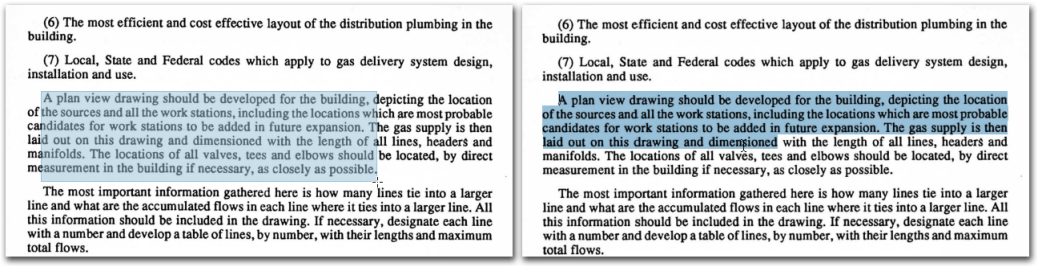
Creating Good Scannable PDFs

1. Why do a Searchable PDF?

If you scan a document and then convert that page into a PDF, what you have is a PDF of an image. The good news is that this PDF will take less storage space than the original image (this is because the image is being compressed with JPEG compression), but there is a one significant lacking feature: you cannot search within the document.

As shown in the image below, one the left is what happens when you try to select text in a PDF that has not been made searchable. When you try to marquee some text, all you end up selecting is a region of the page and no text has been selected. On the right, the PDF has been made searchable. Now, selecting text is possible and the contents can be searched, selected, and even copied. This is now a functional, "live" document.



(6) The most efficient and cost effective layout of the distribution plumbing in the building.

(7) Local, State and Federal codes which apply to gas delivery system design, installation and use.

A plan view drawing should be developed for the building, depicting the location of the sources and all the work stations, including the locations which are most probable candidates for work stations to be added in future expansion. The gas supply is then laid out on this drawing and dimensioned with the length of all lines, headers and manifolds. The locations of all valves, tees and elbows should be located, by direct measurement in the building if necessary, as closely as possible.

The most important information gathered here is how many lines tie into a larger line and what are the accumulated flows in each line where it ties into a larger line. All this information should be included in the drawing. If necessary, designate each line with a number and develop a table of lines, by number, with their lengths and maximum total flows.

(6) The most efficient and cost effective layout of the distribution plumbing in the building.

(7) Local, State and Federal codes which apply to gas delivery system design, installation and use.

A plan view drawing should be developed for the building, depicting the location of the sources and all the work stations, including the locations which are most probable candidates for work stations to be added in future expansion. The gas supply is then laid out on this drawing and dimensioned with the length of all lines, headers and manifolds. The locations of all valves, tees and elbows should be located, by direct measurement in the building if necessary, as closely as possible.

The most important information gathered here is how many lines tie into a larger line and what are the accumulated flows in each line where it ties into a larger line. All this information should be included in the drawing. If necessary, designate each line with a number and develop a table of lines, by number, with their lengths and maximum total flows.

While the process of creating a Searchable PDF is very straightforward, like most things in life, the devil is in the details. And, just like glassblowing, photography, and projects in the garage, the better your original components, the better your final products. If you start a project with dirty glass, you will not end up with a pretty piece of labware. And if you start out the process of creating a searchable PDF with a bad scan, you will not end up with very good output.

The process of creating a searchable PDF is simply running a computer process of OCR (Optical Character Recognition) on the document so that the image of each character is translated into actual text. What happens is that the software examines the document and recognizes shapes. The cleaner the shapes the more accurate the OCR. If the scan is not clean, things like exclamation points "!" may be interpreted as a lowercase ell "l" or a one "1." If you're trying to OCR a page that's been photocopied over and over, it can be faster to simply re-type the page than try to OCR that page and try to correct the errors.

Thus, the road to a great OCR of the document can best be initiated with a great scan of the page. Also, working with the best original document (not a photocopy of a photocopy) is extremely important.

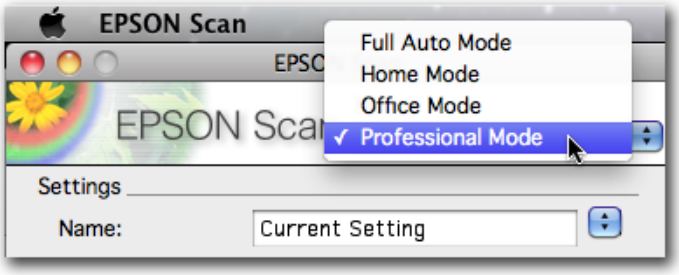
2. How to do a good scan

People have been led to believe that any photograph taken can be as sloppy as one wishes and just "fix it in Photoshop." The is not true and getting a good scan is just as critical as getting a good initial photograph. The good news is that if you follow some simple rules, getting good quality scans is pretty simple. It all starts with the software.

Whether you have a standalone scanner, or one that's part of a printer, fax machine, etc., you also have some software to run the scanner. Most of the time, this software is set to "Auto" because the manufacturer wants you to think that their machine is so good, you don't have to do a thing but click a button. This of course is nonsense: while a basic scan is certainly possible, the nuances of scanning a document require more hands-on activity. Ironically, scanning a photo is more forgiving than scanning a document. Therefore setting up your scanner to create a good document scan does require more initial interaction by the user.

The first thing you need to do after starting your scanning software is to look for User Levels. As shown in the image below, look for something that provides the most control. [Note: I have an Epson scanner and therefore all of my screenshots are taken with that software. What you have may very likely be different but it's been my experience that most scanning software on a reasonable quality scanner will provide most of what I'm asking you to look for. If your scanning software does not provide the features mentioned below, consider either purchasing a newer/better quality scanner ($100 to $150 should cover it) or see if the software "ViewScan" is compatible with your scanner <http://www.hamrick.com/>. ViewScan is compatible with over 1750 different scanners on both PC, Unix, and Mac platforms.]

Now that you are in the setting with the most options, which ones do you need to be concerned about?



As shown to the right, there are four main settings to be concerned with:

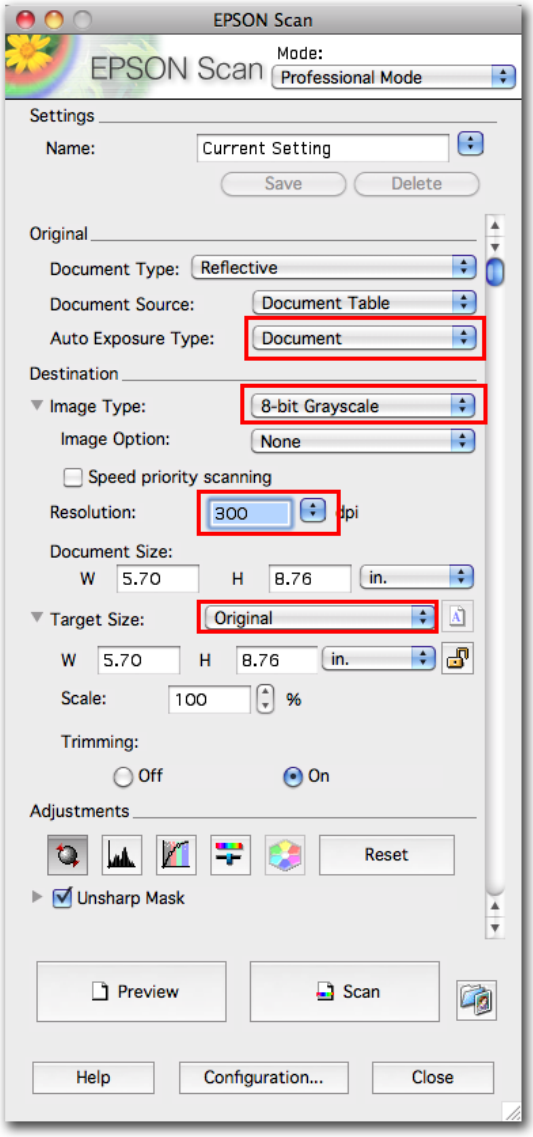
- **Exposure Type**: Scanners have different internal settings depending on the type of item being scanned. The two most common are for scanning photos or documents. Since you are scanning from Fusion or Proceedings, set this to Documents.

- **Image Type**: Scanners can scan either in color or grayscale. If the page is black and white, select Grayscale. [This involves "image depth." An 8-bit image is 2^8 power or 256 bits of data (gray). A 24 bit image is $2^8 \times 2^8 \times 2^8$ (red, green, and blue) or (over) 16 million colors. Thus, a color scan takes (at least[1]) 3 times the storage space as a grayscale image. Note that if you take a color scan of a black and white page, it will still be as large as if the image had color in it. So, if you are taking a scan of a black and white image, use "8-bit Grayscale." If there is color in the image and you wish to capture the color, select "24-bit Color."] (Ignore 48-bit color and 16-bit Grayscale, those are different issues.)

- **Resolution**: 300 dpi (dots per inch) (or ppi (pixels per inch) or lpi (lines per inch)) is a good standard. Fax machines scan at generally 200 dpi, and is at the bottom range of good consistent quality OCR results. Stay with 300 dpi.

- **Target size**: We want the final size to be the same as the original, so we are looking for "100%," or "Original," or something like that. [Note: this is directly related to the scanning's resolution. If you scan at 300 ppi and print at 300 dpi, your final images should be at 100% of your original image.]

When you first start up your scanning software, you will probably have to do what's called a Preview scan. This lets the scanner (and you) see what the scanner sees. It lets you define the edges of the scanned region (the edges of the page) as well as set the next subject: Levels. Once you've set everything after the Preview, then you can click on the Scan button.



3. *Setting levels*

There's one more setting that warrants attention IF you have access to it, that's Levels. Levels lets you define what's white and what's black. Below is a scan from Fusion. On the left is a scan done without any setting of levels. In this image you can see the ghost image of the text on the reverse side and the white page comes out somewhat gray. On the right is the same image after adjustments for levels. The ghost text is gone and the page is now white.

1 The reason this is "at least" is because the more detail in an image the more storage space it uses. Thus, an image of a blank wall will be smaller than an image of a field of grass.

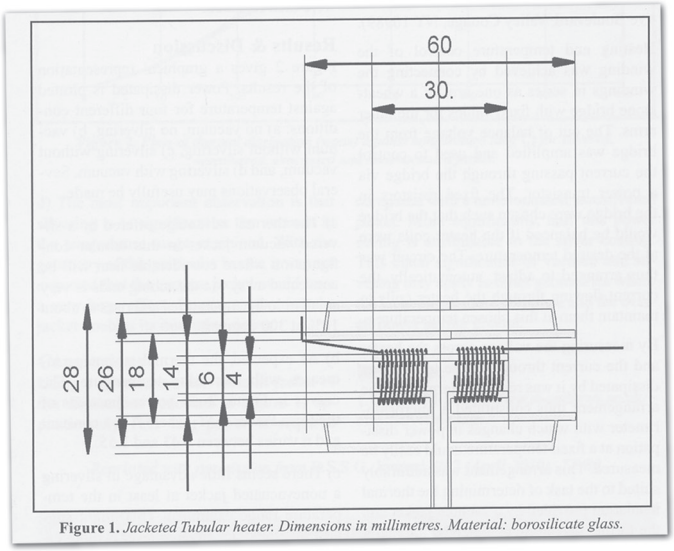


Figure 1. *Jacketed Tubular heater. Dimensions in millimetres. Material: borosilicate glass.*

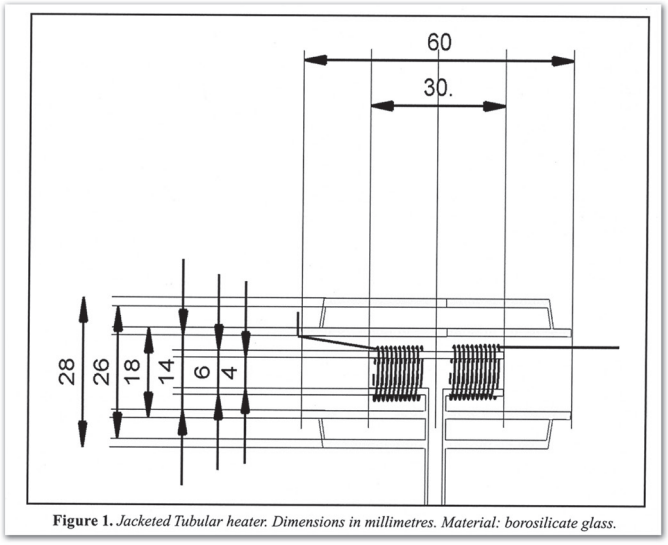
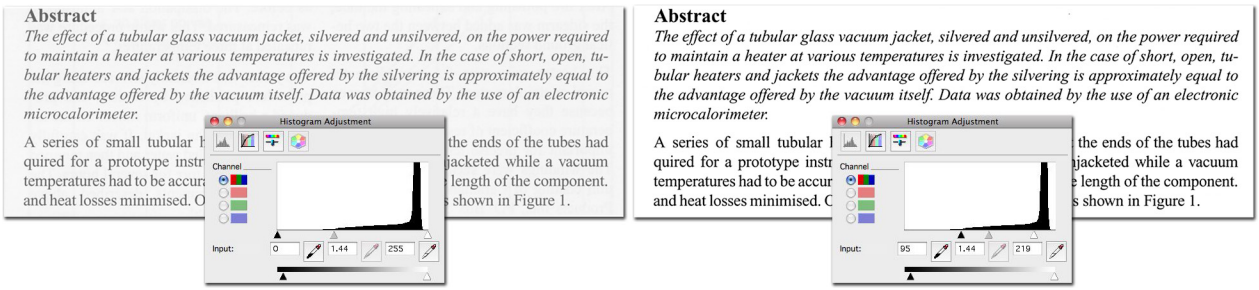


Figure 1. *Jacketed Tubular heater. Dimensions in millimetres. Material: borosilicate glass.*

For reasons I do not wish to get into in this monograph, you can do the Levels operation in either an application like Photoshop or at the scanner. Without going into details, there are substantial reasons that you should do the Levels corrections in the scanner.

Levels is done by selecting something that (in your software) looks something like this:

Once you click or select Levels, you then have something that may look like this:



Levels will always show some black hills, a black arrowhead on the left side of the hills and a white arrowhead on the right side of the hills and a gray arrowhead in the middle. While there will be a number of other settings, that's all you need to be concerned with. What Levels presents is a bar graph of the number of pixels at every level of gray from complete black (on the left) to complete white (on the right). Where the bars (hills) are tall means that there are a greater number of pixels of that shade of gray and where there are no hills means that there are no pixels of that shade of gray. What you see in the graph above on the left shows is that at complete white, there are no pixels. In fact the graph shows that as far as this image is concerned, the lightest pixels in the image are somewhat gray. What the "white" arrowhead lets you do is to tell the software what you want the software to consider as white. Thus, on the right side I've dragged that arrowhead a bit to the left so that what the scanner saw as gray, I want the scanner to consider white. You can see the results of this in the text above the Levels window. On the left hand side, the page is showing rather gray. On the right side, you can see that what is gray, I've defined as white and therefore the page is showing white. Depending on the nature of the original document, you may or may not wish to slide the black and/or gray sliders as well. (This is a visual call and you will probably need to play some with your scanner's software and learn what takes place. After a few scans you will have leaned what to expect and know what you should need to tweak.)

You can now scan your document

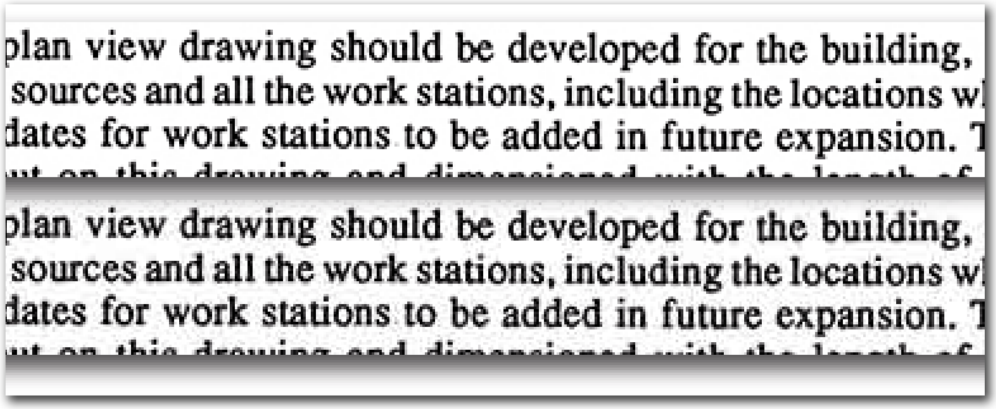
4. *Saving your documents*

If you ask 100 people which format they should save their image in, 98%[2] of them will probably select JPEG. That format is selected because it's so ubiquitous. People see it in images from their camera and images on the web. Simply, it's something they recognize. For many purposes, JPEG is a great format and very useful for many images. But not so much for scans that you want to convert into searchable PDFs.

2 I'm making this number up, don't quote me.

The better format is TIF (or TIFF). It doesn't make a difference about bite order or any other custom setting you may see. (If you see an option for LZG compression, that's a good one. It will not make any difference on the TIF images functionality, but will decrease the images' storage size.)

The big difference is that JPEG is a "lossy" format while TIF is not. What lossy means is that it compresses by deleting data from the image. When data is removed, the quality of the image can be compromised. The image below shows a TIF image on top and a JPEG (that's had considerable compression done) on the bottom. Notice the artifacts all over the text in the bottom image. This is called JPEG degradation and can significantly effect OCR processing. Normally you seldom see this, but if you look closely at an image with something like a telephone pole against a clear sky, you very well might.



Most of the time compression isn't great enough and/or you need something of high contrast against a flat surface for this to be seen: text is perfect for this to be displayed. If you have an image of something very busy, such as grass, you'd likely never see any degradation.

The advantage of JPEG is that it does a great job of compression and thereby has a small storage size. TIF images, on the other hand, are typically very large. The good news is that once the image has been converted into a PDF, the size will go down substantially. The size of a Searchable PDF varies depending on the amount of text (the more text the greater the size reduction) and there are a number of ways to generate the OCR results. In this test I use Acrobat Pro and did the OCR two different ways: Searchable Text and ClearScan.

Document type	Storage size (TIF)	PDF Size (as image)	Searchable PDF Size†	OCR Errors*	Searchable PDF Size‡	OCR Errors*
Photo setting, color scan, bad levels	13.5 MB	201 KB	193 KB	3	127 KB	4
Photo setting, color scan, good levels	13.5 MB	242 KB	233 KB	2	135 KB	2
Photo setting, grayscale scan, bad levels	4.5 MB	184 KB	188 KB	4	123 KB	3
Photo setting, grayscale scan, good levels	4.5 MB	225 KB	221 KB	1	127 KB	1
Document setting, color scan, bad levels	13.5 MB	201 KB	188 KB	3	127 KB	5
Document setting, color scan, good levels	13.5 MB	238 KB	229 KB	7	135 KB	4
Document setting, grayscale scan, bad levels	4.5 MB	184 KB	193 KB	3	127 KB	1
Document setting, grayscale scan, good levels	4.5 MB	221 KB	221 KB	0	127 KB	3

† "Searchable Image" setting from Acrobat Pro (Generally more accurate scans, required for Govt. work.)

‡ "ClearScan "setting from Acrobat Pro (Smaller scans, removes most of the text layer and substitutes a text layer over the original.)

** Please note the last column of OCR errors. This was done purely subjective based on either misidentified letter (e.g., "vacuwn" instead of "vacuum") or unnecessary added space (e.g., "l ength" instead of "length"). Due to the way the page was laid out, hyphenated words across two lines were split into two words and this was ignored (e.g., "tubular" was split into "tu-" and "bular"). Generally, the errors on documents with good quality levels were entirely either dropped or added spaces while some of the errors with bad quality levels changed letters in the words. In addition, original misspellings were also ignored.*

This was a small test and in no way can be considered definitive. Nonetheless, in general, good scans provided batter results. Setting Levels properly provides better scans and significantly improves the page visual quality. Following are two example pages, which would you rather be looking at?

Heat Loss From Silvered and Unsilvered Tubular Glass Vacuum Jackets

by

Dr. Kevin F. Scott[a]

Abstract

The effect of a tubular glass vacuum jacket, silvered and unsilvered, on the power required to maintain a heater at various temperatures is investigated. In the case of short, open, tubular heaters and jackets the advantage offered by the silvering is approximately equal to the advantage offered by the vacuum itself. Data was obtained by the use of an electronic microcalorimeter.

A series of small tubular heaters were required for a prototype instrument in which temperatures had to be accurately maintained and heat losses minimised. Other design constraints meant that the ends of the tubes had to be open and unjacketed while a vacuum jacket extended the length of the component. The arrangement is shown in Figure 1.

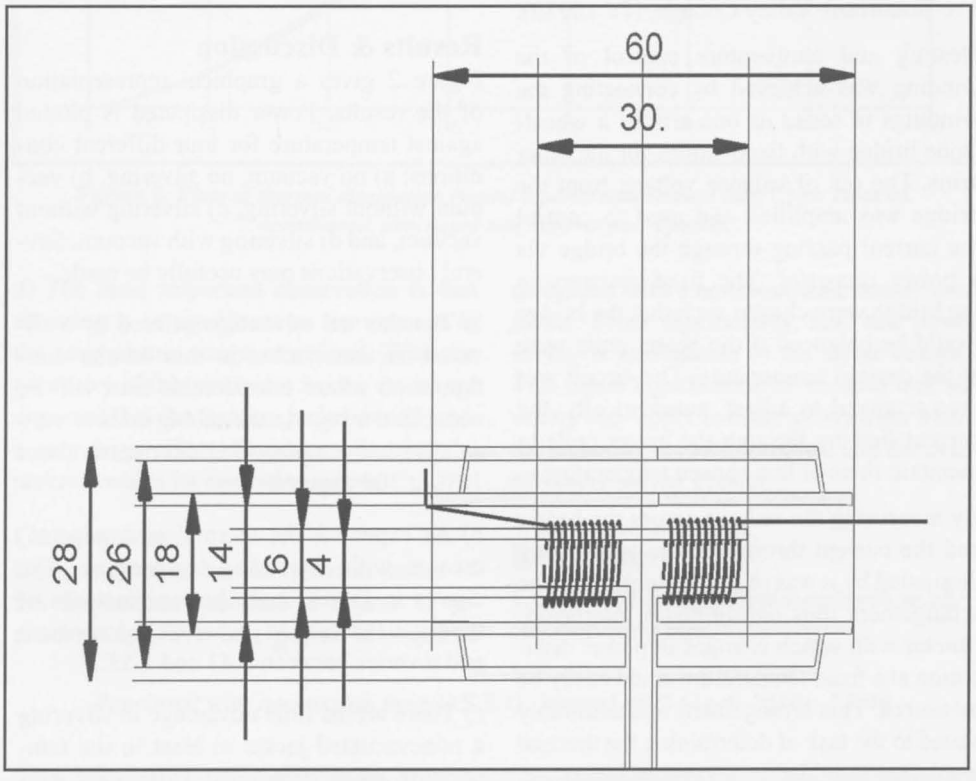


Figure 1. *Jacketed Tubular heater. Dimensions in millimetres. Material: borosilicate glass.*

[a] Meteorites Limited, Kirklands, Craigend Road, The Stow of Wedale, Galashiels, Selkirkshire, TD1 2RJ Scotland. E-mail: kevin.scott@meteormetrics.com.

Heat Loss From Silvered and Unsilvered Tubular Glass Vacuum Jackets

by

Dr. Kevin F. Scott[a]

Abstract

The effect of a tubular glass vacuum jacket, silvered and unsilvered, on the power required to maintain a heater at various temperatures is investigated. In the case of short, open, tubular heaters and jackets the advantage offered by the silvering is approximately equal to the advantage offered by the vacuum itself. Data was obtained by the use of an electronic microcalorimeter.

A series of small tubular heaters were required for a prototype instrument in which temperatures had to be accurately maintained and heat losses minimised. Other design constraints meant that the ends of the tubes had to be open and unjacketed while a vacuum jacket extended the length of the component. The arrangement is shown in Figure 1.

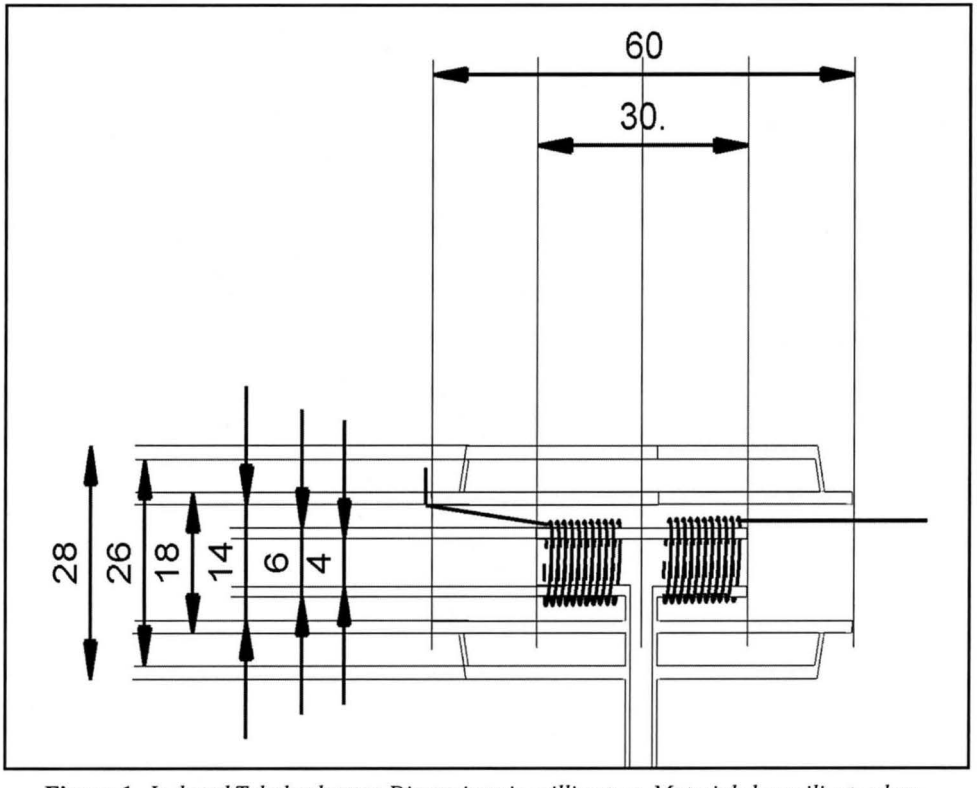


Figure 1. *Jacketed Tubular heater. Dimensions in millimetres. Material: borosilicate glass.*

[a] Meteorites Limited, Kirklands, Craigend Road, The Stow of Wedale, Galashiels, Selkirkshire, TD1 2RJ Scotland. E-mail: kevin.scott@meteormetrics.com.