1. Write a script to collect all websites mentioned in the two PDFs, Part15 PX2 Declaration WIllis GoDaddy.pdf and Domains (Name.com).pdf

In [1]:

```
import PyPDF2
import idna
import uritools
import appdirs
import urlextract
from urlextract import URLExtract
```

In this question, I will use Python url extractor to help me extract urls from the original PDF. The following is a small example:

In [2]:

```
# Build extractor object to extract url from string

extractor = URLExtract()
urls = extractor.find_urls('in the custody of GoDaddy.com, Inc')
print(urls)
```

```
['GoDaddy.com']
```

In [72]:

```
# Build pdf object for Domains.pdf

pdfFileObj = open('Domains.pdf', 'rb')
pdfReader = PyPDF2.PdfFileReader(pdfFileObj)
```

In [84]:

```
print('This pdf contains {} pages in total.'.format(pdfReader.numPages))
```

```
This pdf contains 26 pages in total.
```

In [85]:

```
# res is a list that contains website urls from each page

res = []

for i in range(pdfReader.numPages):
    pageObj = pdfReader.getPage(i)
    extractor = URLExtract()
    urls = extractor.find_urls(pageObj.extractText())
    res.append(urls[:])
```

```
In [86]: res
```

```
Out[86]: [[],
          [],
          [],
          [],
          [],
          [],
          [],
          [],
          [],
          [],
          [],
          [],
          [],
          [],
          [],
          [],
          [],
          [],
          [],
          [],
          [],
          [],
          [],
          [],
          [],
          []]
```

As we can see, there is no urls extracted from Domains.pdf.

```
In [3]: # Build pdf object for Part15 PX2 Declaration WIllis GoDaddy.pdf

        pdfFileObj = open('Part15 PX2 Declaration WIllis GoDaddy.pdf', 'rb')
        pdfReader = PyPDF2.PdfFileReader(pdfFileObj)
        print('This pdf contains {} pages in total.'.format(pdfReader.numPages))
```

```
This pdf contains 26 pages in total.
```

```
In [4]: # res is a list that contains website urls from each page

        res = []

        for i in range(pdfReader.numPages):
            pageObj = pdfReader.getPage(i)
            extractor = URLExtract()
            urls = extractor.find_urls(pageObj.extractText())
            res.append(urls[:])
```

```
In [5]: res
```

```
Out[5]: [[],
         [],
         [],
         [],
         [],
         [],
         [],
         [],
         [],
         [],
         [],
         [],
         [],
         [],
         [],
         [],
         [],
         [],
         [],
         [],
         [],
         [],
         [],
         [],
         [],
         []]
```

As we can see, there is no urls extracted from Domains.pdf.

After I manually check the two PDFs, I find there are acutally some website urls that cannot be extracted by the script. For example, the page 3-6 in WIllis GoDaddy.pdf contains some webiste urls. I think the script cannot extract the urls for two reasons: the first one is the pdf is vertically wrote instead of horizontally like what we normal read. The second one is PyPDF2.PdfFileReader is not suitable for reading the information of certain pdfs. I am still trying to work on this.

```
In [ ]:
```