I accomplish the simple change of pixel density on PS or EPS files? You just use a feature of PostScript as a graphic programming language. EPS is a close relative of PS which is a programming language. A PS file contains a program source code written in the PS language, so it is (almost) human writable. (Sorry this may be a way to explain the idea to computer scientists and mathematicians; a seasoned Photoshop user must be able to do the same thing in a different way.) In particular, you should never need to leave the ps, eps or pdf file format. E.g., for [62b] with eps originals, I would just change the line

```
354.6 301.59 scale
```

in 62bp612.eps to

```
343.4 292.07 scale
```

and do the similar changes to the other eps files for [62b]. You can do this editing with a text editor that can handle binary files. (PS files and some EPS files are text (ASCII) files so such caution isn't needed. But your EPS files are binary, so while the line you want to change is in the "text part", you have to use an editor that can be happy to see a binary file.)
Then you achieve the reduction from 4.75 inch text width of the original article to 4.6 inch required text width for Springer's monograph style. (Ideally you need to change a few more lines, but for reduction (not for enlargement) this suffices.) This is a minimalistic approach and may not guarantee the best quality printout, but the printout should be at least as good as the printout from SPi's version 149784_1_En_2_Chapter_PrintPDF.PDF , it lets you get a pdf file which is a lot smaller than SPi's, and it will look much better than SPi's (indeed, just as good as the original eps files) when magnified (pixel-peeped) on your computer screen.

In this case, the printer does the necessary scaling so in a sense at least one of the basic problems still remains, but the ill effect of scaling will be less visible when you use a better printer (a higher resolution one like Springer's 2400 dpi CTP machine). So I think this is a better approach. I don't have the faintest idea why SPi decided to go through the TIFF format, to spoil all the benefits of the PS, EPS or PDF formats.

Although we haven't discussed it, SPi has also revealed its ignorance of the way LaTeX works, and the LaTeX part of the book is still heavily damaged, even after they fixed the strange font problem. After seeing their inability to handle LaTeX as well as PostScript material, how can we trust them?

All the best,
Takahiro Shiota

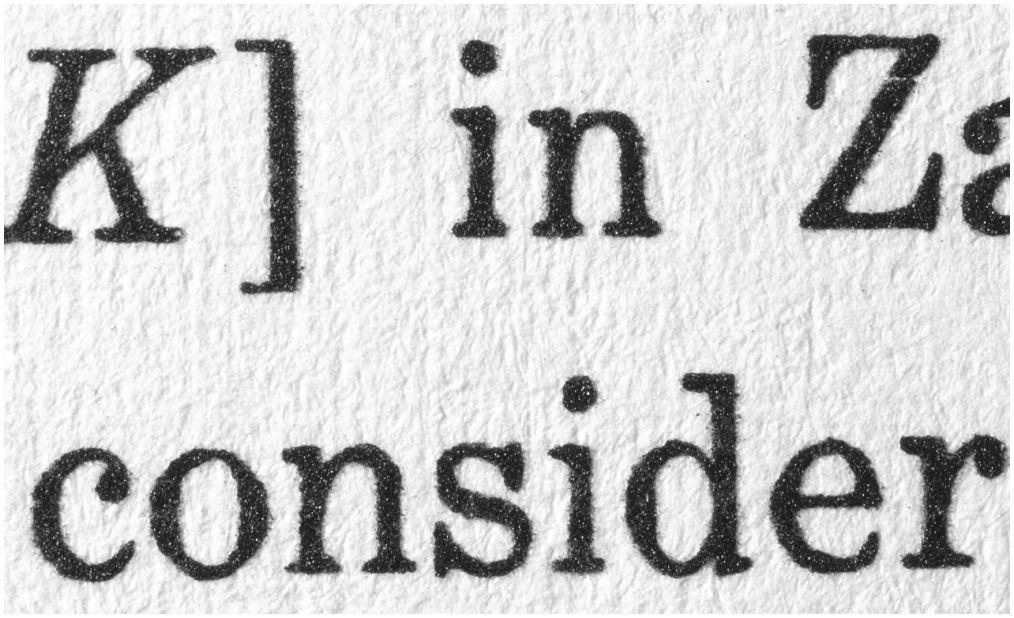

Figure 1: SPi1.jpg

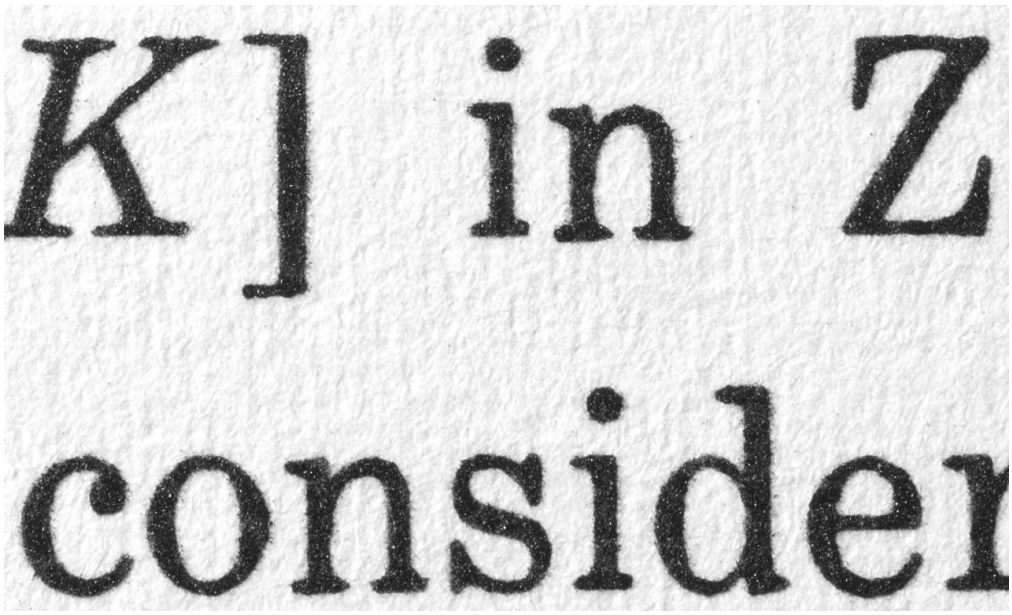

Figure 2: JSTOR1.jpg