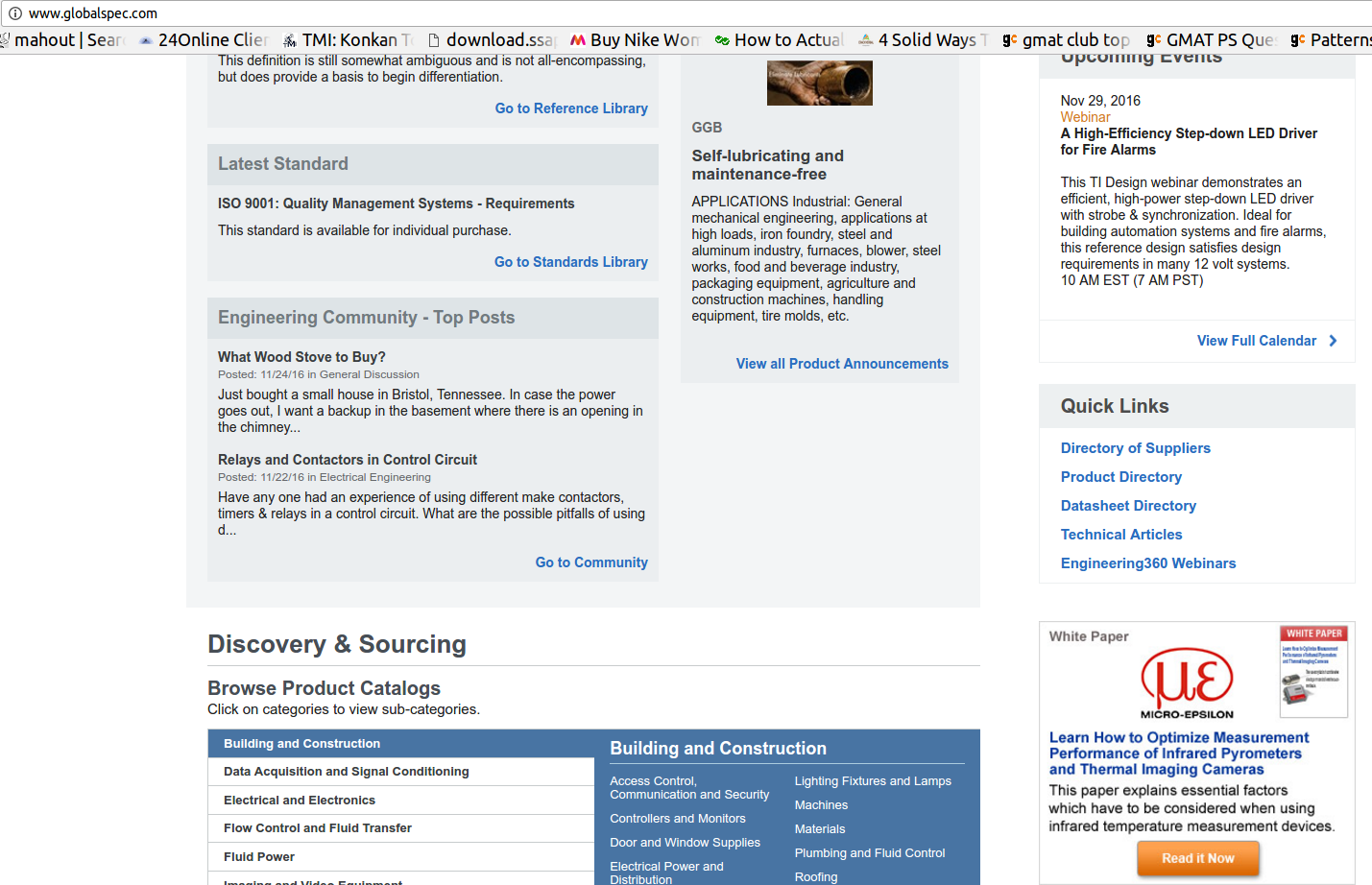
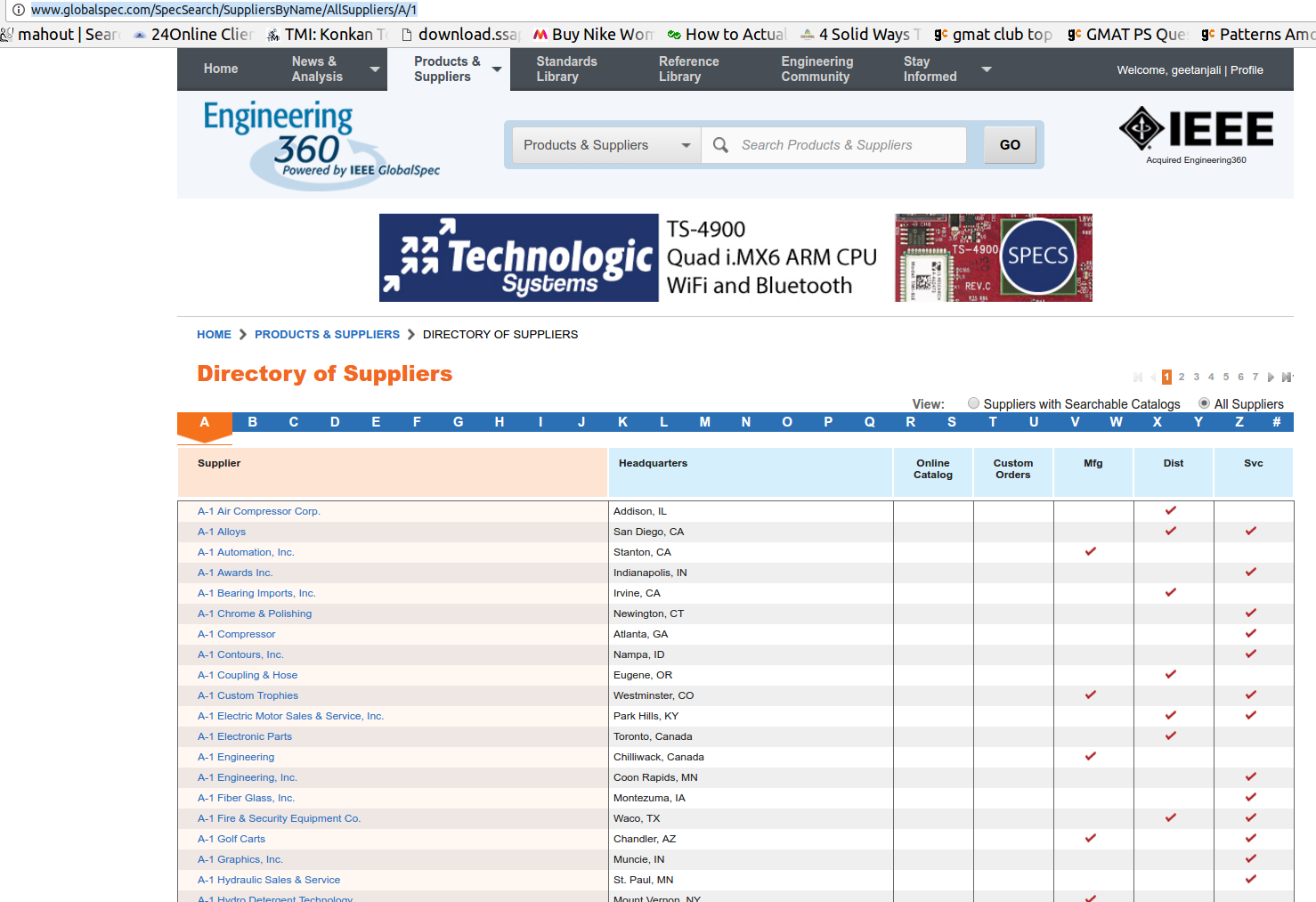
1. [www.globalspec.com](http://www.globalspec.com/)
   1. From the home page, we need to extract the two highlighted links
      1. Directory Of Suppliers
      2. Product Directory



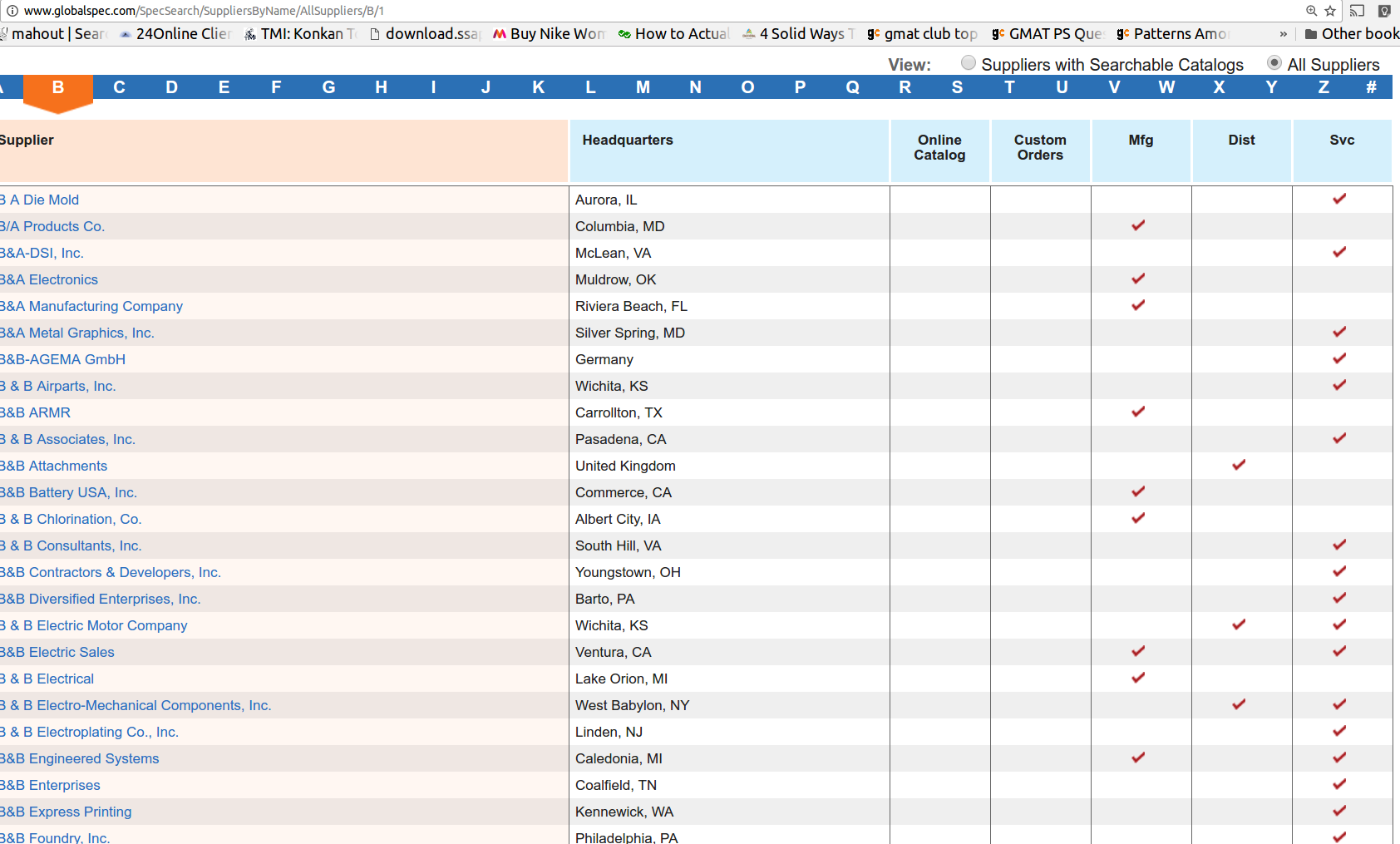
**Directory Of Suppliers**

1. <http://www.globalspec.com/SpecSearch/SuppliersByName/AllSuppliers/1/1>



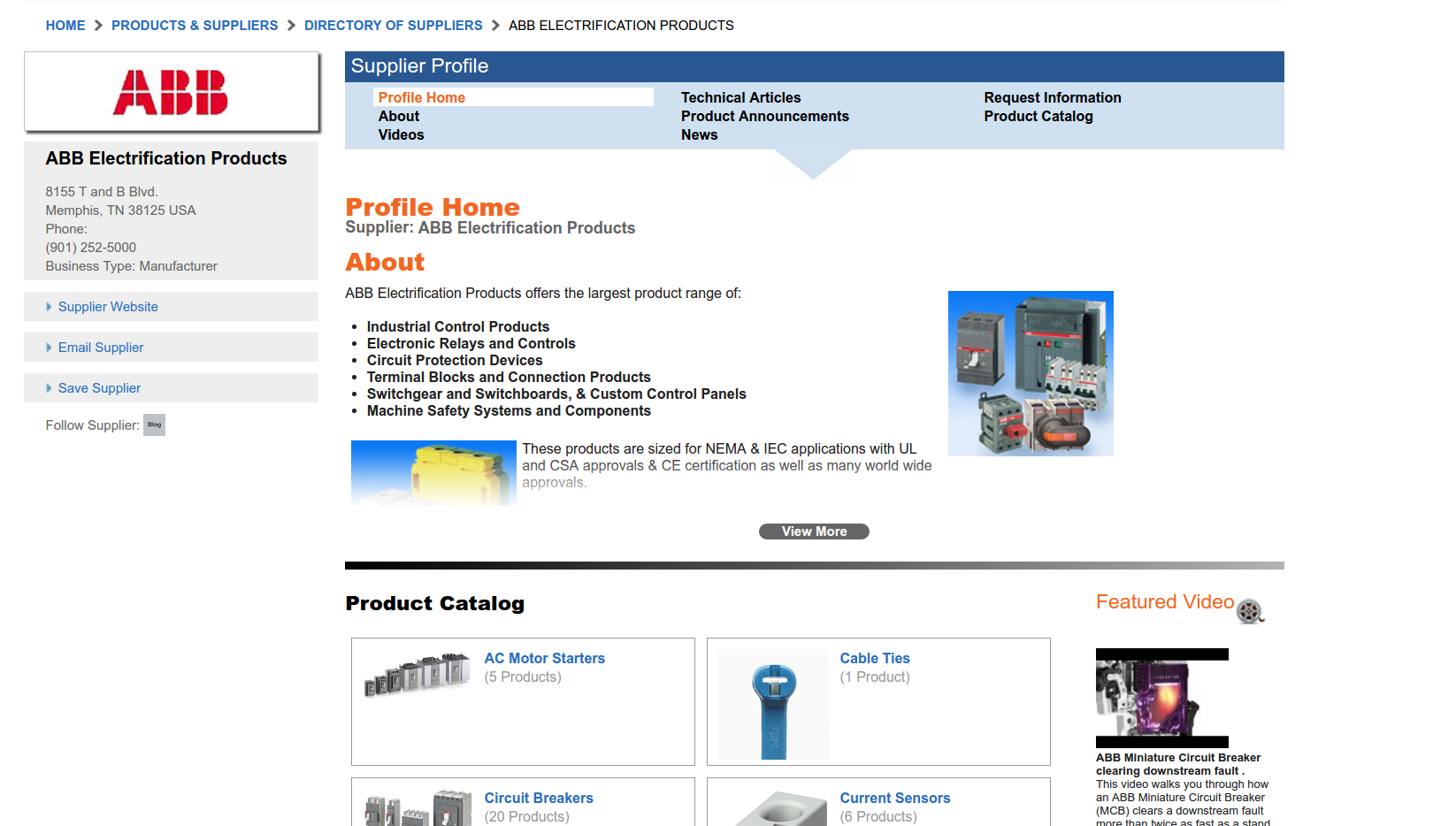
1. <http://www.globalspec.com/industrial-directory>

In the directory of suppliers, all links from A to # have to be scrapped



1. <http://www.globalspec.com/supplier/profile/ABBElectrificationProducts>
   1. This is the final HTML page to be scrapped for Directory Of Supplier
   2. **The Link present at “Email Supplier” needs to be extracted, decoded and kept in a seperate text file for further scrapping**

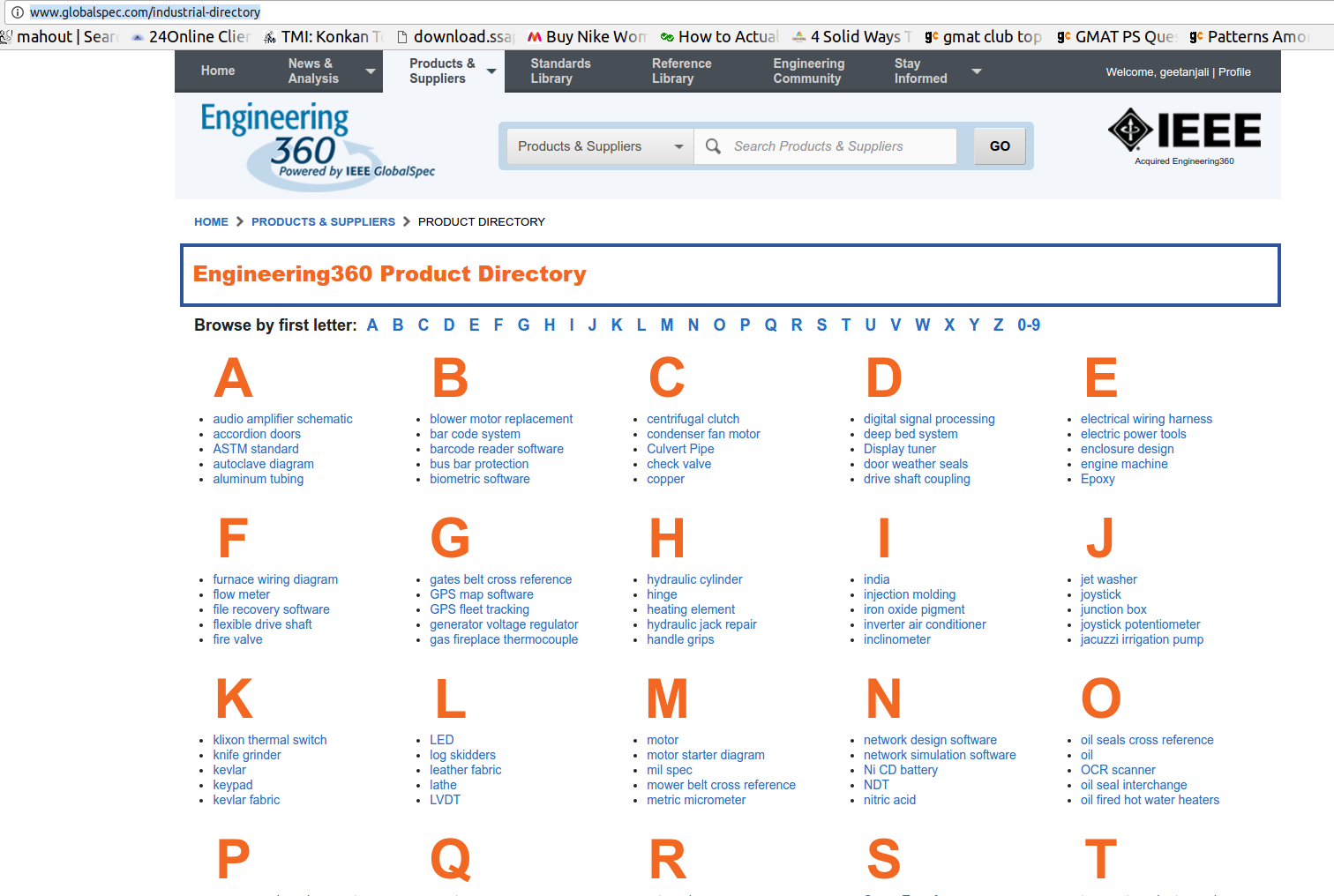
Unclear how to proceed, wait for reply



* All the media files (Image, video) and all PDFs have to be extracted and saved locally
* All the tabs under “Supplier Profile” have to be scrapped:
  + Profile Home, Technical Articles, Request Information, Product Catalog
  + About, Product Announcement, Videos, News

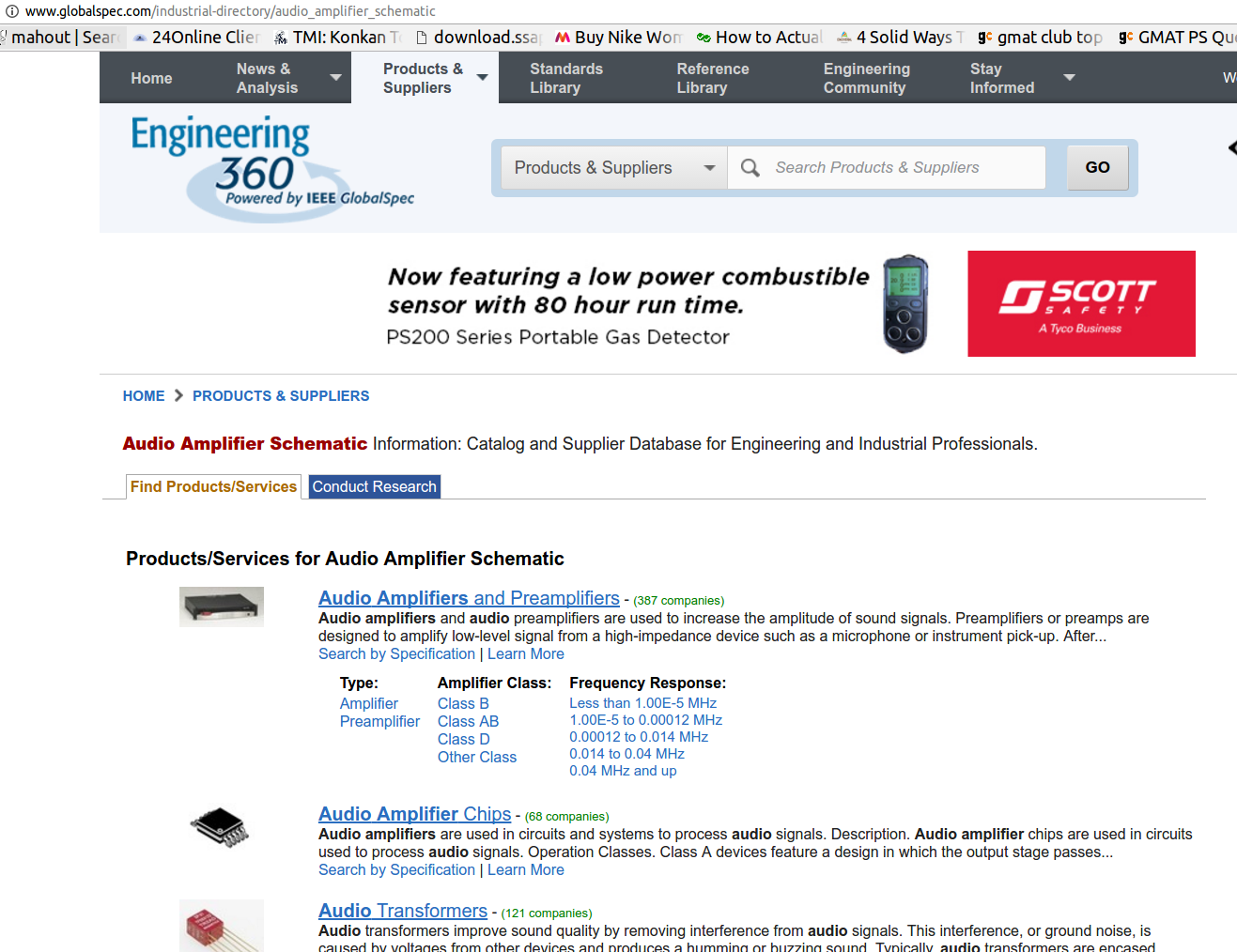
**Product Directory**

1. [http://www.globalspec.com/industrial-directory98924](http://www.globalspec.com/search/products?page=mi" \l "sqid=18975000&comp=2113&vid=98924)
   1. All the product links need to be scrapped from the directory
   2. The final page required is similar to the profile page as indicated above for supplier



E.g: Link scrapped:

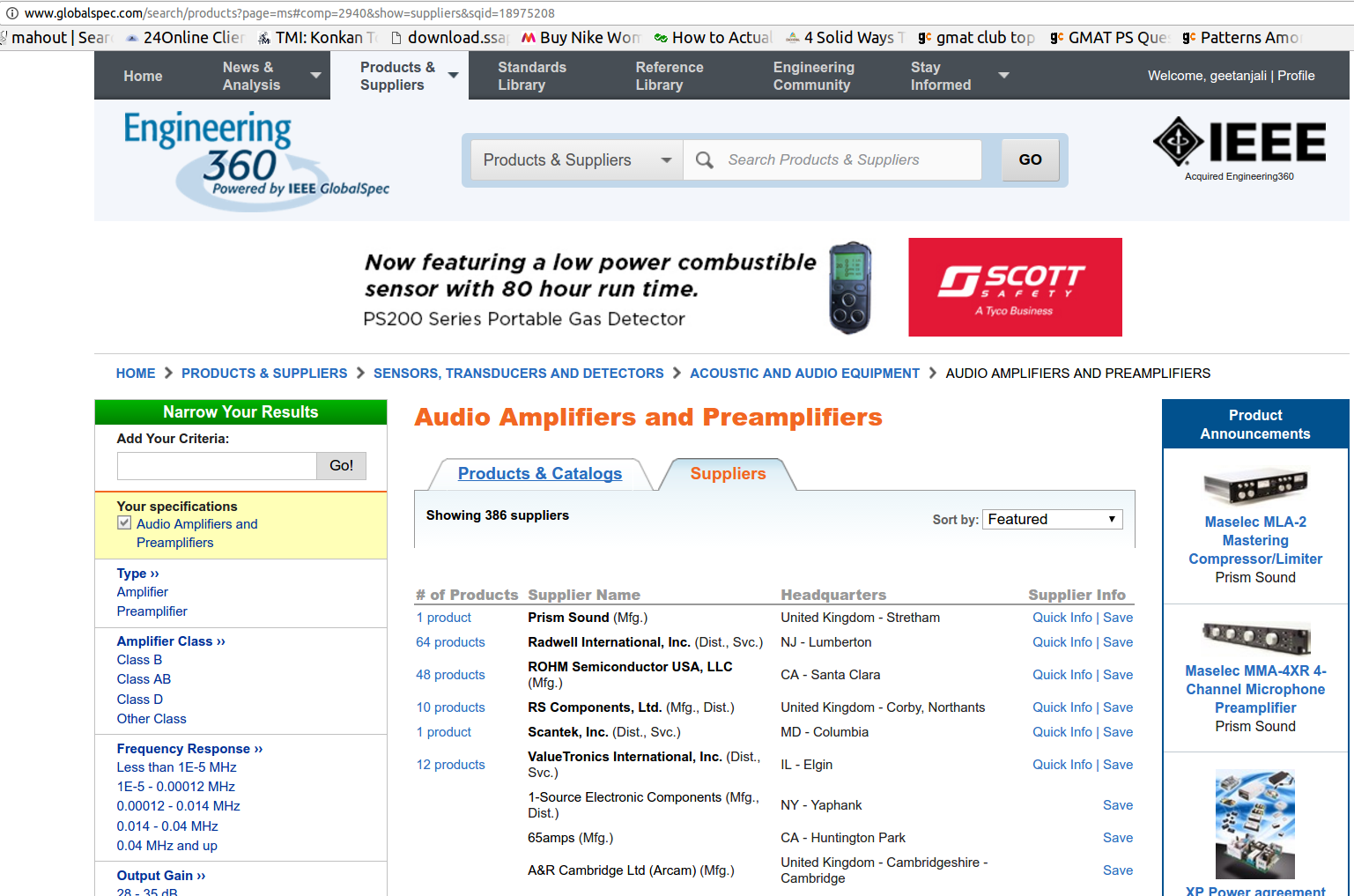
* <http://www.globalspec.com/industrial-directory/audio_amplifier_schematic>



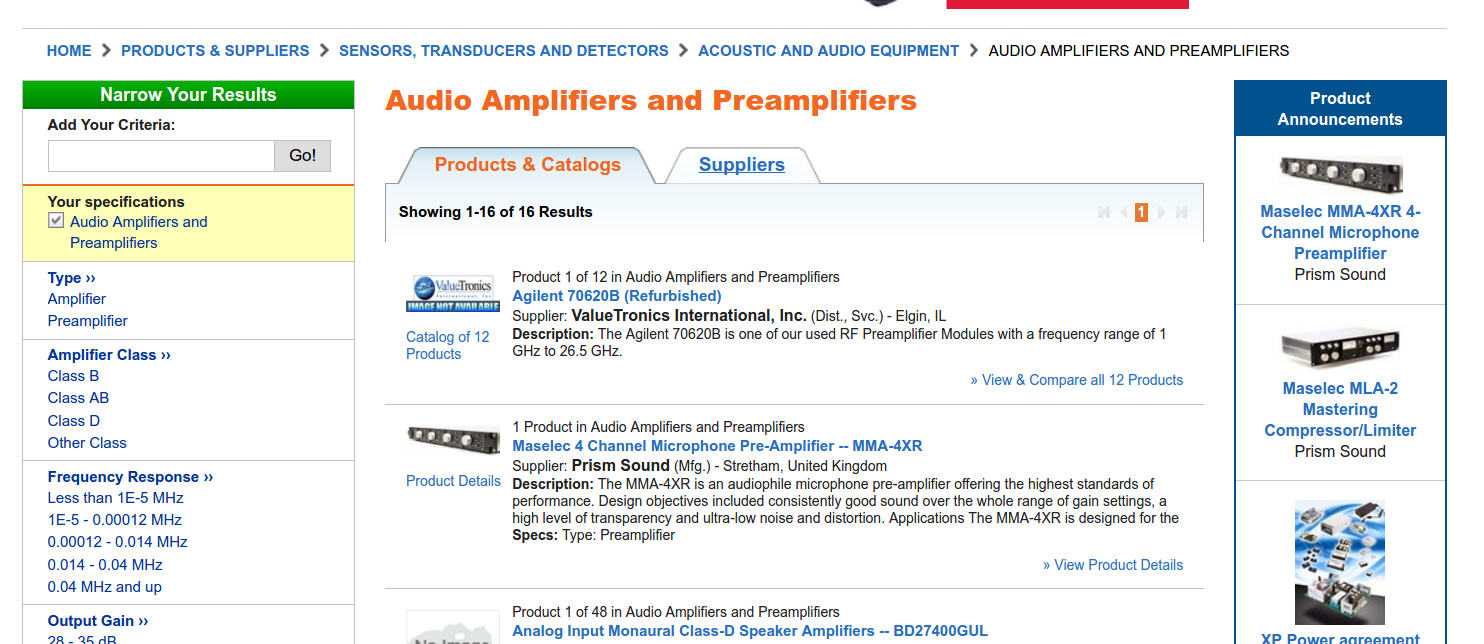
We will scrape all the product links here: Audio Amplifires and Preamplifiers, Audio Amplifier Chips, Audio Transformers, etc..

Inside ‘Audio Amplifires and Preamplifiers’, we get

* [http://www.globalspec.com/search/products?page=ms#comp=2940&show=products&sqid=18975208](http://www.globalspec.com/search/products?page=ms" \l "comp=2940&show=products&sqid=18975208)

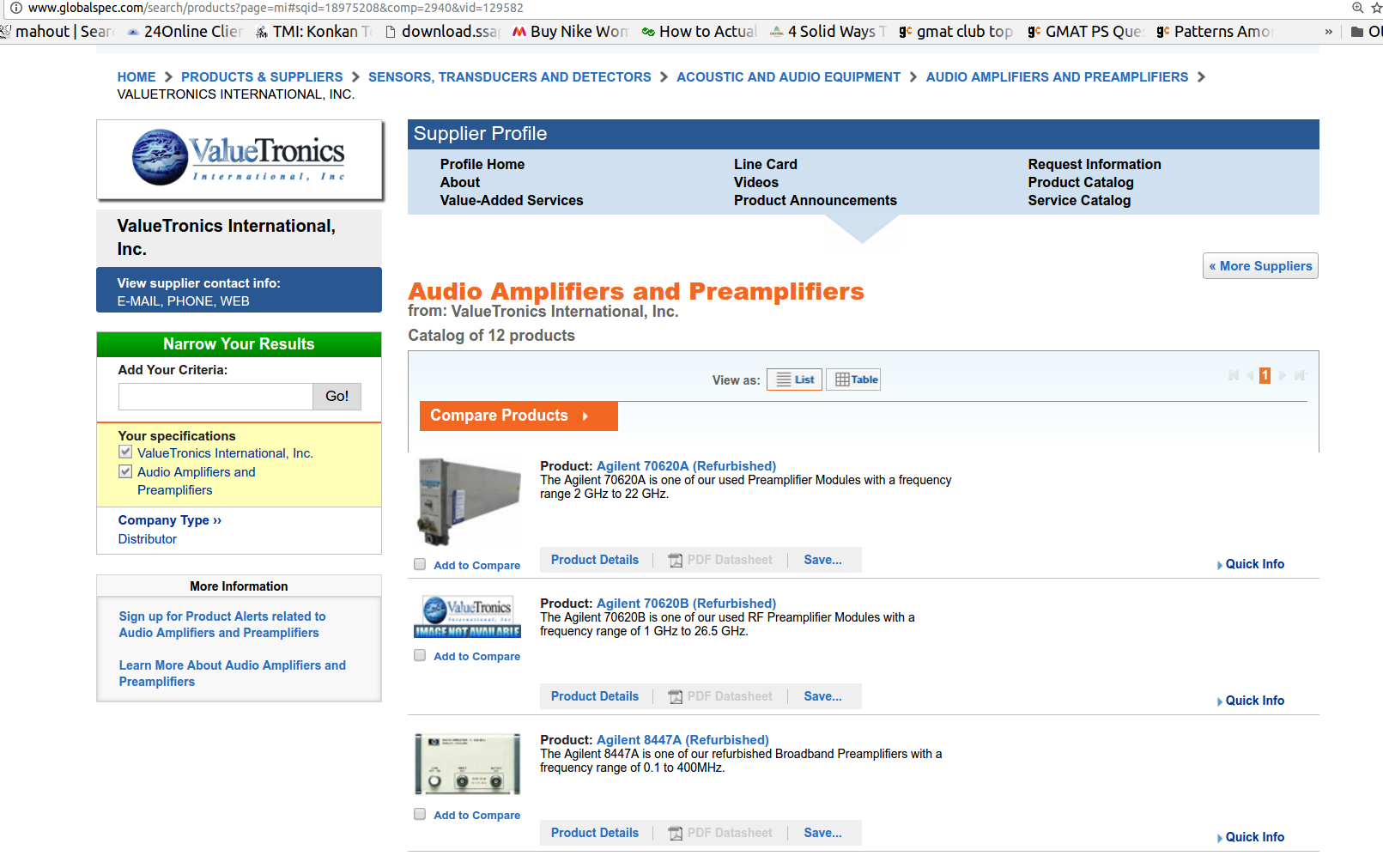


By default, it takes us at the Supplier link. We will need to get the Products & Catalogs link:

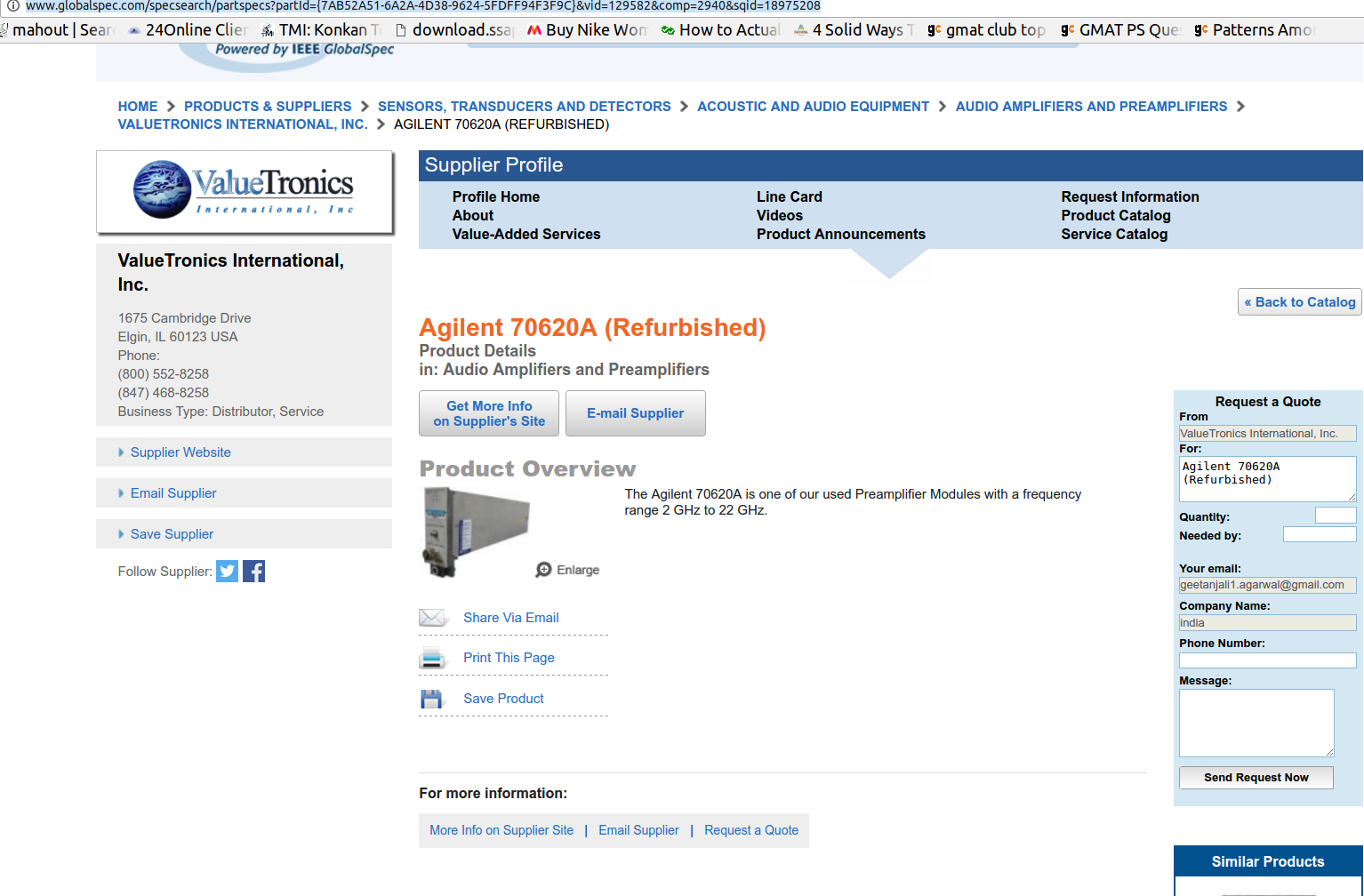


Here, we extract link to all the products links:

* [http://www.globalspec.com/search/products?page=ms#comp=2940&show=products&sqid=18982954](http://www.globalspec.com/search/products?page=ms" \l "comp=2940&show=products&sqid=18982954)



After extracting the link to final product pages, we get:

* <http://www.globalspec.com/specsearch/partspecs?partId={7AB52A51-6A2A-4D38-9624-5FDFF94F3F9C}&vid=129582&comp=2940&sqid=18975208>
* 
* This is the final page to be extracted as per the product directory
* All the media files, PDF, video, etc has to be extracted.
* The “Supplier website” needs to be extracted, decoded and input in a text file for further extraction.

**Where will we save all these HTMLs and media files**

* We will use MongoDb to save them
* All media files must be saved by the name = URL of the particular file
* These will be saved in the same collection as their corresponding HTMLs
* The HTML have to be arranged in MongoDB as per their Breadcrumb:

Selection_089

* This type of Breadcrumb is present on every page, and will be useful for us to keep track of every HTML page in MongoDB