



Better annotations for Text Mining: using a knowledge server

Search Engine Meeting, Boston, April 23th, 2007

**Pascal Coupet, CTO
TEMIS, Philadelphia**



- Entities extraction is popular
 - Facet Navigation
 - Augmented documents
 - Smart links
 - Text Analytics
- Users satisfaction is strongly related to the extraction quality
 - Better perception of misses than errors
 - Text Analytics require a strong extraction quality
 - Normalization is needed
 - Disambiguation is important for some domains
- Users have the domain knowledge
 - They want to be in control
 - It must be easy for them to correct/update the system
- Migrations and evolutions should be easy

Text Analytics

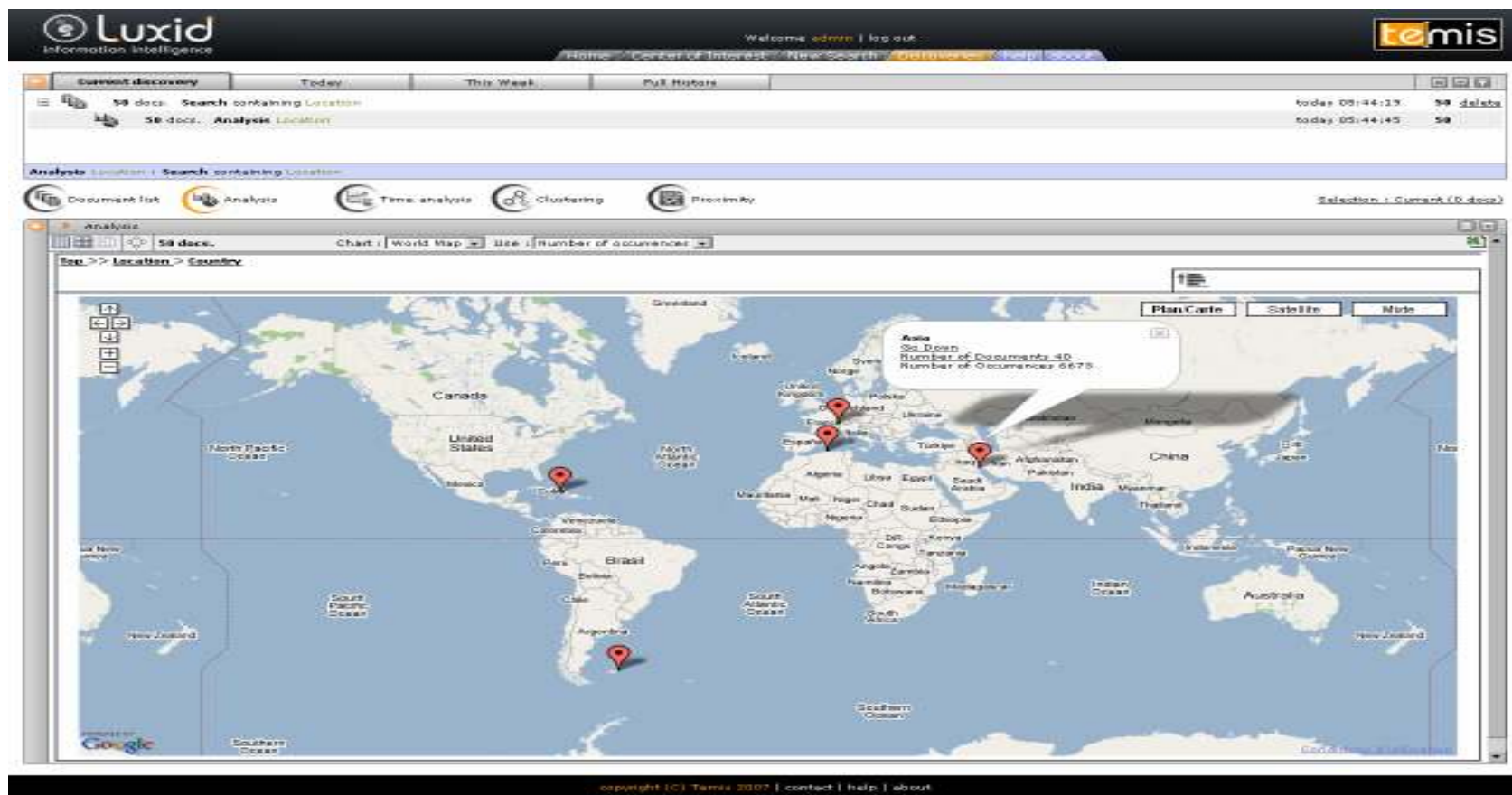
The screenshot displays the Luxid Information Intelligence interface with the following components:

- Header:** Luxid Information Intelligence logo, navigation links (Home, Center of Interest, New search, Discoveries, Knowledge browser, Help, About), and the TEMIS logo.
- Left Panel:** A network graph titled "Sub graph ADIPOQ1 into Expand ADIPOQ1 into Relation GeneExpression having Object contain ADIPOQ". The graph shows relationships between entities like "hep-1", "hepatocytes", "PPARG", "adiponectin receptor", "E-cad", "mouse", and "GK". A table on the right lists entity types: Cells And Tissues, Chemical Entity, Tools, Find, and Proof.
- Top Right Panel:** Search results for "atherosclerosis" with 1047 documents. It lists related topics such as "[Osteopontin and atherosclerosis]", "[Atherosclerosis: butter on the arteries?]", and "Prediabetes & atherosclerosis: what's the connection?".
- Bottom Left Panel:** A time analysis bar chart for "Publishing Data for atherosclerosis" from Jan 03 to Jan 06. The chart shows a peak in publications around late 2005.
- Bottom Right Panel:** A horizontal bar chart titled "Labelled Protein Gene" showing the number of occurrences for various genes. The data is as follows:

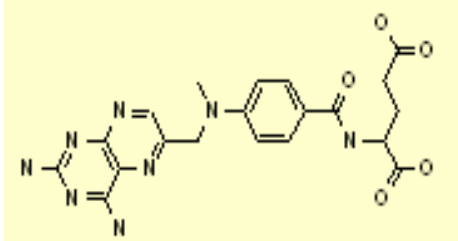
Gene	Number of occurrences
INS	717
ADIPOQ	250
IL6	145
CRP	129
TNF	94
PPARG	92
RETN	88
CCL2	59
PON1	53
AGT	51
AHSG	49
STN	48
APOE	42

Example: Locations

- Highly ambiguous: Cambridge, San Jose ...



Example: Chemical structures

Chemical name	structure
<p>Amethopterin 4-amino-N10-methylpteroylglutamic acid MEXATE METHOTREXATE 2-{4-[(2,4-Diamino-pteridin-6-ylmethyl)-methyl-amino]-benzoylamino}-pentanedioic acid</p>	 <p>The image shows the chemical structure of Methotrexate, which consists of a 2,4-diaminopteridine ring system connected via a methylene bridge to a 4-methylamino-2-benzoylamino group, which is further linked to a glutamic acid moiety. The structure is highlighted in a yellow box.</p>

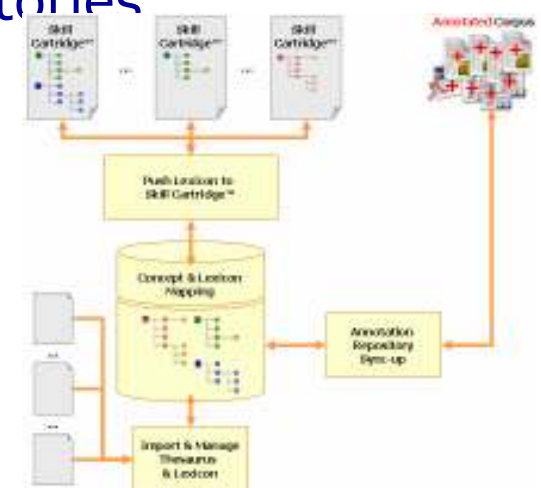
- **REGSTRING**
 - Unique fingerprint
 - Allow struct. search
 - Visualization

Example: People names

- We want to identify the network of people and companies related to Mr. Burns, head of Roche
 - Mr. Burns can be cited in different ways (Mr. Burns, William Burns, William H. Burns, Bill Burns ...)
 - there is different Mr. Burns, in particular Mr. William Burns, US Assistant Secretary of State.

Cluster /			
Id	Descriptors	Documents	Sub clusters
/1	US ; secretary ; state ; assistant ; assistant secretary ; affair ; Jordan ; assistant secretary of state ; east ; eastern ;	56	10
/2	Palestinian ; Israeli ; envoy ; Arafat ; U.S. ; Israel ; meet ; Palestinian ; Palestinian leader ;	54	10
/3	Roche ; Roche ; patient ; product ; pharmaceutical division ; development ; division ; head ; disease ; therapy ;	46	10
/4	Roche ; Roche ; sale ; drug ; head ; growth ; division ; pharmaceutical ; continue ;	46	9
/5	chairman ; president ; executive ; CEO ; Roche ; vice ; board ; group ; vice president ;	37	9
/6	Burns ; daughter ; born ; St. Petersburg ; William ; husband ; grandchildren ; late ; home ; assist ;	28	
/7	service ; funeral ; Johnson Rice ; Johnson Rice ; Minrad ; technology ; company ; analyst ; cemetery ; go ;	25	
/8	manufacture ; south ; Fuzeon ; Queensferry ; reduce ; capacity ; Queensferry ; price ; T-20 ; head of Roche pharmaceutical ;	24	
/9	school ; high ; John ; teacher ; celebrate ; river ; fall ; family ; national ; director ;	14	
/10	hospital ; ounce ; boy ; girl ; victim ; medical ; name ; new ; unit ; system ;	13	

- Goals & business benefits
 - Rapidly enrich Annotators w/o development skills
 - Bring agility & flexibility to annotation processes
 - Provide a unified knowledge representation (entities, attributes, relationships, categories)
- How
 - Extend extraction models with lexicons & thesauri
 - Normalize entities (synonyms, display forms, ...)
 - Enrich entities with attributes (zip code, id, ...)
 - Receive feedback from annotation repositories
 - Push thesaurus changes to repositories
- Key features
 - Import & manage lexicons & thesaurus
 - Integrate & propagate end-user changes
 - Test extractions
 - Automatically update Annotators



Knowledge Manager[®] Preview



Thesaurus hierarchy

Entity definition

The screenshot displays the Knowledge Manager interface with several key components:

- Thesaurus Hierarchy:** A tree view on the left showing a hierarchy of entities, with 'Company (200)' selected.
- Entity Definition Table:** A table listing various entities with columns for 'Secteur Coface', 'Secteur Nace', 'Sigle', 'Ville', and 'language'. A red circle highlights the 'Sigle' column.
- Extraction Testing:** A window titled 'Extraction' showing a 'Text-extraction' tab. It contains an 'Input Text Sample' with the text 'ACTEBIS is a company.' and an 'Extraction result' section showing the extracted text: 'ACTEBIS is a company.' under 'Extraction from: Basic' and 'Extraction from: Entities/Relations'.
- Cartridge List:** A window titled 'Cartridges' showing a list of cartridges, including 'Entity/Company', 'Entity/Product', 'Noun/Phrase', and 'Verbs'. A red circle highlights the 'Entity/Company' cartridge.

List of Skill Cartridge[™] to update

Annotation testing & check

- Strategies are used for normalization and disambiguation
- Attached to entity types
- Access to information from
 - Knowledge Base
 - Annotated corpus
 - Current document
- Ex: ProperName Strategy
 - `<ProperNameStrategy
parameterId="p1"
macro="/UserDefined/Company"
amacro="/UserDefinedAmbiguous/Company">`
 - `<ProperNameParameters id="p1">`
 - `<Set language="English" ignoreDash="true">`
 - `<MultiToken caseOnFirst="preserveFirst"
caseOnNext="ignore"/>`
 - `<SingleToken case="preserve"
caseOnAmbiguous="preserveFirst"
ambiguousIfLessThan="3"
useXeLDA="yes"/>`
 - `</Set>`
 - `<Set language="French" ...>`
 - `</Set>`
 - `</ProperNameParameters/>`

- Goals & business benefits
 - Let **business users** adapt Luxid® to their needs
 - Constantly increase system quality
- How
 - Let **business users** report annotation suggestions
 - Propagate changes to documents & repository
- Key features
 - Merge 2 entities [Novadel = Novadel Pharma Inc]
 - Rename entities [Novadel Corp & Novadel Inc become Novadel]
 - Remove entity [BUT is not a company (although a french one)]
 - Add entity [XyyyZ is a protein]

Dynamic Mapping Preview



Adjust thesauri

Adjust document

The screenshot shows the Luxid E.T. interface in a Windows Internet Explorer browser window. The address bar shows the URL: <http://localhost:8080/Luxid42/document.jsp?idDoc=549&idLuxlet=8>. The search bar contains the text "Search BASF in section all sections and their concepts" with a page indicator "2/18".

On the left side, there is a "Quick Highlight" section with the text "basf" and a "Highlight" section showing a tree view of entities. The tree is expanded to show "Company (9)" with sub-items: "Abbott (1)", "BASF (2)", "G D Searle (1)", "Novadel (1)", "Novadel Pharma Inc (1)", "Pfizer (1)", and "Pharmacia (2)". Below the tree are buttons for "Merge", "Delete", and "Move".

The main content area displays a document titled "Personnel; Specialty pharmaceutical company names vice president and general counsel". The document text includes:

Personnel; Specialty pharmaceutical company names vice president and general counsel

Obesity, Fitness & Wellness Week - 2004-06-26

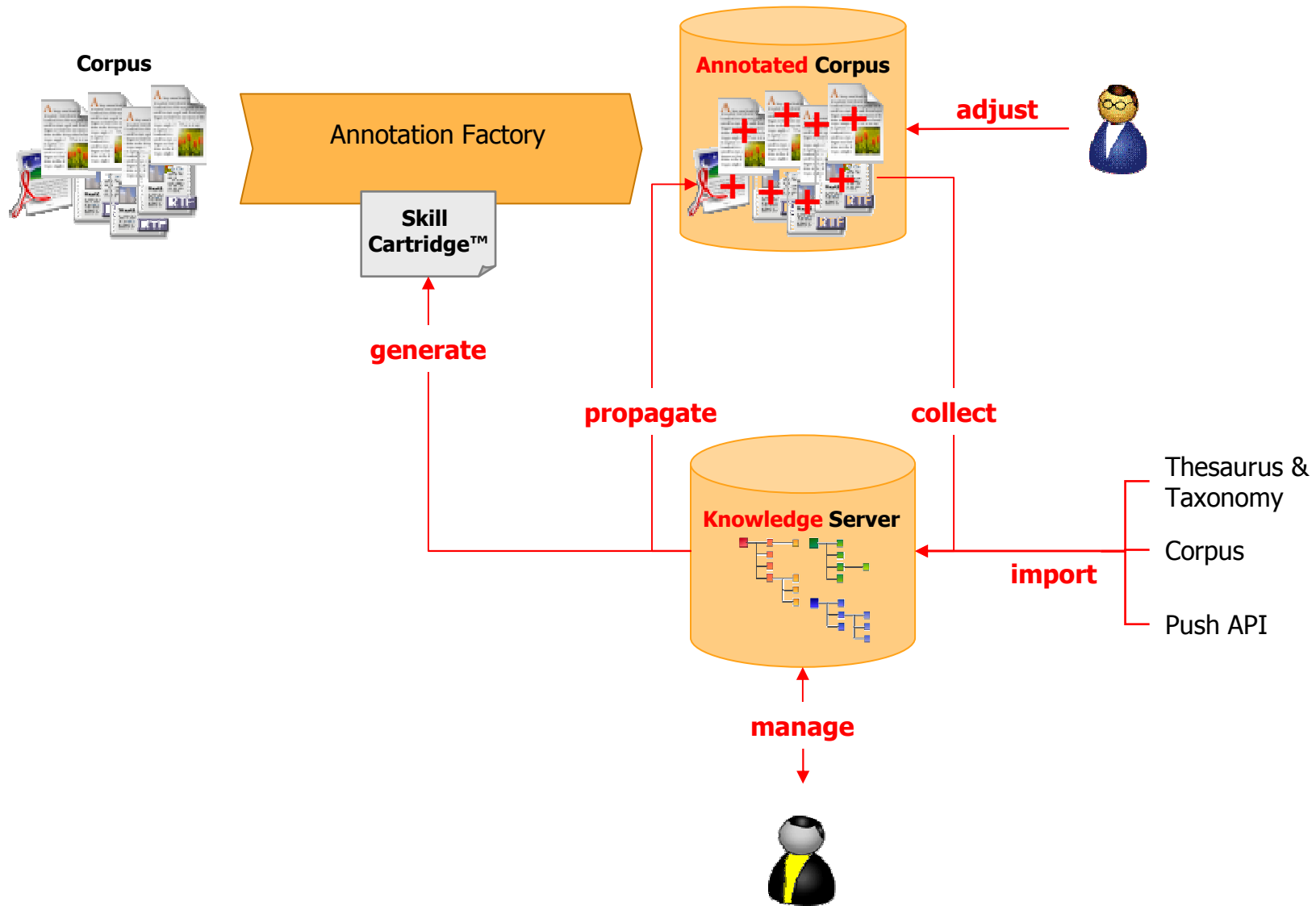
2004 JUN 26 - (NewsRx.com & NewsRx.net) -- NovaDel Pharma, Inc., (NVD) announced that Jean Whitehead Frydman joined the company as vice president and general counsel. Frydman had been associate general counsel of Pfizer, Inc., for the past 17 months after serving 5 years in senior legal posts at Pharmacia Corp., which was acquired by Pfizer, and G. D. Searle, which was acquired earlier by Pharmacia & Upjohn.

She previously spent 13 years in the legal departments of Abbott Laboratories, with a special focus on regulatory compliance, and of BASF Corporation, providing legal services to the company's Knoll Pharmaceutical division. "Jean joins NovaDel at an important juncture for the company, a time when we are moving rapidly with focus on the development of new products and preparing to file our first product approval submission." She is a highly qualified and immediately valuable member of our team." During her tenure at BASF, she provided legal guidance and support in a variety of areas including general litigation, and IP life cycle matters. Major products include Celebrex, Camptosar, Depo-Provera, Healon and others. She was a member of a crisis team formed to defend a \$6.5 million lawsuit. She also authored two ethics and compliance policy guidelines for Pharmacia. At Pharmacia she managed a complex manufacturing compliance issue and managed a major regulatory matter. She earned a Juris Doctorate degree from Chicago-Kent College of Law and completed her undergraduate studies at Ohio State University's School of Medical Technology, earning a BS degree. NovaDel Pharma Inc. is a specialty pharmaceutical company engaged in the development of novel drug delivery systems for prescription and over-the-counter drugs. This article was prepared by Obesity, Fitness & Wellness Week editors from staff and other reports. Copyright 2004, Obesity, Fitness & Wellness Week via NewsRx.com & NewsRx.net.

(c) Copyright 2004 Obesity, Fitness & Wellness Week via NewsRx.com

A context menu is open over the word "BASF" in the text, showing options: "Export into Entity list", "Delete in this discovery", "Delete in this document", "Delete in all knowledge", "Move...", "Advanced edit...", and "Rename".

Knowledge Server™ Architecture



- Users are in control
- Complexity is hidden
- The system is dynamic
- Evolutions and migrations are easy
- Different level of customization to handle any cases