IBM LanguageWare

Text Mining in Life Sciences

UIMA Framework and Knowledge Discovery at IBM

Alex Nevidomsky, IBM Dublin Software Lab.
alex_nevidomsky@ie.ibm.com

© 2006 IBM Corporation

- UIMA Concept
- UIMA reference implementation
- IBM OmniFind: UIMA-enabled platform
- Text analytics and domain customization
- UIMA applications in HCLS
  - Clinical trials search
  - Chemical search
  - Knowledge discovery

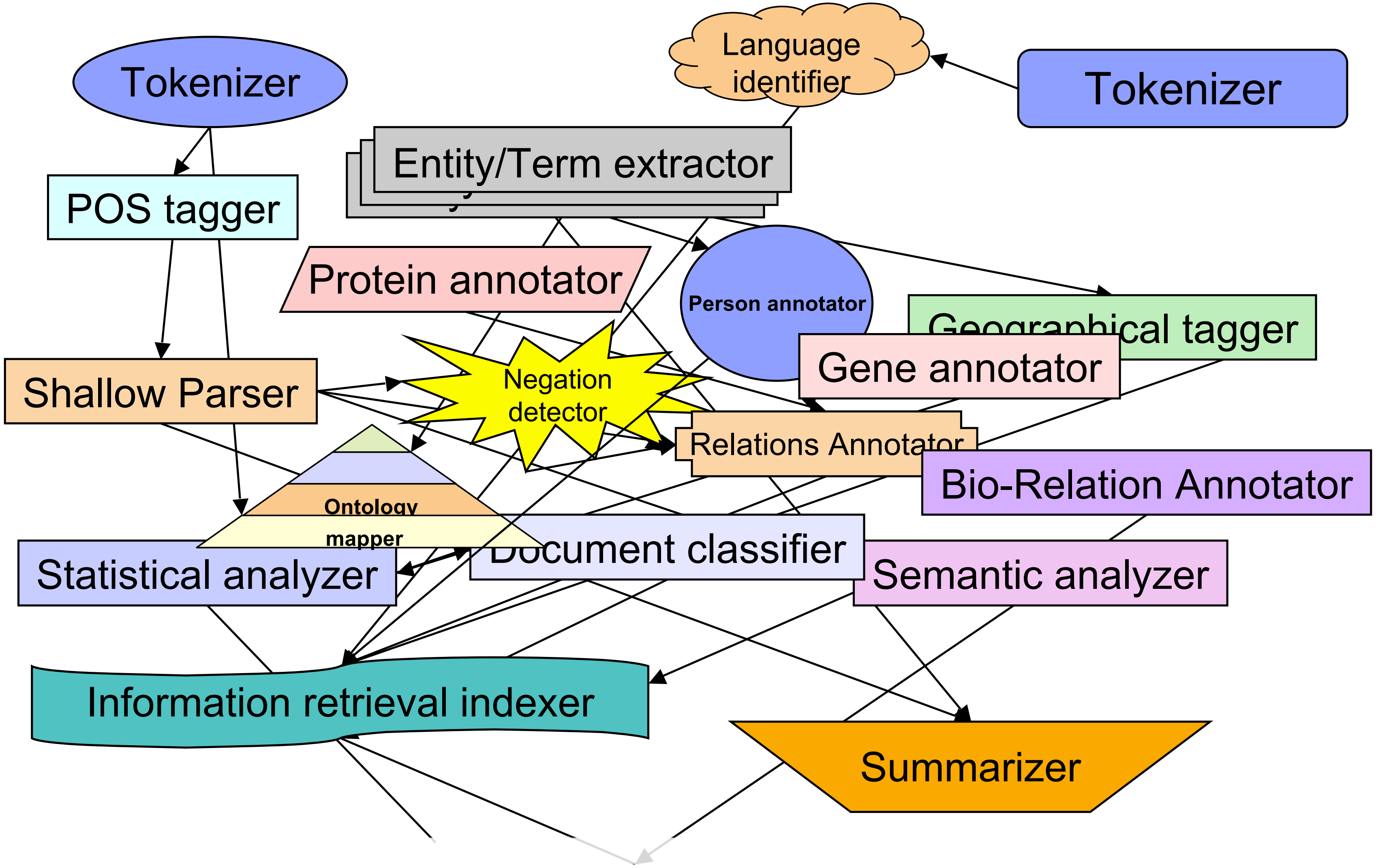
© 2006 IBM Corporation

UIMA Concept

- Analytics
  - Specialized
  - Independently developed
  - Using different techniques
- Combination Hypothesis
  - Combination of different techniques (even based on mutually exclusive premises) produces better results
- Issue of integration/interoperation

© 2006 IBM Corporation

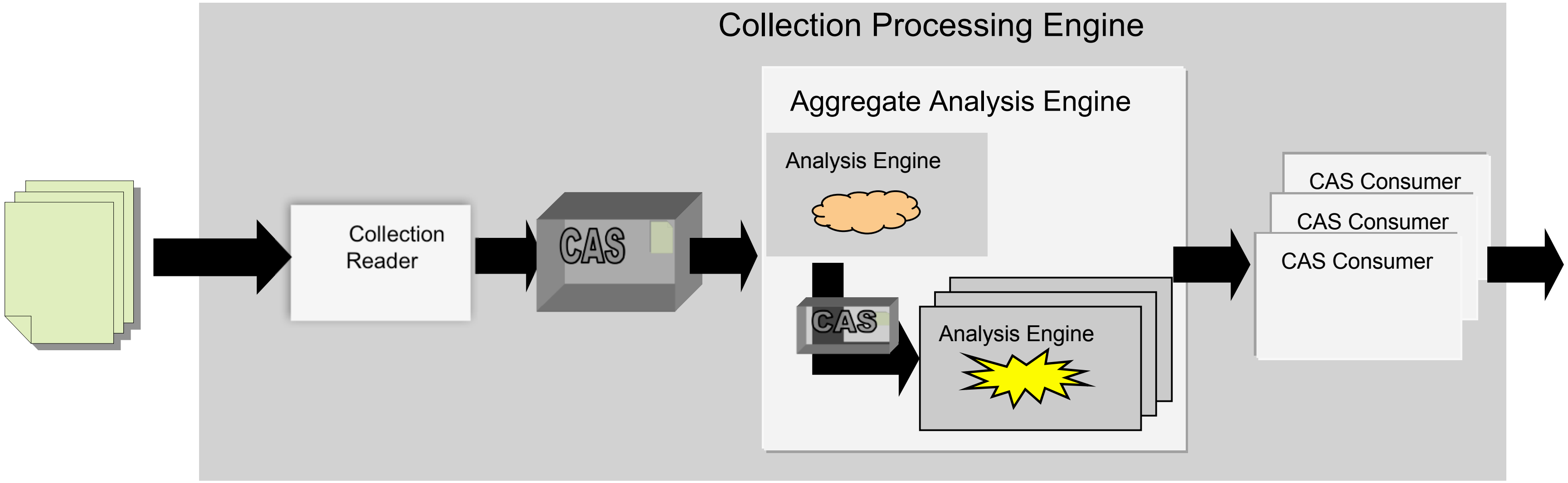
Combination and Integration



© 2006 IBM Corporation

UIMA Concept

- CAS (Common Annotation Structure)
  - Container that is passed between analysis engines
  - Allows creating and consuming annotations in a standard way



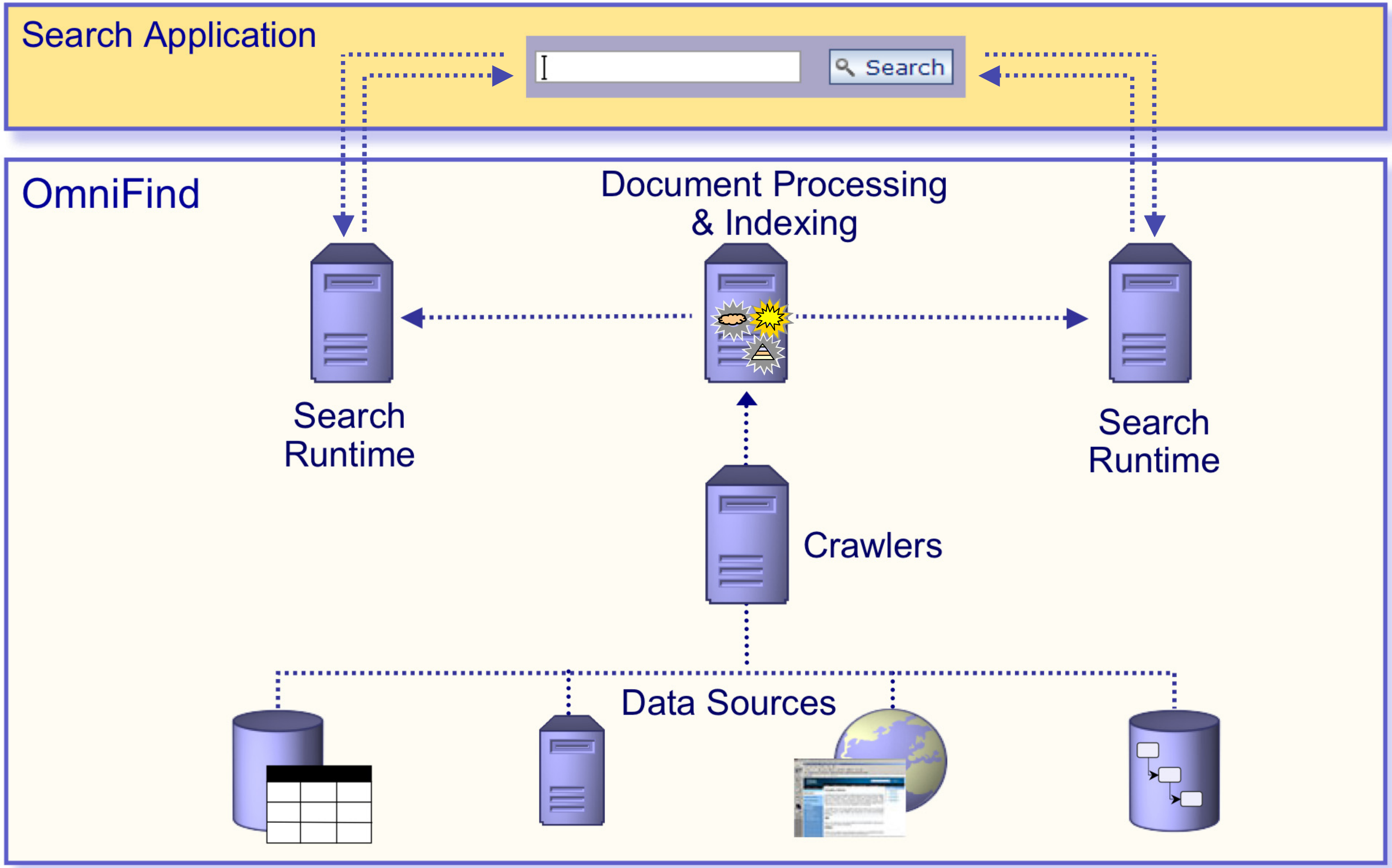
IBM Text Analytics: UIMA and LanguageWare Platform

- UIMA: Component Software Architecture
  - Specifies component interfaces, design patterns, data representations
- UIMA SDK
  - Java (C++ , Python)
  - Available from IBM AlphaWorks site (www.alphaworks.ibm.com)
- IBM LanguageWare: multi-lingual NLP technology
  - Provides tools for text analytics
  - Provides base UIMA text annotator

© 2006 IBM Corporation

UIMA Application in Information Retrieval: OmniFind

(WebSphere Information Integrator OmniFind Edition)



© 2006 IBM Corporation

Text Analytics in Life Sciences

- Unstructured data is major part of all available data
- Text Analytics
  - Annotation
    - Leveraging domain language
  - Relational information
    - Co-occurrence
    - Connections between nouns, through verbs
    - Sentence parsing
- Moving past Information Retrieval

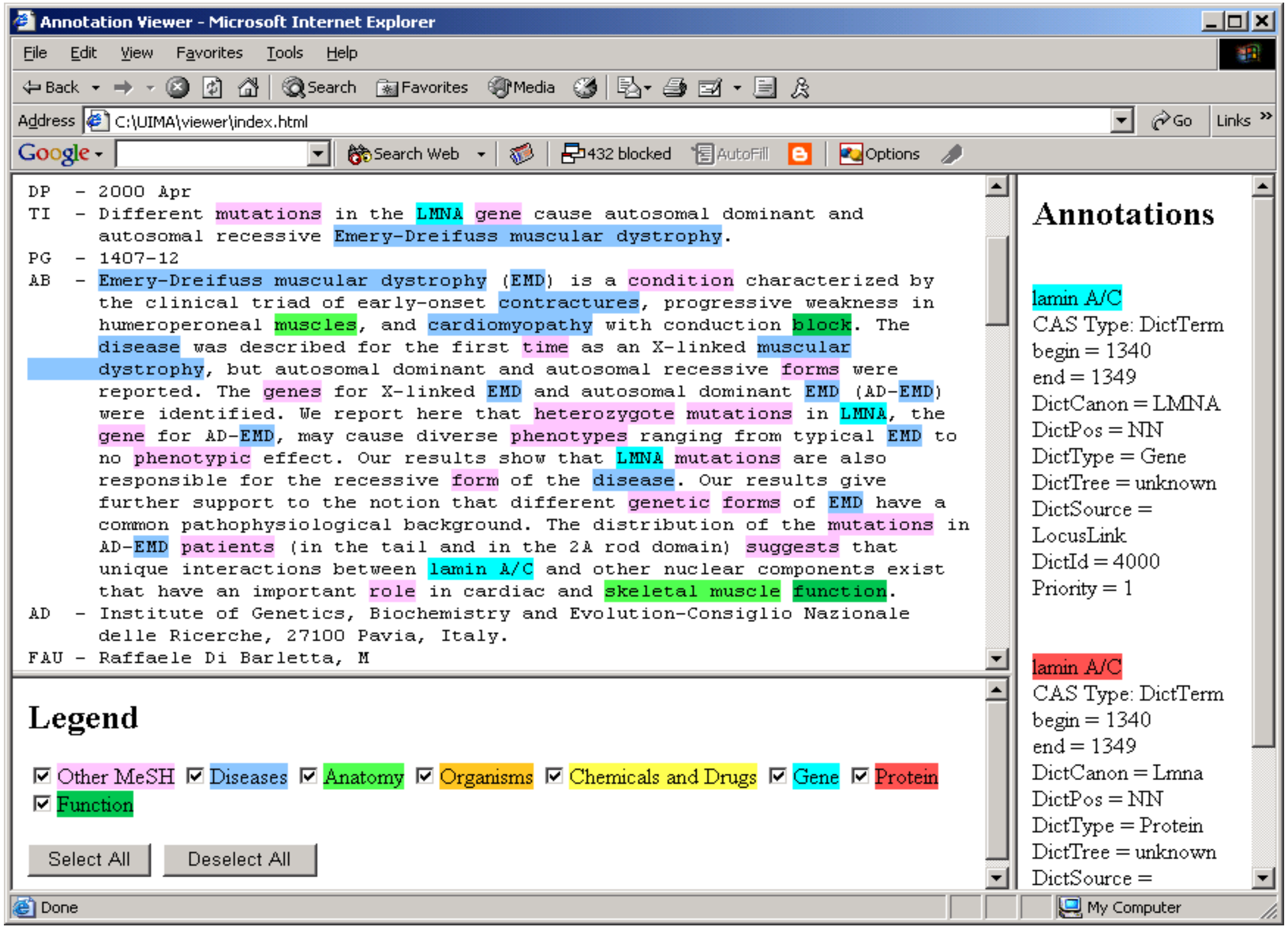
© 2006 IBM Corporation

Domain Knowledge

- Domain knowledge is very important
- Domain customization
  - Any information that describes the language of a domain can prove extremely valuable in customizing the analysis
    - Vocabulary, terminology, relationships, spelling variants, abbreviations, prefixes & suffixes, rules & regular expressions
  - Semantics
- Using existing data
- IBM LanguageWare Workbench
  - Semi-automatic generation of UIMA annotators
  - Java/Eclipse based

© 2006 IBM Corporation

UIMA and Domain Customization


Annotation Viewer - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address C:\UIMA\viewer\index.html
Go
Links
Google
Search Web
432 blocked
AutoFill
Options
Done
My Computer

```
DP  - 2000 Apr
TI  - Different mutations in the LMNA gene cause autosomal dominant and
      autosomal recessive Emery-Dreifuss muscular dystrophy.
PG  - 1407-12
AB  - Emery-Dreifuss muscular dystrophy (EMD) is a condition characterized by
      the clinical triad of early-onset contractures, progressive weakness in
      humeroperoneal muscles, and cardiomyopathy with conduction block. The
      disease was described for the first time as an X-linked muscular
      dystrophy, but autosomal dominant and autosomal recessive forms were
      reported. The genes for X-linked EMD and autosomal dominant EMD (AD-EMD)
      were identified. We report here that heterozygote mutations in LMNA, the
      gene for AD-EMD, may cause diverse phenotypes ranging from typical EMD to
      no phenotypic effect. Our results show that LMNA mutations are also
      responsible for the recessive form of the disease. Our results give
      further support to the notion that different genetic forms of EMD have a
      common pathophysiological background. The distribution of the mutations in
      AD-EMD patients (in the tail and in the 2A rod domain) suggests that
      unique interactions between lamin A/C and other nuclear components exist
      that have an important role in cardiac and skeletal muscle function.
AD  - Institute of Genetics, Biochemistry and Evolution-Consiglio Nazionale
      delle Ricerche, 27100 Pavia, Italy.
FAU - Raffaele Di Barletta, M
```

Annotations

lamin A/C
CAS Type: DictTerm
begin = 1340
end = 1349
DictCanon = LMNA
DictPos = NN
DictType = Gene
DictTree = unknown
DictSource = LocusLink
DictId = 4000
Priority = 1

lamin A/C
CAS Type: DictTerm
begin = 1340
end = 1349
DictCanon = Lmna
DictPos = NN
DictType = Protein
DictTree = unknown
DictSource =

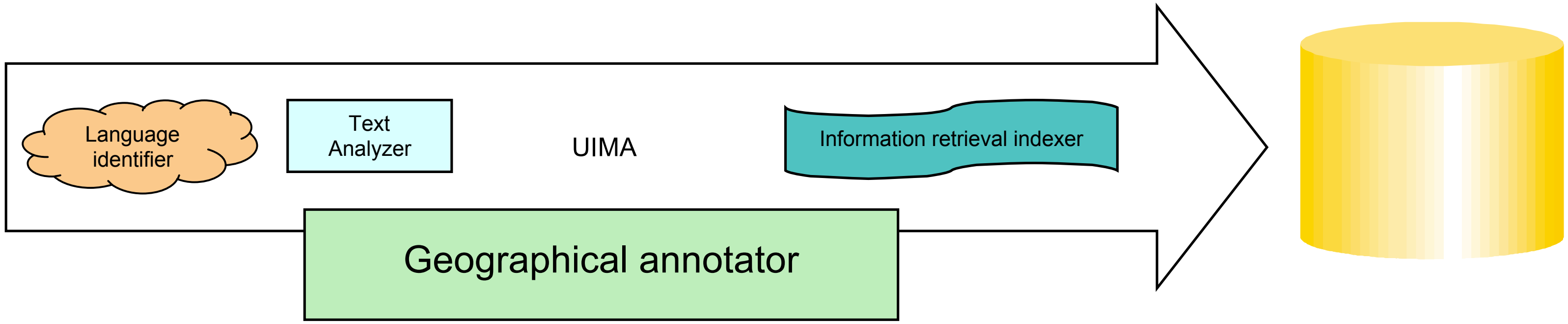
Legend

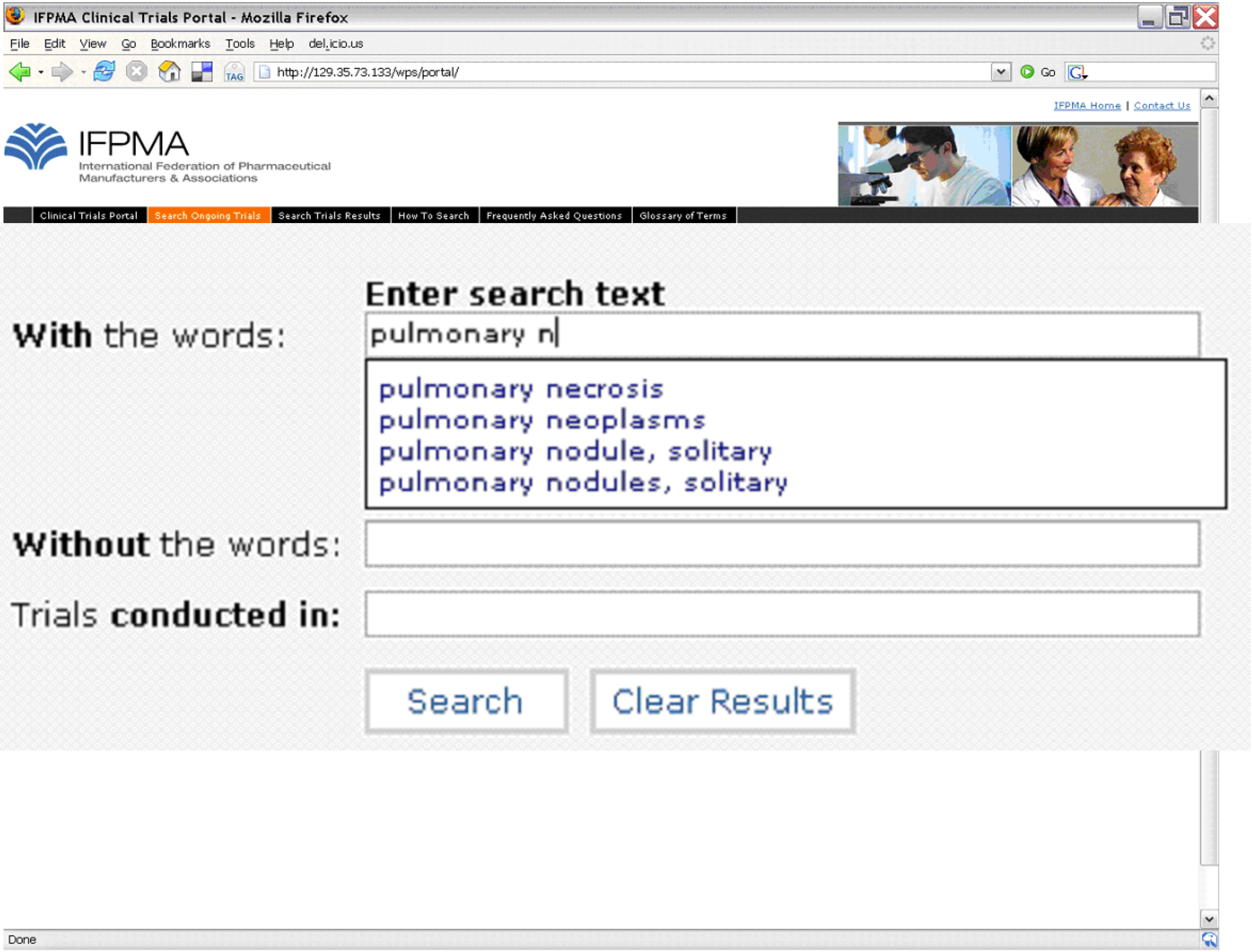
☑ Other MeSH ☑ Diseases ☑ Anatomy ☑ Organisms ☑ Chemicals and Drugs ☑ Gene ☑ Protein
☑ Function

Select All Deselect All

Examples of UIMA Applications

Clinical Trials Search





IFPMA Clinical Trials Portal - Mozilla Firefox

File Edit View Go Bookmarks Tools Help del.icio.us
http://129.35.73.133/wps/portal/
Go
IFPMA Home | Contact Us
IFPMA
International Federation of Pharmaceutical Manufacturers & Associations
Clinical Trials Portal
Search Ongoing Trials
Search Trials Results
How To Search
Frequently Asked Questions
Glossary of Terms

Enter search text

With the words: pulmonary n

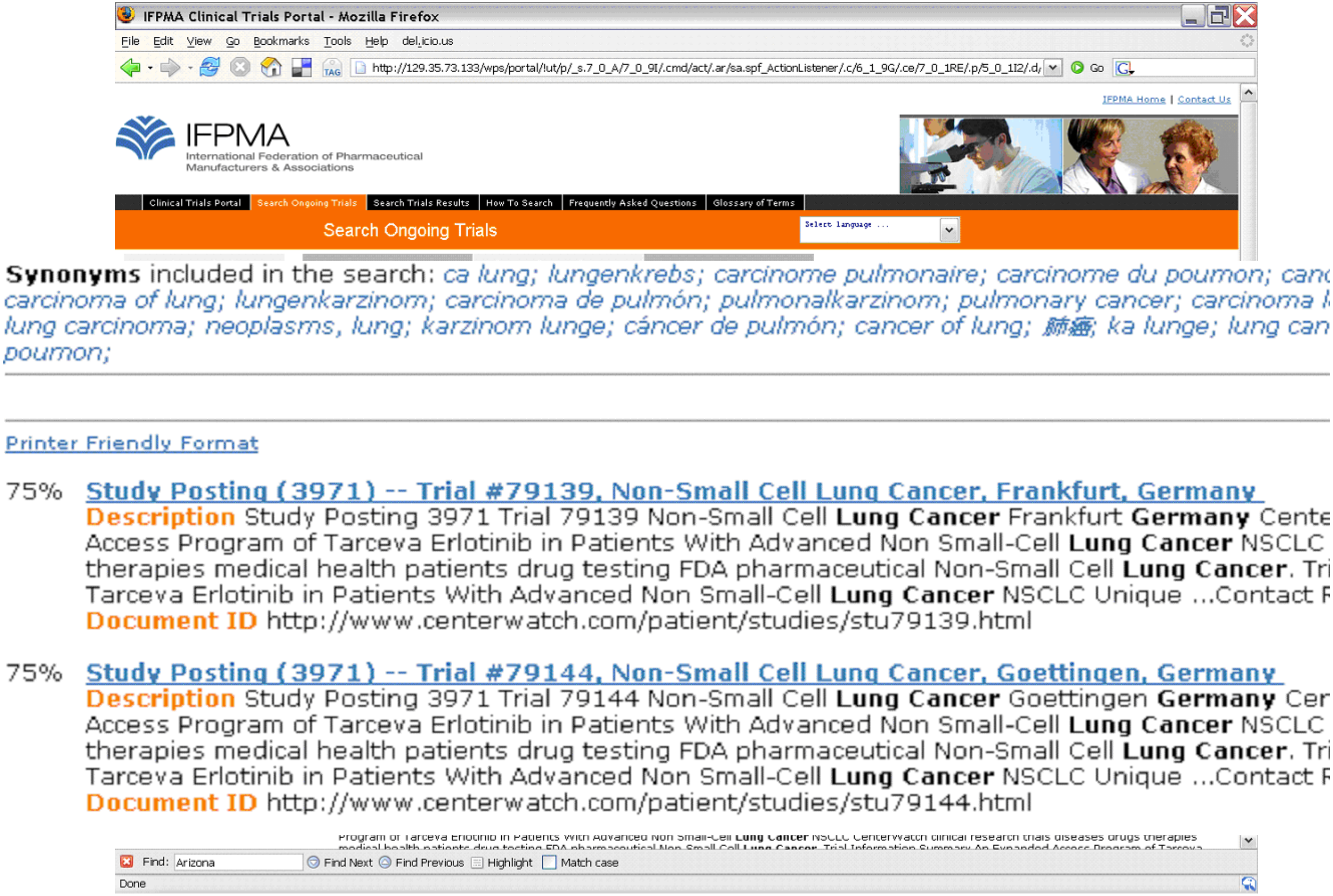
pulmonary necrosis
pulmonary neoplasms
pulmonary nodule, solitary
pulmonary nodules, solitary

Without the words:

Trials conducted in:

Search
Clear Results
Done

© 2006 IBM Corporation


IFPMA Clinical Trials Portal - Mozilla Firefox
File Edit View Go Bookmarks Tools Help del.icio.us
http://129.35.73.133/wps/portal/!ut/p/_s.7_0_A/7_0_9I/.cmd/act/.ar/sa.spf_ActionListener/.c/6_1_9G/.ce/7_0_1RE/.p/5_0_1I2/.d/
Go
IFPMA Home | Contact Us
IFPMA
International Federation of Pharmaceutical Manufacturers & Associations
Clinical Trials Portal
Search Ongoing Trials
Search Trials Results
How To Search
Frequently Asked Questions
Glossary of Terms
Search Ongoing Trials
Select language ...

Synonyms included in the search: *ca lung; lungenkrebs; carcinome pulmonaire; carcinome du poumon; canc carcinoma of lung; lungenkarzinom; carcinoma de pulmón; pulmonalkarzinom; pulmonary cancer; carcinoma l lung carcinoma; neoplasms, lung; karzinom lunge; cáncer de pulmón; cancer of lung; 肺癌; ka lunge; lung can poumon;*

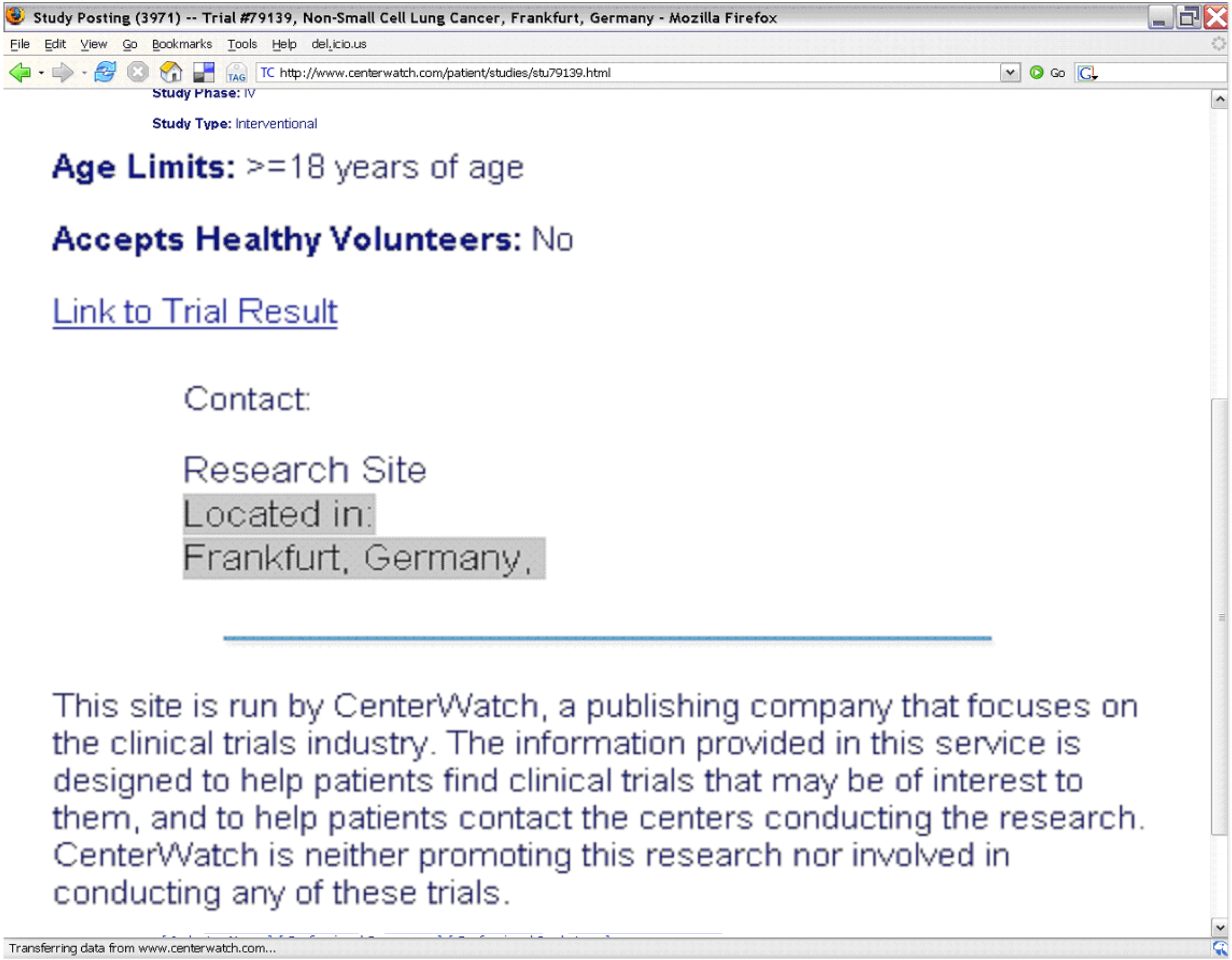
Printer Friendly Format

75% Study Posting (3971) -- Trial #79139, Non-Small Cell Lung Cancer, Frankfurt, Germany
Description Study Posting 3971 Trial 79139 Non-Small Cell **Lung Cancer** Frankfurt **Germany** Cente Access Program of Tarceva Erlotinib in Patients With Advanced Non Small-Cell **Lung Cancer** NSCLC therapies medical health patients drug testing FDA pharmaceutical Non-Small Cell **Lung Cancer**. Tri Tarceva Erlotinib in Patients With Advanced Non Small-Cell **Lung Cancer** NSCLC Unique ...Contact F
Document ID http://www.centerwatch.com/patient/studies/stu79139.html

75% Study Posting (3971) -- Trial #79144, Non-Small Cell Lung Cancer, Goettingen, Germany
Description Study Posting 3971 Trial 79144 Non-Small Cell **Lung Cancer** Goettingen **Germany** Cer Access Program of Tarceva Erlotinib in Patients With Advanced Non Small-Cell **Lung Cancer** NSCLC therapies medical health patients drug testing FDA pharmaceutical Non-Small Cell **Lung Cancer**. Tri Tarceva Erlotinib in Patients With Advanced Non Small-Cell **Lung Cancer** NSCLC Unique ...Contact F
Document ID http://www.centerwatch.com/patient/studies/stu79144.html

Program of Tarceva Erlotinib in Patients With Advanced Non Small-Cell **Lung Cancer** NSCLC CenterWatch clinical research trials diseases drugs therapies medical health patients drug testing FDA pharmaceutical Non Small Cell **Lung Cancer**. Trial Information Summary An Expanded Access Program of Tarceva

© 2006 IBM Corporation



Study Posting (3971) -- Trial #79139, Non-Small Cell Lung Cancer, Frankfurt, Germany - Mozilla Firefox
File Edit View Go Bookmarks Tools Help del.icio.us
http://www.centerwatch.com/patient/studies/stu79139.html
Go

Study Phase: IV

Study Type: Interventional

Age Limits: >=18 years of age

Accepts Healthy Volunteers: No

Link to Trial Result

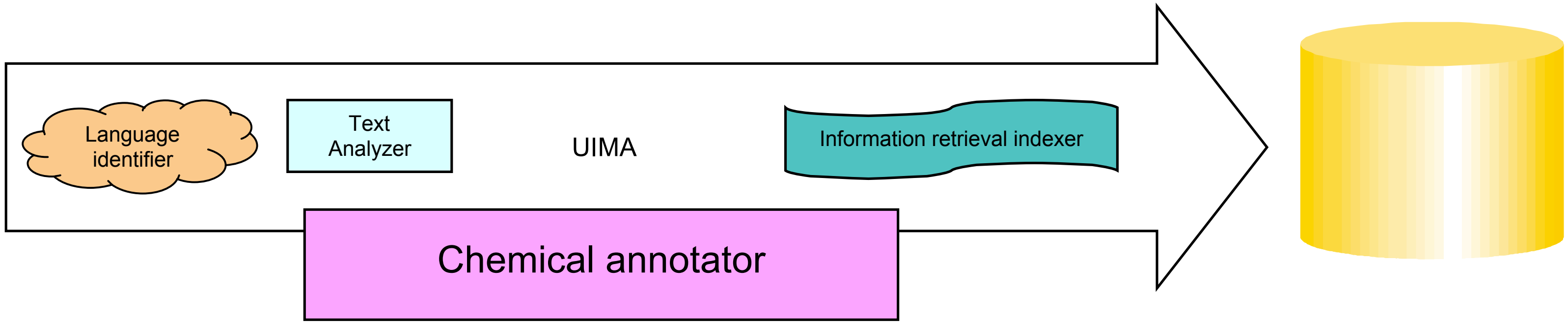
Contact:

Research Site
Located in:
Frankfurt, Germany,

This site is run by CenterWatch, a publishing company that focuses on the clinical trials industry. The information provided in this service is designed to help patients find clinical trials that may be of interest to them, and to help patients contact the centers conducting the research. CenterWatch is neither promoting this research nor involved in conducting any of these trials.

Examples of UIMA Applications

Chemical search



© 2006 IBM Corporation

The Problem: Search Patents for Chemical Structures

- Chemicals have multiple names, trivial and official

Names for Valium

ALBORAL, ALISEUM, ALUPRAM , AMIPROL ,ANSIOLIN , ANSIOLISINA , APAURIN, APOZEPAM, ASSIVAL , ATENSINE , ATILEN , BIALZEPAM , CALMOCITENE, CALMPOSE , CERCINE, CEREGULART, CONDITION, DAP, DIACEPAN, DIAPAM , DIAZEMULS , DIAZEPAM , DIAZETARD , DIENPAX, DIPAM , DIPEZONA, DOMALIUM , DUKSEN, DUXEN, E-PAM, ERIDAN, EVACALM, FAUSTAN, FREUDAL , FRUSTAN, GIHITAN, HORIZON, KIATRIUM, LA-III , LEMBROL, LEVIUM, LIBERETAS , METHYL DIAZEPINONE, MOROSAN , NEUROLYTRIL, NOAN, NSC-77518, PACITRAN, PARANTEN, PAXATE, PAXEL, PLIDAN, QUETINIL, QUIATRIL, QUIEVITA, RELAMINAL, RELANIUM, RELAX, RENBORIN, RO 5-2807, S.A.R.L., SAROMET, SEDAPAM, SEDIPAM, SEDUKSEN, SEDUXEN, SERENACK, SERENAMIN, SERENZIN, SETONIL, SIBAZON, SONACON, STESOLID, STESOLIN, TENSOPAM, TRANIMUL, TRANQDYN, TRANQUASE, TRANQUIRIT, TRANQUO-TABLINEN , UMBRIUM, UNISEDIL, USEMPAX AP, VALEO, VALITRAN, VALRELEASE, VATRAN, VELIUM, VIVAL, VIVOL, WY-3467

  - Straightforward text search does not work
  - Synonym expansion is insufficient
  - Searching by structure is required
- There is no explicit list of all organic chemicals

© 2006 IBM Corporation

Chemical Annotator

- Accurately recognizes organic chemicals in text
  - Uses a small number of common chemical morphemes (fragments)
    - Names, prefixes, suffixes, endings, etc.
  - Uses pattern rather than a dictionary approach
    - Rules for accepting and combining fragments
  - Not restricted to “known chemicals”

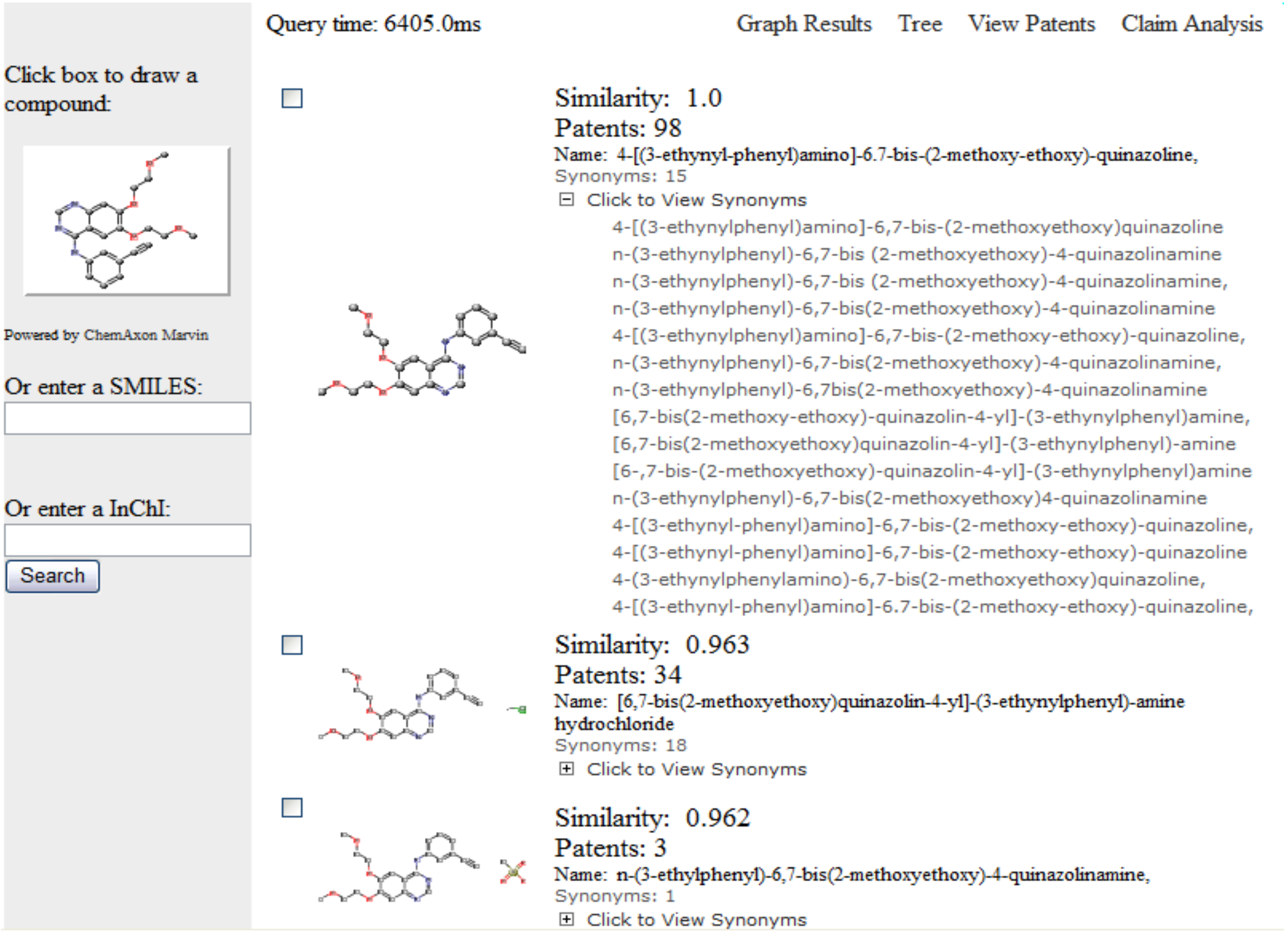
o-vinylbenzyl glycidyl ether

Chemical Search Application

- To create chemical search application
  - Taken >4M US patents
  - Extracted structured information
    - Title, authors, assignee, date, abstract, brief description, full description, claims…
  - Extracted chemical names
  - Converted chemical names into SMILES and InChI strings
  - Indexed

© 2006 IBM Corporation

Chemical Search Application



Click box to draw a compound:

Powered by ChemAxon Marvin

Or enter a SMILES:

Or enter a InChI:

Query time: 6405.0ms

□

Similarity: 1.0

Patents: 98

Name: 4-[(3-ethynyl-phenyl)amino]-6.7-bis-(2-methoxy-ethoxy)-quinazoline,

Synonyms: 15

⊟ Click to View Synonyms

4-[(3-ethynylphenyl)amino]-6,7-bis-(2-methoxyethoxy)quinazoline
n-(3-ethynylphenyl)-6,7-bis (2-methoxyethoxy)-4-quinazolinamine
n-(3-ethynylphenyl)-6,7-bis (2-methoxyethoxy)-4-quinazolinamine,
n-(3-ethynylphenyl)-6,7-bis(2-methoxyethoxy)-4-quinazolinamine
4-[(3-ethynylphenyl)amino]-6,7-bis-(2-methoxy-ethoxy)-quinazoline,
n-(3-ethynylphenyl)-6,7-bis(2-methoxyethoxy)-4-quinazolinamine,
n-(3-ethynylphenyl)-6,7bis(2-methoxyethoxy)-4-quinazolinamine
[6,7-bis(2-methoxy-ethoxy)-quinazolin-4-yl]-(3-ethynylphenyl)amine,
[6,7-bis(2-methoxyethoxy)quinazolin-4-yl]-(3-ethynylphenyl)-amine
[6-,7-bis-(2-methoxyethoxy)-quinazolin-4-yl]-(3-ethynylphenyl)amine
n-(3-ethynylphenyl)-6,7-bis(2-methoxyethoxy)4-quinazolinamine
4-[(3-ethynyl-phenyl)amino]-6,7-bis-(2-methoxy-ethoxy)-quinazoline,
4-[(3-ethynyl-phenyl)amino]-6,7-bis-(2-methoxy-ethoxy)-quinazoline
4-(3-ethynylphenylamino)-6,7-bis(2-methoxyethoxy)quinazoline,
4-[(3-ethynyl-phenyl)amino]-6.7-bis-(2-methoxy-ethoxy)-quinazoline,

□

Similarity: 0.963

Patents: 34

Name: [6,7-bis(2-methoxyethoxy)quinazolin-4-yl]-(3-ethynylphenyl)-amine hydrochloride

Synonyms: 18

⊞ Click to View Synonyms

□

Similarity: 0.962

Patents: 3

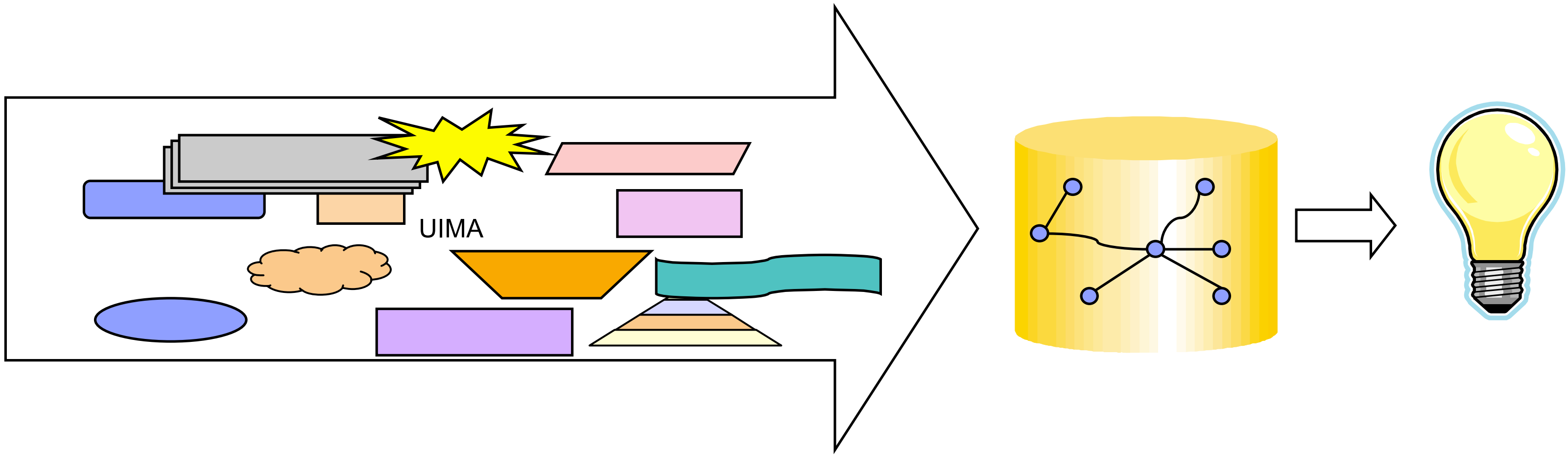
Name: n-(3-ethylphenyl)-6,7-bis(2-methoxyethoxy)-4-quinazolinamine,

Synonyms: 1

⊞ Click to View Synonyms

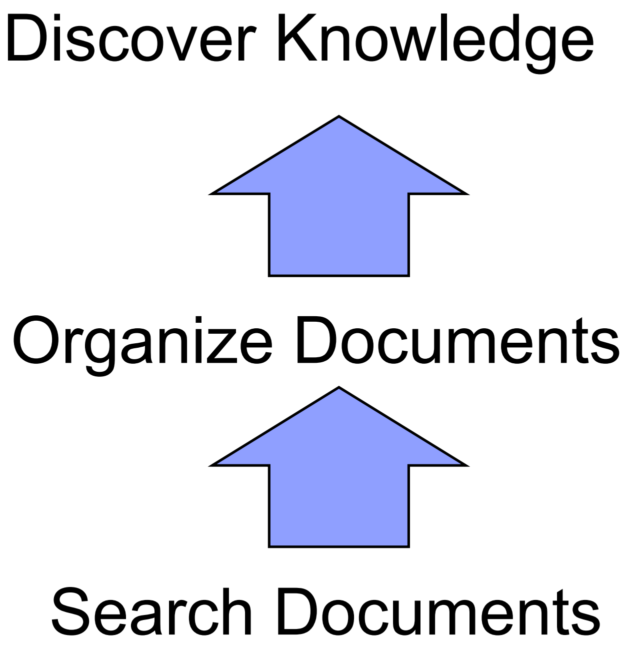
Examples of UIMA Applications

Knowledge discovery



© 2006 IBM Corporation

Text Mining and Knowledge Discovery



Discover Knowledge

Organize Documents

Search Documents

Text Mining and Knowledge Discovery

- TAKMI, BIW

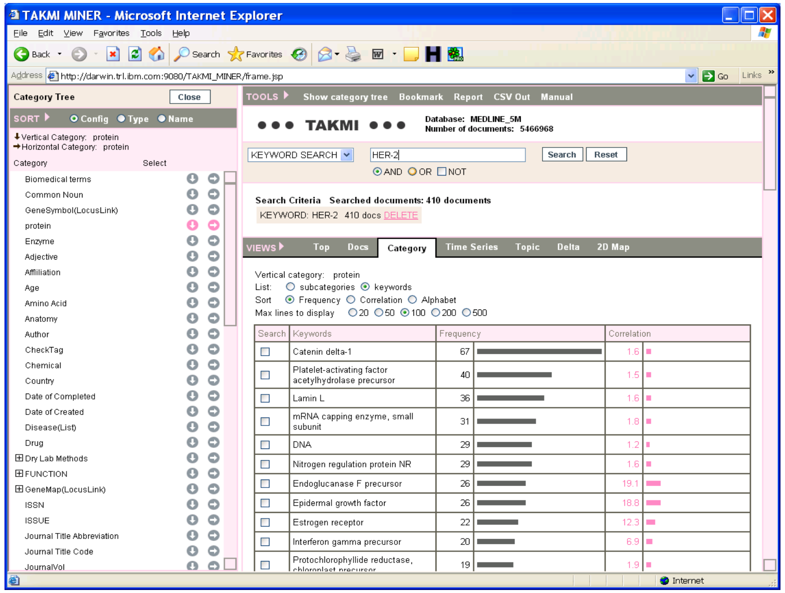
Stack of UIMA components

- NLP
- Domain-specific lexical and semantic data

Collection-level analysis

Interactive (selection-level) analysis

- Relationship mining
- Trend analysis
- Visualization



Address: http://darwin.trl.ibm.com:9080/TAKMI_MINER/frame.jsp

SORT: ● Config ○ Type ○ Name

Vertical Category: protein
Horizontal Category: protein

Category Select

Biomedical terms
Common Noun
GeneSymbol(LocusLink)
protein
Enzyme
Adjective
Affiliation
Age
Amino Acid
Anatomy
Author
CheckTag
Chemical
Country
Date of Completed
Date of Created
Disease(List)
Drug
⊞ Dry Lab Methods
⊞ FUNCTION
⊞ GeneMap(LocusLink)
ISSN
ISSUE
Journal Title Abbreviation
Journal Title Code
JournalVol

●●● TAKMI ●●●

Database: MEDLINE_5M
Number of documents: 5466968

KEYWORD SEARCH: HER-2 Search Reset
● AND ○ OR □ NOT

Search Criteria Searched documents: 410 documents
KEYWORD: HER-2 410 docs DELETE

Vertical category: protein
List: ○ subcategories ● keywords
Sort ● Frequency ○ Correlation ○ Alphabet
Max lines to display ○ 20 ○ 50 ● 100 ○ 200 ○ 500

Search	Keywords	Frequency	Correlation
□	Catenin delta-1	67	1.6
□	Platelet-activating factor acetylhydrolase precursor	40	1.5
□	Lamin L	36	1.6
□	mRNA capping enzyme, small subunit	31	1.8
□	DNA	29	1.2
□	Nitrogen regulation protein NR	29	1.6
□	Endoglucanase F precursor	26	19.1
□	Epidermal growth factor	26	18.8
□	Estrogen receptor	22	12.3
□	Interferon gamma precursor	20	6.9
□	Protochlorophyllide reductase, chloroplast precursor	19	1.9

Conclusions

- UIMA framework facilitates integration and interoperation of text analysis modules
  - Essential for large scale development
    - Allows clear requirements specification
    - Easy to debug and verify
  - Useful for research cooperation
    - Simplifies integration
    - Simplifies deployment
  - Creates development community
- UIMA simplifies development and deployment of Knowledge Discovery applications
  - Industrial KD applications benefit most from componentized process

© 2006 IBM Corporation

- The presented works are by
  - IBM T.J. Watson Research Center
  - IBM Almaden Research Center
  - IBM Tokyo Research Laboratory
  - IBM Software Group
- Clinical Trials Portal developed in cooperation with IFPMA

© 2006 IBM Corporation