



# “The Importance of Being Ernest”

Why gathering and cleaning all the relevant data matters for  
patent analysis

Anthony Trippe – Science IP/CAS

229<sup>th</sup> ACS National Meeting, San Diego, March 13<sup>th</sup> 2005

# Topics to be discussed

- Email
- First attempt
  - USPTO Web site
- Fundamentals
  - GIGO
  - Good analysis starts with a good search
  - Linear law of patent analysis
- 2<sup>nd</sup> try
  - Online research
  - Data cleanup
- Conclusions



You receive the following email at  
2pm on Friday from the VP of R&D  
for your company

- We have developed a new formulation for sustained release of Amoxicillin. We are interested in learning of the top 15 companies that have US patents covering Amoxicillin.



# So, you rush off to the USPTO Website

Patent Database Search Results: amoxicillin in 1976 to present - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Print Mail

Address http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetacgi/nph-search-bool.html&r=0&f=5&l=50&TERM1=amoxicillin&FIELD1=&co1=AND&TERM2=&FIELD2=&d=pbxt

## USPTO PATENT FULL-TEXT AND IMAGE DATABASE

[Home](#) [Quick](#) [Advanced](#) [Pat Num](#) [Help](#)  
[Next List](#) [Bottom](#) [View Cart](#)

Searching 1976 to present...

Results of Search in 1976 to present db for:  
**amoxicillin:** 1415 patents.  
*Hits 1 through 50 out of 1415*

PAT. NO.	Title
1 6,858,407	<a href="#">Human leucine-rich repeat containing protein expressed predominately in small intestine, HLRRS11</a>
2 6,852,713	<a href="#">Lactam derivatives and methods of their use</a>
3 6,852,526	<a href="#">Transdermal assay with magnetic clamp</a>
4 6,852,485	<a href="#">Method for identifying a compound for the treatment of microorganism infections by inhibiting energy storage and utilization in pathogens</a>
5 6,849,719	<a href="#">Antibody to an IL-17 receptor like protein</a>
6 6,849,608	<a href="#">Macrolide antibiotics</a>
7 6,849,413	<a href="#">PGRP-L polynucleotides, polypeptides, and antibodies</a>
8 6,849,271	<a href="#">Microcapsule matrix microspheres, absorption-enhancing pharmaceutical compositions and methods</a>
9 6,846,670	<a href="#">Genetically engineered herpes virus for the treatment of cardiovascular disease</a>
10 6,846,478	<a href="#">Promoting whole body health</a>



# Search and analysis

- You open the simple search page and search all fields for the term amoxicillin
- Over 1,400 records are retrieved (quite a large number you think)
- You extract out the Patent Assignee information and find the following...



# Records	Percentage	Patent Assignee
44	6.85	HUMAN GENOME SCIENCES, INC.
21	3.27	THE PROCTER & GAMBLE COMPANY
14	2.18	ADVANCIS PHARMACEUTICAL CORP.
14	2.18	TAKEDA CHEMICAL INDUSTRIES, LTD.
14	2.18	THE BIO BALANCE CORPORATION
12	1.87	ASTRAZENECA AB
11	1.71	CORIXA CORPORATION
10	1.56	ADOLOR CORPORATION
9	1.40	ARCHER-DANIELS-MIDLAND COMPANY
8	1.25	BASF AKTIENGESELLSCHAFT
7	1.09	THE UNIVERSITY OF BRITISH COLUMBIA
6	0.93	RESEARCH CORPORATION TECHNOLOGIES, INC.
6	0.93	VANDERBILT UNIVERSITY
5	0.78	ABBOTT LABORATORIES
5	0.78	BYK GULDEN LOMBERG CHEMISCHE FABRIK GMBH
5	0.78	NEW HORIZONS DIAGNOSTICS CORPORATION
5	0.78	NEW YORK UNIVERSITY
5	0.78	NYCOMED IMAGING AS
5	0.78	RUTGERS, THE STATE UNIVERSITY OF NEW JERSEY



# So, problem solved...right?

- Not so fast...
  - What about pending US patent applications?
  - What about PCT applications designating the US?
  - Further down the list, several of the companies are listed a second, third or fourth time because of misspellings
  - Some companies have merged since the patents were first issued
  - Are we really happy with this search?
    - Are all the documents on target?



# Fundamentals – GIGO

- GIGO – Garbage In Garbage Out
  - If you throw in the kitchen sink you will always be worried about how accurate your analysis results are
  - Good analysis is like any other scientific experiment – follow the scientific method
    - Hypothesis
    - Experiment
    - Conclusion





# Fundamentals – Good analysis starts with good search

- The search is your hypothesis stage
  - You learn about the area
  - You begin to understand what is reasonable and what is not
  - You formulate your strategy for how you will conduct your experiments



# Fundamentals – Linear law of patent analysis

- The experimental stage
  - Understand the need behind the need
  - The need drives the question
  - The question drives the data
  - The data drives the tool
- It is critical to follow these steps in this order



## 2<sup>nd</sup> try – search

- Use multiple searching techniques
- Use indexed databases for focused retrieval and accuracy
- Restrict full-text sources to presence in claims only
- Use multiple data sources for comprehensiveness



# Multiple searching techniques

- Structure searches
  - 123 substances found in CAS Registry file (20 more than if the Amoxicillin RN was run as CRN)
  - Additional records found in Marpat
  - 5 substances found in Derwent Chemical Resource
- Chemical name synonyms
  - Over 300 synonyms for Amoxicillin
- Controlled vocabulary
  - Use of IFI Uniterms
  - Derwent Manual codes (subscribers only)
  - US/pcs





# Indexed databases for accuracy

- In order for a document to be indexed by a database produce it traditionally needs to play a significant role in the document
  - Avoid laundry lists while gist of document is focusing on something different or a different substance all together
  - Additional indexing allows for focused retrieval on the role or use of the substance



# Restrict full-text sources to claims only

- Claims are the portion of the document which is legally binding
- If it is not in the claims but is in the specification than it may be a minor element of the document
- Portions of the specification may be misleading
  - The Background of the Invention section details what the invention is not about



# Use multiple data sources for comprehensiveness

- Each component added to the comprehensiveness of the search
  - Name searching found additional records from structure search
  - Derwent and IFI Uniterm searches discovered unique records which were not in the CPlus database
    - This is not the fault of the individual indexing staffs but simply a difference in editorial policy





# But be careful...

- It is important when using multiple sources and multiple patent offices to reduce to one document per invention
  - Duplicate removal was employed
  - In the full-text files the US applications were removed if there was a subsequent granted patent
  - PCT applications were removed if there was a corresponding US application or granted patent



# Use multiple data sources for comprehensiveness

L84 1005 DUP REM L32 L38 L48 L53 L69 L83 (6  
DUPLICATES REMOVED)

ANSWERS '1-700' FROM FILE HCAPLUS

ANSWERS '701-704' FROM FILE MARPAT

ANSWERS '705-871' FROM FILE WPINDEX

ANSWERS '872-943' FROM FILE IFIUDB

ANSWERS '944-964' FROM FILE USPATFULL

ANSWERS '965-1005' FROM FILE PCTFULL



## 2<sup>nd</sup> try – analysis

- Merge names with misspellings
- Where possible factor in mergers
- Download references to confirm results



# Single Column Analysis

and us/ds  
1005

**STN Analyze Plus**

Analyze this L-number on 1 or 2 fields. Select "Group similar terms" to be able to group terms within a field. The results will be charted in Microsoft Excel.  
Click Analyze to process the information. Click Cancel to exit.

**1005 answers are available to analyze.**

One field analysis  
Select first field  
 Author/Inventor Name  
 Corporate Source/Patent Assignee  
 Company Name  
 Publication Year  
 Controlled Terms  
 Patent Country  
 National Classification  
 WIPO International Classification  
 by SubClass  by Main group  All  
 Group similar terms

Two field analysis  
Select second field  
 Author/Inventor Name  
 Corporate Source/Patent Assignee  
 Company Name  
 Publication Year  
 Controlled Terms  
 Patent Country  
 National Classification  
 WIPO International Classification  
 by SubClass  by Main group  All  
 Group similar terms

Options

< Back Analyze Cancel

Select Discover! Wizard

Choose a search wizard:  
Select Database  
Author  
CAS Registry Number  
Chemical Name  
Corporate Source  
Subject  
Edit alert

Select an L-number:  
L67 1  
L68 21  
L69 22  
L70 278  
L73 69  
L75 414  
L77 99  
L79 5  
L81 4  
L82 64  
L83 41  
L84 1005

Choose a results wizard:  
Analyze Plus  
Display  
Refine  
Save  
Save R-group data  
Save for SciFinder  
Get Related Polymers  
Get Related Sequences  
Create Single-file SDI  
Go to L-number

```
'ATFULL' ENTERED AT 16:49:26 ON 01 MAR 2005
ING COPYRIGHT (C) 2005 AMERICAN CHEMICAL SOCIETY (ACS)

'FULL' ENTERED AT 16:49:26 ON 01 MAR 2005
' (C) 2005 Univentio
'O COST=ACS_TALK
IG COMPLETED FOR L32
IG COMPLETED FOR L38
IG COMPLETED FOR L48
IG COMPLETED FOR L53
IG COMPLETED FOR L69
IG COMPLETED FOR L83
1005 DUP REM L32 L38 L48 L53 L69 L83 (6 DUPLICATES REMOVED)
ANSWERS '1-700' FROM FILE HCAPLUS
ANSWERS '701-704' FROM FILE MARPAT
ANSWERS '705-871' FROM FILE WPINDEX
ANSWERS '872-943' FROM FILE IFIUDB
ANSWERS '944-964' FROM FILE USPATFULL
ANSWERS '965-1005' FROM FILE PCTFULL
```



# Data cleanup



Copyright © 2005 American Chemical Society.  
Copying or further distribution requires permission of CAS.

# Final results

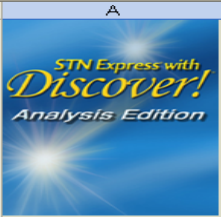
Microsoft Excel - Book1

File Edit View Insert Format Tools Data Window Livelink Help Adobe PDF

STN Express with Discover! Analysis Edition

100% Magnification Sort Row or Column Delete Row Combine

A3 Pfizer

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2	GlaxoSmithKline	139										
3	Pfizer	66										
4	USA	59										
5	Bristol-Myers Squibb	22										
6	Gist-Brocades N. V., Neth.	22										
7	Human Genome Sciences	17										
8	Merck	15										
9	Bayer	12										
10	F. Hoffmann-La Roche	12										
11	Ruben, Steven, M.	12										
12	Dsm Ip Assets B.V., Neth.	10										
13	Novartis	10										
14	Biochemie G.M.B.H., Austria	9										
15	Procter and Gamble	9										
16	Depomed	8										



# Compare the approaches

1 <sup>st</sup> Attempt	2 <sup>nd</sup> Attempt
HUMAN GENOME SCIENCES, INC.	GSK
THE PROCTER & GAMBLE COMPANY	PFIZER
ADVANCIS PHARMACEUTICAL CORP.	US GOVERNMENT
TAKEDA CHEMICAL INDUSTRIES, LTD.	BMS
THE BIO BALANCE CORPORATION	GIST-BROCADES
ASTRAZENECA AB	HUMAN GENOME SCIENCES, INC.
CORIXA CORPORATION	MERCK
ADOLOR CORPORATION	BAYER
ARCHER-DANIELS-MIDLAND COMPANY	HOFFMANN LA ROCHE
BASF AKTIENGESELLSCHAFT	RUBEN, STEVEN
THE UNIVERSITY OF BRITISH COLUMBIA	DSM LP ASSETS
RESEARCH CORPORATION TECHNOLOGIES, INC.	NOVARTIS
VANDERBILT UNIVERSITY	BIOCHEMIE GMBH
ABBOTT LABORATORIES	THE PROCTER & GAMBLE COMPANY
BYK GULDEN LOMBERG CHEMISCHE FABRIK GMBH	DEPOMED



# Display only the records you need

PRIORITY APPLN. INFO.: GB 1975-152210 19750414

**STN Analyze Plus**

*STN Express with Discover!*

To display retrieved answers, select an Answer range and a Display format, then click Display.

Click Finish to return to STN.

From file: HCAPLUS

Answer range: All Answers

Display format:  Basic  Advanced

TI-----Title only  
IBIB-----Bibliographic Information  
IBIB IABS-----Bibliographic plus Abstract  
IALL-----Complete Record

Issue cost warning

< Back Display Finish

COPYRIGHT (C) 2005 AMERICAN CHEMICAL SOCIETY (ACS)  
CHARGED TO COST=ACS\_TALK





# Conclusions

- Good analysis requires good search
- Good analysis requires:
  - Accurate data
  - Cleaned data
  - Focus on objective
- Take the road less taken and be Ernest in your efforts for the best result

