PDG Partnering in Patent Information

PDG Chemistry Task Force's PubChem Project

Mark Ede; James Rudman

Tuesday November 5th 2024

PubChem

PubChem - How and Why did the Project Start?

- PubChem - Hosted by the NCBI (part of the US Government)
- PubChem too big to ignore – over 317 million substances indexed from 991 data sources and includes over 50 million patents.
  - Includes **118 million** “unique” substances
- PubChem interfaces with Reaxys.com and Minesoft CHEMX so it can be accessible to Patent Examiners
- TF members were aware of quality issues
  - More chemical compounds indexed than CAS and Reaxys – why?
  - Poor depositor supplied synonyms e.g. Patent numbers for synonyms
  - “Old” patent numbers indexed for “new” active pharmaceutical ingredients – nonsensical!
  - Unknown quality control process for producing PubChem

Reaxys

minesoft CHEMX

PubChem

PubChem Errors

- Depositor name errors

PubChem Glycerol (Compound)

3.4.2 Depositor-Supplied Synonyms

	NCGC00090950-04	Glycerin, meets USP testing specifications	32-EP2270004A1
	NCGC00090950-05	32-EP2269610A2	32-EP2270007A1
	NCGC00253975-01	32-EP2269978A2	32-EP2270008A1
	NCGC00259626-01	32-EP2269985A2	32-EP2270011A1
riol	BP-31039	32-EP2269991A2	32-EP2270895A2
	E422	32-EP2269994A1	32-EP2272510A1
6	Glycerol, for molecular biology, >=99%	32-EP2269996A1	32-EP2272516A2
	Glycerol, JIS special grade, >=99.0%	32-EP2270001A1	32-EP2272537A2
	Glycerol, Vetec(TM) reagent grade, 99%	32-EP2270002A1	32-EP2272822A1

2.4.2 Depositor-Supplied Synonyms

: Alcohol, ACS reagent	DTXSID9020584	Ethanol, technical grade, 93.8%	Ethanol 2
determination--alcohol	Ethanol, technical grade, 93%	Ethanol, technical grade, 99.5%	Ethanol, 9
ting Fluid 100C.NPA	Ethanol, technical grade, 99%	Ethanol, >=99.5%, for HPLC	Ethanol, p
p.a., 99.8%	UNII-7528N5H79B	NSC85228	Ethanol, t
: Alcohol, reagent grade	CHEBI:17246	STR05604	Ethanol, U
). 37 (Salt/Mix)	Ethanol, 95.1-96.9%	Ethyl Alcohol 95% ACS/USP Grade	Ethanol, U
99	poly(vinyl alcohol) macromolecule	Tox21_202510	AKOS009
dehydrated, >=85.0%	Ru-Tuss Expectorant (Salt/Mix)	6869AF	Ethanol, R
4028331	Ethyl alcohol (6CI,7CI,8CI)	STL264245	7528N5H

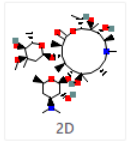
Record for Ethanol

PubChem

PDG Partnering in Patent Information

PubChem Errors

Azithromycin

PubChem CID	447043
Structure	 2D Find Similar Structures
Molecular Formula	$C_{38}H_{72}N_2O_{12}$
Synonyms	azithromycin Zithromax 83905-01-5 Sumamed Hemomycin More...
Molecular Weight	749.0
Dates	Modify 2023-05-05 Create 2005-06-24

Azithromycin is an antibacterial prescription medicine approved by the U.S. Food and Drug Adm

Various bacterial respiratory diseases, including community-acquired pneumonia, acute sinus an

Pelvic inflammatory disease

15.1 Depositor-Supplied Patent Identifiers

Page 17,351 of 86,755 items View More Rows & Details Download

SORT BY Priority Date

Publication Number	Title	Priority Date	Grant Date
US-1306217-A	Gas-meter		1919-06-10
US-1317160-A	of basel		1919-09-30
US-1328570-A	Asbiowob		1920-01-20
US-1416273-A	Power transmission and control		1922-05-16
US-1544924-A	Hafen-on-the-rhtne		1925-07-07

Previous 1 ... 17,349 17,350 17,351

PubChem

Azithromycin was invented in 1980 but has patent references from 1919 onwards!

PubChem

TASK FORCE Chemistry

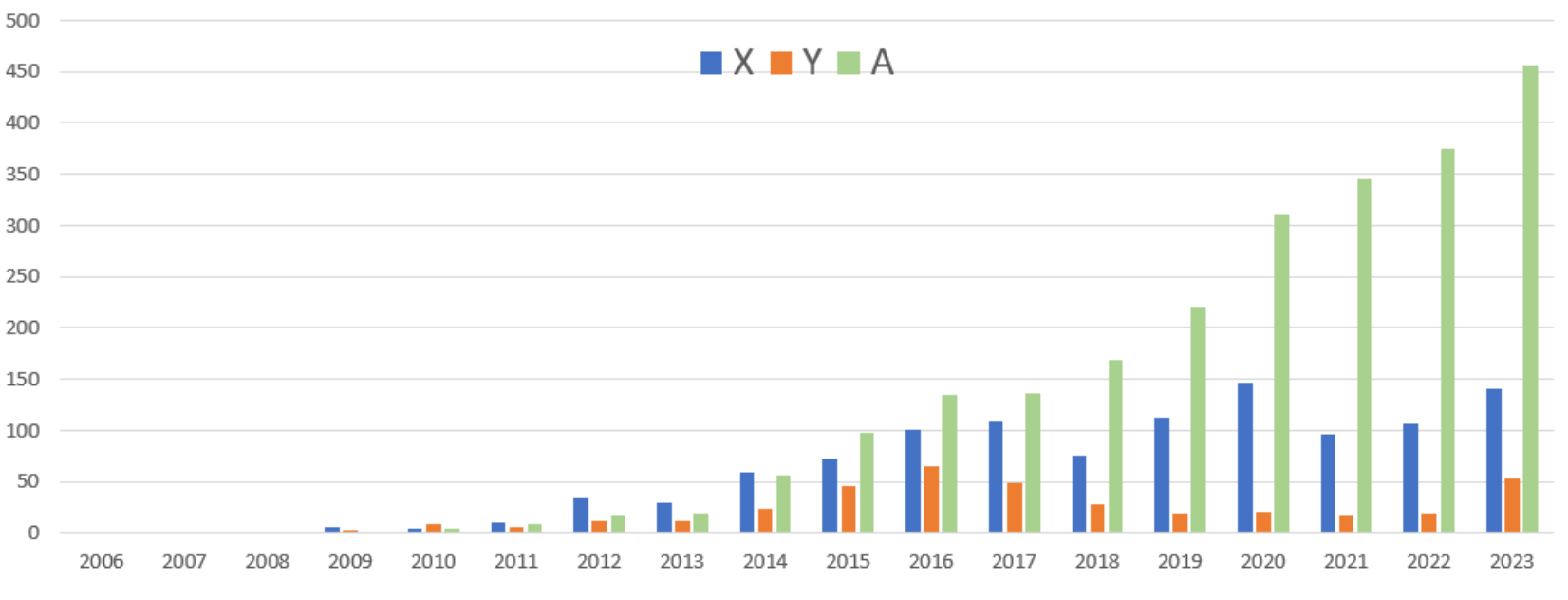
PDG Partnering in Patent Information

PubChem in Examiner Citations and EP Oppositions

ISR Examiner citations: 3519 total

Document categories: 1102 [X], 379 [Y], and 2354 [A]

Occurrence of PubChem citations by earliest publication year and examiner code (sourced from Orbit)



US examiner citations: 26 total

Citation categories: 5 [Sec. 102] and 21 [Sec 103]

EP Opposition citations: 28 Third-party oppositions

PubChem

PubChem cited in Legal Proceedings

- **PubChem has been cited by courts** as a chemical information source
- US Federal courts have cited PubChem in **30** decisions
  - See Google Scholar Case Law
  - PubChem was cited for compound names or identification, physiological and safety data, established “teachings”, etc.
- US PTAB cited PubChem in **27** decisions, proceedings, or documents
- Non-US courts have also cited PubChem

TASK FORCE
Chemistry

PubChem

PDG Task Force – PubChem Project Launched in September 2023

- Objective was to seek improvements to PubChem
  - Plan was to survey PDG Chemistry TF on PubChem usage; analyse PubChem records ; detail and collate observed errors; reach conclusions and seek a meeting with the NCBI
- September 2023
  - A new publication!
  - Validity of PubChem compounds supplied by Patentscope or SureChEMBL by Dr Joerg Ohms
    - A game changer for the direction of the project!
    - Focused on automated structure errors sourced from WIPO Patentscope and SureChEMBL
    - 60% error rates in automated chemical indexing
  - Change of plan to partner with Dr Ohms and bring attention to his work
    - (Chemical compounds are important to us!)

TASK FORCE Chemistry

PubChem

Validity of PubChem compounds supplied by Patentscope or SureChEMBL

- Dr Ohms presented to:
  - PDG – January 2023
  - PIUG - June 2023
  - NCBI – September 2023
  - CEPIUG – September 2023
  - Google/OntoChem – November 2023
- Dr Ohms added Google Patents to his analysis (not previously published)
- Publicised the PubChem quality issues amongst Patent Information Professionals, the PubChem hosts and to 3rd parties that send data for Pubchem Indexing!

TASK FORCE
Chemistry

PDG Partnering in Patent Information

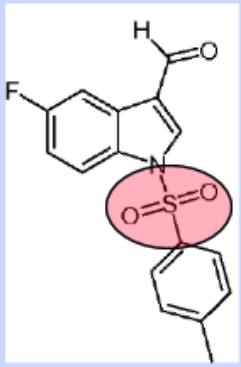
Dr Ohms' Most Important slide - **Citation of CID 88951786 in International Search Report (ISR)**

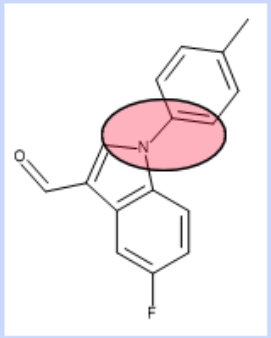
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X – A	"Pubchem CID 88951786" Create Date: 13 February 2015 (13.02.2015) Date Accessed: 08 November 2021 (08.11.2021); pg. 2	1 – 2-3, 19

ISR to WO 2021253013

For this compound, PubChem references WO 2012101487 as the underlying patent publication.

Source	PubChem Record
WO 2012101487, Page 64	CID 88951786





TASK FORCE Chemistry

=> ISR cites a PubChem compound incorrectly extracted by SureChEMBL from WO2012101487.

Vendor Communication Phase

- Organisations/Sources relying on automated image recognition indexing include:
  - **WIPO Patentscope** – Contributes 21.05 million substances
    - ❖WIPO admit to a **55%** error rate for compounds indexed from image recognition!
  - **SureChEMBL** – Contributes 24.46 million substances
  - **Google Patents** – Contributes 21.42 million substances
- The parties that could improve Chemical indexing were **Google/OntoChem**; **EBI** and **WIPO/deepmatter**

- The NCBI supported our approaches to these 3rd Parties to discuss improvements/solutions/reductions in errors – PDG agreed to brief the NCBI on the outcomes
- NCBI reloaded depositor supplied synonyms to reduce these errors

WIPO SureChEMBL Google Patents ontochem

Meeting with Google and OntoChem

- They received the presentation from Dr Ohms covering his study
- The two vendors were impressively proactive!
- Google/OntoChem indicated they are already changing production processes to use the better applications
- They tested 8 OCSR applications described in this article linked here
- CWU files (produced by applicants or patent offices) not considered a way forward
- A parallel filtering process (removing errors from final production)….not discussed at the meeting

Meetings with the EBI

- **Meetings with EBI (SureChEMBL) on September 20th , October 16th and January 24th**
  - **First Meeting** was demonstration of SureChEMBL 2.0 - Fast processing, simplified interface
    - SureChEMBL 2.0 now released in November 2023
    - We introduced the EBI to the Dr Ohms paper linked here
  - **Second Meeting:** different considerations e.g. filtering, indexing, using CWU files etc. They determined that OCSR application they used called OSRA had issues
  - **Third meeting:** Decimer seemed to be better than OSRA and therefore they are evaluating this tool and whether it would be easily integrated into their production system. They also wanted to read and analyse Google/Ontochem article linked here

PDG Partnering in Patent Information

WIPO deepmatter®

Meetings with WIPO and deepmatter

- **Meeting late in December 2023** when the PatentScope developmental budget would be refreshed for 2024: they were concerned about chemical indexing error rates & agreed to further meeting with deepmatter (They were using **OSRA** for chemical indexing and no update since 2017)
- **Meeting 24th January:** WIPO & deepmatter envisaged an error filtering project – were aware of organisations choosing **Decimer** ahead of **OSRA** – were keen to read Google/Ontochem article which the PDG forwarded. They qualified our expectations *"we cannot shoot for the moon*"
- **Meeting 11th March:** reducing Chemical Indexing Errors is now top priority – Agree that OSRA version 2017 was not good enough – Agreed that there should be wider application of Decimer for chemical indexing – They are considering their own testing and still devising a plan
  - ➢ Can the PDG provide funds for testing? –Not possible! – outside of PDG's remit!

TASK FORCE Chemistry

Meeting with the NCBI

- Thanked us for meeting and influencing the 3rd parties
- Request made to us - Please carry on meeting with them!
- Google/OntoChem batch submission in November 2023 contained significantly less compound indexing (hints at change in production methods)
- NCBI sees itself as the host of the data only
  - No anticipated role for NCBI in reducing and filtering errors
  - They have developed a list of trusted depositors (reduces errors received)

PubChem

TASK FORCE Chemistry

Following the Vendor Meetings…

- Kept Dr Joerg Ohms fully briefed (done in April 2024)
- Paused meetings from April 2024
- Continued email exchanges when required
- **October 2024 – WIPO –** Awaiting a study from deepmatter on replacing “Imagetostructure” components”
- Google/Ontochem have been emailed on the very latest production method/policy for Google Patents… awaiting a reply.
- Will consider introducing new discussion issue – Errors in **Chemical text** to structure indexing (another source of PubChem errors) if the vendors seem amenable
- Will contact NCBI for closing discussion

PDG Partnering in Patent Information

Any Questions?

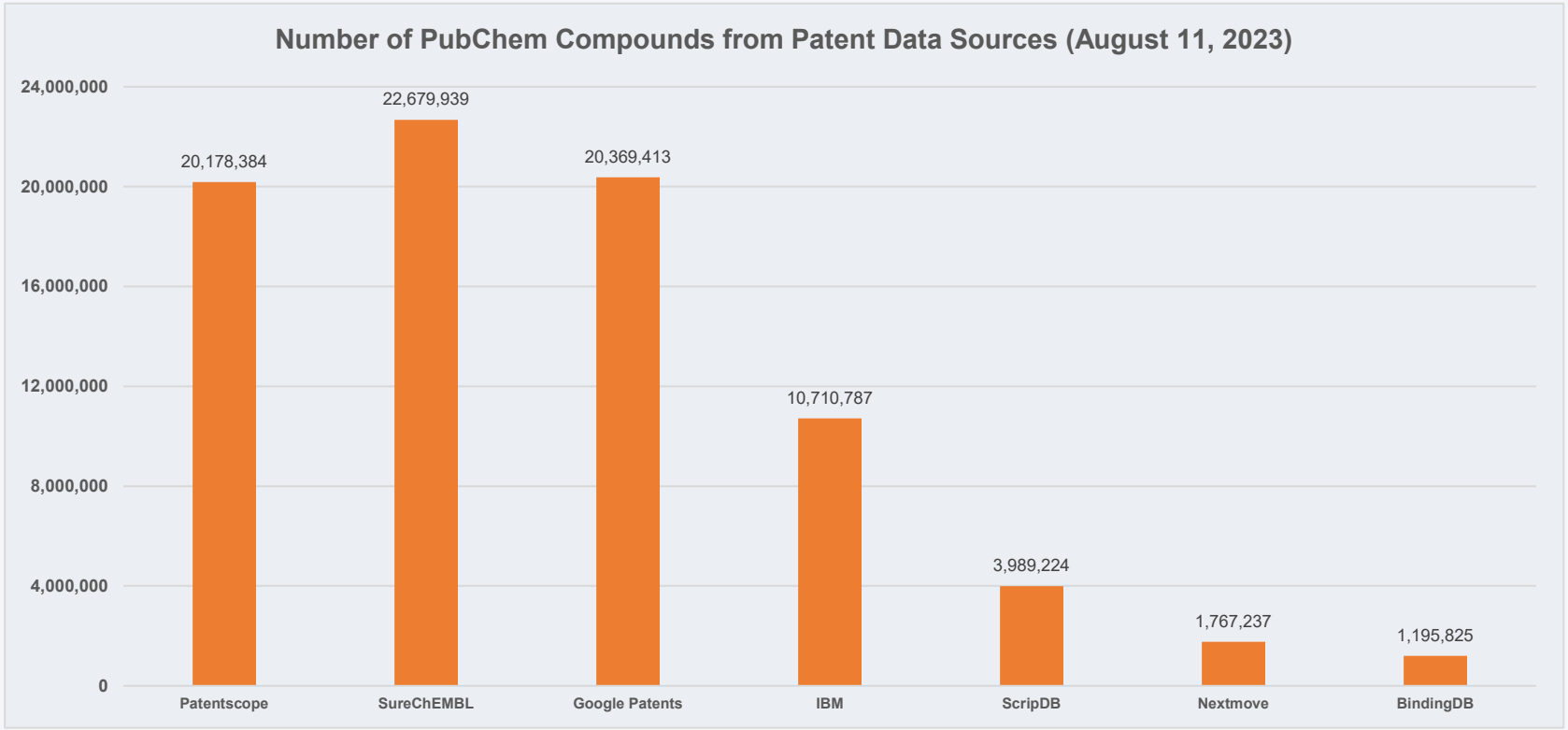
TASK FORCE Chemistry

Appendix 1

- PatCID by IBM…..
- PatCID: an open-access dataset of chemical structures in patent documents | Nature Communications
- “PatCID high-quality data outperforms currently available automatically-generated patent-databases. PatCID even competes with proprietary manually-created patent-databases.”
- 81 million compounds indexed
- 14 million unique compounds

PDG Partnering in Patent Information

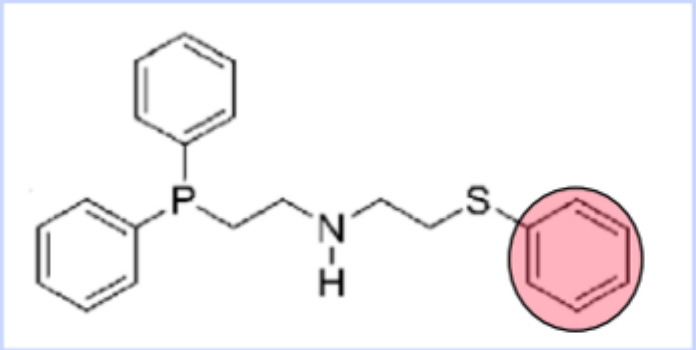
Appendix 2

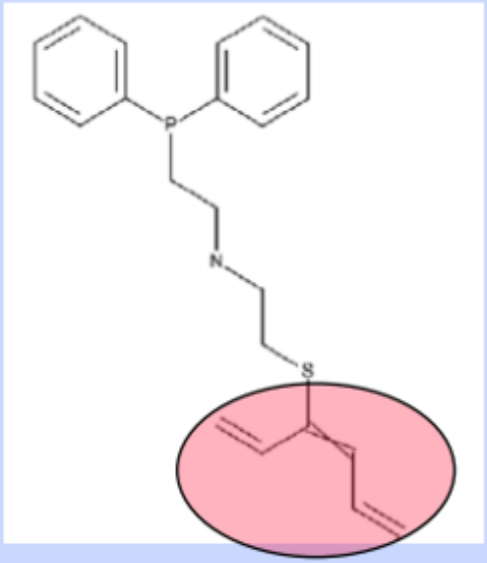


Appendix 3 - Dr Ohms Identified Several Types of Errors

Error – Benzene made acyclic

Source	PubChem Record
WO 2016031874, Page 75	CID 145024446





And many other implausible structures uncovered!

TASK FORCE Chemistry

Appendix 4 - Dr Ohms Identified Several Types of Errors

Error Categorisation – Fragmentation

Source	PubChem Record
WO 2017179883, Page 18	CID 145384411