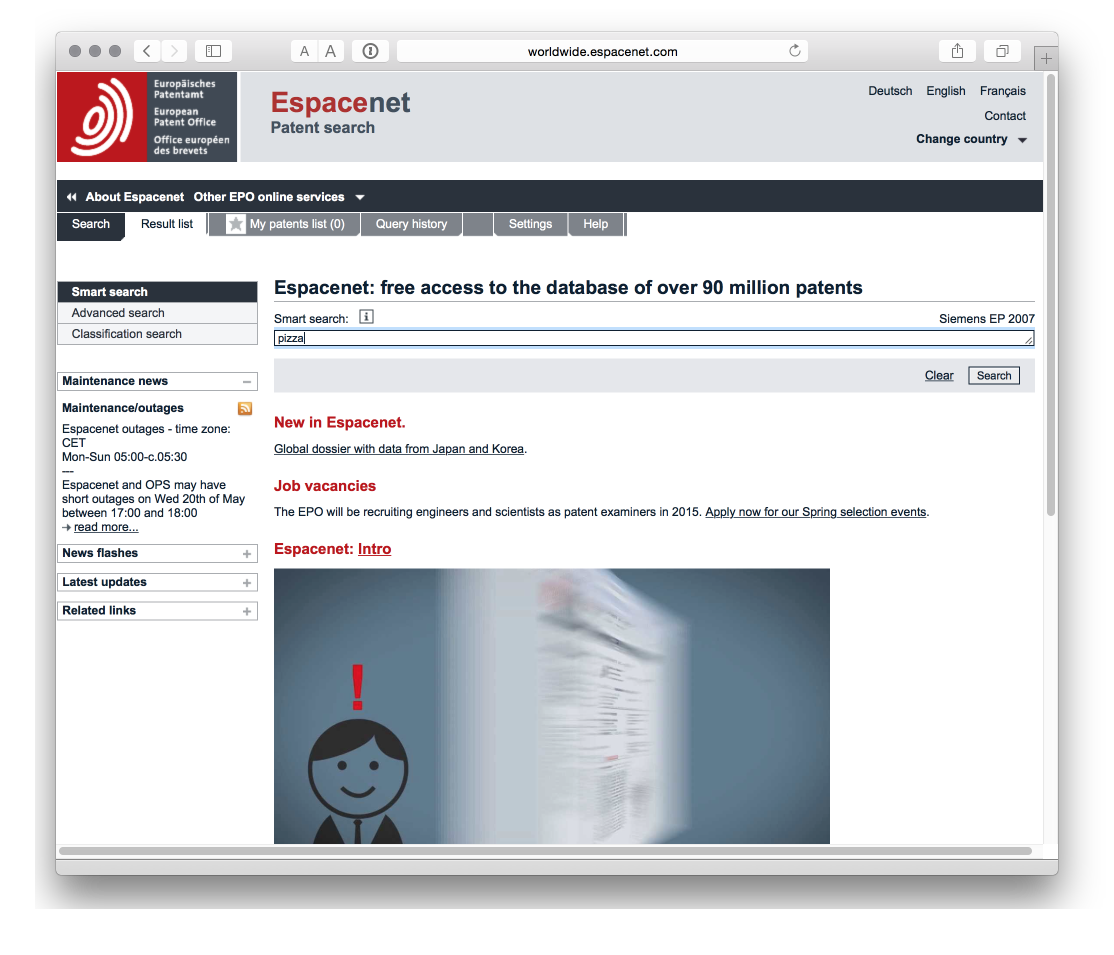
# Obtaining Patent Data: Patent Databases

This chapter provides a quick overview of some of the main sources of free patent data. It is intended for quick reference and points to some free tools for accessing patent databases that you may not be familiar with.

It goes without saying that getting access to patent data in the first place is fundamental to patent analysis. There are quite a few free services out there and we will highlights some of the important ones. Most free sources have particular strengths or weaknesses such as the number of records that can be downloaded, the data fields that can be queried, the format the data comes back in or how clean data is in terms of the hours required to prepare for analysis. We won't go into all of that but will provide the odd pointer.

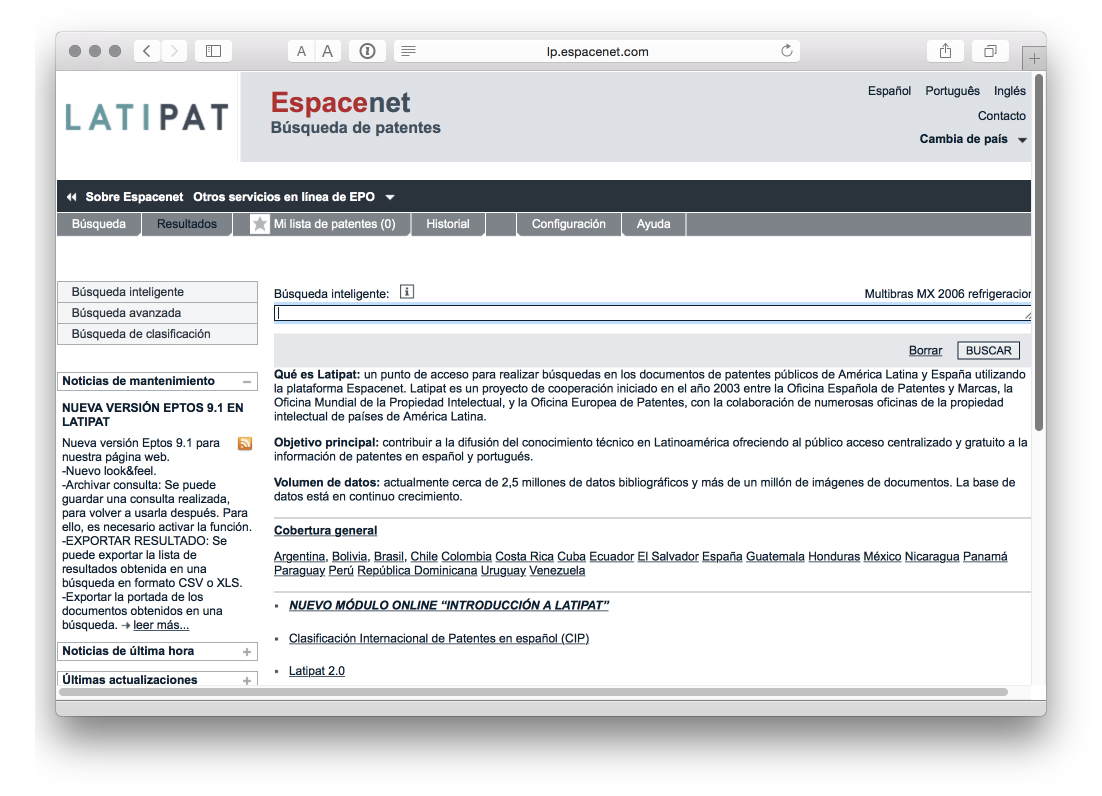
**1. espacenet**[[1]](#footnote-1)

Probably the best known free patent database from the European Patent Office.



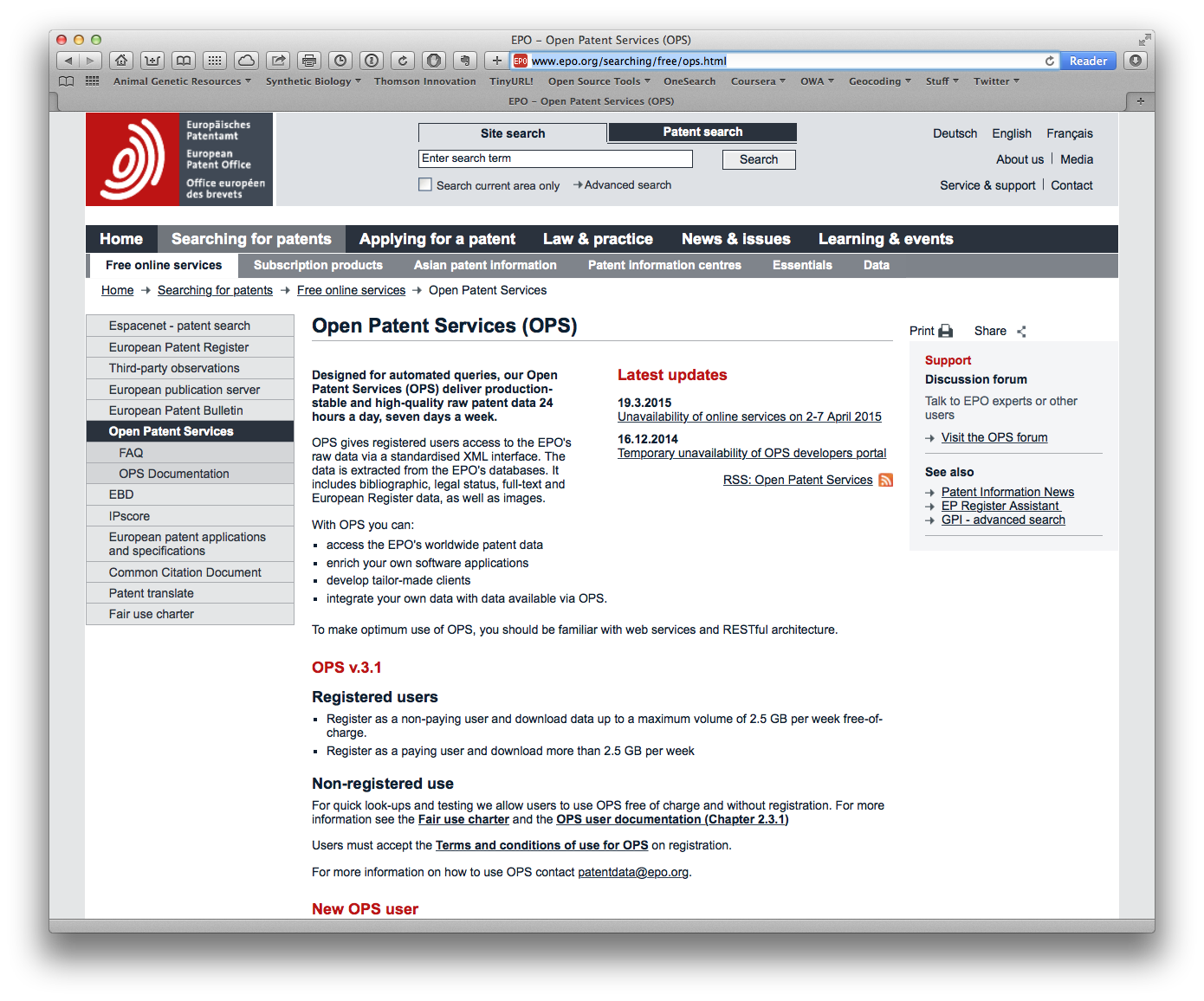
**2. LATIPAT**[[2]](#footnote-2)

For readers in Latin America (or Spain & Portugal) LATIPAT is a very useful resource.

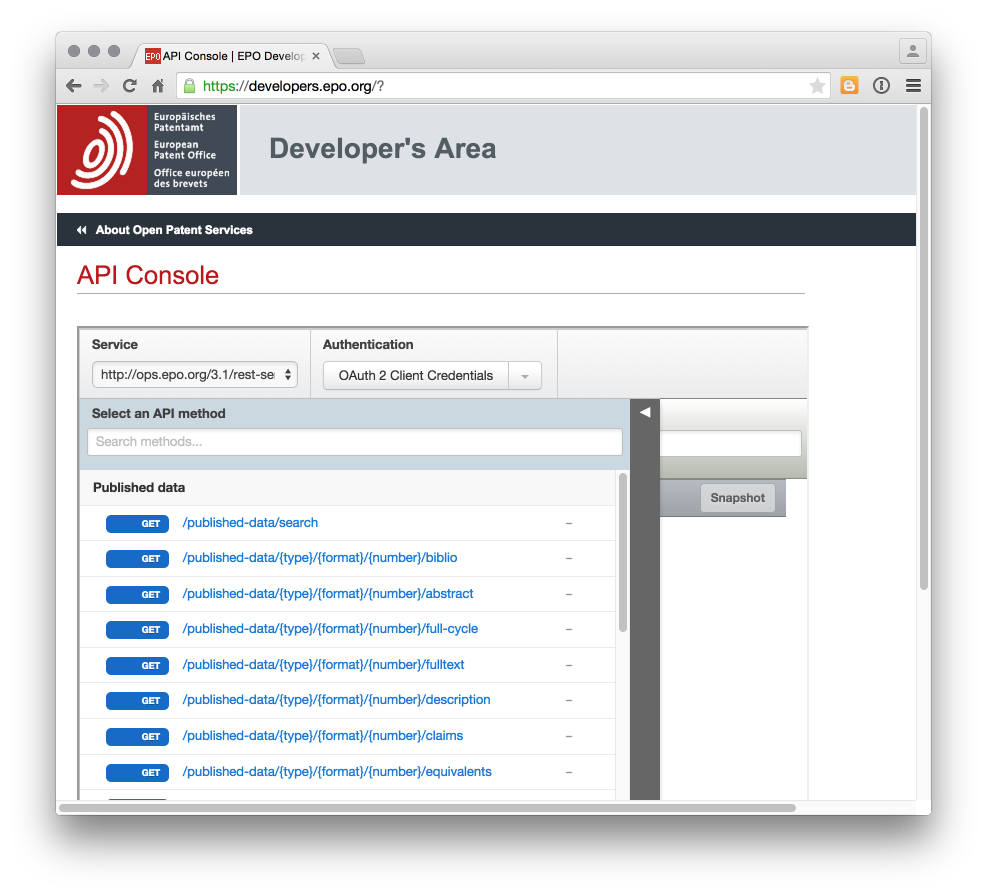


**3. EPO Open Patent Services**[[3]](#footnote-3)

Access patent data through the EPO API free of charge.



The developer portal allows you to test your API queries and is recommended.

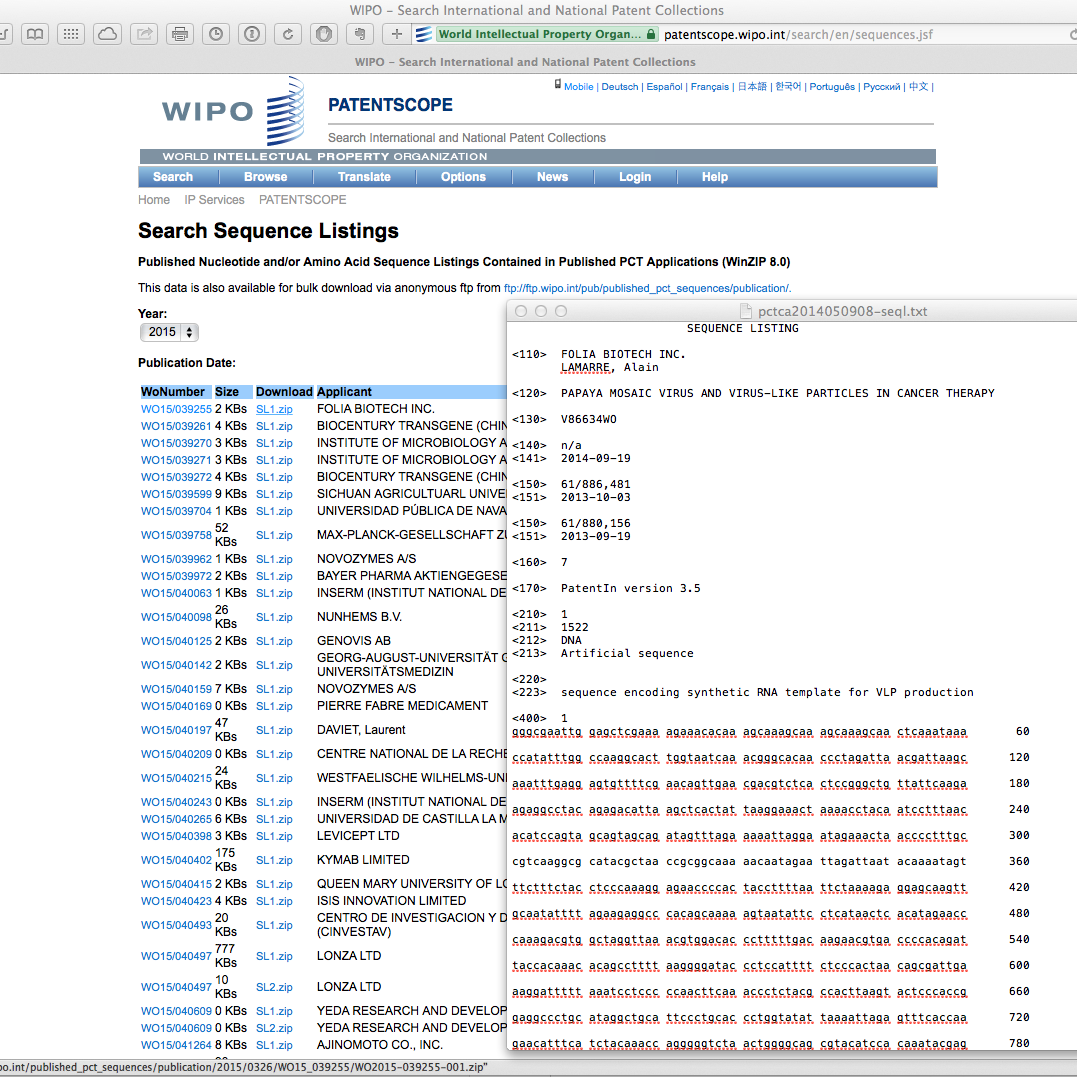


**4. Patentscope**[[4]](#footnote-4)

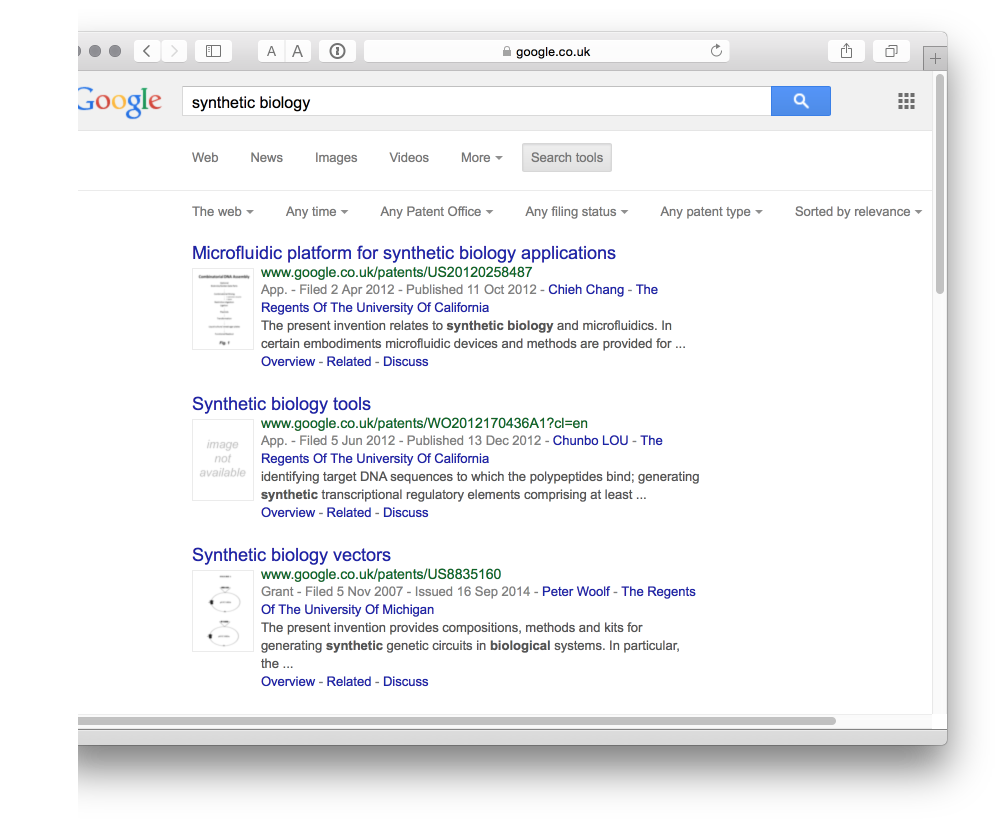
The WIPO Patentscope database provides access to Patent Cooperation Treaty data including downloads of a selection of fields (up to 10,000 records), a very useful search expansion translation tool[[5]](#footnote-5), and translation[[6]](#footnote-6).



Obtaining sequence data from Patentscope[[7]](#footnote-7). Note that this rapidly becomes gigabytes of data.



\*\*5. Google Patents[[8]](#footnote-8)

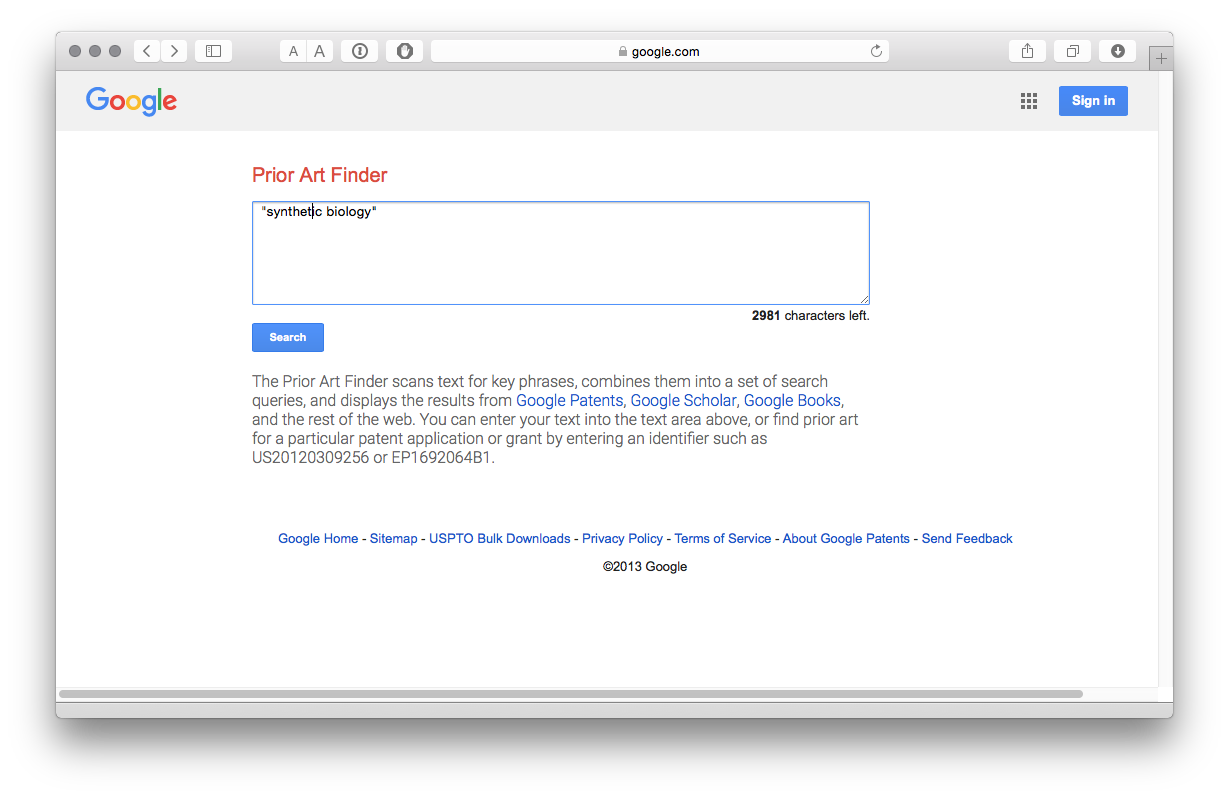


The Google Patent Search API[[9]](#footnote-9) has been deprecated. Access through the Google Custom Search API with the API flag for patents reported[[10]](#footnote-10) to be &tbm=pts with example code for using the API in Python.

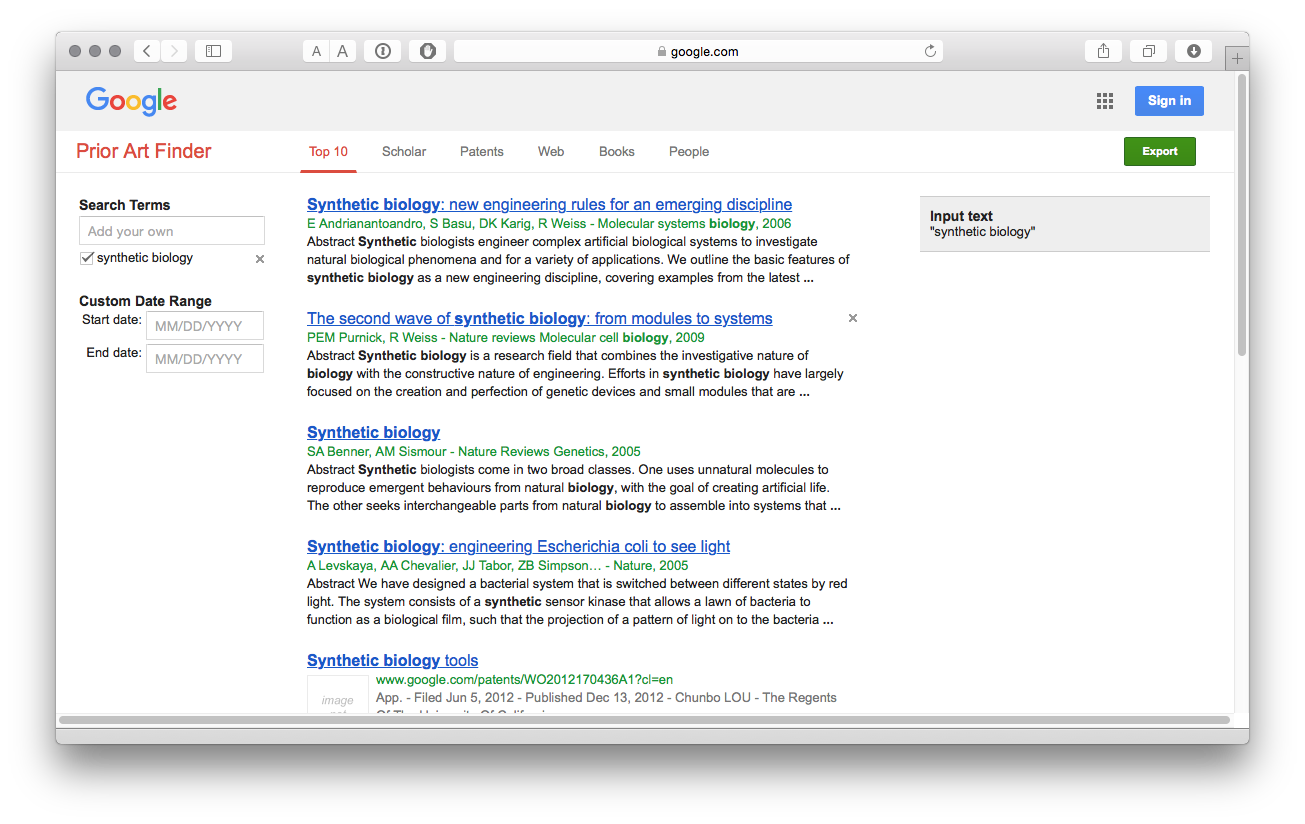
In the free version of the Google Custom Search API data retrieval is limited and the patent field headings are unclear (that is they use non-standard names). For free patent analytics, Google Custom Search is presently of very limited use.

**6. Google Prior Art Finder**[[11]](#footnote-11)

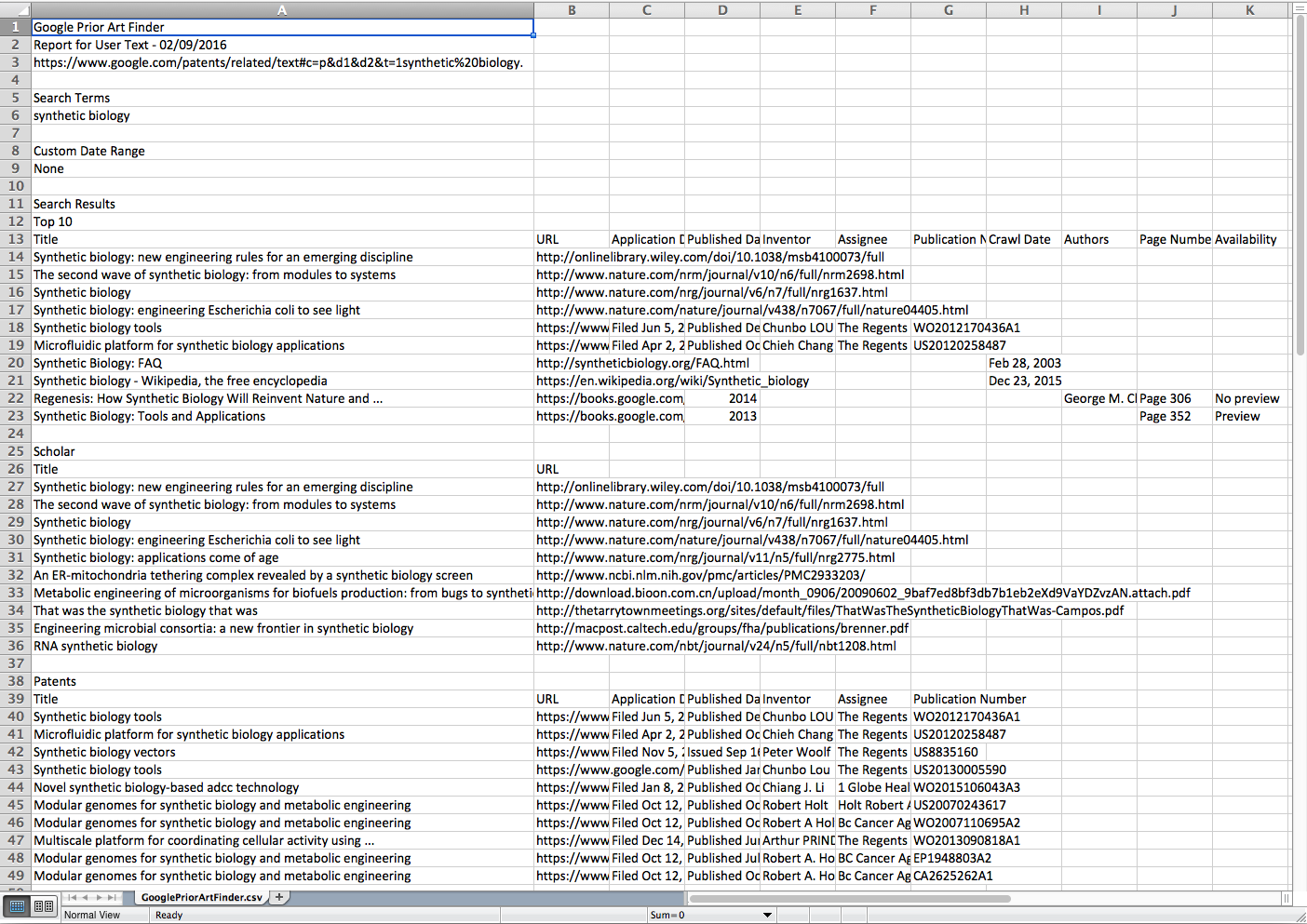
The Google Prior Art Finder is a relatively recent development that allows you to enter search terms or patent numbers and to view and export results.



The results include a Top Ten and are broken down into sections including Google Scholar, Patents etc.



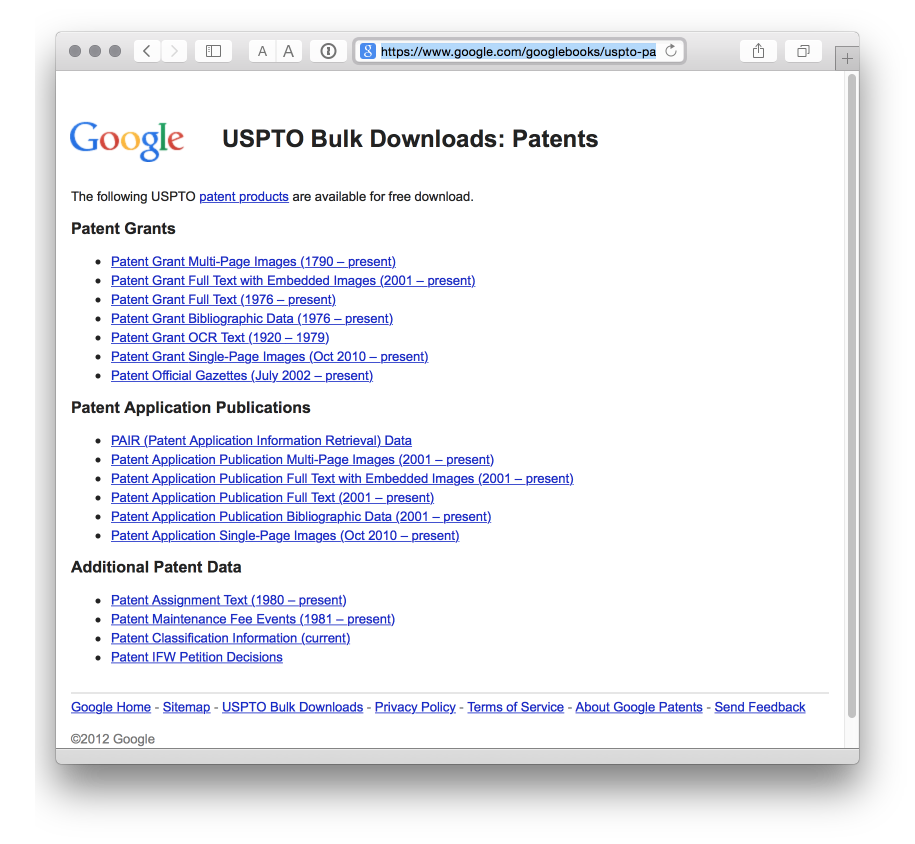
The Export button will export the top ten results for each section in a .csv file.



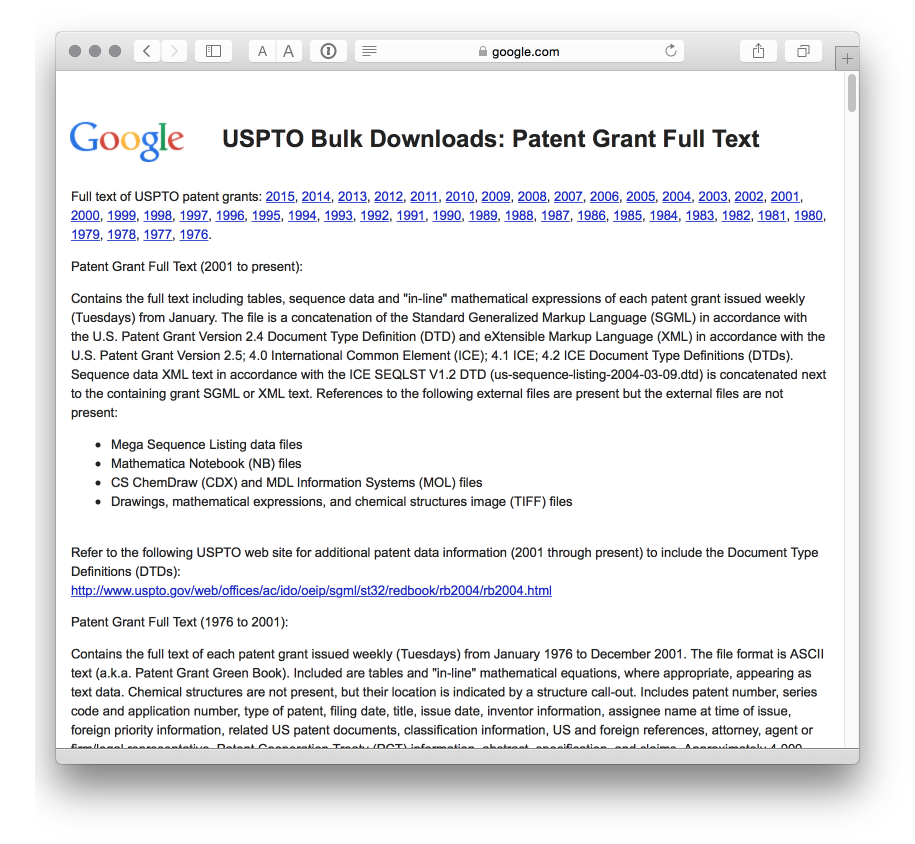
It is possible to load more results for a section (e.g. see More Patent Results at the bottom of the results) and then export them (e.g. 20 patent documents rather than 10). In a test we managed to export 140 patent results but this could rapidly become laborious. An additional issue is that the data will need transposing. At the time of writing we had not identified an API route to Prior Art Finder.

**7. Google USPTO Bulk download**[[12]](#footnote-12)

The USPTO patent databases[[13]](#footnote-13) may be archaic but you can download the entire US collection from the Google USPTO Bulk download service[[14]](#footnote-14).



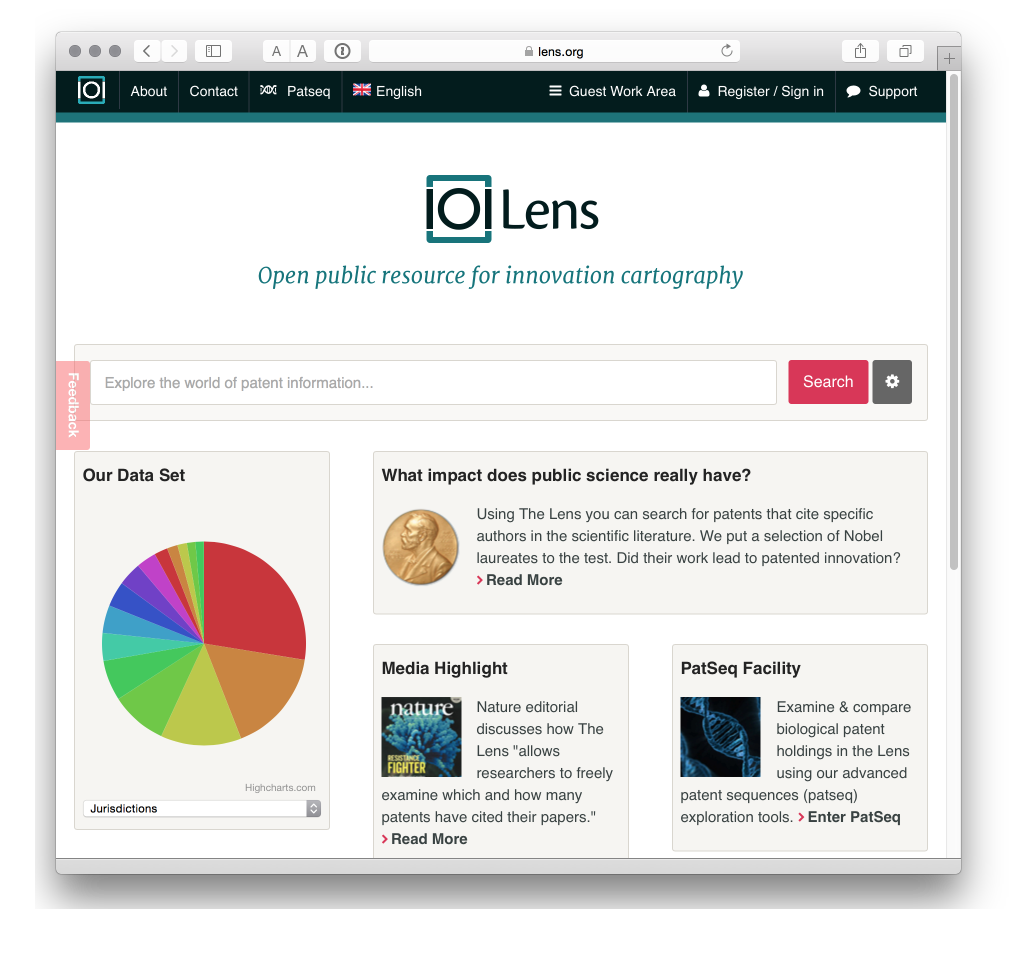
It is a fantastic service, and an example to patent offices everywhere on freeing up patent data. If you have a good broadband connection and the hard drive space, it is quite good fun to suddenly have access to millions of patent records. The authors used the service to text mine the collection for millions of biological species names as reported here[[15]](#footnote-15).



However, one important issue to note is that the XML delimiting individual documents is not always well demarcated. This means that any code that will work for one bulk set of files may fail on another set. While it is possible to address this, be prepared to spend time working on this and/or seek assistance from a professional programmer. For an insight into these issues see this Stackoverflow discussion[[16]](#footnote-16) on parsing the data in R.

**8. The Lens**[[17]](#footnote-17)

Previously known as the Patent Lens this is a well designed site with quite a few visualisation options and access to sequence data. It is possible to share data in collections and in 2015 the ability to download up to 10,000 records was added. When combined with interactive charts that allow the user to drill down into results set, this has transformed the Lens into a useful and innovative database and visualization tool.



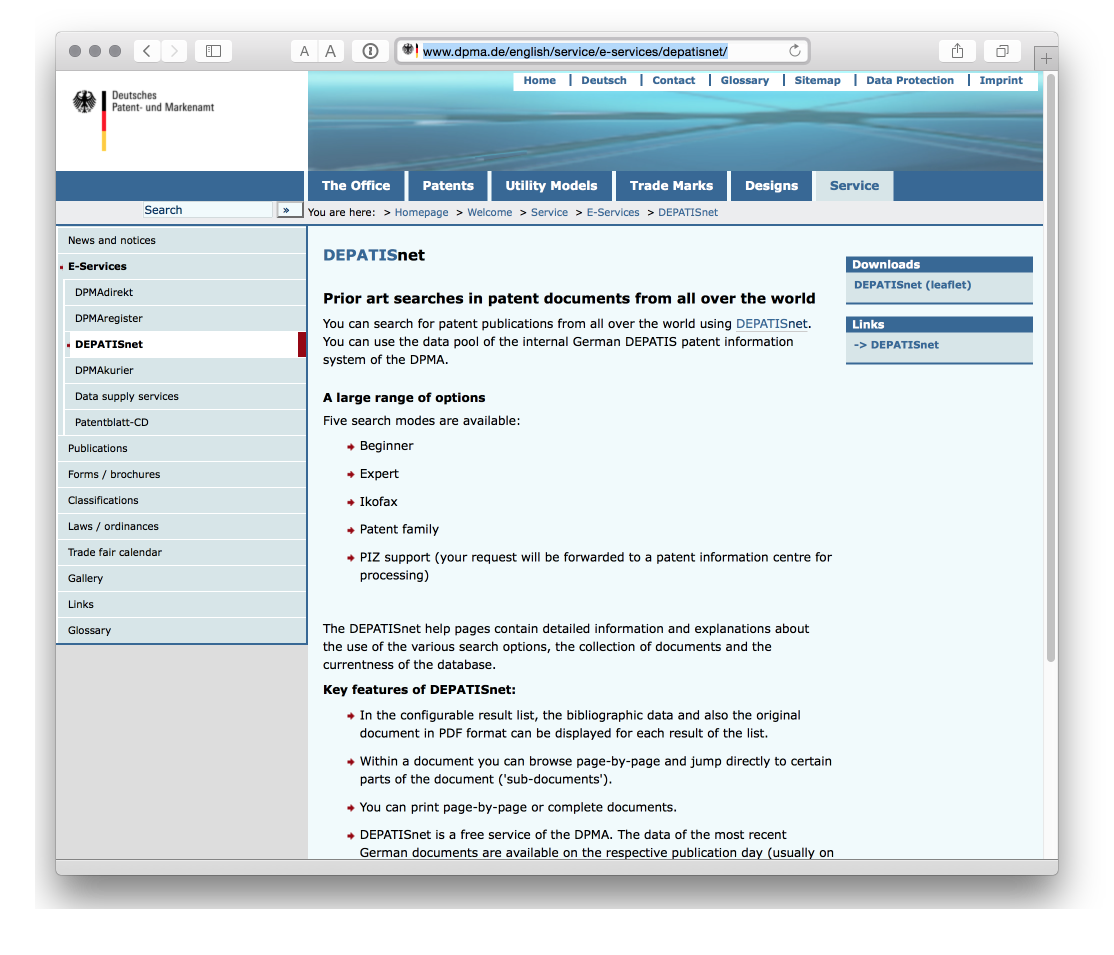
**9. Free Patents Online**[[18]](#footnote-18)

Sign up for a free account for enhanced access and to save and download data. It has been around quite a while now and while the download options are limited we rather like it.



**10. DEPATISnet**[[19]](#footnote-19)

We are not covering national databases. However, the patent database of the German Patent and Trademark Office struck us as potentially very useful. It allows for searches in English and German and has extensive coverage of international patent data, including the China, EP, US and PCT collections. The coverage details are here[[20]](#footnote-20). Worth experimenting with.



**11. OECD Patent Databases**[[21]](#footnote-21)

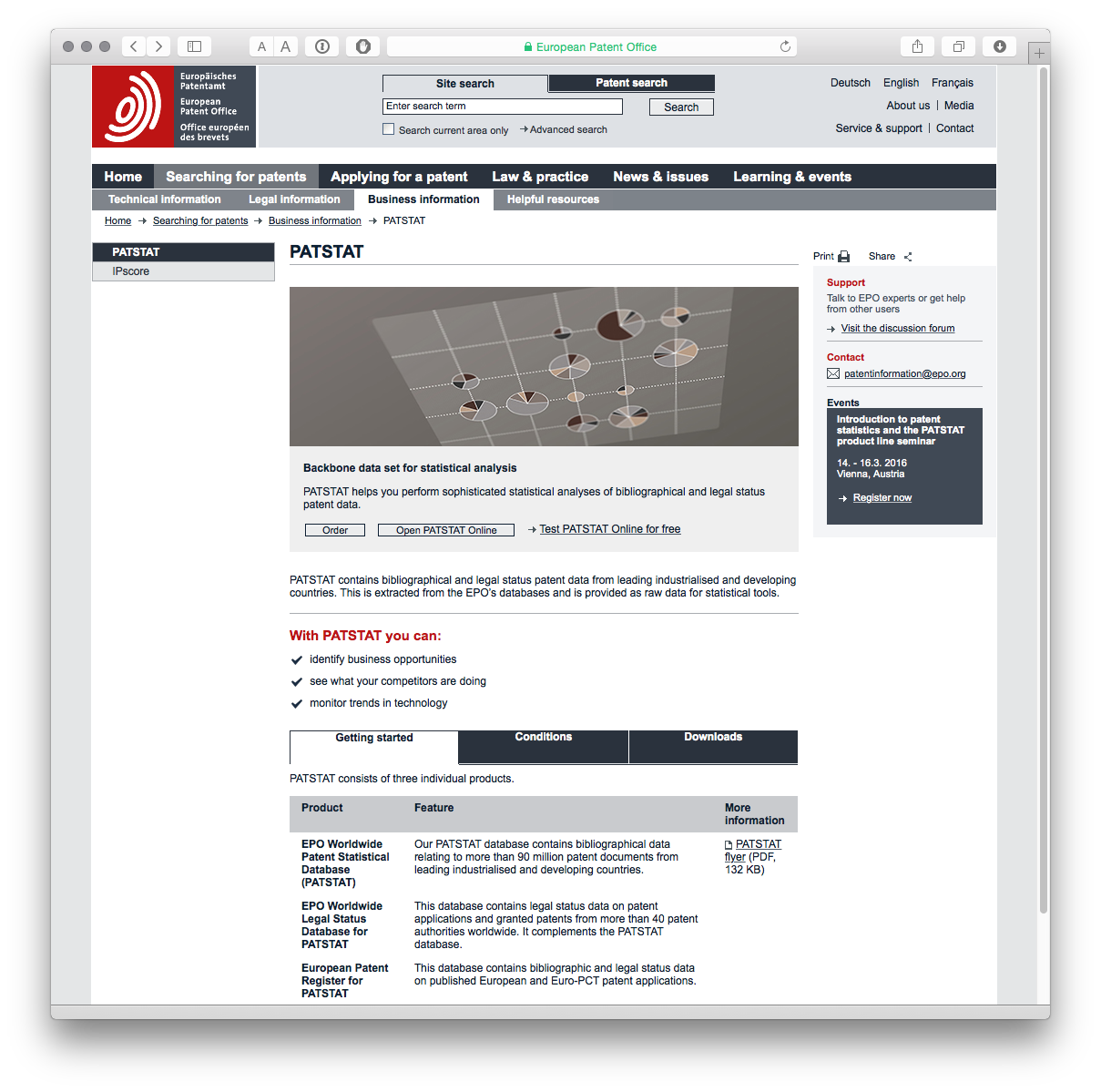
One that is more for patent statisticians. The OECD has invested a lot of effort into developing patent indicators and resources including citations, the Harmonised Applicants names database HAN database[[22]](#footnote-22), mapping through the REGPAT database[[23]](#footnote-23) among other resources that are available free of charge.



Along the same lines the US National Bureau of Economic Research NBER US Patent Citations Data File[[24]](#footnote-24) is an important resource.

\*\*12 EPO World Patent Statistical Database[[25]](#footnote-25)

The most important database for statistical use is the EPO World Patent Statistical Database (PATSTAT) and contains around 90 million records. PATSTAT is not free and costs 1250 Euro for a year (two editions) or 630 Euro for a single edition. The main barrier to using PATSTAT is the need to run and maintain a +200 Gigabyte database. However, there is also an online version of PATSTAT that is free for the first two months if you wish to try it by signing up for the trial (knowledge of SQL required).

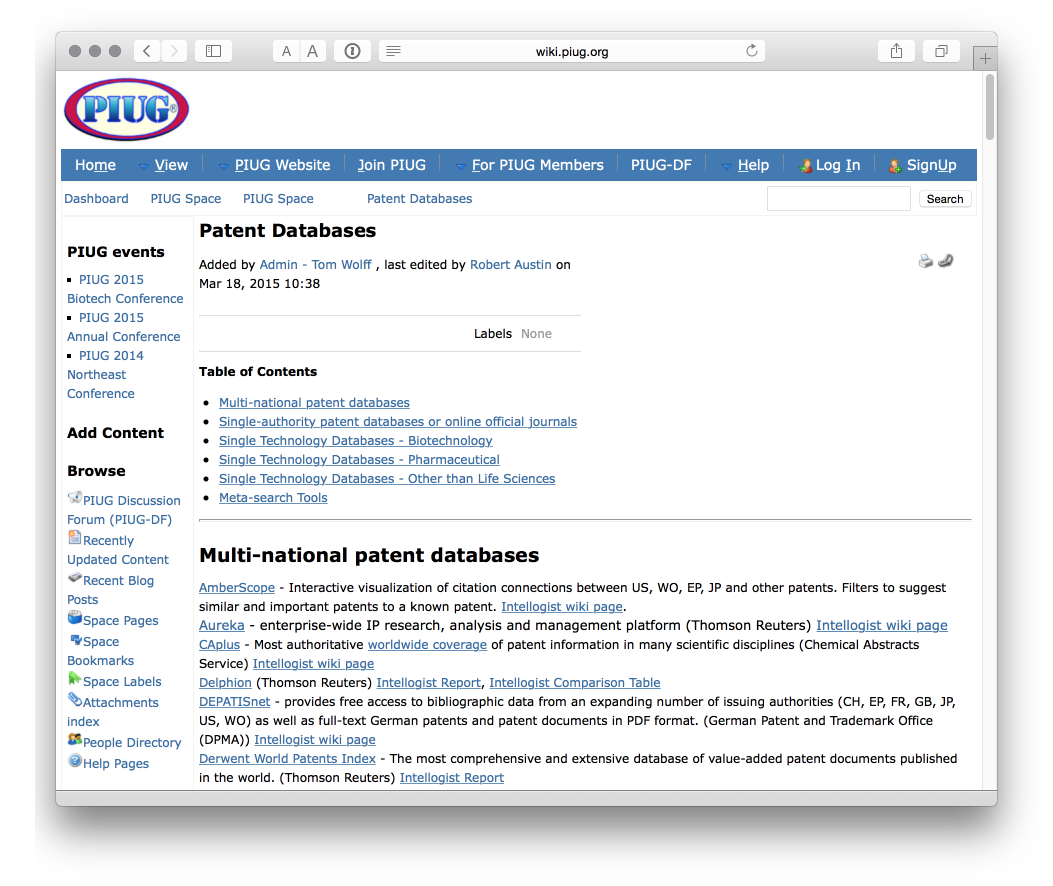


For users seeking to load PATSTAT into a MySQL database Simone Mainardi provides the following code on Github[[26]](#footnote-26).

**12. Other data sources**

A number of companies provide access to patent data, typically with tiered access depending on your needs and budget. Examples include Thomson Innovation[[27]](#footnote-27), Questel Orbit[[28]](#footnote-28), STN[[29]](#footnote-29), and PatBase[[30]](#footnote-30). We will not be focusing on these services but we will look at the use of data tools to work with data from services such as Thomson Innovation.

For more information on free and commercial data providers try the excellent Patent Information User Group[[31]](#footnote-31) and its list of Patent Databases[[32]](#footnote-32) from Tom Wolff and Robert Austin.



Also worth mentioning is the Landon IP Intellogist[[33]](#footnote-33) blog which maintains Search System Reports[[34]](#footnote-34).

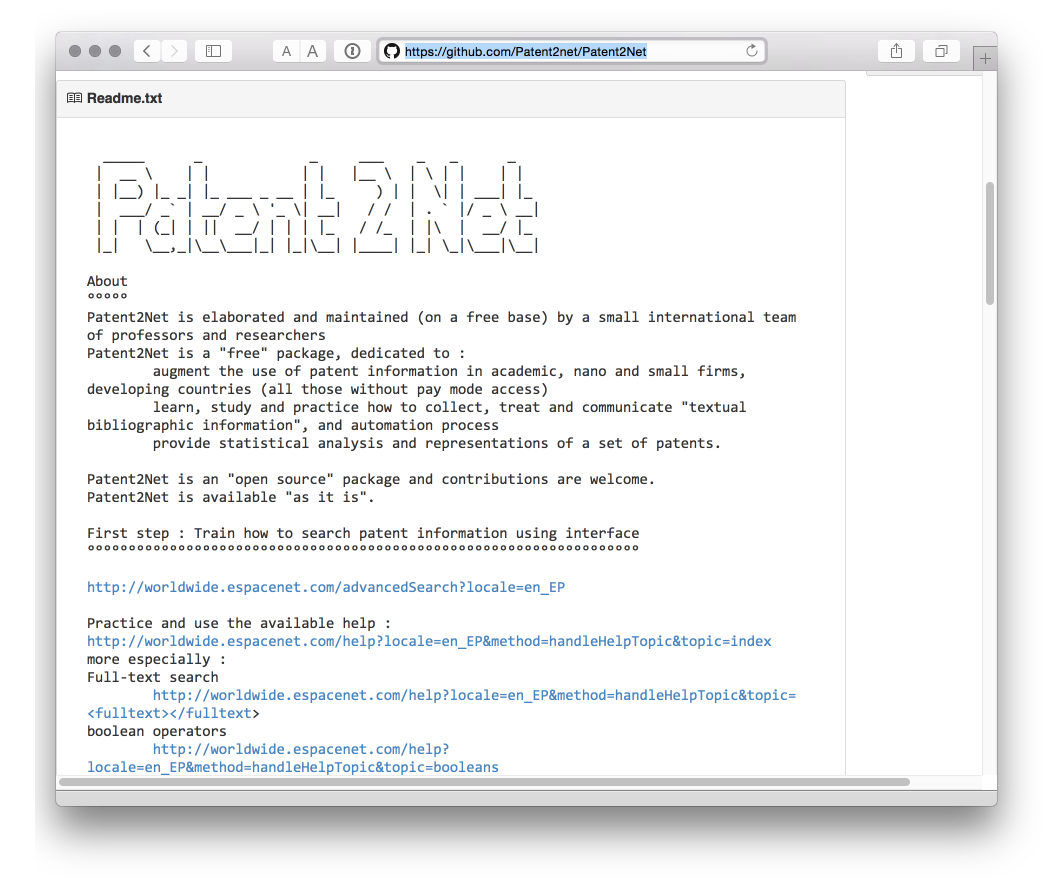


### Tools for Accessing Patent Data

In closing this chapter we will highlight a couple of tools for accessing patent data, typically using APIs and Python. We will come back to this later and are working to try this approach in R.

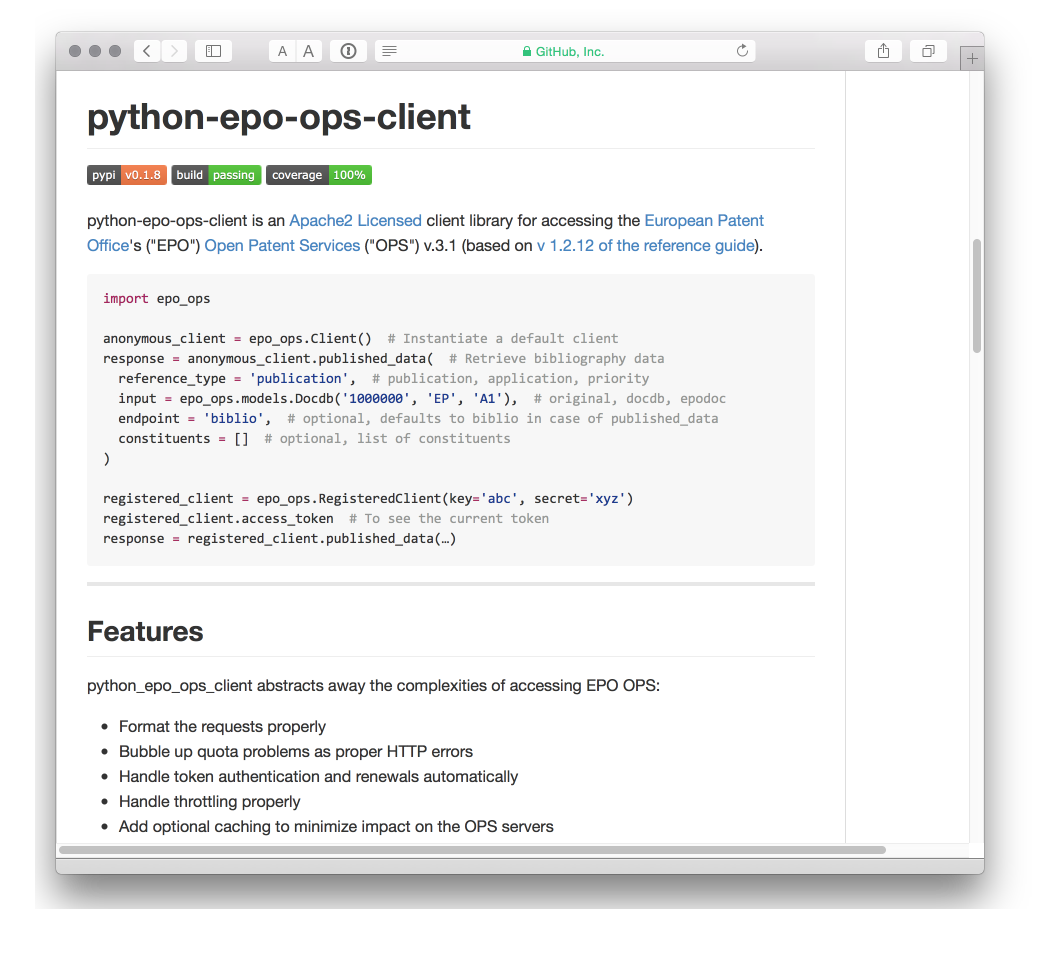
**13. Patent2Net**[[35]](#footnote-35) **in Python**

A tool to access and process the data from the European Patent Office OPS service.



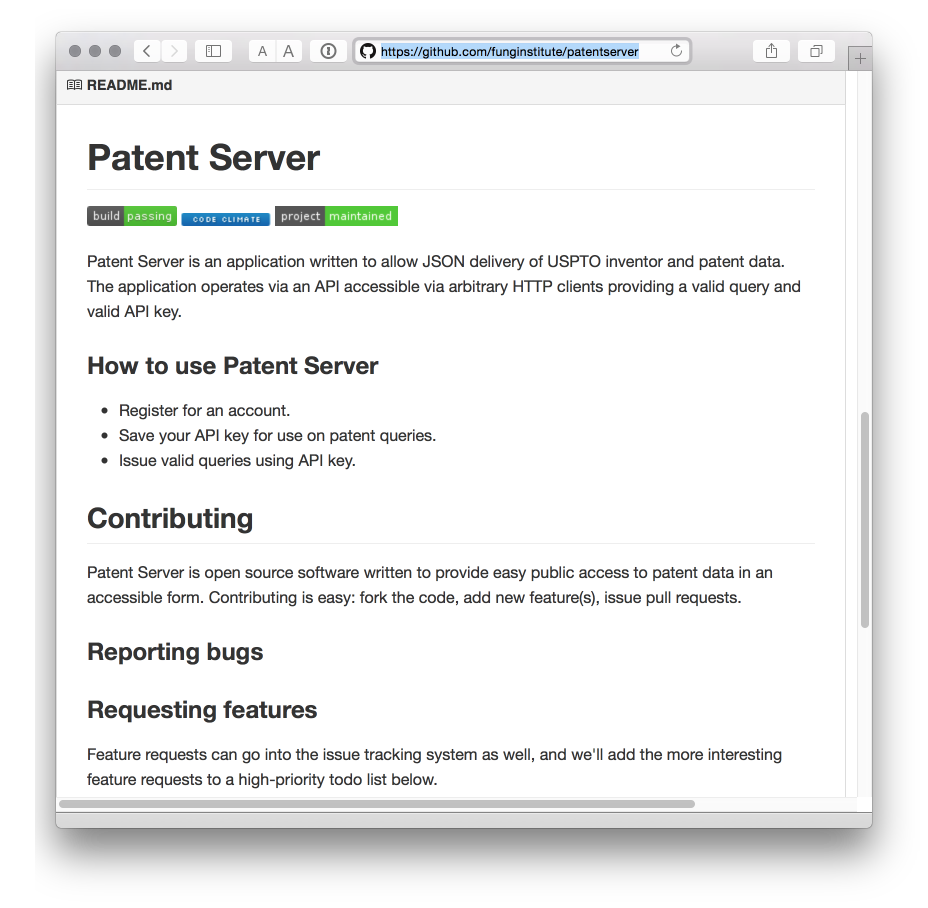
**14. Python EPO OPS Client**[[36]](#footnote-36) **by Gsong**

A Python client for OPS access developed by Gsong and freely available on GitHub. Used in Patent2Net above.



**15. Fung Institute Patent Server**[[37]](#footnote-37) **for USPTO data in JSON**

Researchers at the Fung Institute have also been active in developing open source resources for accessing and working with patent data. We highlight patentserver but it is worth checking out other resources in the repository such as patentprocessor[[38]](#footnote-38), a set of Python scripts for processing USPTO bulk download data. Note that development of these tools no longer appears to be active.



## Round Up

One problem confronting patent analysts is access to data in a form that is suitable for more detailed analysis. Typically this involves hundreds or many thousands of records. Recent years have increasingly opened up patent data through the ability to download 1,000 or 10,000 records at a time. However, access to downloads of titles, abstracts and claims or descriptions and full text remains limited when this is what is needed. Patent offices such as the USPTO have taken a leading role in making bulk patent data available and this is very much to be welcomed for those working on patent analytics. However, it is reasonable to say that the present situation is one of improvements in access (through Patentscope, the Lens and the EPO OPS service) but not quite in the quantitities or with the data fields patent analysts would like.

1. <http://worldwide.espacenet.com/?locale=en_EP> [↑](#footnote-ref-1)
2. <http://lp.espacenet.com> [↑](#footnote-ref-2)
3. <http://www.epo.org/searching/free/ops.html> [↑](#footnote-ref-3)
4. <https://patentscope.wipo.int/search/en/search.jsf> [↑](#footnote-ref-4)
5. <https://patentscope.wipo.int/search/en/clir/clir.jsf?new=true> [↑](#footnote-ref-5)
6. <https://www3.wipo.int/patentscope/translate/translate.jsf?interfaceLanguage=en> [↑](#footnote-ref-6)
7. <https://patentscope.wipo.int/search/en/sequences.jsf> [↑](#footnote-ref-7)
8. <http://www.google.com/patents> [↑](#footnote-ref-8)
9. <https://developers.google.com/patent-search/terms> [↑](#footnote-ref-9)
10. <http://stackoverflow.com/questions/15028166/python-module-for-searching-patent-databases-ie-uspto-or-epo> [↑](#footnote-ref-10)
11. <https://www.google.com/patents/related> [↑](#footnote-ref-11)
12. <https://www.google.com/googlebooks/uspto.html> [↑](#footnote-ref-12)
13. <http://patft.uspto.gov> [↑](#footnote-ref-13)
14. <https://www.google.com/googlebooks/uspto-patents.html> [↑](#footnote-ref-14)
15. <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0078737> [↑](#footnote-ref-15)
16. <http://stackoverflow.com/questions/25107557/parseing-xml-by-r-always-return-xml-declaration-error> [↑](#footnote-ref-16)
17. <https://www.lens.org/lens/> [↑](#footnote-ref-17)
18. <http://www.freepatentsonline.com> [↑](#footnote-ref-18)
19. <http://www.dpma.de/english/service/e-services/depatisnet/> [↑](#footnote-ref-19)
20. <https://depatisnet.dpma.de/DepatisNet/depatisnet?action=datenbestand> [↑](#footnote-ref-20)
21. <http://www.oecd.org/sti/inno/oecdpatentdatabases.htm> [↑](#footnote-ref-21)
22. <http://www.oecd.org/sti/inno/43846611.pdf> [↑](#footnote-ref-22)
23. <http://www.oecd.org/sti/inno/40794372.pdf> [↑](#footnote-ref-23)
24. <http://www.nber.org/patents/> [↑](#footnote-ref-24)
25. <https://www.epo.org/searching-for-patents/business/patstat.html> [↑](#footnote-ref-25)
26. <https://github.com/simonemainardi/load_patstat> [↑](#footnote-ref-26)
27. <https://www.thomsoninnovation.com/login> [↑](#footnote-ref-27)
28. <https://www.questel.com/index.php/en/> [↑](#footnote-ref-28)
29. <http://www.stn-international.de/index.php?id=123> [↑](#footnote-ref-29)
30. <https://www.patbase.com/login.asp> [↑](#footnote-ref-30)
31. <http://www.piug.org> [↑](#footnote-ref-31)
32. <http://wiki.piug.org/display/PIUG/Patent+Databases> [↑](#footnote-ref-32)
33. <http://www.intellogist.com/wiki/Main_Page> [↑](#footnote-ref-33)
34. <http://www.intellogist.com/wiki/Category:Intellogist_Reports> [↑](#footnote-ref-34)
35. <https://github.com/Patent2net/Patent2Net> [↑](#footnote-ref-35)
36. <https://github.com/55minutes/python-epo-ops-client> [↑](#footnote-ref-36)
37. <https://github.com/funginstitute/patentserver> [↑](#footnote-ref-37)
38. <https://github.com/funginstitute> [↑](#footnote-ref-38)