About the Open Source Patent Analytics Manual Project

The aim of this project is to provide an overview and practical manual on the use of open source and free software tools for patent analysis. The main outcome of the project will be an Open Source Patent Analytics Manual as a reference guide and a set of electronic resources including walkthroughs, videos and code that can be freely used by anyone. The development of the Manual is supported by the [World Intellectual Property Organisation (WIPO)](http://www.wipo.int/portal/en/) as part of its [Patent Landscapes Project](http://www.wipo.int/patentscope/en/programs/patent_landscapes/) which has produced a range of patent landscape reports to inform policy debates and decision-making on issues of interest to developing countries. The Manual is mainly intended for researchers, patent professionals and patent offices in developing countries. However, we expect that it will be of wider interest to researchers and patent professionals.

Patent data is important because it is a valuable source of technical information that can inform decision-making on whether or not to pursue a particular avenue of research and development, whether to license a particular technology, or whether to pursue product development in particular markets. Patent data is also important in economic and policy terms because it provides a key indicator and insight into trends in science and technology. Patent data is commonly used by organisations such as the [OECD](http://www.oecd.org/sti/inno/oecdpatentdatabases.htm), [EUROSTAT](http://ec.europa.eu/eurostat/statistics-explained/index.php/Patent_statistics) and others to report on trends in research and development. Researchers, including the authors of this manual, use patent data to investigate issues such as the role of genetic resources in innovation and new and emerging areas of science and technology such as synthetic biology and geoengineering [see About].

Patent activity can also be controversial. Important controversies over the last 20 years include DNA patents, software patents, patents on business methods, the rise of patent 'trolls' and the implications of the internationalisation of patent activity for developing countries. The free software and open source movements (based on the flexibilities in copyright law) are in part a response to the controversies that have arisen around proprietary software models involving copyright and patents and a desire to do things differently. This has led to new models for sharing data, cooperation in innovation and new business models. In particular, a wide range of free software tools are now available for research and analysis. This project provides an overview of the available tools for patent analysis and explores a small number in greater depth.

### Topics

We will mainly focus on answering three main questions:

* How to obtain patent data in a form that is useful for different types of analysis?
* How to tidy, analyse, visualize and share patent data using open source and free software?
* How to promote confidence that the results of analysis are transparent and robust for informed decision-making?

In approaching these issues we will organise the Manual and materials into seven main topics:

1. Approaching Patent Data
2. Obtaining Patent Data
3. Cleaning and Tidying Patent Data
4. Analysing Patent Data
5. Visualising Patent Data
6. Sharing Patent Data
7. Tools and Resources

As a project focusing on open source and free tools, all data and tools developed for the manual will be made available through the [GitHub project repository](https://github.com/poldham/opensource-patent-analytics). We encourage you to take a look at the repository and it will fill up as the project develops. To get started with GitHub and join in with the project sign up and then install [GitHub](https://github.com). It's actually quite easy.

We will now take a quick look at the background to the topics.

### Approaching Patent Data

One feature of the project is that it assumes no prior knowledge of patent data and the emphasis is on learning by doing. We are preparing a series of short walkthough videos for people new to patent data and patent analytics through the project [YouTube Channel](https://www.youtube.com/channel/UCwFhEASbKdm6WYoax73XsOw). Since it is quite hard to know everything about specific areas of the patent system, particular software tools or programming languages, we are hoping that specialists will contribute with tips and information on tools or approaches they find useful. You can contribute, ask questions, or raise issues through the [GitHub project repository](https://github.com/poldham/opensource-patent-analytics/issues) or start a discussion by commenting at the bottom of blog posts such as this one.

### Obtaining Patent Data

One major challenge in understanding the implications of patent activity, either in fields such as climate change technologies, software or pharmaceuticals is accessing and understanding patent data. Recent years have witnessed a major shift towards the use of open source research tools and the promotion of open access to scientific data along with the promotion of open science. One of the main purposes of the patent system is to make information on inventions available for wider public use. The patent system has responded to this through the creation of publicly accessible databases such as the [European Patent Office espacenet database](http://worldwide.espacenet.com/?locale=en_EP) containing millions of patent records from over 90 countries and organisations. Others initiatives to make patent data available include [Patent Scope](https://patentscope.wipo.int/search/en/search.jsf), [Google Patents](https://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCIQFjAA&url=http%3A%2F%2Fwww.google.com%2Fpatents&ei=7LFIVeyiIInmau2rgMgL&usg=AFQjCNG_XlAI_9dSaH28NeN5O6bXJSSuSw&sig2=lgFU0x6MQCnaWZBXgPYDAA&bvm=bv.92291466,d.d2s) and [Patent Lens](http://www.lens.org/lens/) and [Free Patents Online](http://www.freepatentsonline.com) .

In the case of the United States it is possible to bulk download the entire USPTO collection through the [Google Bulk Download of USPTO patents](https://www.google.com/googlebooks/uspto.html). A range of commercial providers such as [Thomson Innovation](http://info.thomsoninnovation.com) and [PatBase](https://www.patbase.com/login.asp), among others, provide access to patent data and, in the case of Thomson Innovation, add additional information through the [Derwent World Patent Index](http://thomsonreuters.com/en/products-services/intellectual-property/patent-research-and-analysis/derwent-world-patents-index.html). As such there is an ecosystem of patent information sources and providers out there. As we will see, obtaining patent data in the quantity and with the coverage needed, and with the desired fields for analytics purposes, can be a significant headache. This project will walk through the different information servies and go into detail on those free services that are the most useful for patent analytics.

### Cleaning and Tidying Patent Data

Anyone familiar with working with data will know that the majority of the work is taken up with cleaning data prior to analysis. In particular data from different patent databases typically involves different cleaning challenges. Most of these challenges involve cleaning inventor and applicant names or cleaning text fields prior to analysis.

In the course of the project we will be following the ideas contained in two recent publications on data science.

1. Jeff Leek's [The Elements of Data Analytic Style](https://leanpub.com/datastyle)(available free of charge if required)
2. Hadley Wickham [Tidy Data](http://vita.had.co.nz/papers/tidy-data.pdf) and this [video](https://vimeo.com/33727555)

We suggest that you take a look at these papers because they contain core ideas for effective approaches to working with patent data.

### Analysing Patent Data

Analysis of patent data is the core event but is closely linked to visualizing patent data. The core questions in patent analysis are: who, what, where, when, how, and with what? The way in which we approach these questions will depend on the goal of the patent analysis. However, in almost all circumstances realising that goal will depend on combinations of answers to the core questions.

We will focus on a range of tools, from the simple such as Open Office and Excel to the complex. In particular we will be using [RStudio](http://www.rstudio.com) as a key tool because of the range of analysis and visualization tasks that can be performed using R as well as the huge amount of free resources and help that is out there. We do not expect anyone to have any prior knowledge of R and again our approach is one of learning by doing. When it comes to R we are doing the same thing. If you would like to get ahead we suggest downloading the [RStudio preview edition for your platform](http://www.rstudio.com/products/rstudio/download/preview/) and trying [DataCamp](https://www.datacamp.com) as an introduction and [R-Bloggers](http://www.r-bloggers.com) for an overview of what's hot and tutorials.

### Visualising Patent Data

In recent years, the growing availability of open source software tools has witnessed an increasing shift towards the visualisation of complex data. The reason for this is that humans are better at processing and understanding visual representations (such as graphs, networks or maps) than simply staring at spreadsheets or words. We will cover important tools in the visualisation toolbox including network visualisation and geographic mapping and interactive visualisation using [Tableau Public](https://public.tableau.com/s/gallery), [Plotly](https://plot.ly/) and [Shiny](http://shiny.rstudio.com/gallery/) among other tools.

### Sharing Patent Data

Here we will be looking at options for storing and sharing patent data with others, taking into account issues around confidentiality.

### Tools and Resources

One of our aims is to generate an open access list of tools and resources that others working on patent analysis find useful. The tools and resources section is likely to grow considerably as the project develops. We won't be able to review them all, but we will be able to make information on them available

Please feel welcome to submit information onother free tools that you think will be useful to me at the project email address [email](mailto:%20openpatentanalytics@gmail.com). We also welcome contributions from programmers interested in exploring the potential of patent data.