# 40 million open chemical structures from patents: treasure trove? junk yard? or both?

**C Southan:** Medicines Discovery Catapult. Alderley Park, Macclesfield, Cheshire, SK10 4ZF, UK.

md.catapult.org.uk

## Introduction

Compared to the literature, the patent document corpus has both pros ("treasure trove") and cons ("junk yard ") with just a selection below;

- A "Cinderella" difficult to get to grips with data source (especially for academics)
- Paradoxically, completely open for text mining and entity extraction
- Contains ~ 3x to 5x more medicinal chemistry SAR than published papers
- Massively redundant document corpus from patent families and Kind codes
- Discloses pursued drug targets and chemotype landscape years ahead of papers
- Challenging to extract chemistry < > data relationships via curation or automatically
- Examples of deliberate obfuscation as well as virtual compounds
- Small "gold nuggets" of experimental data entombed in 100+ page PDF "junk yards"
- Global resource of execute synthesis protocols and analysis data
- Monopoly of commercial curation is now broken by open chemistry extraction

While grappling with these pros and cons, MDC engages intensively with patents, a) in various informatics data extraction and integration projects b) in applications developed from these for our customers and collaborators c) for Competitive Intelligence (CI) landscaping in support of research projects with a bioactive chemistry component. We thus endeavour to keep up with developments in both commercial and open sources. Consequently, this work was undertaken to give an update of open extractions in general and the expanding integration of these within PubChem in particular.

## Top PubChem patent chemistry submitters

**These are listed below by substance (SID) counts and the last submission date**

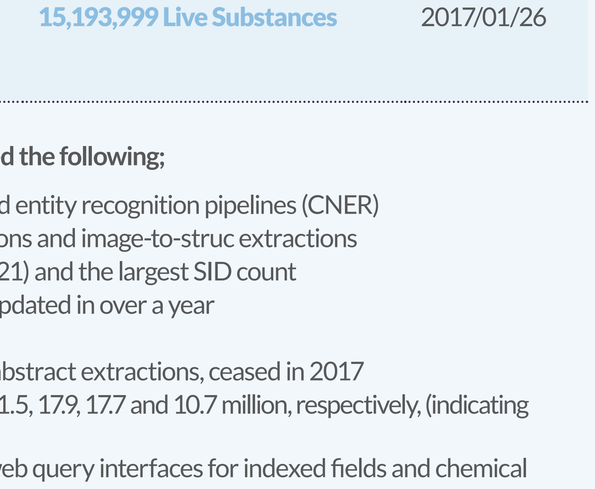
| | | |
|---|---|---|
| **SureChEMBL**<br>Research and Development Curation Efforts<br>*United Kingdom* | **21,641,384 Live Substances** | 2021/08/21 |
| **Google Patents**<br>Research and Development<br>*United States* | **18,964,777 Live Substances** | 2020/08/20 |
| **PATENTSCOPE (WIPO)**<br>Governmental Organizations<br>*Switzerland* | **17,448,098 Live Substances** | 2021/02/22 |
| **IBM**<br>Research and Development<br>*United States* | **15,193,999 Live Substances** | 2017/01/26 |

**Analysis of these four major sources established the following;**

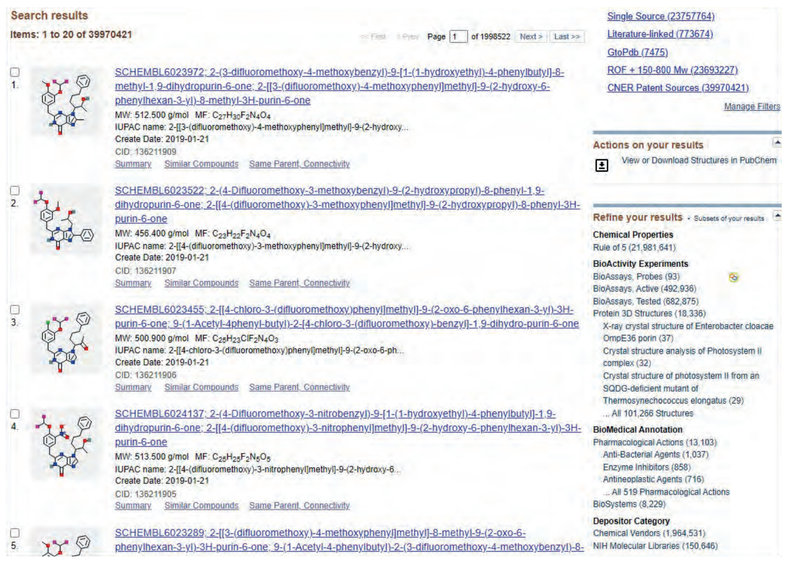
- They all use similar, automated chemical named entity recognition pipelines (CNER)
- These include name look-ups, IUPAC conversions and image-to-struc extractions
- SureChEMBL is both the most recent (Aug 2021) and the largest SID count
- Google patents is ranked second but has not updated in over a year
- WIPO, ranked 3rd updated in Jan 2021
- The IBM submissions, that included PubMed abstract extractions, ceased in 2017
- For these sources compound (CID) counts are 21.5, 17.9, 17.7 and 10.7 million, respectively, (indicating some duplication in Google and IBM SIDs)
- The top three have active, open, stand alone web query interfaces for indexed fields and chemical structure searches
- SureChEMBL and WIPO extract, update and index their chemistry for searching *in situ* within a week or so of publication (unfortunately with long lag times in PubChem)
- The chemistry indexing speed for Google patents is more like a month
- Google patents also search-indexes chemistry from Google Scholar papers
- Google patents indexes and counts gene names in documents
- SureChEMBL patent documents can be toggled for the SciBite Termite engine that generates entity mark-up, including gene names (but not search-indexed)

## Total CNER patent extractions

The compound CID query includes the four major sources, a legacy source (SCRIPDB) of 3.9 mill and a 1.8 mill chemical synthesis set from NextMove Software. These add up to just under 40 mill from the (Oct 2021) total of 111 mill. Considering this started with the pioneering first IBM submission of 2.5 mill in 2012, progress has been remarkable. In addition, the PubChem team are congratulated on their efforts not only in wrangling and integrating these sources but also linking and search-indexing the chemistry linked to the patent documents they were extracted from (see PMID: 33151290 and try the new search interface).




## Statistics of the 40 million




The right-hand facets above indicate both the default PubChem CID statistics plus a set of five custom filters (at the top). There are many details that can be explored but salient points can be discerned (in descending order) as follows.

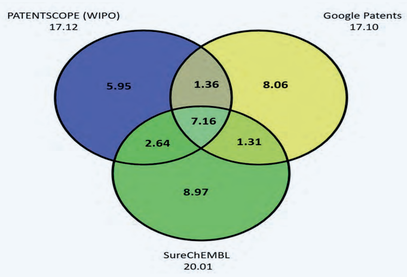
- 2.3 million are unique to individual patent sources (but the overall patent-unique proportion is much higher since many will be only in one or two patent sources)
- Literature-linked is the PubMed, MeSH plus ChEMBL intersect of 0.6 mill
- Of the Guide to Pharmacology 8,974 curated ligand CIDs 83% have a patent extraction match (includes first-filings for most leads and approved drugs)
- The lead-likeness property filter covers 60%
- Only 12% have positive results in PubChem BioAssay (that includes ChEMBL assays)
- Not unexpectedly, the vendor intersect is below 2 mill

While not widely known, a public "treasure trove" of SAR from patents has been manually curated since 2013 by BindingDB. This comprises 5,082 USPTO documents, 761,001 binding values from 7,371 assays, 379,307 CIDs and 2,130 target proteins.

## The "junk yard" problem

This accrues mainly from the following extraction associated caveats:

- Automated extraction quality is lower than expert curation
- CNER produces erroneous structures from IUPAC splitting via poor document OCR
- The "treasure trove" of exemplified structures with SAR constitutes only ~ 5 million
- This questions the IP and scientific value of the 35 million (junk yard?)
- CNER leads to extensive "over-mapping" of common chemistry to 1000s of documents (e.g. aspirin CID2244 is linked to 410,666 patent numbers in PubChem)
- As indicated by the Venn major sources are discordant in the chemistry they extract from nominally the same document corpus




## Conclusions

This survey provides an update of contemporary open patent chemistry for internal MDC exploitation including the extensive PubChem linking that commercial patent data sources do not offer. It also helps us to advise customers who may not have commercial subscriptions. The extracted chemistry in PubChem is, of course, both "treasure trove" and "junk yard" which presents the challenge of discriminating between the two.

**References**

*Opening up connectivity between documents, structures and bioactivity.* PMID: 32280387, 2020, OA

*Expanding opportunities for mining bioactive chemistry from patents.* PMID: 26194581, 2015, OA

*Examples of SAR-centric patent mining using open resources* (book chapter 2017) https://www.research.ed.ac.uk/en/publications/examples-of-sar-centric-patent-mining-using-open-resources (OA link)

*Tracking 20 years of compound-to-target output from literature and patents.* PMID: 24204758, 2013, OA

Innovate UK