

Text Mining Patents with I2E and Intellixir Software: Differences and Synergies

Piotr Masiakowski

PIUG 2012 Biotechnology Meeting

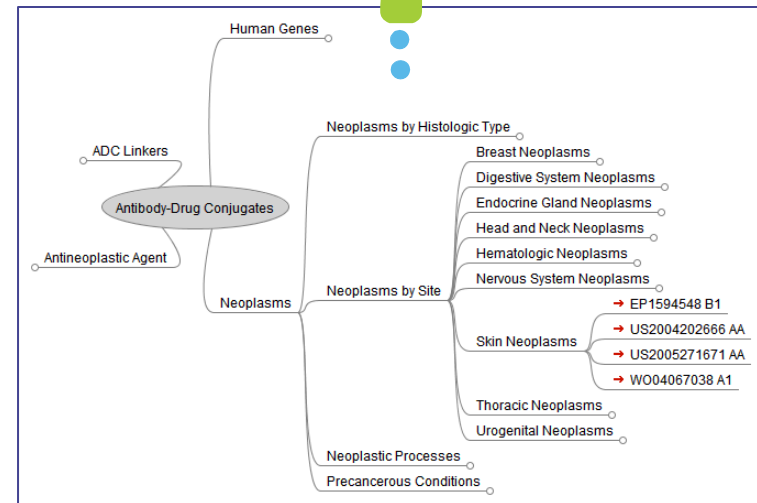
AGENDA

- Why do we need text mining in patent analysis?
- Brief introduction to two commercial platforms
 - **Intellixir System**
 - **Linguamatics I2E**
- Side-by-side comparison
- Combining the strengths of I2E and Intellixir
 - **Patmob prototype: a hub for patent resources**
 - **Information processing workflow**
 - **An example: Antibody-Drug Conjugate patent landscape**
- Conclusions

WHY TEXT MINING?

“As machines revolutionized the industrial age in the 19th century, text mining/visualization tools have the potential to revolutionize the information age and drive innovative solutions in the 21st century by providing facile mechanisms for tackling the ever-growing deluge of patent information.”

Yang Y, Akers L, Klose T, Barcelon Yang C (2008) Text mining and visualization tools – Impressions of emerging capabilities. World Patent Inform. 30:280-93



IN THEIR OWN WORDS



INTELLIXIR System is a hosted and secure web application -a SaaS solution (Software As A Service)- allowing its users to analyze patent and non-patent literature exported from commercial or private databases. **Statistical measures are graphically represented through dynamic and interactive web pages.**

[\[http://intellixir.com/ \]](http://intellixir.com/)

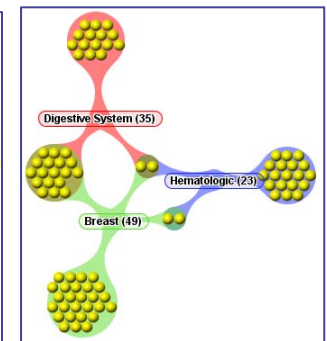
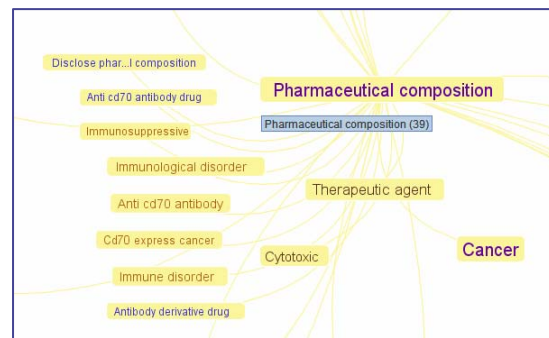
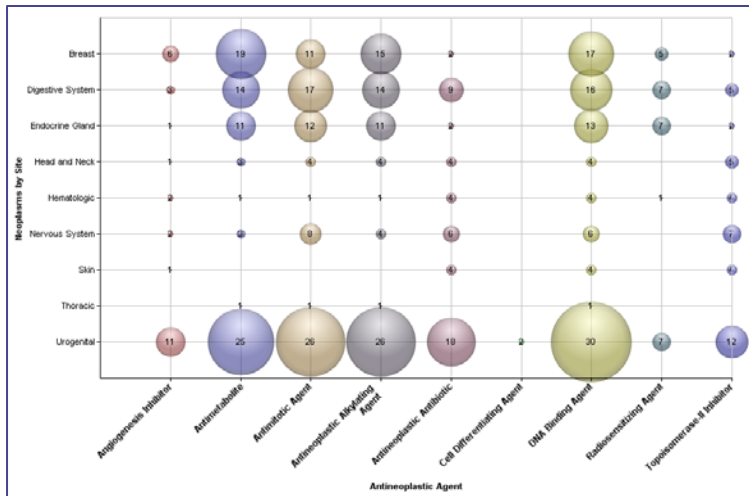
Linguamatics I2E provides **agile, high performance enterprise text mining software, enabling rapid discovery of new intelligence from text.** I2E lets you answer questions and discover knowledge that would be intractable or impossible to find using conventional methods.

[\[http://linguamatics.com/ \]](http://linguamatics.com/)

INTELLIXIR GRAPHICAL USER INTERFACE

The screenshot shows the INTELLIXIR web application interface. The top navigation bar includes 'Home', 'Search', 'Global', 'Concepts', 'Players', 'Data', 'Experts', 'My Docs', 'Shortcuts', and 'Help'. A sidebar on the left contains 'Database information' (last updated 28/12/2011 08:09:23, 268 Patents, 205 Affiliations, 27 Concepts, 658) and 'New Import' options. The main content area displays 'Main concepts (Full screen)' with a list of terms and their associated counts, such as 'Antibody drug' (19), 'Cancer' (15), and 'Linker' (25).

The screenshot displays the 'Statistics on thesaurus' page. It features a hierarchical tree diagram on the left showing the structure of the thesaurus, with 'Antibody-Drug Conjugates' highlighted in red. On the right, there is a search bar and a list of filters including 'Geography' (Originating countries, Protected countries, Extended countries), 'Dates', 'Organizations', 'Authors', 'Custom Fields', 'Thesaurus', 'Patents', 'IPCs', and 'ECLA'.



I2E GRAPHICAL USER INTERFACE

The screenshot displays the I2E graphical user interface for a Sanofi Pasteur trial. The main window is titled "citrate - AB-formulation - Linguamatics I2E Pro". It features a menu bar (File, Edit, View, Index, Tools, Window, Help) and a toolbar with various icons. The interface is divided into several panes:

- Left Pane:** A tree view showing the trial structure, including "Sanofi Pasteur Trial", "Standard", "Custom", and "Fairview Patents". The "AB-formulation" node is selected.
- Query Editor:** A central pane showing a query for "CITRATE" with various filters and operators. The query is structured as follows:
 - Operator: ≤1w (unordered)
 - Operator: ≤0w (ordered)
 - Field: [2]
 - Field: [2] Unit
 - Operator: Numerics + / ...
 - Operator: Numerics + / - x
 - Operator: Amount Volume
 - Operator: Concentration
- Right Pane:** A hierarchical tree of "Human Genes" with the following structure:
 - Human Genes
 - Cell adhesion molecule
 - Cell junction protein
 - Gap junction
 - GJA1
 - GJA10
 - GJA3
 - GJA4

- Bottom Pane:** A table of search results with the following columns: [PT], PatBase Family, PatBase Family [citrate], [sucrose], [polysorbate80], Doc, and Hit. The table contains four rows of results.

INTELLIXIR MAIN FEATURES

- In Intellixir, user works within single “topic” (database) of imported patents and/or journal articles
- The program automatically extracts main concepts and displays them as a cloud
- Users can define Boolean queries to group similar documents
 - Thesaurus
 - Custom Fields
- Large selection of interactive graphs available to analyze and display information
- Multiple levels of users
 - From 0 – Access to a simplified interface (Easylixir) only, used to publish a selection of graphs on the Web
 - To 5 – Administrator can create new databases and add users

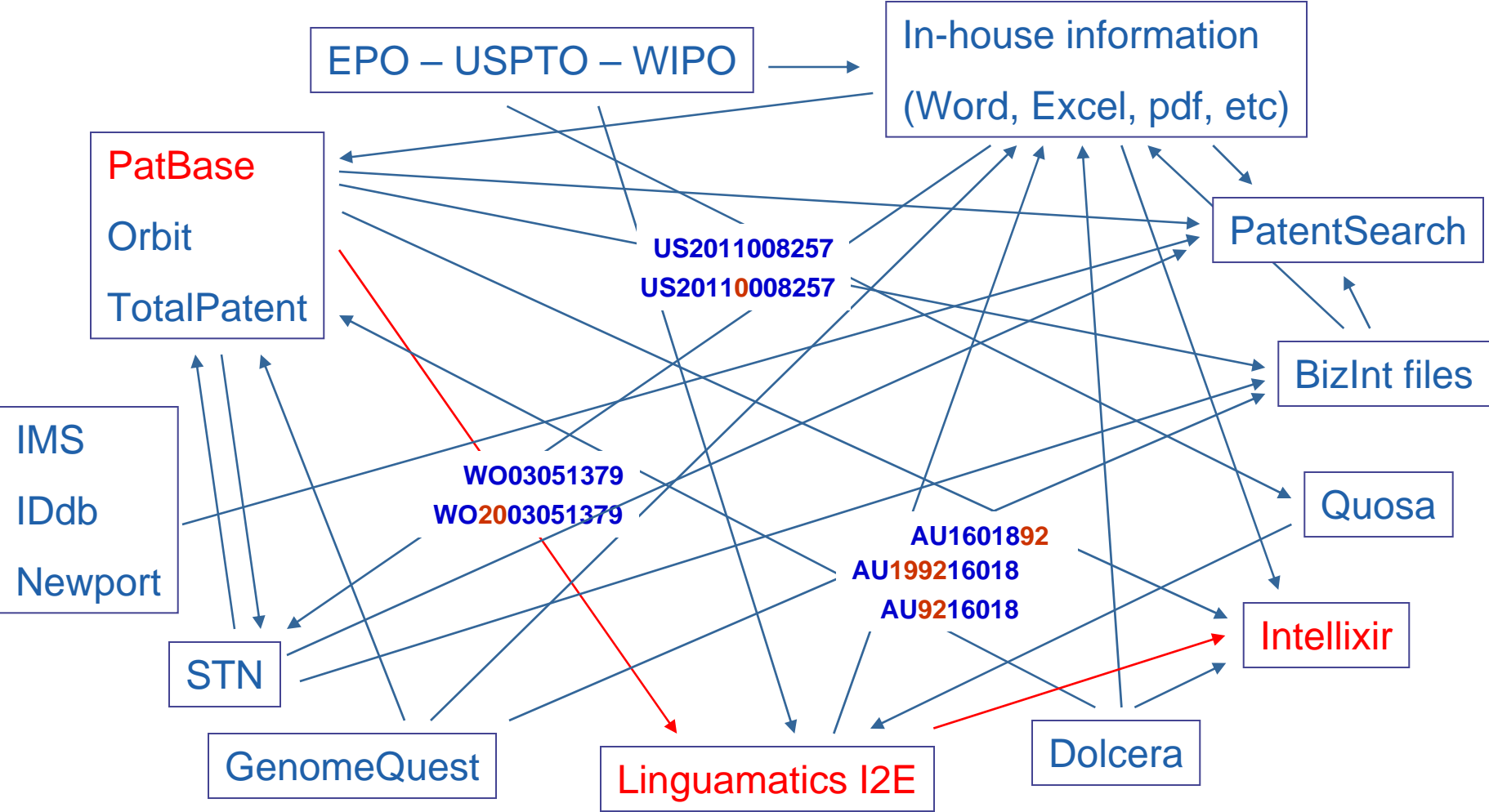
I2E MAIN FEATURES

- I2E Query Editor and Output Editor offer powerful tools to extract information from automatically indexed text
 - **Classes**
 - Hierarchies of predefined concepts (i.e. Entrez Genes class will find any gene synonym)
 - Default linguistic classes can be used as linguistic wildcards for things and actions
 - **Multi Queries combine queries using various operators**
- Administrators create indexes and manage the system
- Users have a choice of 3 interfaces for creating and executing queries
 - I2E Pro
 - I2E Express
 - Smart Queries

INTELLIXIR VS. I2E SUMMARY

	Intellixir	Linguamatics I2E
Deployment	Web application	I2E OnDemand or I2E Enterprise
Content	Imports	Imports; Medline, Fairview Patents
Indexing	Keywords only	Ontologies included
Text analysis	Statistical analysis, Boolean queries	NLP-based queries using vocabularies and linguistic wildcards
Result display	A variety of graphs to visualize the information	HTML table of hits; (export to Excel, Cytoscape, FreeMind)
Main strength	Visualization tools	Powerful querying
Main weakness	Limited querying	Limited editing and displaying results

COMPLEXITY OF PATENT RESOURCES



PATMOB CORE MAIN FEATURES

- A Java desktop application designed to facilitate rapid development of plug-ins for converting patent data formats
- Advantages:
 - **Simplicity: runs on Windows, Mac OS or Linux – no special installation.**
 - **Designed specifically to work with patent information**
 - Interactive editing of patent collections.
 - Automatic conversions of patent number formats
- The core infrastructure provides:
 - **Relational database (JavaDB)**
 - Permanently store and search patent collections and notes
 - **Internet access protocols**
 - Instantaneously review most current information on the Web
 - **XML support**
 - **Tree editor**

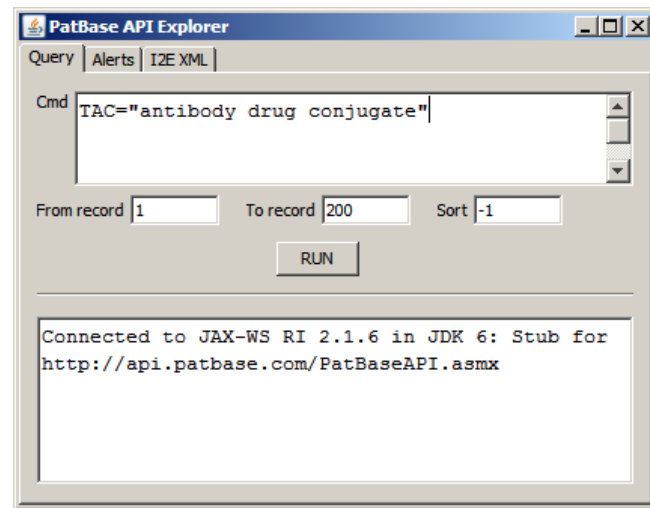
PATMOB TREE BRANCH EDITOR

Real-time bibliographic data, with the full publication cycle of the patent application. Fetched from the Open Patent Services (OPS) at the European Patent Office.

- Any branch of the database tree can be opened in the editor
- Plug-ins can create new trees and open them in the editor
- The information can be interactively edited
- The edited information can be saved in the database or written to an XML or text file

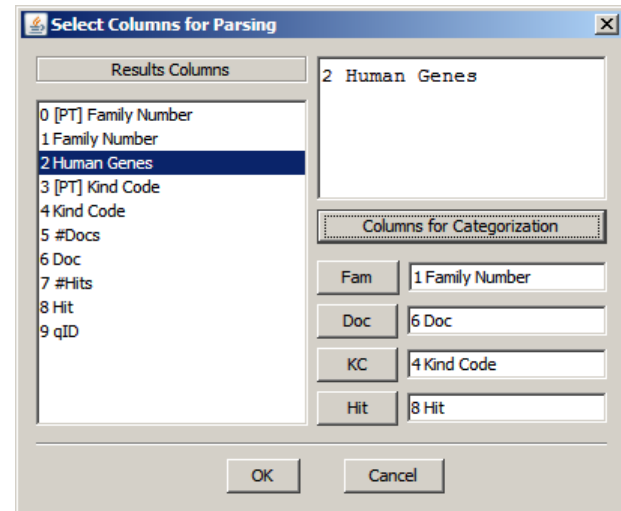
PATMOB PLUG-INS DEVELOPMENT

- Plug-ins are small programs with their own GUI, rapidly developed using the core API
- Plug-ins have full access to all core resources
- For example, PatBase API Access plug-in
 - Connects to PatBase API
 - Allows querying
 - Downloads selected patent families
 - Converts PatBase XML to a format suitable for indexing in I2E



IMPORTING I2E RESULTS TO PATMOB

- Import and index the corpus in I2E
- Run the search in I2E with a query using the relevant classes or macros
- Export I2E search results in XML format
- Open the I2E results XML files from the I2E Tools plug-in, defining the terms to be used for categorization



EDITING I2E RESULTS IN PATMOB

The screenshot displays the 'Tree Branch Editor: I2E Results' application on the left and a web browser on the right. The browser shows the URL https://www.i2eondemand.com/i2e/sanofi/server/tmp/antigene_f_00026569_120124_165300-1_-69035976 - Micros... and displays patent claim text for an anti-Robo4 antibody.

Tree Branch Editor: I2E Results

Tree Node

- PatBase 43791879
- PatBase 43791880
- PatBase 43791906
- PatBase 43791909
- PatBase 43791930
- PatBase 43959836
 - US2008247951 AA
 - US2011059013 AA
 - US2011059112 AA
 - US7834154 BB
 - ROBO4
- WO08100805 A2
- PatBase 44208713
 - US2008050310 AA
 - US2011142859 AA
 - CD22
- WO07140371 A2
- PatBase 44248510
- PatBase 44967998
- PatBase 45965016
- PatBase 47027875
- PatBase 47098517
- PatBase 47898993
- PatBase 48505298
- PatBase 48505312

INFO Properties Documents

US7834154 Biblio:

PN: US2008247951 A1

PD: 20081009

TI: ANTI-ROBO4 ANTIBODIES AND

PA: GENENTECH, INC;

AB: The invention provides an antibody comprising a light chain variable domain comprising an amino acid sequence selected from the group consisting of: SEQ ID NOS:72-97 as shown in FIGS. 1A and 2A.

PN: US7834154 B2

PD: 20101116

TI: ANTI-ROBO4 ANTIBODIES AND

PA: GENENTECH, INC;

AB: The invention provides an antibody comprising a heavy chain variable domain comprising an amino acid sequence selected from the group consisting of: SEQ ID NOS:140-165 as shown in FIGS. 1B and 2B.

Save to DB Write to XML

Web Browser: https://www.i2eondemand.com/i2e/sanofi/server/tmp/antigene_f_00026569_120124_165300-1_-69035976 - Micros...

We claim: <p>1. An isolated anti-Robo4 antibody comprising: (i) an HVR-L1 comprising sequence A1-A11 wherein A1-A11 is RASQDVSTAVA (SEQ ID NO:1) (ii) an HVR-L2 comprising sequence B1-B7 wherein B1-B7 is SASFLYS (SEQ ID NO:2) (iii) an HVR-L3 comprising sequence C1-C9 wherein C1-C9 is QQSYTTPPT (SEQ ID NO:3) (iv) an HVR-H1 comprising sequence D1-D10 wherein D1-D10 is GFTINGYYIH (SEQ ID NO:17) (v) an HVR-H2 comprising sequence E1-E18 wherein E1-E18 is GFYYPAGGDDYADSVKVG (SEQ ID NO:18); and (vi) an HVR-H3 comprising sequence F1-F17 wherein F1-F17 is ARLIGNKFGWSSYGMDY (SEQ ID NO:19).

2. An isolated anti-Robo4 antibody comprising a light chain variable domain comprising an amino acid sequence selected from the group consisting of: SEQ ID NOS:72-97 as shown in FIGS. 1A and 2A.

3. An isolated anti-Robo4 antibody comprising a heavy chain variable domain comprising an amino acid sequence selected from the group consisting of: SEQ ID NOS:140-165 as shown in FIGS. 1B and 2B.

4. An isolated anti-Robo4 antibody comprising (i) a light chain variable domain comprising an HVR-L1 comprising SEQ ID NO:1 an HVR-L2 comprising SEQ ID NO:2 and an HVR-L3 comprising QQSRSDHPT (SEQ ID NO:20); and (ii) a heavy chain variable domain comprising an HVR-H1 comprising SEQ ID NO:17 an HVR-H2 comprising SEQ ID NO:18 and an HVR-H3 comprising SEQ ID NO:19.

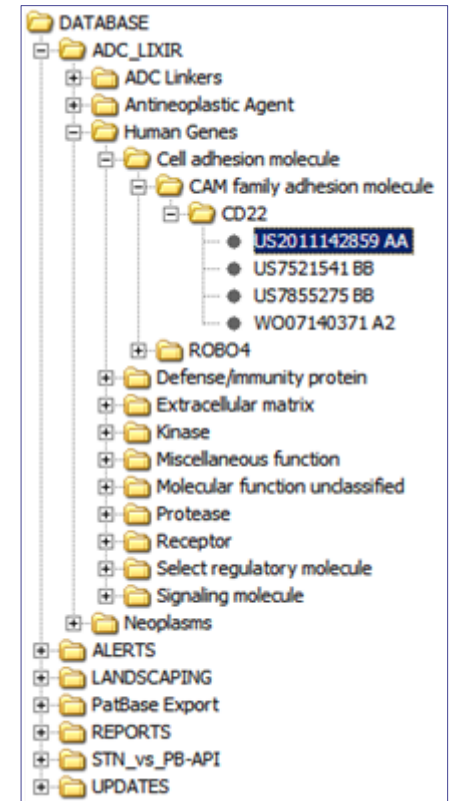
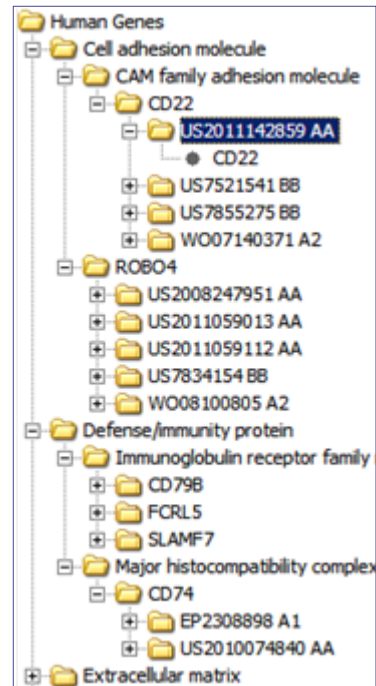
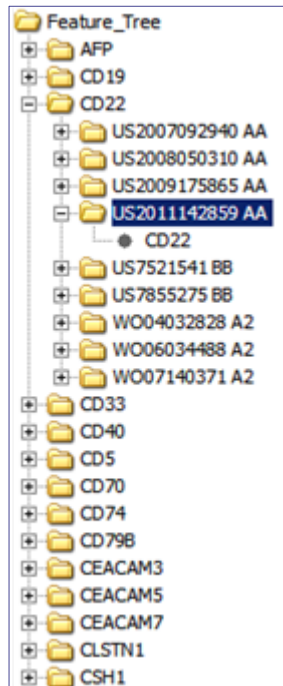
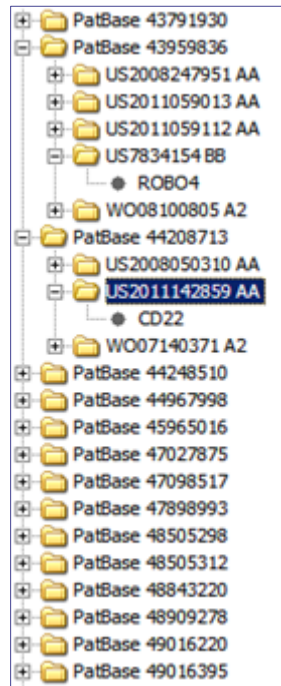
5. The antibody of any one of claims 1-3 and 4 wherein the antibody is humanized.

6. The antibody of any one of claims 1-3 and 4 wherein the antibody is selected from the group consisting of a Fab a Fab' and a (Fab').sub.2.

7. The antibody of any one of claims 1-3 and 4 further comprising a cytotoxic agent.

8. The antibody of claim 7 wherein the cytotoxic agent is selected from the group consisting of: N.sup.2'-deacetyl-N-.sup.2'(3-mercapto-1-oxopropyl)-maytansine (DM1) monomethyl auristatin E (MMAE) monomethyl auristatin F (MMAF) and combinations thereof.

FORMATTING I2E RESULTS IN PATMOB



In Patmob Tree Branch Editor:

Manually edit results in the context of patent families



From the menu, re-organize the tree by I2E search hits



From the menu, organize the tree by importing ontology structure

Save the final (or any intermediate) tree in the database

PATMOB AUTOMATICALLY GENERATES FILES FOR UPLOAD TO INTELLIXIR

```

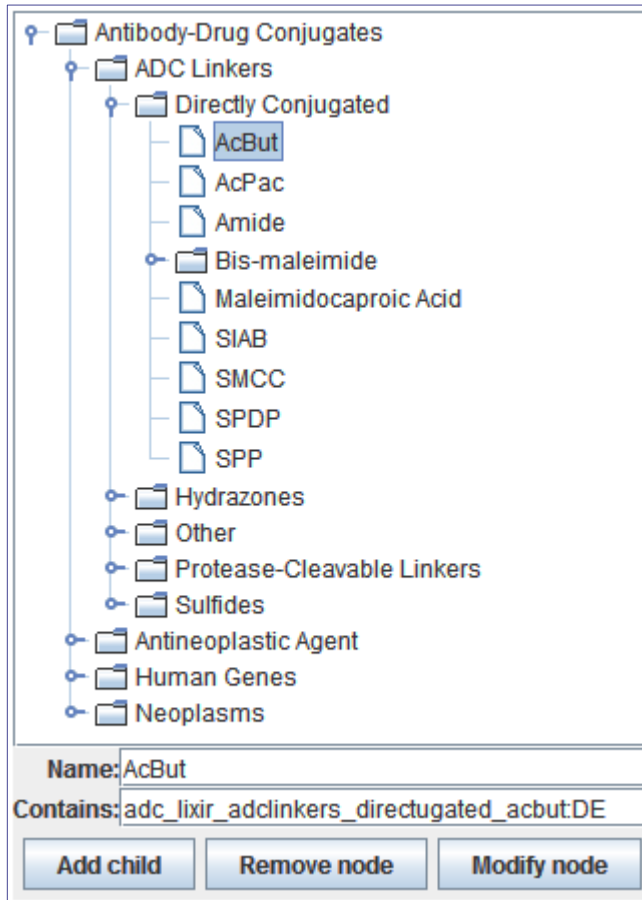
<DOCUMENT>
- <DOCUMENT>
  <DT>Brevet</DT>
  <TI>Humanized anti-5T4 antibodies and anti-5T4/calicheamicin
  conjugates</TI>
  <PY>2006</PY>
  - <AF-LIST>
    <AF>WYETH;</AF>
  </AF-LIST>
  - <PN-LIST>
    <PN>US2006088522 A1</PN>
  </PN-LIST>
  - <DE-LIST>
    <DE>adc_lixir_adclinkers_directugated_acbut</DE>
    <DE>adc_lixir_adclinkers_directugated_acpac</DE>
    <DE>adc_lixir_adclinkers_directugated_amide</DE>
  </DE-LIST>
  <AB>Chimeric and humanized anti-5T4 antibodies and
  antibody/drug conjugates and methods for preparing and using
  the same.</AB>
  
```

- XML file for upload of documents to Intellixir
 - A unique string loaded in the descriptors (DE) field

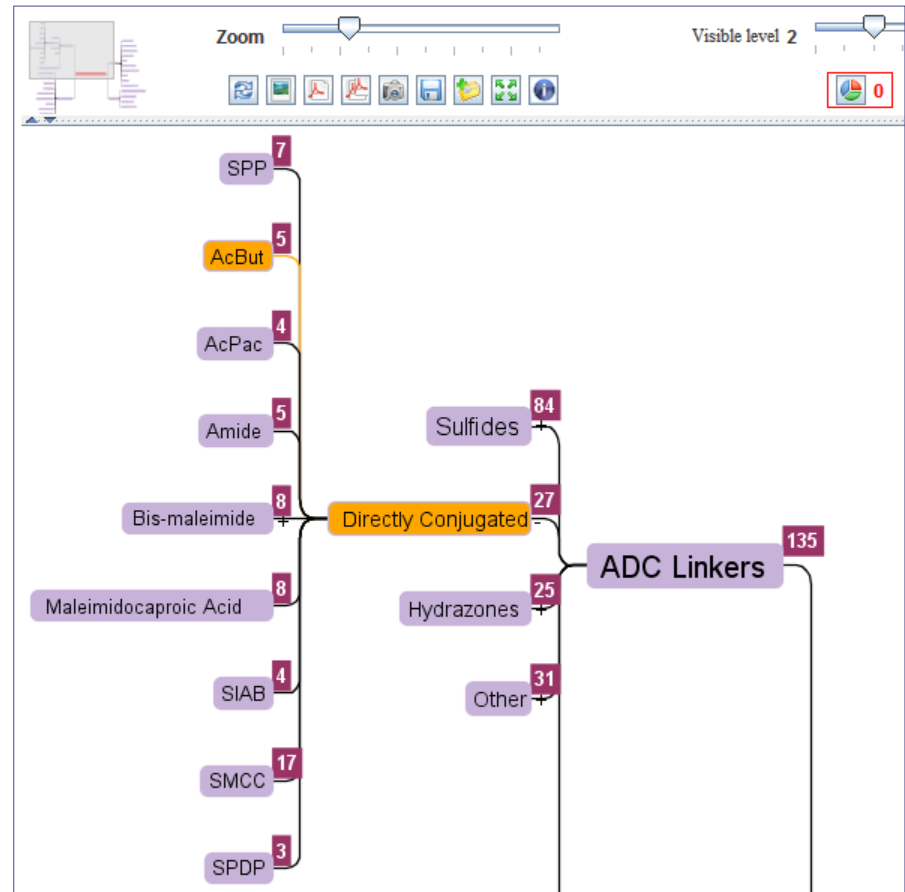
Antibody-Cl	adc_lixir:DE				
	ADC Linke	adc_lixir_adclinkers:DE			
		Directly Co	adc_lixir_adclinkers_directugated:DE		
		AcBut	adc_lixir_adclinkers_directugated_acbut:DE		
		AcPac	adc_lixir_adclinkers_directugated_acpac:DE		
		Amide	adc_lixir_adclinkers_directugated_amide:DE		
		Bis-malein	adc_lixir_adclinkers_directugated_bismaleimide:DE		
		BM(PEO)3	adc_lixir_adclinkers_directugated_bism		
		BM(PEO)4	adc_lixir_adclinkers_directugated_bism		
		BMB	adc_lixir_adclinkers_directugated_bism		
		BMDB	adc_lixir_adclinkers_directugated_bism		

- TXT file defining Intellixir thesaurus
 - Node name
 - Boolean query to search for a string in the descriptors field

INTELLIXIR THESAURUS



Editor



Interactive Graph

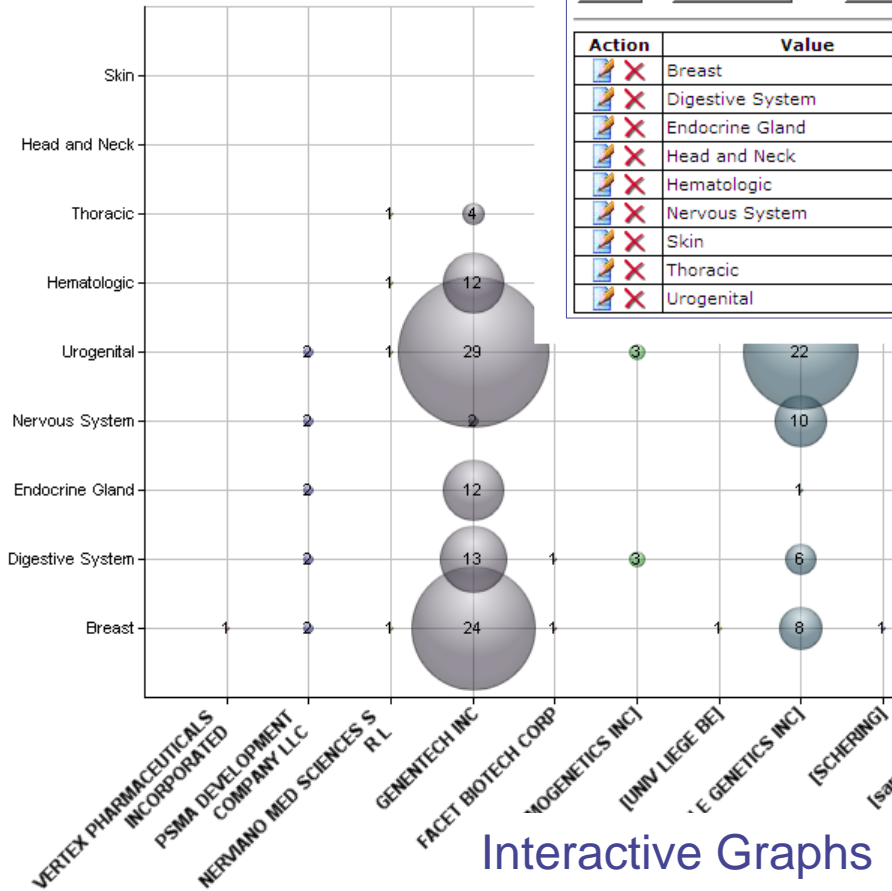
INTELLIXIR CUSTOM FIELDS

Default search into: Descriptors

Back Add values Fill in Delete Export values to Excel Clear

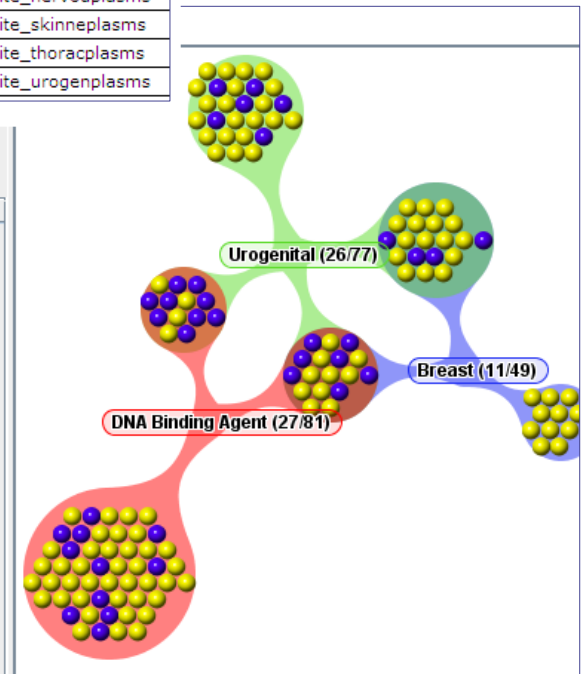
Action	Value	Contains
	Breast	adc_lixir_neoplasms_neoplabsite_breastplasm
	Digestive System	adc_lixir_neoplasms_neoplabsite_digestplasm
	Endocrine Gland	adc_lixir_neoplasms_neoplabsite_endocrplasm
	Head and Neck	adc_lixir_neoplasms_neoplabsite_headanplasm
	Hematologic	adc_lixir_neoplasms_neoplabsite_hematoplasm
	Nervous System	adc_lixir_neoplasms_neoplabsite_nervouplasm
	Skin	adc_lixir_neoplasms_neoplabsite_skinneplasm
	Thoracic	adc_lixir_neoplasms_neoplabsite_thoracplasm
	Urogenital	adc_lixir_neoplasms_neoplabsite_urogenplasm

Editor



Interactive Graphs

- Root 183
 - Antineoplastic Ag... 116
 - DNA Binding Ag... 81
 - Antineoplastic Ant... 71
 - Antimetabolite 46
 - Antimitotic Agent 44
 - Antineoplastic Alk... 43
 - Topoisomerase-II I... 43
 - Angiogenesis Inhi... 23
 - Radiosensitizing ... 13
 - Cell Differentiating ... 3
 - Asparaginase 2
 - Heat Shock Protein ... 1
 - Urokinase-Derived ... 0
 - Neoplasms by Site 112
 - Urogenital 77
 - Breast 49
 - Digestive System 35
 - Hematologic 23



CONCLUSIONS

- Text mining and data visualization software is crucial for efficient patent analysis – especially large ‘patent landscaping’ projects
- Different programs have different strengths
 - Linguamatics I2E offers agile, high performance text mining using Natural Language Processing-based queries
 - Intellixir System allows data visualization through dynamic and interactive web pages
- For optimal results, combine the strengths of multiple programs!

ACKNOWLEDGMENTS

- Special thanks to Sunny Wang, Jim Calkins, May Peng, Sunita Saluja, Mohani Sukhdeo and my other colleagues at Sanofi for their interest, encouragement and advice
- Grateful acknowledgment is made for permission to use the screenshots of the following software programs:
 - I2E, copyright © 2002-2011 by Linguamatics
 - Intellixir System, copyright © 2002 by Intellixir
 - PatBase, copyright © 2003-2012 RWS and Minesoft Ltd.
 - PatMOB, copyright © 2006-2012 by GeneMOB

Thank you