Parsing Analysis

This notebook is trying to analyse the accuracy of the data parsed from the raw download files, to test its accuracy and show that it has been done to a good standard; and identify ways in which this data needs to be cleaned. A comparison of our results to valverde can be found in "Analysis_Reproducing_Valverde.Rmd".

We aim to:

- Parse the 2001 data both through sgml, and .txt formats and compare the differences.
- Compare patent counts to uspto summary statistics and "...lst.txt" files which list patents supposed to be present in the files which have been parsed.

In the original data there are 3 different data formats for different periods of time (txt, sgml, xml), only one year 2001 has a crossover of two of those formats, as txt was prioritised we parse the sgml files for that year to compare.

```
# Parse
source("Parse_directory3.R")
parse_directory(2001, doc_type = "sgml", write.dir = "../DataFiles/SampleFiles/2001sgml")
```

Comparing Patent Files

First we apply the basic cleaning of **removing duplicated entries** and adding a date measure converted into **posixct date format**.

```
# sgml
sgml2001 <- read_csv("../DataFiles/SampleFiles/patent/2001.csv", col_names = c("Patent", "Date", "Order
sgml2001 <- unique(sgml2001)
sgml2001$Date2 <- lubridate::ymd(sgml2001$Date)

# Txt
txt2001 <- read_csv("../DataFiles/Processed/patent/2001.csv", col_names = c("Patent", "Date", "Order"))
txt2001 <- unique(txt2001)
txt2001$Date2 <- lubridate::ymd(txt2001$Date)
```

```
## Warning: 1 failed to parse.
```

```
head(sgml2001)
head(txt2001)
```

```
## Source: local data frame [6 x 4]
##
##     Patent     Date Order      Date2
##      <chr>    <int> <int>     <date>
## 1 D0435713 20010102     8 2001-01-02
## 2 D0435714 20010102     6 2001-01-02
## 3 D0435715 20010102     7 2001-01-02
## 4 D0435716 20010102     6 2001-01-02
## 5 D0435717 20010102    15 2001-01-02
## 6 D0435718 20010102    17 2001-01-02
## Source: local data frame [6 x 4]
##
##      Patent     Date Order      Date2
##       <chr>    <chr> <chr>     <date>
## 1    Patent     Date Order       <NA>
```

```
## 2 D04357132 20010102     8 2001-01-02
## 3 D04357140 20010102     6 2001-01-02
## 4 D04357159 20010102     7 2001-01-02
## 5 D04357167 20010102     6 2001-01-02
## 6 D04357175 20010102    15 2001-01-02
```

Looking at the head of each file we can see a difference, the **text parsing has an additional digit** at the end of each patent number. If we use the patent search feature on the uspto website. We can see find that the actual patent numbers do not have this extra digit, also they **do not have any '0's at the start** (after the type character).

```
norm <- grepl("^(D|RE|PP|H|T)*0", txt2001$Patent)
sum(norm)
length(txt2001$Patent)
```

```
## [1] 184172
## [1] 184173
```

We can see that all the patents processed follows the pattern of a class tag (D,RE,PP,H,T) followed by at least one 0 followed by the patent number. It seems the purpose of the 0 is to pad the patent number to be the same total number of characters.

We remove the final digit in the txt parsing as an additional cleaning step then compare the two.

```
# Remove the extra trialing character from text patent numbers
txt2001$Patent <- sapply(txt2001$Patent, function(x) substring(x, 1, nchar(x) - 1))

# Remove the 0s from text patent numbers
txt2001$Patent_Raw <- txt2001$Patent
type_txt <- stringr::str_extract(txt2001$Patent, "^(D|RE|PP|H|T)")
type_txt[is.na(type_txt)] <- ""
rem0_txt <- sub("^(D|RE|PP|H|T)*0+", "", txt2001$Patent)
txt2001$Patent <- paste0(type_txt, rem0_txt)

# Remove the 0s from sgml patent numbers
sgml2001$Patent_Raw <- sgml2001$Patent
type_sgml <- stringr::str_extract(sgml2001$Patent, "^(D|RE|PP|H|T)")
type_sgml[is.na(type_sgml)] <- ""
rem0_sgml <- sub("^(D|RE|PP|H|T)*0+", "", sgml2001$Patent)
sgml2001$Patent <- paste0(type_sgml, rem0_sgml)
```

```
## [1] "Patents present in sgml but not txt: 0"
## [1] "Patents present in txt but not sgml: 95"
## [1] "Number of observations not identical in both: 95"
## Source: local data frame [95 x 5]
##
##     Patent     Date Order      Date2 Patent_Raw
##      <chr>    <chr> <chr>     <date>      <chr>
## 1    Paten     Date Order       <NA>      Paten
## 2  6173583 20010116    52 2001-01-16   06173583
## 3  6175216 20010116    29 2001-01-16   06175216
## 4  6175306 20010116     3 2001-01-16   06175306
## 5  6177664 20010123     7 2001-01-23   06177664
## 6  6179725 20010130    14 2001-01-30   06179725
## 7  6180216 20010130    42 2001-01-30   06180216
## 8  6181581 20010130    10 2001-01-30   06181581
## 9  6181986 20010130    10 2001-01-30   06181986
```

```
## 10 6182450 20010206    12 2001-02-06   06182450
## ..     ...      ...   ...        ...        ...
```

- We find that the text parsing has 95 more entries 184078 vs. 184173
- We find there are 94 patents present in the text parsing but not in the sgml parsing but none vice versa.
  - Looking at a few of these individually we can confirm that they are patent records rather than accidental parsing of the wrong information.
  - **Todo: talk about entries not being the same after reparsed txt with bug fix**

Patent Counts

For this section we will use the txt data as it is slightly more reliable. The uspto publish summary statisics for patent counts on their website and in later years datafiles are partnered with text files listing the patents present within them (in addition pdfs summarising that week include lists of patents absent from those files, however this is not easily parsable)

```
our_data <- read_csv("../DataFiles/Cleaned/patent_cat.csv", progress = FALSE)
```

```
our_data$Year <- lubridate::year(our_data$Date2)
our_data_summary <- our_data %>% group_by(Year) %>% summarise(count = n()) %>% filter(Year %in% 1976:20
```

Unfortunately the data is not easily parsable so has been hard copied from the website, we can compare this with our data.

```
yearcounts <- rev(c(325979,326032,302948,276788,247716,244341,191927,185224,182899,196405,157718,181299
```

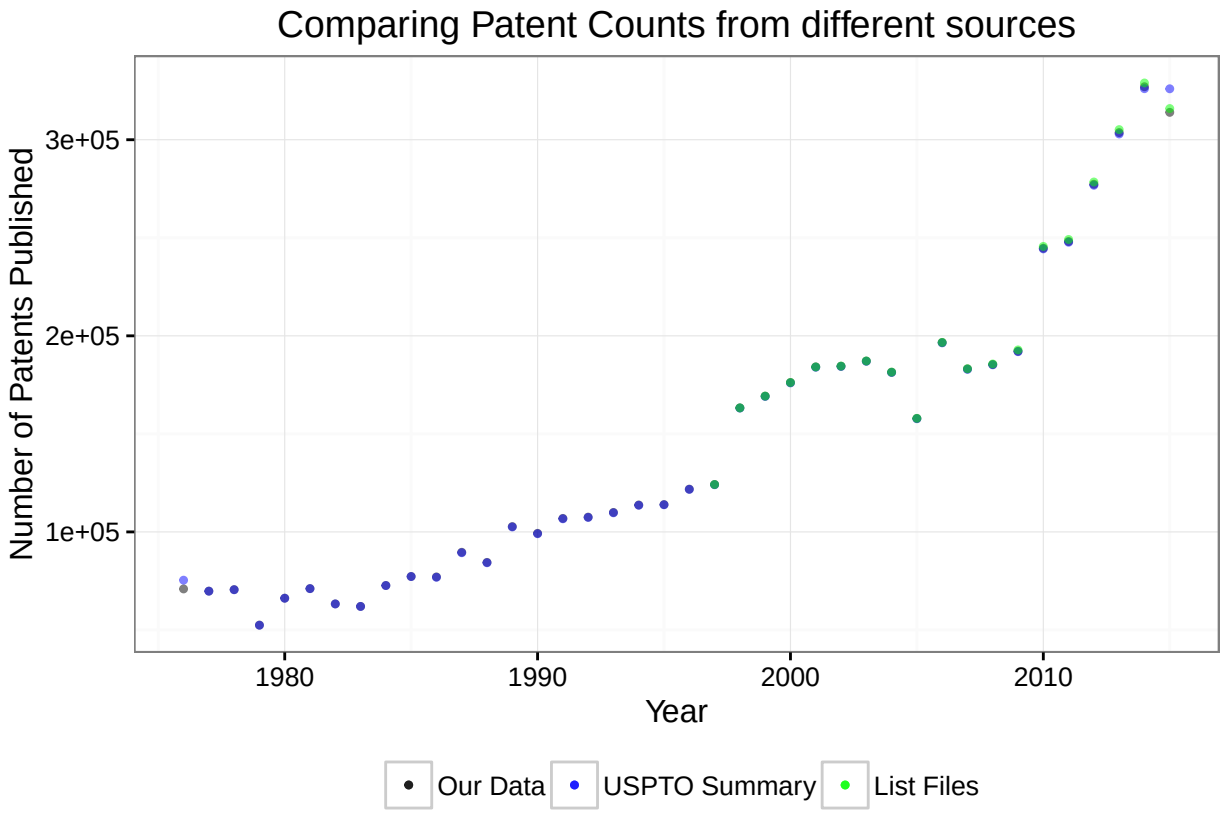
Comparing patent counts to uspto list files

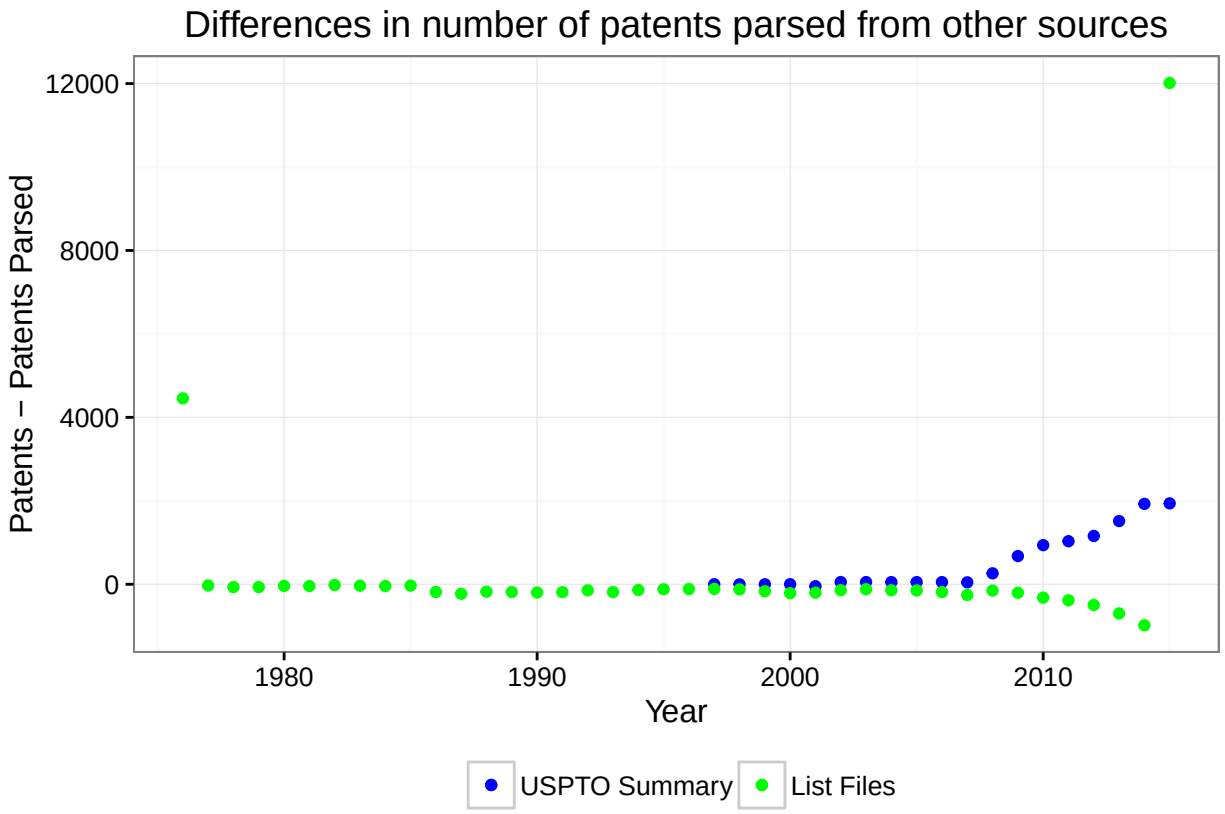
The list files are present in the raw data from 1997. They are a mirror of a data file containing a list of the patent numbers recorded in that file.

```
yrs <- (1997:2015)
lst.full <- NULL
for (yr in yrs) {
    path <- paste0("../DataFiles/Raw/", yr)
    files <- list.files(path, full.names = TRUE)
    files <- stringr::str_subset(files, "lst")
    cat <- stringr::str_subset(files, "cat")
    if (length(cat) > 0 ) files <- cat

    lst.cat <- NULL
    for (file in files) {
        lst <- readr::read_lines(file)
        lst.cat <- c(lst.cat, lst)
    }
    lst.full[[make.names(yr)]] <- lst.cat
}
saveRDS(lst.full, "Dat/lst.rda")
```

```
lst.full <- readRDS("Dat/lst.rda")
yrs <- sapply(names(lst.full), function(x) as.numeric(substring(x, 2)))
npats <- sapply(lst.full, length)
lst.df <- data.frame(year = yrs, count = npats)
```


Comparing Patent Counts from different sources
Number of Patents Published
3e+05
2e+05
1e+05
1980
1990
2000
2010
Year
Our Data
USPTO Summary
List Files



- There are two large errors, one in 1976 for unknown reasons, one in 2015 due to the 2 weeks of data not being parsed, this is why the error exists between the summary statistics but not between the listed values.
- With the exception of these two years our parsed data contains more patents than the reported sumamries.
- In recent years (since 2008) the difference between these summaries and the parsed data increases substantially. In addition list files begin to describe larger numbers of patents which are not parsed up to ~2000. This is less than 1% of the data but a worrying trend.

```
## Source: local data frame [1 x 8]
##
##    Year count.txt count.lst count.summary diff.lst diff.summary count.sgml
##   <dbl>     <int>     <int>         <dbl>    <int>        <dbl>      <int>
## 1  2001    184172    184125        183970      -47         -202     184078
## Variables not shown: diff.sgml <int>.
## [1] "Patents in list file but not txt parse: 47"
## [1] "Patents in txt parse but not list file: 95"
## [1] "Patents in sgml parse but not list file: 0"
```

Comparing Citation Files

Document where NA values are found

```
sgml2001_cit <- read_csv("../DataFiles/SampleFiles/2001sgml/citation/2001.csv",
                         col_names = c("Patent", "Citation", "Date", ""),
```

```
                          progress = FALSE, col_type = "cccc")

# For some reason date is split between two columns, merge them
index <- is.na(sgml2001_cit[,4])
sgml2001_cit[!index,3] <- sgml2001_cit[!index, 4]
sgml2001_cit[,4] <- NULL
# Add one to date
sgml2001_cit$Date <- as.character(as.numeric(sgml2001_cit$Date) + 1)
```

```
## Warning: NAs introduced by coercion
```

```
sgml2001_cit <- unique(sgml2001_cit)
sgml2001_cit$Date2 <- lubridate::ymd(sgml2001_cit$Date)
```

```
## Warning: 746 failed to parse.
```

Sources of NA or missing values:

- File Parsing
- Errors in parsing the raw data causes missing entries rather than NA values, this effect is analysed in the next section.
- E.g. There are 2 weeks in 2015 where the uspto website didn't have the data availible (denied access error code).
- Parsing, parsing date as an integer when reading the data creates a small number of NA values.

Conclusion

- Parsing Text is slightly more complete than parsing sgml.
  - We found using the 2001 data (which has both txt and sgml) that the text parsing has more entries 184078 vs. 184173, and that all the entries in sgml are present in txt.
- There are patent files parsed which are not listed by the uspto summary statistics.
  - Differences in parsed vs summary statistics or lst files only appear after ~ 2008.
  - 1976 and 2015 are anomylous in terms of patent numbers parsed (2015 due to 2 weeks data un-obtainable, 1976 for unknown reasons)
- Cleaning steps required includles:
  - Removing duplicated entries
  - Removing extra digit at end of txt patent numbers
  - Removing "filler" 0s in all patent patent numbers so that they match citation patent numbers and website.