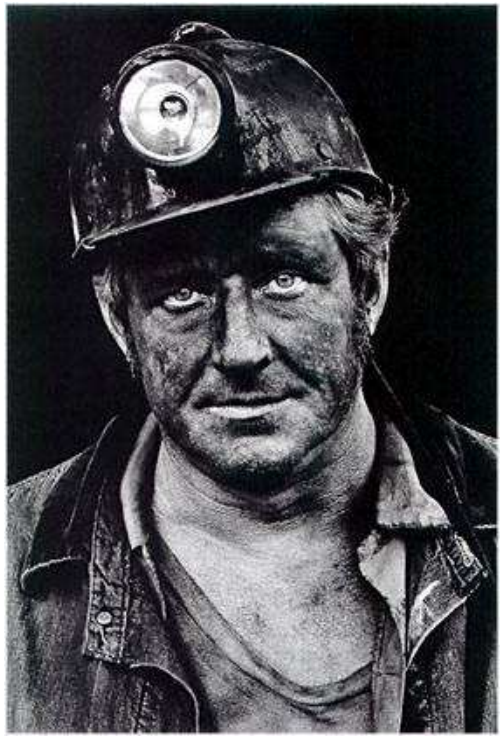
ICIC-Conference 2007
October, 21st – 24th Barcelona

Information Extraction from Full-Text

Challenges and Opportunities



Wolfgang Thielemann

Bayer HealthCare
Bayer Schering Pharma

The Standard: Added-value Databases

For decades added-value databases with their manual or semi-automatic created indices have been the primary source for information professionals in life sciences

For good reasons:

- **Focus on core aspects**
- **Standardized vocabulary (one language, controlled terms)**
- **Chemical structures are made searchable**
- **Relationships are structured (e.g. Registry No + Role)**
- **Structured indices can be used for analysis of large document sets**

Availability was an obvious reason when access to electronic full text was not available...

Added-value databases

Some examples of well established life science focused added-value databases:

THOMSON

Biosis

WPI

Dolphin

DCR

A division of the American Chemical Society

Registry

CAPlus

SciFinder Part of the process.™

Marpat

ifi CLAIMS® Patent Services THE PATENT RESEARCH EXPERTS

ELSEVIER EMBASE.com

PubMed www.pubmed.gov

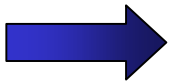
Full-Text & Text Mining

Full-text opens the way to the complete information

Patents of the main countries are available as full-text.
Powerful Text Mining applications can be used to search and analyze unstructured text

Advantage 

- **all aspects present in full-text**
- **the exact wording (claims!) can be analyzed**
- **commercial and proprietary thesauri/ontologies allow structured access to many aspects relevant to life science**
- **patent full-text available for a fraction of cost compared to indexed databases**

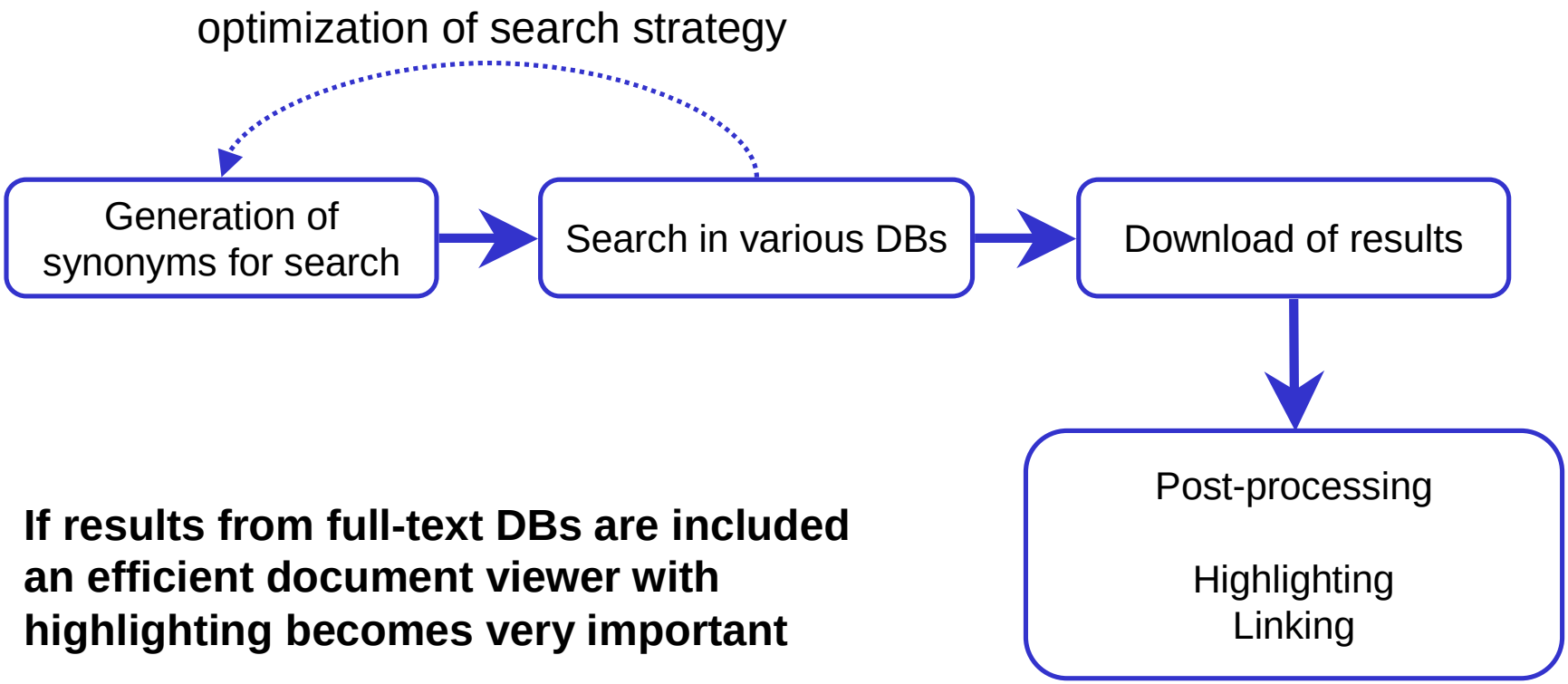


Wonderful! Full-text is the solution for all information retrieval problems!

... really? Let's have a closer look

Classical Search vs Text Mining approach

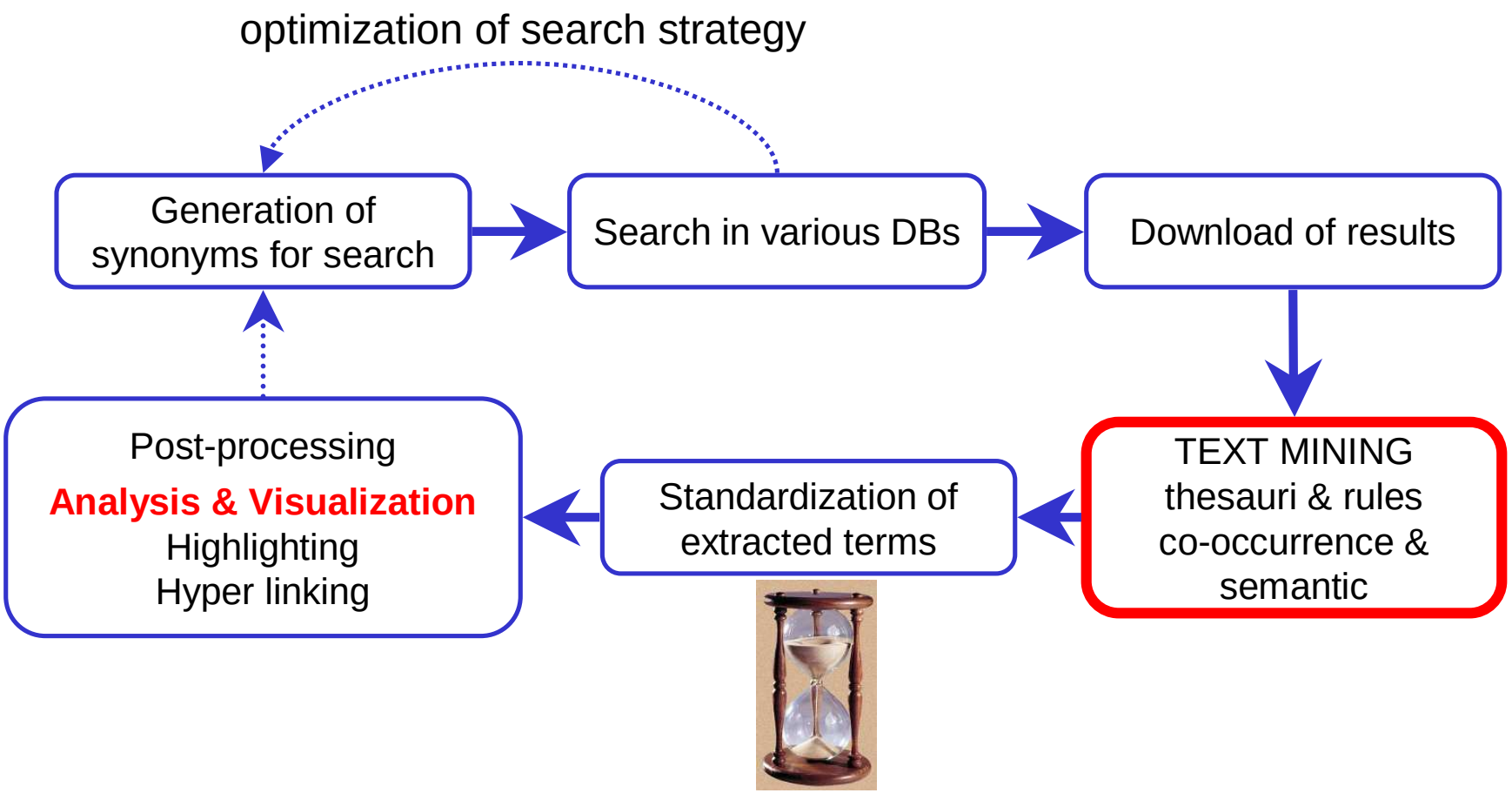
Typical workflow of a “classical” search



If results from full-text DBs are included an efficient document viewer with highlighting becomes very important

Classical Search vs Text Mining approach

Typical workflow of a thesaurus based text mining approach:



Text Mining Applications

Some examples of Text Mining (and related) applications:

LINGUAMATICS

temis

STN® ANAVIST™

aureka®
Search • Organize • Analyze • Collaborate

metalife

Medscan Reader™

Basically two different mining approaches:

1. **Topic maps using all terms and statistical methods / fingerprints (excluding stop words)**

2. **Mining and analysis of topics using thesauri / ontologies**

Searching in full-text patents

Many request to pharmaceutical R&D focus one or more of the following topics:

- **Compounds / Drugs**
- **Drug actions**
- **Indications**
- **Formulations**

Are unstructured full-texts useful sources for these topics?

Searching compounds in full-text patents

National Library of Medicine - Medical Subject Headings

2007 MeSH

MeSH Descriptor Data

Return to Entry Page

Standard View. Go to Concept View; Go to Expanded Concept View

MeSH Heading	Simvastatin
Tree Number	D04.615.638.400.900
	A derivative of LOVASTATIN and potent competitive inhibitor of 3-hydroxy-3-methylglutaryl coenzyme A reductase (HYDROXYMETHYLGLUTARYL COA e in cholesterol biosynthesis. It may also interfere with steroid hormone production. Due to the induction of hepatic LDL HOLESTEROL.
Entry Term	MK-733
Entry Term	Synvinolin
Entry Term	Zocor
	OU EC HI IM IP ME PD PK PO RE SD ST TO TU UR
Action	Anticholesteremic Agents
Pharm. Action	Antilipemic Agents
Pharm. Action	Hydroxymethylglutaryl-CoA Reductase Inhibitors
Registry Number	79902-63-9
Previous Indexing	Lovastatin/analogs & derivatives (1988-1997)
Previous Indexing	Naphthalenes (1986-1987)
History Note	98; use SIMVASTATIN (NM) 1986-97
Date of Entry	19970620
Unique ID	D019821

Searching compounds in full-text patents

Searching full-text patents (WO, EP, US, FR, GB, DE, JP) for the term "Simvastatin" yields 9030 patents (3666 INPADOC families).

But there are 392 more patents which are not found due to typos and ORC errors:

SIMVASTATINE	71
SIMVESTATIN	43
SIMVISTATIN	33
SINVASTATIN	28
SIMVARSTATIN	26
SYMVASTATIN	14
SIMAVASTATIN	13
SLMVASTATIN	9
SIMBASTATIN	8
SIMVASTSTIN	8
SIMVATATIN	7
SIMVASTATINA	6
SIMIVASTATIN	5
SIMVASTATION	4

S1MVASTATIN	3
SIMASTATIN	2
SIMNVASTATIN	2
SIMVASTATIV	2
SIMVASTITIN	2
SIVASTATIN	2
IMVASTATIN	1
S IMVASTATIN	1
S IMVASTATINA	1
SIIMVASTATIN	1
SIM ASTATINE	1
SIMVASTACIN	1
SIMVASTAFIN	1
SIMVASTALIN	1

SIMVASTANINA	1
SIMVASTATI NE	1
SIMVASTATIN7	1
SIMVASTATING	1
SIMVASTATINM	1
SIMVASTATINO	1
SIMVASTATIU	1
SIMVASTATJN	1
SIMVASTATN	1
SIMVASTAT'N	1
SIRVASTATIN	1
YSIMVASTATINE	1

and more...

Searching compounds in full-text patents

If you think that was bad... look at the IUPAC names:

WO2007096753	6(R)-[2-(8'(S)-2",2"-dimethylbutyryloxy-2'(S),6'(R)-dimethyl- l',2',6',7,'8',8a'(R)-hexahydronapthyl-l'(S))-ethyl]-4(R)-hydroxy -3,4-5,6-tetrahydro- 2H-pyran-2-one
WO2005095374	6(R)-[2-[8(5)-(2,2-dimethyl.butyyloxy)-2 (S), 6 (R)-dimethyl-1, 2, 6, 7, 8, 8a(R)-hexahydro-l (S)-napthylelhyl/-4(R)-hydroxy-3, 4, 5, 6-tetrahydro-2H-pyran-2 one
WO2005095374	6(R)-[2-[8(S)-(2, 2-dimethylbulyryloxy)-2 (S), 6 (R)-dimethyl-1, 2, 6, 7, 8, 8a(R)-hexabydro-l (S)-napthylethyl/-4(R)-hydroxy-3, 4, 5, 6-tetrahydro-2H-pyran-2 one
WO2003018570	6(R)-[2-[8(S)-(2,2 10 dimethylbutylyloxy)-2(S),6(R)-dimethyl-1,2, 6,7,8,8a(R) hexahydronaphthyl]-l(S)ethyl]-4(R)-hydroxy-3,4,5,6 tetra hydro-2H-pyrane-2-one
WO2003048149	6(R)-[2-[8(S)-(2,2- dimethylbutylyloxy)-2(S),6(R)-dimethyl-1,2,6,7,8,8a(R)-hexahydronaphthyl]-l(S)ethyl]-4(R)-hydroxy-3,4,5,6 20 tetrahydro-2H-pyrane-2-on
WO2003018570	6(R)-[2-[8(S)-(2,2-dimethylbutylyloxy)-2(S),6(R)-dimeth yl-1,2,6,7,8,8a(R)-hexahydronaphthyl]-l(S) ethyl]-hydrox y-3,4,5,6-tetrahydro-2H-pyrane-2-one
WO2005095374	6(R)-[2-[8(S)-(2,2-dimethylbutyrylaxy)-2 (S),6 (R)-dimethyAl, 2, 6, 7, 8, 8a(R)-hexahydro-l (S)-napthylJethyl)-4(R)-hydroxy-3, 4, 5, 6-tetrahydro-2H-pyran-2 one
WO2006072963	6(R)-{2[8(S)-(2,2dimethylbutyryloxy)2(5),6(R).. dimethyl.. 1,2,6,7,8,8a(R)-hexahydro-1 (S)-naphthylJethy1J-4(R)hydroxy3,4,5, 6 tetrahydro-2H-pyran-2-one

Searching compounds in full-text patents

In 141 patents containing the IUPAC name of Simvastatin not one (!) contained the correct name:

6(R)-[2-[8(S)-(2,2-dimethylbutyryloxy)-2(S),6(R)-dimethyl-1,2,6,7,8,8a(R)-hexahydronaphthyl]-1(S)ethyl]-4(R)-hydroxy-3,4,5,6-tetrahydro-2H-pyran-2-one

After removing all characters which are not a letter or number:

6R28S22dimethylbutyryloxy2S6Rdimethyl126788aRhexahydronaphthyl1Sethyl4Rhydroxy3456tetrahydro2Hpyran2one

13 out of 141 patents were found...

Searching compounds in full-text patents

Searching for (long) IUPAC names in patents full-text will miss most hits

This is very relevant for all applications which convert IUPAC names into chemical structures!

Nevertheless, searching for brand names or generic names will for sure find additional relevant hits!

Searching compounds in full-text patents

Examples of hits found in title, abstract & claims of PCT applications but not found with CAS search:

WO0017166: A method for treating atherosclerosis as recited in claim 32 wherein the second compound is lovastatin, simvastatin, pravastatin, fluvastatin, atorvastatin or rivastatin.

WO2006089309: A pharmaceutical composition in accordance with claim 3 further comprising about 10mg, 20mg or 40mg of simvastatin.

WO2005051451: The medical device of claim 1 wherein the agent is a HMGCoA reductase inhibitor, wherein the HMGCoA reductase inhibitor is -- simvastatin or an analogue or derivative thereof.

WO2004100857: The tablet of claim 2, wherein the active ingredient is a drug selected from the group consisting of ....atorvastatin, lovastatin, pravastatin, simvastatin, nitroglycerin, phenobarbital, phenytoin...

WO2003059192: The stent of claim 11, wherein the inhibitor is one or more of lovastatin, simvastatin, pravastatin, fluvastatin, atorvastatin, or cerivastatin.

Why were these relevant and correct hits not found in the indexed database?

Searching compounds in full-text patents

Synonyms, synonyms...

All available synonyms have to included to be reasonable comprehensive

1944	SIMVASTATIN*
235	ZOCOR*
225	LIPEX
105	VELOSTATIN
37	SYNVINOLIN*
5	ZOCORD
3	LIPOVAS
3	MK 733*
3	SIMOVIL
2	LODALES

Search in PCT full (IPCs a61k or a61p or a01n or c07c or c07d or c07e or c07f or c07g or c07h or c07i or c07j or c07k or c12n or c12P or c12q or g01n-033)

*contained in MeSH Thesaurus

Searching compounds in full-text patents

There are > 300 patents which refer to Pharmaproject numbers instead of the compound name:

US7101875 B2: Methods for treating arthritic disorders

Suitable cyclooxygenase 2 (COX-2) inhibitors for use with the methods of this invention include, but are not limited to, ... 1-(7-tert-butyl-2,3-dihydro-3,3-dimethyl-5-benzo-furanyl)-4-cyclopropyl butan-1-one, **Pharmaprojects No. 6089** (Kotobuki Pharmaceutical), RS-113472, RWJ-63556, S-2474, S-33516, SC-299...

US6967204 B2: Treatment of insulin resistance syndrome and type 2 diabetes with PDE9 inhibitors

Still other type cGMP PDE5 inhibitors useful in conjunction with the present invention include: ... 1-[4-[(1,3-benzodioxol-5-ylmethyl)amino]-6-chloro-2-quinazolinyl]-4-piperidineca rboxylic acid, monosodium salt; **Pharmaprojects No. 4516** (Glaxo Wellcome); **Pharmaprojects No. 5051** (Bayer); **Pharmaprojects No. 5064** (Kyowa Hakko; see WO 96/26940); **Pharmaprojects No. 5069** (Schering Plough); GF-196960 (Glaxo Wellcome); E-8010 and E-4010 (Eisai); ...

Searching compounds in full-text patents

Sepracor INC used 64 times the name “Sildenophil” (in 18 patents) without once mentioning the correct name “Sildenafil”:

US6974837 B2 “Compositions comprising sibutramine metabolites in combination with phosphodiesterase inhibitors”
SEPRACOR INC

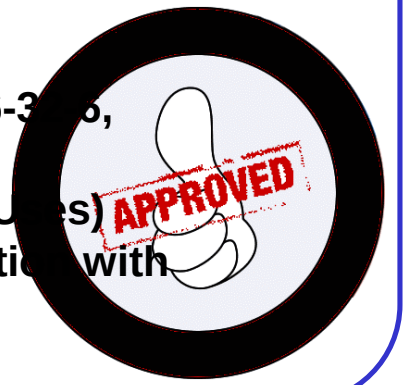
....Particular phosphodiesterase inhibitors include, but are not limited to, sildenophil (Viagra®), desmethylsildenophil, vinopocetine, milrinone...

Could these patents be found using indexed databases?

Searching compounds in full-text patents

CAplus combines the 18 patents into 6 families; 3 of these families could be found with the name "**Sildenafil**". "**Sildenophil**" is not mentioned once

US 2002010198
IT ...81840-15-5, Vesnarinone 139755-83-2, Sildenafil 391936-3[illegible]6,
Peroximone
RL: THU (Therapeutic use); BIOL (Biological study); USES (U[illegible]es)
(compns. comprising sibutramine metabolites in combinati[illegible] with phosphodiesterase inhibitor)



WPI combines the 18 patents into 9 families; none (!) of these families could be found with the correct name "**Sildenafil**" but 6 with "**Sildenophil**":

US 2002010198
TECH
PHARMACEUTICALS - Preferred Component: The phosphodiesterase inhibitor is sildenophil, desmethylsildenophil, vinopocetine, milrinone, amrinone, pimobendan, cilostamide, enoximone, peroximone, vesnarinone,

Searching for Drug Actions

National Library of Medicine - Medical Subject Headings

2007 MeSH

Name of Substance	cyclic nucleotide phosphodiesterases, type 5
Entry Term	PDE 5 enzyme
Entry Term	PDE5 protein
Entry Term	PDE5A protein, human
Entry Term	PDE5A1 protein, human
Entry Term	Pde5 protein, mouse
Entry Term	Pde5a protein, mouse
Entry Term	Pde5a protein, rat
Entry Term	cGMP-binding, cGMP-specific 3',5'-cyclic nucleotide phosphodiesterase
Entry Term	phosphodiesterase 5
Entry Term	phosphodiesterase 5A, cGMP-specific protein, human
Entry Term	phosphodiesterase 5A, cGMP-specific protein, mouse
Entry Term	phosphodiesterase 5A, cGMP-specific protein, rat
Entry Term	phosphodiesterase V
Entry Term	phosphodiesterase-5

Searching for Drug Actions

In a set of 1422 full text patents on PDE 5 inhibitors over 500 (!) variations for "PDE 5 inhibitor" were found!

Some examples:

- **inhibitor of** cyclic guanosinemonophosphate (cGMP)-specific **phosphodiesterase type 5**
- **inhibitors of** monophosphate (cGUT)-specific **phosphodiesterase type 5**
- **inhibit Type 5** isoenzyme form of **phosphodiesterase**
- **Inhibitory** Effects of Compounds on **PDE V** Activity
- **type V** cyclic GMP-specific **phosphodiesterase inhibitor**
- **Inhibitoren** von cGMP-artigen **V- Phosphodiesterasen**
- **inhibition** of **PDE** 3 and **5**

Searching for Drug Actions

A search string like:

(phosphodiest? or PDE?) (5A) (5 or V) (5A) (inhib? or antago?)

will also retrieve these false positives:

- Beispiel **5** [0072] Protokolle fuer die **PDE**-IV-**Inhibitionsaktivitaet**
- 3‘,**5**'-monophosphate (cAMP) **PDE** type IV (**PDE** IV) **inhibitors**
- **inhibitors** on the activity of **PDE9A** are shown in Figure **5**
- also **5** -lipoxygenase **inhibitors**, **phosphodiesterase inhibitors**, or
- 0. **5** mM of the **phosphodiesterase inhibitor**
- **phosphodiesterase inhibitor** was used the i.**v**
- **inhibits** the type 4 **PDE** with an ED.sub.50 of 0.1-0.**5**

Searching for Indications

Indications are typically well covered in the MeSH thesaurus

But what if paraphrases are used instead of established indication names?

The following searches were done in title, abstract & claims

US6752986B2: “Composition and methods for affecting metallocorrinoid uptake”

12. A method for increasing TCII-R activity in a subject to treat, ameliorate, or diagnose **a condition characterized by cellular proliferation** comprising the step of administering to a subject in need of such treatment interferon-beta; in and amount effective to increase TCII-R activity in the subject and administering a cobalamin drug conjugate.

17. The method of claim 12, wherein said **condition is unwanted cellular proliferation**.

Searching for Indications

...and it gets worse:

US6482798B2

3. A method of treating **a disease state characterized by the activation of plasmatic enzyme systems in a patient suffering therefrom**, said method comprising administering to said patient an amount of the compound according to claim 1 which is effective to inhibit said enzyme.

US20010048930A1

17. A method of treating a patient suffering from a **disease characterised by inappropriate lymphocyte activity** which method comprises administering to the patient a lymphocyte produced by the method of any one of claims 1 to 9.

Searching for Indications

US6358928B1

31. A pharmaceutical composition useful in the treatment of **disease states characterized by an over-activity of one or more serine proteases** comprising a pharmaceutically acceptable carrier containing a therapeutically effective amount of the serine protease inhibitor compound of claim 1.

US6277061B1

6. A method according to claim 5, wherein the mammal has **a condition characterized by excessive MT-MMP activity**. | 7. A method according to claim 5, wherein the mammal has **a condition characterized by excessive osteoclast activity**.

EP1044987B1

7. A macromolecule according to any one of claims 1 to 5 **for use in medicine**. | 8. Use of a macromolecule according to any one of claim to 5 in the preparation of a composition for the **treatment of a diseas characterised by the undesired production of a protein**.

Pharmaceutical formulations

Testset: 100 patents on Cyclosporin A (published 2002 – 2007)

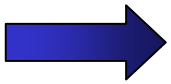
Coverage of formulation information in DOLPHIN, CAS, WPI and Full-Text (Text Mining)

Patent number	DOLPHIN	CAS	WPI Administration	Text Mining from Full-text
US2007087962	Ophthalmic liquid	(carriers; pharmaceutical	Administration is	emulsion; eye; oil;
WO2007042037	Tablet formulation	(beads; combination	The composition is in solid	capsule; delayed release;
WO2007016073	Ophthalmic liquid	(ophthalmic; pharmaceutical	The pharmaceutical	emulsion; eye; oil
WO2007010045	Formulation			
WO2007008894	Ophthalmic formulation;	(emulsions; composition	The composition is	emulsion; eye; ocular; oil
WO2006123354	Formulation	(capsules; oral		capsule; oil-in-water
WO2006108637	Formulation		The particles or suspension	capsule; cream; dispersion;
WO2006110802	Controlled release	(aerosols; nanoparticulate	The compositions are	aerosols; coating; controlled
WO2006105980				capillary; injectable;
WO2006108130				
WO2006102612				
WO2006073786	Ophthalmic formulation	(solns., ophthalmic;		eye; ophthalmic
US07070802	Formulation	(capsules; pharmaceutical		blister; drug delivery system;
WO2006066870	Oral formulation	(solid dispersions; solid		disintegrant; dispersion; oral;
US2006140860				
WO2006055418	Transmucosal formulation;	(emulsions; cyclosporin A	Administration of (I) topical	mouth; mucosal; oral
WO2006055417	Rectal formulation, local	(rectal; cyclosporin A	(I) is present in the rectal	lung; mouth; nasal; oral
WO2006051067	Formulation	Lipophilicity (process for	The composition is	freeze-dried; spray
WO2006050836	Ophthalmic liquid	(gels, ophthalmic; ophthalmic	Administration of (I) is	delivery device; drops;
WO2006050123	Emulsion formulation	(liqs., dispersions;	Administration of	dispersion; drug delivery
US2006094646			Cyclosporine is	
WO2006039558	Sustained release	(gels, ophthalmic; sustained-	Administration can be	delivery system; dispersion;
WO2006036729			The composition is	
WO2006023388	Formulation	(carriers; biodegradable		bucal; controlled release;

Pharmaceutical formulations

Analysis of results:

- **Number of patents with formulation information seems to be comparable for all 4 sources**
- **Information patterns looks quite similar for DOLPHIN, CAS and Full-text**
- **Details sometimes differ a lot!**



A comprehensive formulation thesaurus in combination with text mining can retrieve more details than added-value databases!

Pharmaceutical formulations

WO2006023388

DOLPHIN: Formulation

CAS (*Drug Delivery System*): carriers; biodegradable polyoxyalkylene-polyester diblock copolymers having temperature-dependent reversible sol-gel properties

WPI (*Administration*): -

Full-Text: A method for the administration of a drug to a warm-blooded animal in a controlled release form which comprises: I (1) providing an aqueous biodegradable polymeric drug delivery composition....
The method according to Claim 7 wherein said administration is by parenteral, ocular, topical, inhalation, transdermal, vaginal, buccal, transmucosal, transurethral, rectal, nasal, oral, pulmonary or aural means.

Pharmaceutical formulations

WO2003094840

DOLPHIN: -

CAS (*Drug Delivery System*): -

WPI (*Administration*): The compounds are administered orally, buccally, rectally, parenterally (e.g. intraperitoneally), intradermally, transdermally or intracheally. The dosage of toleragen is at least 0.005 (preferably at least 0.01, especially at least 0.05, particularly not more than 1) mg/kg/week.

Full-Text: may be fornnulated into preparations in solid, semi-solid, liquid or gaseous forms, such as tablets, capsules, powders, granules, ointments, solutions, suppositories, injections, inhalants, gels, microspheres, and aerosols.
As such, administration ofthe compounds can be achieved in various ways, including oral, buccal, rectal, parenteral, intraperitoneal, intradermal, transdermal, intracheal, etc., administration.

Opportunities of Full-Text & Text Mining

Further examples where full-text search (esp. in combination with Text-Mining can be of great advantage:

- **Pharmaceutical combinations (not specifically indexed in added value databases; the position of the compound in the text is important)**
- **General use claims (exact wording of the claims has to be analyzed)**
- **Details of Experiments e.g. in-vivo (only found in description part of patents; typically not indexed)**

Conclusions

Searching in full-text patents makes sense:

- **if you search for details typically ignored in indexed added-value databases or typically found in the description part of a patent**
- **if you are willing to deal with all the hurdles (different languages, OCR errors, semantic variations, unusual paraphrases etc.)**
- **if you have an effective full-text viewer**

But the full power of full-text can only be unleashed if you use Text Mining Technology!!

Thank you!