Practical applications of linguistic technology

compiled by Dagobert Soergel

Reading 3 (required). Note: This TOC also provided as a separate file to view side-by-side Look over these pages to get a general idea and dig deeper if something piques your interest.

This is a compilation of materials on the LUXID system (and its forerunner Insight [not to be confused with Inxight]) produced by TEMIS Corp., and on the SAP BusinessSystems ThingFinder system. ThingFinder originated at Inxight corporation, a spinoff from Xerox PARC (Palo Alto Research Center of mouse and the MAC and Windows interface fame); Inxight was bought by BusinessSystems, which was bought by SAP, a major business software vendor.

Both systems extract information from text. They process text to extract entities and statements that connect entities through relationships. They also assign metadata to documents, in other words, they assist in (or take over) cataloging. The marketing hype says that these systems convert unstructured data (text, better called data with complex structure) into structured data, entity-relationship statements often represented in tables and graphs, structures that are simpler and easier to search than text.

The compilation includes old documents from Inxight because they give better examples and a much better explanation of the process.

Reading 3a (optional) gives more detail. Only for those really interested in this topic

Luxid®

Structure, manage and exploit your unstructured content.

Based on patented and award-winning natural language processing technology, Luxid® Content Enrichment Platform is a powerful and scalable semantic content enrichment solution that recognizes and extracts relevant items of information hidden in plain text and enriches document metadata. By revealing the intimate nature of your informational assets, it helps optimize their management, distribution, access and analysis.

Applications in Professional Publishing

Boost the usage of your content by making it more compelling

- Attract new visitors with structured SEO tags
- Make your content easier to navigate with smart facets based on semantic metadata
- Provide context and perspective with personalized content recommendations and links to structured knowledge
- Enable topic matter insights with metadata-driven analytic widgets

Deliver innovative products & formats

- Create Topic Pages by selecting content based on semantic metadata
- Assemble Knowledge Bases with structured information extracted from your existing content
- Embed your content in customer workflow applications with metadata-rich Content APIs

Increase your editorial productivity

- Increase the flexibility, scalability and consistency of tagging, categorization and information extraction through automation
- Repurpose archives faster by revealing their contents with analytics
- Facilitate the maintenance of your taxonomy by automating the identification of candidate terms

Applications in Enterprise Content Management

Add structure to your content

- Reduce the need for end-user metadata contribution by automating content tagging
- Consistently align metadata to your taxonomy or controlled vocabulary

Manage content hypergrowth

- Increase the efficiency of document auditing and migration tasks
- Accelerate e-discovery tasks
- Enable smart archival workflows with domain-specific metadata
- Enrich your analytical reporting with indicators extracted from unstructured content

Exploit your content and the data it contains

- Boost the speed and effectiveness of search
- Enrich your intranets with semantic facets and content recommendations
- Combine structured & unstructured data in Search-Based Applications and Big Data analytics

STRUCTURE THE UNSTRUCTURED TEMIS

Overview and key components

Luxid® Content Enrichment Platform is subdivided into three key functional areas :

Luxid® Annotation Factory

Luxid® Annotation Factory is the flagship natural language processing pipeline that extracts structured information from unstructured documents by recognizing the key topics, entities and relations mentioned in text. Built as a robust and scalable platform, it embeds syntactical, statistical, taxonomy-based and machine-learning driven information extraction engines and supports 20 languages, enabling it to power high-throughput information extraction applications across a wide range of use cases and geographies.

Luxid® Skill Cartridge® Library

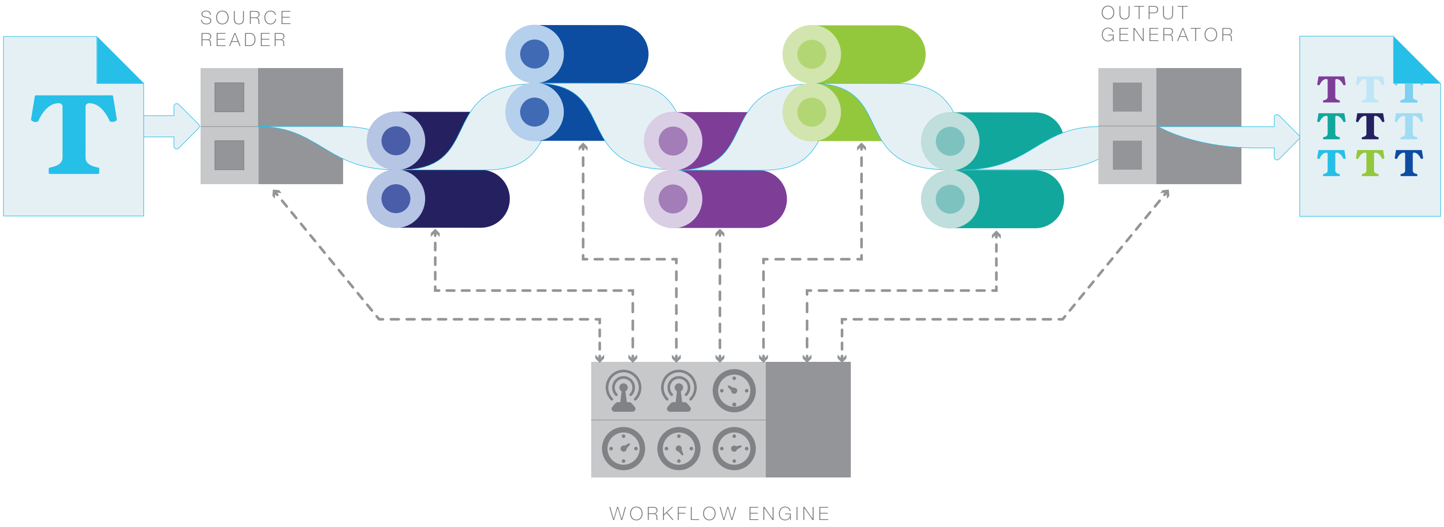
Skill Cartridges® are the specialized modules that focus the information extraction mechanics provided by Luxid® Annotation Factory on domain- or use-case specific entities or relations. Luxid® Skill Cartridge® Library is a range of off-the shelf Skill Cartridges® that address areas of recurring interest such as people names, locations, corporate information and relationships, news categorization, biology, medicine, chemistry, homeland security, and others. Skill Cartridges® can be easily customized and/or developed from scratch with Luxid® Content Enrichment Studio.

Luxid® Content Enrichment Studio

Luxid® Content Enrichment Studio is a suite of four development tools that enable Luxid® licensees to create new Skill Cartridges®, to customize and extend existing ones, and to track and optimize quality and performance. The four tools are Knowledge Editor, Skill Cartridge® Builder, Category Workbench and Annotation Workbench.

Luxid®
Annotation Factory

Luxid® Annotation Factory provides a robust and scalable natural language processing pipeline that supports fast and reliable content enrichment services.



OVERVIEW

Based on UIMA, Luxid® Annotation Factory's distributed architecture automatically balances the workload to all available processing units, and implements native fail-over capabilities sustaining the continuous processing required in mission-critical applications.

AT THE CORE: SKILL CARTRIDGES®

At the core of this distributed pipeline, specialized extraction modules called Skill Cartridges® flexibly support information extraction needs across multiple domains and applications. Depending on the context, Skill Cartridges® may focus on entities of general interest such as companies, people, locations, or dates, or on those that are more domain-specific– such as proteins or genes in biology. Beyond entities, Skill Cartridges® are also able to extract more structured information in the form of the relations that link these entities (for example a merger between two companies or a chemical reaction between two compounds), including the roles played by each entity in the relation, as well as their attributes or other contextual information mentioned in the document.

A BROAD RANGE OF EXTRACTION TECHNIQUES

To optimize its performance across use cases, Luxid® Annotation Factory provides a market-leading part-of-speech tagging layer supporting 20 languages and embeds a wide range of information extraction techniques including morpho-syntactic reasoning, statistics, thesaurus- and taxonomy-based extraction, machine learning and rules-based extraction. Luxid® Annotation Factory is also capable of performing corpus-level operations such as **Categorization** (the classification of documents in predefined categories) or **Clustering** (the grouping of similar documents into dynamically created clusters). Adding even more flexibility to these native capabilities, Annotation Factory also supports the integration of third-party UIMA-compliant annotators in annotation plans.

DEPLOYMENT AND INTEGRATIONS

A key component of Luxid® Annotation Factory, its Workflow Engine is in charge of the end-to-end management of complex workflows that can involve multiple cascaded Skill Cartridges® as well as distributed multi-CPU and multi-node processing. To streamline platform deployment and support the robust operation of such workflows, the Workflow Engine also implements advanced fault-tolerance mechanisms and centralized log management. A web-based administration interface provides users control over all platform parameters and centralized access to its logs. The platform is furthermore JMX compliant, easily integrating into common monitoring consoles.

While the core of its pipeline is natively based on XML, Luxid® Annotation Factory supports more than 200 file formats on the ingress side, and provides a flexible range of output options including XML and RDF. To ease its deployment within the major content management systems, a growing range of off-the-shelf integrations is available, in particular for MarkLogic, Microsoft SharePoint 2010, EMC Documentum, Alfresco Enterprise, and Nuxeo. The Web Services offered by Luxid® Content Enrichment Platform facilitate its integration within other content management systems, applications and information management workflows.

Luxid® Skill Cartridge® Library

Luxid® Skill Cartridge® Library is a range of off-the shelf extractors developed by TEMIS to address recurring use cases and domains.

Here is an overview of some of the most representative of these Skill Cartridges®.

RTF RELEVANT TERM FINDER SKILL CARTRIDGE®

The RTF Skill Cartridge® identifies the most characteristic terms of each document (its "fingerprint") by comparing its vocabulary to a statistical model based on a reference corpus. This model can be adjusted through training on any corpus that appropriately reflects your specific domain. RTF requires no pre-defined conceptual structure and can be applied to a wide variety of use cases such as Similar Documents Recommendation, Clustering, and domain-specific Terminology Extraction.

STF SMART TAXONOMY FACILITATOR SKILL CARTRIDGE®

STF acts as a vehicle for applying taxonomies and controlled vocabularies to documents. It embeds technologies that help overcome two key weaknesses associated to taxonomy-based indexing. The first, Fuzzy Term Matching, automatically produces variants of the forms present in the taxonomy, thereby helping to improve recall. The second, Relevance Scoring, applies a range of heuristics to assign a relevance score to each extracted concept and discards the less relevant ones, therefore improving extraction precision. STF can also exploit part-of-speech tagging information to avoid false positives caused by ambiguous taxonomical terms. Knowledge Editor enables you to conveniently package your own taxonomy into STF and adjust its performance to your specific use case by controlling the parameters driving its heuristics.

MLX MACHINE LEARNING EXTRACTION SKILL CARTRIDGE®

The MLX Skill Cartridge® is a versatile extractor that can be trained to extract, subcategorize and/or score virtually any type of entity in text based on a previously annotated corpus where such entities of interest have been highlighted by domain experts. The corpus annotation and Skill Cartridge® training processes requires no natural language processing expertise and can be performed easily with Annotation Workbench.

IPTC IPTC CATEGORIZATION SKILL CARTRIDGE®

This Skill Cartridge® analyzes the vocabulary used in documents and classifies them into approximately 130 categories based on key headings of the top two levels of the IPTC (News) taxonomy. This Skill Cartridge® is available for the English and French languages. Extensions to other languages and more complete coverage of the IPTC headings can be achieved on a project basis.

TM 360 TEXT MINING 360° SKILL CARTRIDGE®

TM360 extracts more than 20 types of common entities among which the names of People, Companies, Organizations, Locations, as well as Measurements, Money and Time expressions, and Contact information. To achieve this task, TM360 embeds a variety of predictive natural language processing methods including morphological and syntactical heuristics. This Skill Cartridge® is available in 8 languages (English, German, French, Spanish, Italian, Dutch, Portuguese and Arabic).

CI COMPETITIVE INTELLIGENCE SKILL CARTRIDGE®

CI extracts the same entities as the TM360° Skill Cartridge®, as well as 10 types of semantic relationships involving companies, their employees, assets or strategy, including Corporate Mergers and Acquisitions, Court Cases, Board and Management Changes, Financial Reporting, Business Development and Strategy. CI leverages syntactical reasoning to establish the role played by each entity in the relationship, enabling it to recognize for instance the acquiring company from the target or the licensee from the licensor. CI is available in English, French, Spanish and Dutch.

OM OPINION MINING SKILL CARTRIDGE®

The Opinion Mining Skill Cartridge® recognizes the subjective expressions that reveal evaluative judgment or emotional states that are typically found in social media or customer feedback corpora. Leveraging a combination of specialized thesauri and syntactical reasoning, OM qualifies these opinions both in terms of polarity (negative/positive) and intensity, and links them to their target (for example products, service features, brands, organizations or people), helping professionals to quantify, track and analyze information that is otherwise entirely unstructured. This Skill Cartridge® is available in English and French.

BER BIOLOGICAL ENTITIES AND RELATIONSHIP SKILL CARTRIDGE®

Available in the English language, BER extracts both 15 key types of biological entities and 17 relationships binding them, as well as rich contextual information. Extracted entities include Anatomical terms, Cells and Tissues, Proteins and Genes, Disorders, Treatments, Species, Genomic and Mutational information. Relationships include Activation, Inhibition, Regulation, Gene expression, Mutation, Diagnosis, and Therapy. BER is powered by a deep syntactical analysis layer that provides an intimate and highly structured understanding of complex biological mechanisms and pathways.

MER MEDICAL ENTITIES AND RELATIONSHIP SKILL CARTRIDGE®

MER extracts 15 key types of medical entities and 10 types of semantic relationships binding them. Extracted entities include Clinical Trial terms, Diagnostics, Disorders, Transmission information, Genomic information, Species information, Ethnic information, Symptoms, Targets, Treatments, and Administration Routes. Examples of relationships include Adverse Effects (binding a treatment with a negative effect), Therapies (binding a treatment with a disease), and Molecular Targets (binding a process with certain specific entity types).

CHEMICAL ENTITIES RECOGNITION SKILL CARTRIDGE®

CER extracts chemical entities such as chemical compounds, chemical classes and molecular formulae and performs name-to-structure conversion. It includes support for International Union of Pure and Applied Chemistry (IUPAC) as well as Chemical Abstracts Service (CAS) registry numbers. CER can be successfully used in conjunction with BER or MER to reveal the chemical compounds at play in biological and medical contexts. It can also be customized with custom lexicons.

Luxid® Skill Cartridge® Library also includes a range of Skill Units that are more specific to certain specialty areas such as military, political and strategic events, legal proceedings and scientific and technical domains beyond the life sciences.

Luxid®
Content Enrichment Studio

Luxid® Content Enrichment Studio is a suite of four development tools that enables you to customize and extend existing Skill Cartridges®, to develop new ones from scratch, and to track and optimize quality and performance.

LUXID® KNOWLEDGE EDITOR

enables the development and optimization of Taxonomy-based Skill Cartridges®. It provides a user-friendly interface that streamlines the process of importing and editing taxonomies, thesauri, and/or controlled vocabularies, and of packaging them into a Skill Cartridge® that can be used within Luxid® Annotation Factory. Luxid® Knowledge Editor also provides access to a wide variety of parameters and training processes that help fine-tune the behavior of Taxonomy-based Skill Cartridges®. Luxid® Knowledge Editor can also be used to access lexical resources present in all types of Skill Cartridges®.

LUXID® SKILL CARTRIDGE® BUILDER

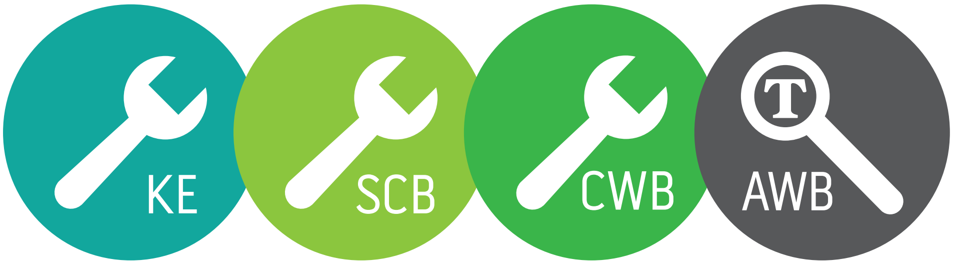
supports the development of semantic rules-based Skill Cartridges®. Such rules, written and optimized by developers, constitute models that leverage the part-of-speech tags and morpho-syntactic properties of text to recognize entities of interest when they appear in documents, as well as their relationships and respective roles in these relationships. Thanks to Luxid® Skill Cartridge® Builder developers can apply the full power of semantics to virtually any type of domain or use case.

LUXID® CATEGORY WORKBENCH

supports both the development and optimization of Categorization Skill Cartridges®. Such Skill Cartridges® leverage statistics to categorize documents according to predefined document classification plans. Luxid® Category Workbench also includes a clustering feature that is able to suggest possible classification plans when no such plan is available for reference.

LUXID® ANNOTATION WORKBENCH

provides a window into the quality and performance of Skill Cartridges® developed with either Luxid® Knowledge Editor or Luxid® Skill Cartridge® Builder. It supports the manual curation (validation, correction, removal and insertion of annotations by the human operator) as well as the automated evaluation and generation of reports regarding Skill Cartridge® extraction quality by comparison to a pre-annotated reference corpus. The reports produced by Luxid® Annotation Workbench can be used to evaluate a Skill Cartridge®'s extraction quality on an absolute basis, to compare it with another Skill Cartridge® (for example subsequent versions of the same, to evaluate progress) or to compare its performance across different corpora. Luxid® Annotation Workbench also enables domain experts to create machine-learning based Skill Cartridges® by simply training them to identify and extract information based on a manually-annotated corpus.



TEMIS

STRUCTURE THE UNSTRUCTURED

About TEMIS

TEMIS helps organizations structure, manage and leverage their unstructured information assets. Its flagship platform, Luxid®, identifies and extracts targeted information to semantically enrich content with domain-specific metadata. Luxid® enables professional publishers to efficiently package and deliver relevant information to their audience, and helps enterprises to intelligently archive, manage, analyze, discover and share increasing volumes of information.

Founded in 2000, TEMIS operates in the United States, Canada, UK, France and Germany, and is represented worldwide through its network of certified partners.

TEMIS' innovative solutions have attracted the business of leading organizations such as AAAS (American Association for the Advancement of Science), Agence France-Presse, BASF, Bayer Schering Pharma, BNA (Bureau of National Affairs), BNP Paribas, CARMA International, Editions Lefebvre-Sarrut, Elsevier, EMC, Europol, French Ministry of Defence, French Ministry of Finance, Gannett, Karger, Invest in France Agency, Merck Serono, Nature Publishing Group, Novartis, Philip Morris International, PSA Peugeot-Citroen, Sanofi-aventis, Simon & Schuster, Springer Science+Business Media, The McGraw-Hill Companies, Thieme, Thomson Reuters, Trinity Mirror plc and the U.S. Department of Agriculture.

www.temis.com
tagline.temis.com

USA

295 Madison Avenue
45th floor
New York, NY 10017 – USA
Tel.: +1 646 392 7717
info.us@temis.com

United Kingdom

London
Tel.: +44 (0)777 474 6278
info.uk@temis.com

Germany

Blumenstraße 15
D-69115 Heidelberg
Tel. : +49 (62 21) 1 37 53 - 0
info.de@temis.com

Canada

Toronto
Tel.: +1 416 493 2486
info.ca@temis.com

France

Tour Mattei,
207 rue de Bercy
75012 Paris
Tel.: +33 (0)1 80 98 11 00
info@temis.com

http://tagline.temis.com/2014/07/01/temis-integrates-ontology-management-and-semantic-enrichment-in-luxid-7/

Downloaded 20140901

TEMIS Integrates Ontology Management and Semantic Enrichment in Luxid® 7

Published July 1, 2014 New product or feature Leave a Comment
Tags: Luxid 7, ontology management, Semantic Content Enrichment, webstudio

Luxid® 7's architecture simplifies key information management workflow for accelerated deployment and lower cost of ownership

New York, NY – July 1st, 2014 – TEMIS, the leading provider of Semantic Content Enrichment solutions for the Enterprise, today announced the launch of Luxid® 7, the seventh generation of its flagship semantic content enrichment platform. With a new, significantly redesigned internal architecture, Luxid® 7 offers an even more scalable and robust semantic enrichment pipeline, and now also includes a dedicated ontology management tool, providing its users with an industry-first integrated workflow which is more powerful and more efficient.

An organization's ontology – or its subset called *taxonomy* – describes the objects that are essential to its business (e.g. products, regions, projects, …) and their relationships. Semantic enrichment provides a scalable mechanism to recognize and capture the mentions of such objects and relationships – so-called *triples* – that are hidden in plain text. In recent years, ontologies and semantic enrichment have become critical to efficiently exploit unstructured content, but they are most often disconnected, requiring intensive, time-consuming, collaboration between domain experts, IT and linguistics experts, to ensure quality results. Luxid® 7 bridges this gap by providing an integrated workflow, thereby considerably reducing the associated time to market and overall costs.

"Luxid® 7 is a powerful and highly scalable semantic enrichment platform, yet it offers simple-to-use, intuitive interfaces for business users and subject matter experts, and easy-to-integrate REST Web Services," said Daniel Mayer, VP Product & Marketing, TEMIS. "This means our customers can enjoy efficient, integrated workflows, all the while deploying the added value of semantics in all their applications in a fraction of the time required with a disconnected workflow," he added. "Our intent is to accelerate adoption within customer organizations and make systems integrators more autonomous in delivering high quality services to their own customers."

NLP-enabled ontology management workflow is simpler, more efficient, and more powerful

Follow

An entirely new component of the Luxid® platform, Luxid® Webstudio is a natively multi-user, collaborative web application enabling users to create, edit and maintain an ontology while governing the way ontological objects are recognized by the Luxid® semantic enrichment pipeline. It also leverages the platform's Natural Language Processing layer to

- **Preview in real time** the results of the semantic enrichment process when applied to users' corpus of documents. This enables users to rapidly see and correct gaps between ontology and real world semantic enrichment, directly within the Webstudio interface, for example by adding variants or adjusting extraction mechanisms.
- **Suggest relevant objects** mentioned in the user's corpus that are not yet included in the ontology. This feature can be extended with any Skill Cartridge® to suggest new objects and relationships based on linguistic patterns, statistics, or machine learning. In this spirit, Webstudio also embeds a Wikipedia-based Skill Cartridge® that suggests synonyms for any given concept. These suggestions can help non Subject Matter Experts to build their own ontologies too.

These powerful features make end-users more autonomous in creating or maintaining an ontology and the corresponding semantic enrichment pipeline. This translates into a considerably simplified and accelerated workflow, enabling improved time-to-market and lower total cost of ownership.

Streamlined, Big Data architecture offers improved scalability and robust integration options

Luxid® Annotation Server is TEMIS's flagship natural language processing pipeline. Supporting 20 languages and leveraging Skill Cartridges® – the specialized information extraction modules popularized by TEMIS – Luxid® Annotation Server lends itself to a broad range of usage scenarios thanks to its comprehensive range of information extraction engines based on syntactical, statistical, taxonomical, and machine-learning algorithms. In its 7th generation, the pipeline has also been rethought from the ground up to offer

- **A simplified internal architecture** exploiting a single data model that reduces overhead, improves CPU utilization and efficiently lends itself to a variety of scale-in and scale-out configurations, for real-time, high-availability or batch content processing applications. Luxid® Annotation Server plays well in both cloud and Big Data (Hadoop) deployments, and flexibly processes ever-increasing volumes of unstructured content.
- **Comprehensive REST Web Services** enabling easy integration into any application or workflow. Beyond information extraction itself, virtually all Luxid® features are accessible through Web Services, enabling rich integrations that work hand-in-hand with other applications such as content management systems, portals, search engines, archival and case management platforms, editorial workflow tools, as well as analytics applications.

Thanks to these evolutions, Luxid® 7 provides users with an unparalleled ability to add the value of semantic intelligence to their existing applications in complete coordination with their ontology or taxonomy management processes.

Luxid® 7 leverages TEMIS's updated Skill Cartridge® Library, a range of off-the shelf, application-specific information extractors, and includes its Content Enrichment Studio suite of tools, ensuring the industrial creation, maintenance and quality evaluation of Skill Cartridges®. Luxid® 7 is a powerful tool for the Luxid® Community, enabling users in a matter of minutes to import any thesaurus or taxonomy and consequently populate the website's Marketplace with new and innovative annotation resources.

http://www.prnewswire.com/news-releases/oecd-chooses-temis-to-semantically-structure-its-knowledge-and-information-management-processes-232818841.html

OECD Chooses TEMIS to Semantically Structure its Knowledge and Information Management Processes

PARIS and NEW YORK, November 21, 2013 /PRNewswire/ --

TEMIS (http://www.temis.com/), the leading provider of Semantic Content Enrichment solutions for the Enterprise, announced today that they have won a call for tender issued by the Organisation for Economic Co-operation and Development (OECD) with their award-winning Semantic Content Enrichment solution Luxid (http://www.temis.com/luxid-6)®.

The OECD provides its expertise, data and analysis to its 34 member governments and 100 other countries to help them support sustainable economic growth, boost employment and raise living standards. To fulfill its vision of increased relevance and global presence, the OECD has launched a Knowledge and Information Management (KIM) Program that establishes an integrated framework for managing and delivering information and improving its accessibility and presentation. The KIM framework is intended as the steward of the OECD's information lifecycle, with a universal knowledge referential at its core facilitating enhanced searching and findability, rationalized content re-use/repurposing processes, and supporting the organisation's Open Data and Linked Data initiatives.

In this context, the OECD has chosen TEMIS's flagship Luxid® Content Enrichment Platform to address all Semantic Enrichment stages of the KIM framework. Luxid® will help OECD to consistently enrich document metadata in alignment with its taxonomies and ontologies, providing a genuinely semantic integration layer across heterogeneous document storage and content management components. This semantic layer will both enable new search and browsing methods and improved relevance and accuracy of search results, as well as progressively build an integrated map of OECD knowledge.

"After careful evaluation, the OECD selected TEMIS Luxid® platform as the key technological component to support the transition from content to semantic information," said Simone Sergi, KIM Senior Program Manager, OECD.

"This mark of trust by OECD represents a new recognition of our ability to address challenges in international organisations. For TEMIS, this is a link between our know-how in the publishing domain and our industrial experience in information systems," said Fabien Gauthier, Sales Director, Enterprise, TEMIS.

Based on patented and award-winning Natural Language Processing technologies, Luxid® exploits off-the-shelf extractors called Skill Cartridges® to extract targeted information from unstructured content and semantically enrich it with domain-specific metadata. This enables professional publishers to efficiently package and deliver relevant information to their audience, and helps enterprises to intelligently archive, manage, analyze, discover and share increasing volumes of information.

About OECD

The OECD is a global economic policy forum. It provides analysis and advice to its 34 member governments and other countries worldwide, promoting better policies for better lives.

Website: http://www.oecd.org (http://www.oecd.org/)

About TEMIS

TEMIS helps organisations structure, manage and exploit their unstructured information assets. Its flagship platform, Luxid®, identifies and extracts targeted information to semantically enrich content with domain-specific metadata. This helps organisations to intelligently archive, manage, package, deliver, access and analyze increasing volumes of information. Founded in 2000, TEMIS operates in the United States, Canada, UK, France and Germany, and is represented worldwide through its network of certified partners.

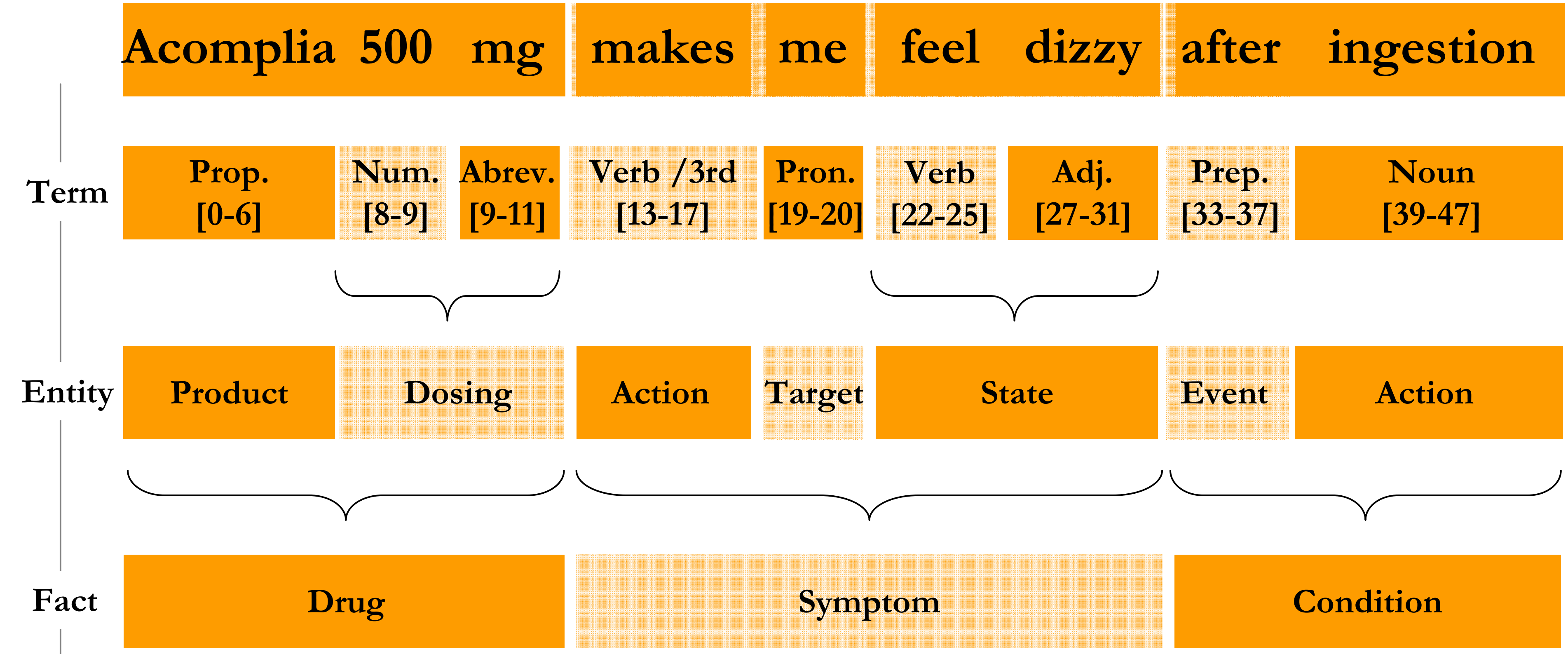
TEMIS' innovative solutions have attracted the business of leading organisations such as AAAS (American Association for the Advancement of Science), Agence France-Presse, BASF, Bayer Pharma, Bloomberg BNA, BNP Paribas, Editions Lefebvre-Sarrut, Elsevier, EMC, Europol, French Ministry of Defence, French Ministry of Finance, Gannett, Karger, Invest in France Agency, Les Echos, Merck KGaA, Nature Publishing Group, Novartis, PSA Peugeot-Citroën, Sanofi, Simon & Schuster, Springer Science+Business Media, The McGraw-Hill Companies, Thieme, Thomson Reuters, the U.S. Department of Agriculture, Volkswagen and Wiley.

Website: http://www.temis.com/ (http://www.temis.com/)

Blog: http://tagline.temis.com (http://tagline.temis.com)

Twitter: https://twitter.com/TEMIS_Group (https://twitter.com/TEMIS_Group)

Innovation? *From Text to Meaning!*



Acomplia 500 mg makes me feel dizzy after ingestion

Term

Prop. [0-6] | Num. [8-9] | Abrev. [9-11] | Verb /3rd [13-17] | Pron. [19-20] | Verb [22-25] | Adj. [27-31] | Prep. [33-37] | Noun [39-47]

Entity

Product | Dosing | Action | Target | State | Event | Action

Fact

Drug | Symptom | Condition

Meaning

Meaning = Potential Adverse Effect

Drug = Acomplia

Dosing = 500mg

Symptom = Dizziness

When = After administration

inxight

Inxight Software, Inc.
U.S. I 500 Macara Avenue, Sunnyvale, CA 94085
Phone: 408.738.6200 I Fax: 408.738.6311
Europe I Centaur House, Ancells Business Park, Ancells Road,
Fleet, Hampshire, GU51 2UN U.K.
Phone: +44 (0) 1252 761314 I Fax: +44 (0) 1252 761315

Inxight LinguistX® Platform

The Most Advanced Natural-Language Processing

Inxight LinguistX intelligently analyzes text in more than 30 languages to quickly deliver accurate and relevant information, enabling software developers to build powerful text analysis features into their products.

Organizations continue to tackle the growing challenge of how to quickly and efficiently retrieve relevant information from the Internet and other electronic text data sources. Inxight's LinguistX® Platform provides advanced text analysis technology that enables software developers to build multi-language text analysis features into their products, while reducing time-to-market and avoiding unnecessary development costs.

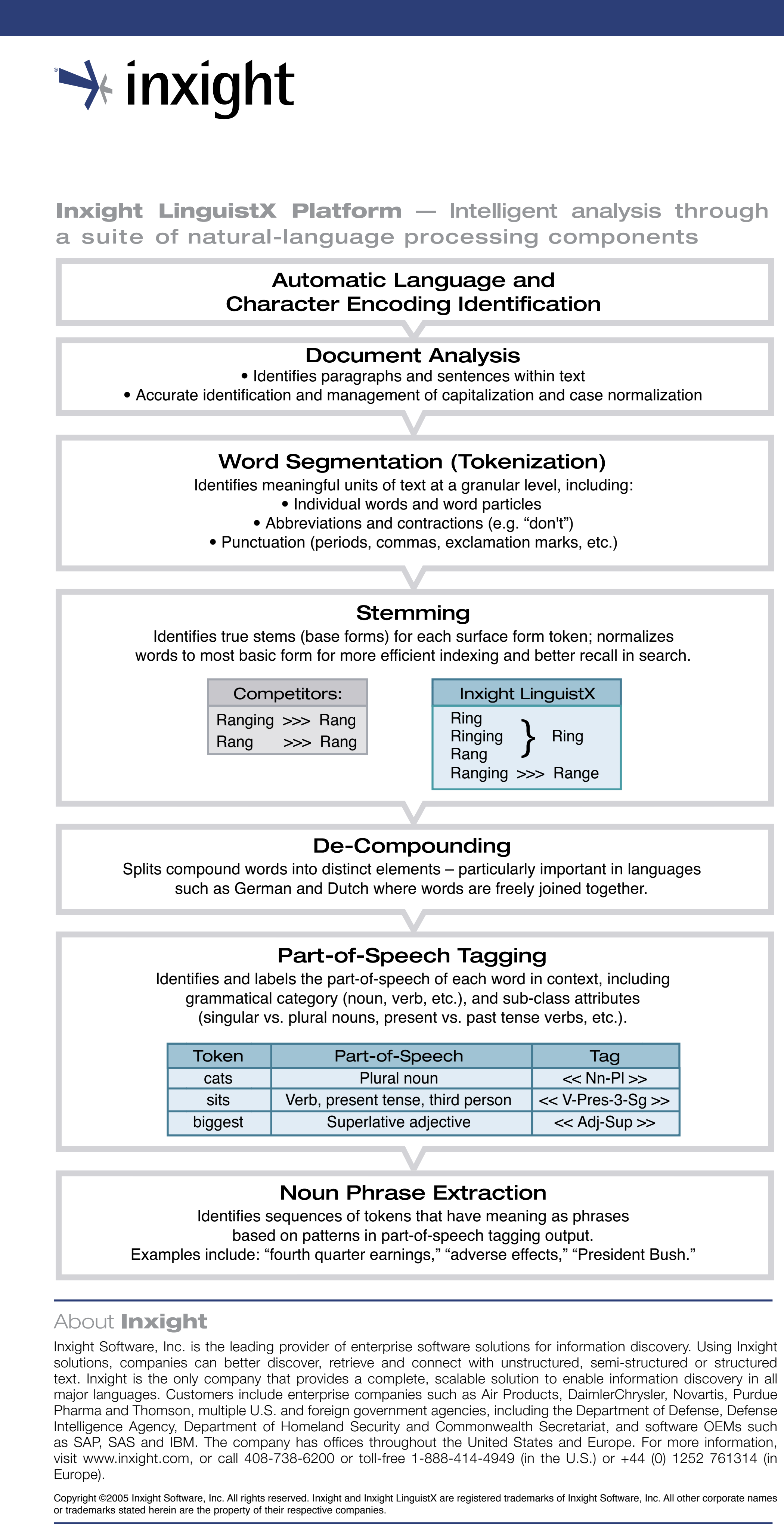
Using a single API for high-performance natural-language processing components, developers can quickly and cost-effectively create applications that enable intelligent, accurate and timely access to information. Inxight LinguistX Platform provides advanced text analysis capabilities in more than 30 languages, making it the solution of choice for search engines, data mining applications, indexing applications, and text categorization and routing tools.

Supporting global organizations and operations

Using a single API, Inxight LinguistX Platform intelligently analyzes text by providing language and character encoding identification, segmentation and stemming in these languages:

- Arabic
- Catalan
- Chinese (Simplified)
- Chinese (Traditional)
- Croatian
- Czech
- Danish
- Dutch
- English
- Farsi (Persian)
- Finnish
- French
- German
- Greek
- Hebrew
- Hungarian
- Italian
- Japanese
- Korean
- Norwegian (Bokmål)
- Norwegian (Nynorsk)
- Polish
- Portuguese
- Romanian
- Russian
- Serbian
- Slovak
- Slovenian
- Spanish
- Swedish
- Thai
- Turkish

Many of these languages also support part-of-speech tagging and noun phrase extraction. Contact Inxight Sales for additional languages.



Inxight LinguistX Platform

inxight

Inxight LinguistX Platform — Intelligent analysis through a suite of natural-language processing components

Automatic Language and Character Encoding Identification

Document Analysis

- Identifies paragraphs and sentences within text
- Accurate identification and management of capitalization and case normalization

Word Segmentation (Tokenization)

Identifies meaningful units of text at a granular level, including:

- Individual words and word particles
- Abbreviations and contractions (e.g. “don't”)
- Punctuation (periods, commas, exclamation marks, etc.)

Stemming

Identifies true stems (base forms) for each surface form token; normalizes words to most basic form for more efficient indexing and better recall in search.

Competitors:		
Ranging	>>>	Rang
Rang	>>>	Rang

Inxight LinguistX		
Ring Ringing Rang	}	Ring
Ranging	>>>	Range

De-Compounding

Splits compound words into distinct elements – particularly important in languages such as German and Dutch where words are freely joined together.

Part-of-Speech Tagging

Identifies and labels the part-of-speech of each word in context, including grammatical category (noun, verb, etc.), and sub-class attributes (singular vs. plural nouns, present vs. past tense verbs, etc.).

Token	Part-of-Speech	Tag
cats	Plural noun	<< Nn-Pl >>
sits	Verb, present tense, third person	<< V-Pres-3-Sg >>
biggest	Superlative adjective	<< Adj-Sup >>

Noun Phrase Extraction

Identifies sequences of tokens that have meaning as phrases based on patterns in part-of-speech tagging output.
Examples include: “fourth quarter earnings,” “adverse effects,” “President Bush.”

About Inxight

Inxight Software, Inc. is the leading provider of enterprise software solutions for information discovery. Using Inxight solutions, companies can better discover, retrieve and connect with unstructured, semi-structured or structured text. Inxight is the only company that provides a complete, scalable solution to enable information discovery in all major languages. Customers include enterprise companies such as Air Products, DaimlerChrysler, Novartis, Purdue Pharma and Thomson, multiple U.S. and foreign government agencies, including the Department of Defense, Defense Intelligence Agency, Department of Homeland Security and Commonwealth Secretariat, and software OEMs such as SAP, SAS and IBM. The company has offices throughout the United States and Europe. For more information, visit www.inxight.com, or call 408-738-6200 or toll-free 1-888-414-4949 (in the U.S.) or +44 (0) 1252 761314 (in Europe).

Temis XeLDA
has the same basic functions

Language identification. Automatically recognizes the language used by each document

Segmentation. Divides a text into sentemces

Tokenization. Splits a text into basic lexical units

Morphological analysis. Returns the normalized form (the lemma) and the potential grammatical categories [parts of speech] for all the words identified during the tokenization stage

Morpho-syntactic disambiguation. Determines the exact grammatical category of a word according to its context.

Extraction of noun phrases

Dictionary lookup. Identifies the context of a word to find the corresponding dictionary entry

Recognition of idiomatic expressions

inxight

Inxight White Paper Series

Linguistics: Adding Value to e-Publishing and e-Content

p.1.408.738.6200 f.1.408.738.6203 500 Macara Avenue, Sunnyvale, CA 94085 US p.+44.1252.761314 f.+44.1252.761315 Centaur House, Ancells Business Park, Ancells Rd, Fleet, Hampshire GU51 2UN, UK
www.inxight.com

Executive Summary

Constant change is *the* overriding factor shared by thousands of electronic publishing companies that provide information through Web sites. Traditional information distribution mechanisms (email, postal services, etc.) must now coexist with the chaos of the Web and its billions of pages. Smart, agile publishers and information aggregators - grouped under the headings e-Publishers and infomediaries - are learning that their success is dependent on monetizing content while protecting its inherent value from creeping “Napsterization” - all the while bucking the “everything for free” culture of the Internet.

So, the problem becomes how to charge a premium for information that the online world assumes should be free of charge. The answer lies in adding significant, obvious and relevant value by delivering specific, targeted information that saves users the time, effort and resources required to sift through the Web's massive data storehouses themselves.

Of the products and technologies targeting this market, only Inxight's e-Content Publishing Platform has the appropriate combination of functionality and underlying architecture to achieve the three key e-Publishing objectives: return on investment, improved productivity and increased value of information.

As the Web's overall content spirals upward into the realm of exabytes and beyond, the sheer volume of information that requires high-speed, intelligent search and retrieval software has presented an opportunity for editors to embrace new technology systems that replace former error-prone methods. Inxight's e-Content Publishing Platform presents e-Publishers with the best available choice.

e-Publishing Industry Overview

Electronic publishing is growing at a staggering rate. Up from $146 billion in 1999 (Outsell, Inc.), industry insiders suggest that the number could now be as high as $300 billion, and the expansion of the Web and the Internet as information distribution engines can only spur further growth.

IDC estimates that 200 million pages are added each month, while an estimated 100 million become obsolete. The latter figure demonstrates clearly why information management on the Web is so difficult. Obsolete pages proliferate misinformation, although there's usually no way for users to gauge the validity of any given page.

Timeliness is a key element in certain e-Publishing areas, notably the distribution of news content. No one is going to pay a premium for yesterday's news. In other areas, though, time is less important.

Looking at the bigger picture, the Web has imposed a set of issues unknown in the publishing industry a decade ago, including the commoditization of information, the explosive growth of content and content aggregators, and the focus on profitability - to the chagrin of those who would contend that the Web must continue to operate at no cost to the user.

While the Internet and Web are generally regarded as free sources of information, business realities have created new mandates to generate revenue and profits, driving the publishing industry to control intellectual property assets more closely. Market factors affecting publishing today include:

- Growth and expansion of e-Publishers/infomediaries
- High expectations for quality and the need for specific content
- User expectations and demand for real-time information
- New players, roles and commoditization lead to slimmer margins
- Information needs to be available anytime, anywhere
- Content vendors need to adapt their content to support expanded media

The Players

These days, the e-Publishing distribution system is dominated by “infomediaries.” These include many of the leading syndicators from the “old economy,” such as Reuters and Dow Jones, along with the new breed of aggregation companies that serve as go-betweens, linking content providers with mass audiences. The latter group includes YellowBrix, iSyndicate, Factiva and Screaming Media. They don't produce their own content, nor do they have the kinds of direct relationships with information consumers that newspapers and magazines create. Rather, they serve as matchmakers for both ends of the distribution chain, linking content with knowledge consumers.

Infomediaries separate content from a specific interface or distribution channel, and make it available through an infinite number of channels. They aggregate content from many primary publishers - just as the traditional syndicators did - but instead of rolling the content into a single one-size-fits-all package, they offer slices of the content to a Web site or intranet that wants to add third-party content to its mix.

The e-Publishing cast looks like this:

- Infomediaries, including syndicators who collect information from selected sources and resell it to other buyers. Leading players include Factiva, The New York Times Syndicate, and iSyndicate. Aggregators are also part of the infomediary group. They don't own the content, but collect multiple information types from multiple sources and re-distribute it to buyers. Key players include Lexis Nexis, Dialog and Screaming Media.
- Buyers, in this case an organization that buys content on behalf of end users and publishes it on an information portal. The buyer could be a private company or a commercial portal like msn.com or yahoo.com.
- End users, or information consumers. Players include individuals browsing Web sites for private use or business users collecting decision support information for commercial use.

Industry Dynamics

As the nature of the publishing business changes - and specifically the move to e-Content - so, too, do the drivers and inhibitors that dictate industry dynamics. As we've already seen, the sheer volume of information is growing exponentially. As a result, applications that categorize, tag and cull Web-based information today must encompass terabytes of data. In the not too distant future, applications must be capable of performing the same set of functions on to ‘exabytes' of data (one million terabytes). The demand, therefore, is for information management architecture that scales to accommodate a virtually infinite set of information derived from multiple sources.

These sources include corporate ERP/CRM applications, document management systems, external research materials, news feeds, interpersonal email, and data housed on personal computers. Metadata - the “rules” by which information is categorized and culled, is scrambled and non-standardized. In addition, commoditization has impacted the value of e-Content and made it difficult to create enough perceived value to charge money for it.

The obvious needs are for highly refined filtration systems and tight, standardized organizational methodologies. In short, information delivery needs to “get personal,” sorting through the maze of online content to zero-in on the requirements of specific users and organizations. Let's now examine the technologies that make that possible.

Technologies

Just as HTML has led the pack in describing Web presentation, XML (eXtensible Markup Language) is growing as the industry standard used by application developers as the mechanism to define key information within documents. Used by a variety of enterprise applications, XML can be used as the common link between data from disparate sources. XML is essential in architecting a robust enterprise system, and allows for reuse of information across multiple, integrated applications.

As *PC Week* (now *eWEEK)* stated in 1999, “Like HTML, XML derives from the granddaddy of all markup languages: SGML (Standard Generalized Markup Language). SGML is a meta-language, or a system for defining markup languages such as HTML. XML is also a meta-language, a subset of SGML designed for use on the Web. As with SGML, you can use XML to define different markup languages for specific uses, particularly for data representation.”

Adding to the confusion, different XML derivatives are constantly being promulgated as standards by various industry groups. Among the markup languages in publishing, whether NewsML, PRISM, NITP, each leverages XML, but modifies it for specific applications (e.g.,
NewsML for news publishers and ICE information and Content Exchange> as a protocol to automate content syndication. Now, a look at how Inxight's underlying technology and application-level systems combine to deliver the most efficient e-Publishing content delivery solution available today.

The Inxight Business Case

Strong technology and a rich feature set is one thing; mapping these capabilities to address real business needs is another. Fortunately, Inxight's e-Content Publishing Platform excels at delivering the benefits that define a successful e-Publishing solution: return on investment; productivity increases; and increased value of information.

Increased ROI

Decrease Operation and Administrative Costs:
Time is money - especially in the case of publishing content. The older the news, the less valuable it becomes. So, e-Publishers/infomediaries invest a significant amount of money on human capital to produce and maintain content. These days, much of that content is derived from the Web and, as discussed earlier in this paper, searching billions of Web pages for pertinent, related information is the single most difficult, time-consuming activity in the e-Publishing industry. By automating the process, most publishers can decrease their operational and administrative costs, and realize significant savings with an investment payout in less than one year.

A small news agency, for example, that manages 300-plus news articles a day, will spend at least $1,000 a day on the time editors will take to categorize, summarize and tag an electronic article. Considering that it takes five minutes to manually categorize, tag and summarize a typical article, the cost added to the article is about $3.00 just to prepare it for electronic distribution. By automating this process, the savings would amount to at least $365,000 each year.

Increased productivity

Another advantage of e-content automation technologies is enhancing productivity and efficiency. For editors using the Inxight e-Content Publishing Platform, the benefit is the time they can save on tagging documents, enabling them to spend their time on more strategic editorial responsibilities.

Editors simply don't scale well - you can only get so much quality work out of a single human. With Inxight's e-Content Publishing Platform, scalability is built in; the engine is more than capable of sifting through billions of entries in minutes and extracting relevant, personalized information. Also, human editors often miss related content.

Take, for example, a business analyst looking to correlate internal sales data with external market research and consumer trend documents. In essence, she is looking for very closely related information that, on the surface, may seem to have little in common. A human editor may miss the connections entirely. Because editorial decisions are highly interpretive, it's predictable that even experienced editors will miss or inaccurately categorize and tag relevant information.

Inxight's e-Content Publishing Platform doesn't miss a thing. It uses the industry's most advanced linguistic algorithms, capable of making valid connections between seemingly disparate items. It's also transnational, capable of culling information in 12 western and four Asian languages.

But productivity savings don't stop at the publisher or aggregator. Rather, the chain accelerates in value as consumers interact with the information.

Increased value of information

Buried under an ever-increasing volume of unfiltered data, readers don't have the time to search and read every piece of news in hopes of finding that rare, relevant kernel of useful information. Using Inxight's e-Content Publishing Platform, however, e-Publishers can boost revenues by increasing overall readership by offering a menu of time saving features such as auto-summarization.

In addition, publishers can add enormous value by personalizing information to meet the demographic profiles or preference lists of their subscribers. Studies show that, even though the Web is viewed as an essentially free medium, people are more than willing to pay subscriber fees to get precise, targeted information that saves them significant search time. E-Publishers wanting to stave off the effects of commoditization on their bottom lines can automatically increase the value of their information through personalization.

Inxight Technology Focus

Inxight's architecture model focuses on the four key requirements involved in identifying, sorting and delivering targeted information from the mass of data residing on both public Web sites and private intranets - creation, collection/aggregation, normalization and distribution.

- Organizing - The process of automatically classifying content into both topics/subjects and entities (companies, people's names, etc.)

- Enriching - The automation of enriching content by applying metatags into the document that embed the characterization of the document's topics, key entities, hyperlinks to related information, and summaries using XML technology.

- Collection/Aggregation - The process of integrating content from multiple, disparate sources, both internal and external, and organizing it into a body of useful information.
- Normalization - The process of processing and refining aggregated information into cohesive search results. Different infomediary sources use different naming conventions and categorizations. For example, a search for auto racing may turn up articles on individual drivers, NASCAR safety regulations, and repaving the Indianapolis Speedway. Normalizing metadata from content means both an intelligent search that recognizes the relationships and contexts of seemingly unrelated articles, as well as rejecting articles that seem to fit search criteria but are only tangentially related.

- Data personalization - The process of sending the right information to the right people, in the right format, according to both search criteria and the format preferences of the user - abstracts and summaries for downloading to mobile devices such as PDAs and Internet appliances; full article with graphics for computer users.

Paying close attention to these core requirements yields a system for searching, categorizing and retrieving information that encompasses the 10 keys that allow users to take full advantage of dynamic content.

The Inxight e-Content Publishing Platform

Inxight delivers a best-of-breed content infrastructure known as the Inxight e-Content Publishing Platform that enables content businesses to fulfill delivery promises for quality over quality, faster access to pertinent information (horizontal and vertical information), and personalization. The Inxight e-Content Publishing Platform enables content businesses to:

- Automatically classify and index content, such as news feeds and web sites, into predefined subject categories.
- Automatically create executive summaries from the context of each article.
- Automatically create and embed hyperlinks of key concepts.
- Provide dynamic annotation on hyperlinks, allowing users to actively see key ‘live' information (such as a company's current stock price or company overview), or jump to related Web sites such as the company home page, SEC (Edgar) database, a financial news page, or a list of related news articles.
- Actively find similar news content to support the “find more like this” function.
- Create a user-definable thesaurus that translates acronyms, industry terms, company names, name aliases, abbreviations and stock ticker symbols into full titles and names.
- Increase accuracy using patented, linguistic pre-processing engines.
- Store category indexes, summaries and key entities via XML output to an industry standard format, allowing for ease of integration with other enterprise applications.

Inxight’s Linguistic Technology

Feature	Description	Benefit
Auto categorization	Automatically categorizes and organizes documents, such as news feeds, word processing files and email, into pre-defined subject categories.	Speeds search by organizing information into logical directories and folders.
Auto-index tagging	Automatically embeds category indexes as metadata into each article or file, or into an index repository.	Creates lasting, reusable value to documents.
Intelligent summaries	Automatically creates intelligent summaries based on the context of an article or file. Configurable - summaries can be defined to have a specific length or genre preference.	Users save time by 'previewing’ documents while searching.
Hyperlinks	Automatically create and embed hyperlinks that identifies and highlights 27 key entities including people, company names and ticker symbols, product names and places.	Users save time by previewing key information (e.g. stock price) or by jumping to a related URL via a hyperlink.
Hyperlink annotation/ menus	Extends the hyperlink function to include dynamic 'annotative” menus, allowing users to "roll-over” a hyperlink and view information or menus related to the hyperlink. Applied to hyperlinks, this can also be used to present executive style "summaries” of Web pages behind each link. Example: The name Oracle Corporation is hyperlinked. The dynamic menu would give the following information: ORCL 30.25 up 0.50 (jump to cnnfn.com) CEO: Larry Ellison (jumps to eWEEK) www.oracle.com (jumps to Web site) Look up on Yahoo! News (jumps to Yahoo!)	Creates a user-friendly dimension to hyperlinks where the user can decide, using a menu, which place he/she wants to jump to, as well as previewing "live” dynamic data, like a stock price, without having to "jump” around.
Article similarity	Compares document with others in the knowledge base to discover related or "similar” documents.	Makes discovery of related articles fast and easy.
Customizable Thesaurus module	Translates acronyms, abbreviations, name aliases and ticker symbols into normalized concepts. Examples: SEC=Securities Exchange Commission, CTO=Chief Technology Officer, President Bill Clinton = William J. Clinton, ORCL=Oracle Corp	Enhances search and hyperlinking functionalities by taking abstract concepts and normalizing them to a common term for easier navigation and search.

Feature	Description	Benefit
Linguistics- based drivers	Uses natural language processing that includes language identification, word stemming, compound word analysis, word and phrase tokenization, noun-phrase concept identification, and part-of-speech tagging. Examples: *Language ID* "Ich bin ein Berliner" = German (Deutsch) "I am a citizen of Berlin" = English *Word Stemming* "selling" = sell; "bought" = buy *Compound Word Analysis* "homeowner" = home owner *Tokenization* "Mr. Kim, an investor's representative, said, 'The stock is undervalued'." [mr kim an investor representative said the stock is undervalued] *Noun-Phrase* "The financial analyst reports are listed on the finance Internet website." [financial analyst report] and [financial Internet website] *Part-of-Speech Tagging* "The merger initiative folded." [merger: adjective][initiative: noun][folded: verb]	When systems know "what you mean", it enhances the quality of categorization, hyperlinking, summarization and finding similar/ related documents. Moreover, the linguistics aspects helps process text in a way that helps a computer better manage information by knowing more than words - but extracting concepts and content.
XML output	Indexes and summaries are created using XML.	For IT administrators, developers and architects, they will benefit by using an standard XML encoding to integrate Inxight applications and output into their system infrastructure.
Java and C/C++ API's	APIs coded in Java and C/C++ exposes key features in Inxight products.	APIs expose Inxight's best- of-breed technologies and core functions. The APIs are designed to allow for ease of integration for all primary features of Inxight products, whether you are building and application or adding functionality into an existing system infrastructure.

About Inxight

Inxight is the leading provider of enterprise software applications for understanding and effectively using unstructured data. Inxight is the only company that provides customers a comprehensive and scalable enterprise solution to organize, analyze and deliver information from any unstructured source in all major languages. Customers include Computer Associates, Factiva, Hewlett-Packard, Inktomi, Intel, Internal Revenue Service, LexisNexis, Lotus, Oracle, Reuters, SAP, SAS, Thomson, Tivoli, Verity and Xerox. The company has offices throughout the United States and Europe. For more information, visit www.inxight.com or call 408.738.6200.

For more information...

To request more information on Inxight products, please contact us at: www.inxight.com/about/request info.html. You may also email Inxight Sales at sales@inxight.com or call 888.414.4949 (US), +44 (0) 1252 761314 (Europe, Africa and the Middle East) or 408.738.6200 (worldwide).

Inxight Software, Inc.
500 Macara Avenue, Sunnyvale, CA 94085
Phone: 408.738.6299 | sales@inxight.com

Inxight Federal Systems
11951 Freedom Drive, Suite 1300, Reston, VA 20190
Phone: 703.251.4429 | sales@inxightfedsys.com

Inxight SmartDiscovery™ Extraction Server (SDX)

Highly Scalable, Distributed Architecture Framework

Inxight SmartDiscovery Extraction Server's comprehensive text analysis tools are designed to meet the needs of both department-level deployments as well as organizations requiring near-real time extraction of terabytes of data via a grid/clustered, multi-blade deployment.

Every day brings with it new demands for text processing. It's now common for clients to need to process hundreds of gigabytes and even terabytes of data — in near real-time.

Inxight SmartDiscovery Extraction Server (SDX) is a complete and highly scalable solution for text understanding in all major languages. Its comprehensive set of advanced text analysis tools include entity, event and relationship extraction, categorization and summarization.

Automatically identifying concepts, people, organizations, places and other information, it provides structure to unstructured text, providing that structure for use in routing, categorization, search and business intelligence applications.

Inxight SmartDiscovery Extraction Server's highly scalable, distributed architecture framework solution is specifically designed to meet the needs of both department-level deployments as well as organizations requiring near-real time extraction of terabytes of data.

Open and flexible SOAP APIs allow for easy integration into any environment, as well as easy integration of third-party crawlers, search engines and metadata repositories.

Publishers can now process massive amounts of data for near real-time financial analysis or news aggregation/portal applications.

Government analysts looking for the needle in the haystack can process millions of classified messages, intelligence reports, blog information, and so forth for further analysis of relationships and events.

ASP applications, including hosted CRM or business intelligence applications can process and classify massive amounts of data according to the products, companies, people, concepts and other entities mentioned in them for near real-time searching availability.

inxight

Inxight SmartDiscovery Extraction Server (SDX) Modules

ThingFinder® Extraction Module

Inxight SmartDiscovery Extraction Server quickly and efficiently reads text to discover the “who,” “what,” “when” and “where” of each document –- creating consistent, useful metadata.

SmartDiscovery Extraction Server’s ThingFinder Entity Extraction Module (also available as an SDK) requires no training, tuning or customer-supplied lists. Leveraging Inxight’s deep linguistic understanding, SmartDiscovery is able to discover entities based on patterns in text, automatically identifying more than 35 named entity types out-of-the-box, including people, places, companies, dates, measurements, currency figures and email addresses.

It can also be extended to recognize customer-specific list-based entity types, such as SKU numbers or project names.

With its deep understanding of natural language, the advanced entity extraction of ThingFinder Professional allows users to further extend their solution by defining custom patterns of tokens in regular expression syntax. It can be used to extract custom entities, relations and events such as chemical compound names or formulae, phrases for sentiment analysis and medication adverse effects. Out-of-the-box entity, relation and event extraction packs are also available for common business and intelligence applications. Contact Inxight for more details.

FAWIZ AL (RABBATI) PURCHASED TEN 1-TON TRUCKS (NFI) AND GETS SMUGGLERS TO CROSS THE BORDER APPROXIMATELY 10-15 KILOMETERS OUTSIDE OF KHASON. RABBATI RECRUITED JAN ANTON KRACZEWKI (AL-KIELBASA) TO WORK FOR HIM. KRACZEWKI IS APPROXIMATELY 53 YEARS OLD, AND 180 CENTIMETERS (CM) TALL. HE DRIVES A FOUR-DOOR 1984 GREEN SUBARU. KRACZEWKI USES HIS BACKGROUND AS AN ELECTRICIAN TO CREATE SOPHISTICATED BOMBS.

Person	FAWIZ AL (RABBATI), RABBATI, JAN ANTON KRACZEWKI (AL-KIELBASA), KRACZEWKI
Vehicle	TEN 1-TON TRUCKS, FOUR-DOOR 1984 GREEN SUBARU
Person_Common	SMUGGLERS, ELECTRICIAN
Measure	10-15 KILOMETERS, 150 CENTIMETERS (CM)
City	KHASON
Weapon	SOPHISTICATED BOMBS
Buy Artifact	FAWIZ AL (RABBATI) PURCHASED TEN 1-TON TRUCKS (NFI)
Travel across Border	SMUGGLERS TO CROSS THE BORDER APPROXIMATELY 10-15 KILOMETERS OUTSIDE OF KHASON
Recruit	RABBATI RECRUITED JAN ANTON KRACZEWKI ((AL-KIELBASA))
Person Appearance: Age	KRACZEWKI IS APPROXIMATELY 53 YEARS OLD
Person Appearance: Height	KRACZEWKI IS 180 CENTIMETERS (CM) TALL
Person Attributes: Vehicle	HE (KRACZEWKI) DRIVES A FOUR-DOOR 1984 GREEN SUBARU
Make Artifact	KRACZEWKI USES HIS BACKGROUND AS AN ELECTRICIAN TO CREATE SOPHISTICATED BOMBS

Categorizer™ Module

Taxonomies can provide a powerful way to browse and retrieve documents based on a company-specific, meaningful, structured representation of information.

Inxight SmartDiscovery Extraction Server’s Categorizer Module addresses every stage of taxonomy creation, management and content categorization to provide a consistent, accurate way to organize and navigate unstructured data, giving users access to the information they need to make informed decisions.

Inxight’s Taxonomy and Categorization Module provides a hybrid approach to taxonomy management, combining learn-by-example with explicit rules creation, leveraging Inxight’s deep understanding of language. An intuitive Taxonomy Workbench makes creation and testing of taxonomies easier. In addition, users can import home-grown and third-party taxonomies with relative ease.

Only SmartDiscovery Extraction Server provides taxonomy management and categorization capabilities within a complete, integrated and powerful information discovery solution.

Summarizer™ Module

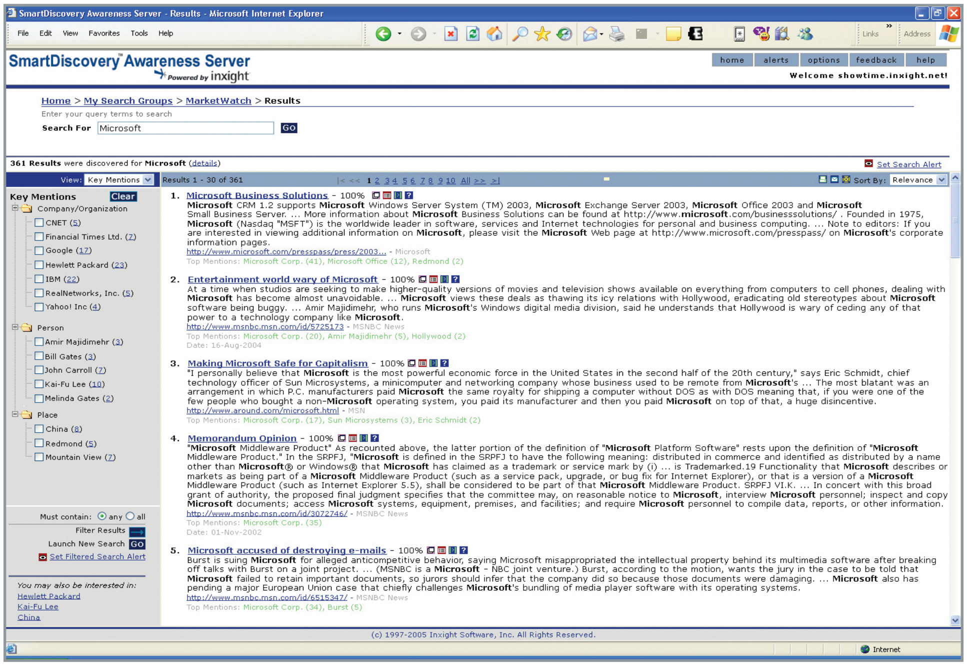
Inxight's Summarizer Module (also available as an SDK) generates accurate abstracts of any document in a fraction of a second, enabling users to scan large sets of information more than 10 times faster than reading the entire text. In a Web-based environment, it automatically summarizes the content of any Web page, so that you can preview a destination before leaving the page. For businesses of all kinds, Summarizer produces increased productivity and substantial cost savings by eliminating hours unnecessarily spent conducting online searches. Summaries can be output for use in alerting, search results display and routing applications.

> Leveraging Inxight's deep linguistic understanding, **SmartDiscovery** is able to discover entities based on patterns in text, automatically identifying more than **35** named entity types out-of-the-box.

Uses for Inxight SmartDiscovery Extraction Server

SmartDiscovery Extraction Server's capabilities enable users to:

- Automatically code documents in near real-time to appropriately route or alert users to relevant documents of interest.
- Create link analysis or business intelligence applications that identify and monitor trends and events mentioned in customer service logs, emails, blogs and other unstructured sources.
- Provide additional value to Documentum and Xerox DocuShare implementations by creating automated attributes for use at check-in time.
- Create applications that augment information search and retrieval operations.
- Add permanent, lasting metadata for future applications and uses.


SmartDiscovery Awareness Server - Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links
Address
SmartDiscovery™ Awareness Server
Powered by inxight
home
alerts
options
feedback
help
Welcome showtime.inxight.net!
Home > My Search Groups > MarketWatch > Results
GO
Set Search Alert
View: Key Mentions
Sort By: Relevance
Key Mentions
Clear
Filter Results
Launch New Search GO
Set Filtered Search Alert
(c) 1997-2005 Inxight Software, Inc. All Rights Reserved.
Internet

Enter your query terms to search

Search For Microsoft

361 Results were discovered for **Microsoft** (details)

Results 1 - 30 of 361 |< << 1 2 3 4 5 6 7 8 9 10 All >> >|

Company/Organization
- □ CNET (5)
- □ Financial Times Ltd. (7)
- □ Google (17)
- □ Hewlett Packard (23)
- □ IBM (22)
- □ RealNetworks, Inc. (5)
- □ Yahoo! Inc (4)

Person
- □ Amir Majidimehr (3)
- □ Bill Gates (3)
- □ John Carroll (7)
- □ Kai-Fu Lee (10)
- □ Melinda Gates (2)

Place
- □ China (8)
- □ Redmond (5)
- □ Mountain View (7)

Must contain: ● any ○ all

You may also be interested in:
Hewlett Packard
Kai-Fu Lee
China

1. Microsoft Business Solutions - 100%
Microsoft CRM 1.2 supports **Microsoft** Windows Server System (TM) 2003, **Microsoft** Exchange Server 2003, **Microsoft** Office 2003 and **Microsoft** Small Business Server. ... More information about **Microsoft** Business Solutions can be found at http://www.**microsoft**.com/businesssolutions/ . Founded in 1975, **Microsoft** (Nasdaq "MSFT") is the worldwide leader in software, services and Internet technologies for personal and business computing. ... Note to editors: If you are interested in viewing additional information on **Microsoft**, please visit the **Microsoft** Web page at http://www.microsoft.com/presspass/ on **Microsoft**'s corporate information pages.
http://www.microsoft.com/presspass/press/2003... - Microsoft
Top Mentions: Microsoft Corp. (41), Microsoft Office (12), Redmond (2)

2. Entertainment world wary of Microsoft - 100%
At a time when studios are seeking to make higher-quality versions of movies and television shows available on everything from computers to cell phones, dealing with **Microsoft** has become almost unavoidable. ... **Microsoft** views these deals as thawing its icy relations with Hollywood, eradicating old stereotypes about **Microsoft** software being buggy. ... Amir Majidimehr, who runs **Microsoft**'s Windows digital media division, said he understands that Hollywood is wary of ceding any of that power to a technology company like **Microsoft**.
http://www.msnbc.msn.com/id/5725173 - MSNBC News
Top Mentions: Microsoft Corp. (20), Amir Majidimehr (5), Hollywood (2)
Date: 16-Aug-2004

3. Making Microsoft Safe for Capitalism - 100%
"I personally believe that **Microsoft** is the most powerful economic force in the United States in the second half of the 20th century," says Eric Schmidt, chief technology officer of Sun Microsystems, a minicomputer and networking company whose business used to be remote from **Microsoft**'s ... The most blatant was an arrangement in which P.C. manufacturers paid **Microsoft** the same royalty for shipping a computer without DOS as with DOS meaning that, if you were one of the few people who bought a non-**Microsoft** operating system, you paid its manufacturer and then you paid **Microsoft** on top of that, a huge disincentive.
http://www.around.com/microsoft.html - MSN
Top Mentions: Microsoft Corp. (17), Sun Microsystems (3), Eric Schmidt (2)

4. Memorandum Opinion - 100%
"**Microsoft** Middleware Product" As recounted above, the latter portion of the definition of "**Microsoft** Platform Software" rests upon the definition of "**Microsoft** Middleware Product." In the SRPFJ, "**Microsoft** is defined in the SRPFJ to have the following meaning: distributed in commerce and identified as distributed by a name other than **Microsoft**® or Windows® that **Microsoft** has claimed as a trademark or service mark by (i) ... is Trademarked.19 Functionality that **Microsoft** describes or markets as being part of a **Microsoft** Middleware Product (such as a service pack, upgrade, or bug fix for Internet Explorer), or that is a version of a **Microsoft** Middleware Product (such as Internet Explorer 5.5), shall be considered to be part of that **Microsoft** Middleware Product. SRPFJ VI.K. ... In concert with this broad grant of authority, the proposed final judgment specifies that the committee may, on reasonable notice to **Microsoft**, interview **Microsoft** personnel; inspect and copy **Microsoft** documents; access **Microsoft** systems, equipment, premises, and facilities; and require **Microsoft** personnel to compile data, reports, or other information.
http://www.msnbc.msn.com/id/3072746/ - MSNBC News
Top Mentions: Microsoft Corp. (35)
Date: 01-Nov-2002

5. Microsoft accused of destroying e-mails - 100%
Burst is suing **Microsoft** for alleged anticompetitive behavior, saying Microsoft misappropriated the intellectual property behind its multimedia software after breaking off talks with Burst on a joint project. ... (MSNBC is a **Microsoft** - NBC joint venture.) Burst, according to the motion, wants the jury in the case to be told that **Microsoft** failed to retain important documents, so jurors should infer that the company did so because those documents were damaging. ... **Microsoft** also has pending a major European Union case that chiefly challenges **Microsoft**'s bundling of media player software with its operating systems.
http://www.msnbc.msn.com/id/6515347/ - MSNBC News
Top Mentions: Microsoft Corp. (34), Burst (5)

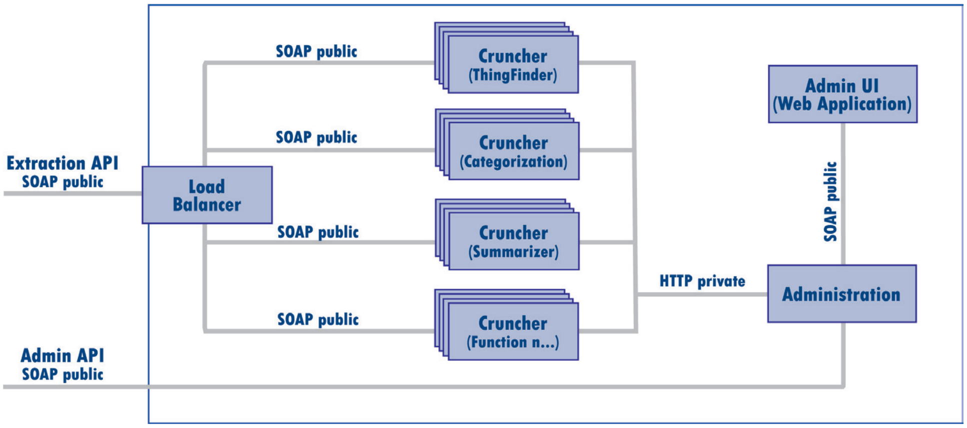
Inxight uses the power of extraction and categorization to power its SmartDiscovery Awareness Server federated search and alert offering.

SDX Architecture

The basic SDX architecture consists of up to 100 or more “crunchers” behind a single load balancer. Administration of the system (managing cruncher nodes, providing some configuration parameters, etc.) is controlled through a simple administration user interface.

These crunchers can provide any linguistic function, including ThingFinder (Entity, relation, event extraction), content filtering, language identification, summarization, and Inxight’s new and improved taxonomy and categorization system.

Open and flexible SOAP APIs allow for easy integration into any environment, as well as easy integration of third-party crawlers, search engines, and metadata repositories.



How SDX Fits Into the SmartDiscovery Ecosystem

SDX is a valuable component of the entire Inxight SmartDiscovery ecosystem, encompassing access, text processing, and text exploration capabilities.

Inxight SmartDiscovery Awareness Server is a federated search solution that enables the SmartDiscovery system to acquire documents of interest on a standing alert or ad-hoc basis from the entire world of available text information – public Web, deep Web (SEC filings, patent databases…), internal sources (Documentum, SharePoint, Google Search Appliances, FAST indexes), and subscription sources (LexisNexis, analyst reports, etc.).

This information can then be fed into SmartDiscovery Extraction Server for further processing, and exposed through SmartDiscovery’s end user applications, including SmartDiscovery Awareness Server Global Scout and Inxight SmartDiscovery Text Analyst (coming soon).

About **Inxight**

Inxight Software, Inc. is the leading provider of enterprise software solutions for information discovery. Using Inxight solutions, organizations can access and analyze unstructured, semi-structured and structured text to extract key information to enable business intelligence. Inxight is the only company that provides a complete, scalable solution enabling information discovery in more than 30 languages. Customers include enterprise companies such as Novartis, Procter & Gamble and Thomson, multiple U.S. and foreign government agencies, including the Department of Defense, Defense Intelligence Agency, Department of Homeland Security and Commonwealth Secretariat, and software OEMs such as SAP, SAS, Oracle and IBM. The company has offices throughout the United States and Europe. For more information, visit www.inxight.com or call 1-408-738-6299 or 703.251.4429.

Technical Specifications

- Industry Standard SOAP APIs for easy integration into any system
- Supports more than 220 file formats, including Microsoft Office documents, PDF, XML, HTML, text and email.
- Browser-based administration
- Reference integrations with several common enterprise systems and repositories, including:
  - -Web, file systems, Microsoft Exchange (Content sources)
  - -Oracle, SQL Server, Mark Logic, and IBM DB2 Viper (metadata storage)

Operating Systems

- Windows 2003
- Solaris 9.0 and 10.0
- Red Hat Linux ES 3.0 and 4.0
- Red Hat Linux AS 3.0

Browsers

- Microsoft Internet Explorer 5.5 and 6.0

Languages

Arabic, Chinese (Simplified), English, Farsi (Persian), French, German, Italian, Japanese, Korean, and Spanish.

Contact Inxight about the availability of other languages.

inxight

Inxight Software, Inc.
U.S. I 500 Macara Avenue, Sunnyvale, CA 94085
Phone: 408.738.6200 I Fax: 408.738.6311
Europe I Centaur House, Ancells Business Park, Ancells Road,
Fleet, Hampshire, GU51 2UN U.K.
Phone: +44 (0) 1252 761314 I Fax: +44 (0) 1252 761315

Inxight ThingFinder®

Automatic Entity Extraction

Inxight ThingFinder automatically identifies and extracts key entities from text data sources, enabling developers to extend the power of software solutions for categorization, link analysis, data mining, business intelligence, customer service, content management and more.

The Inxight ThingFinder SDK (software development kit) “reads” text and automatically identifies and extracts more than 25 key entity types out-of-the-box, such as people, dates, places, companies or other “things” from any text data source, in multiple languages – with no setup or manual creation of rules required.

This ability to automatically identify and classify relevant entities makes Inxight ThingFinder one of the most powerful text analysis and categorization tools on the market.

Providing structure to unstructured information

The Inxight ThingFinder SDK provides robust, open APIs (application programming interfaces) for easy integration into virtually any application that processes textual information, enabling users to:

- ***Find*** all references to products and people in customer service logs and emails – automatically.
- ***Create*** link analysis and business intelligence applications that monitor trends and movements associated with people, places, dates and companies.
- ***Add*** structure to unstructured text documents by identifying and categorizing the most important entities discussed inside the documents.
- ***Mine*** large volumes of text for relevant information and quickly identify trends in data sets.

In addition to the out-of-the-box entity types included with the system, Inxight ThingFinder can be customized to extract other relevant items, such as WatchLists or project names.

The optional Inxight ThingFinder Professional module further extends the power of ThingFinder by allowing developers to define custom entity types using regular expression patterns. For example, developers can add chemical names or bank account numbers, and more.

All of these capabilities have made Inxight ThingFinder an essential tool for hundreds of applications, such as publishing, categorization, link analysis, data mining, business intelligence, customer service, content management and more.

inxight

Inxight ThingFinder — Requires no training sets or manually created rules

Extraction and Categorization

ThingFinder leverages Inxight's true understanding of natural language – language-aware tokenization, part-of-speech tagging and noun phrase identification – to automatically extract and classify all entities.

The proposed merger between Mega, Inc. and CNA Systems, Incorporated, has been postponed, Mega CEO Joe Smith said in an analyst call. "CNA's 1st quarter revenue dropped by 32%, and they lost 23 million dollars," Smith explained. CNA Systems sources blame weak sales in China. CNA shares (CNAI) fell 47 percent to $9.84 on May 12, the first trading day after the announcement.

Company	Mega, Inc., CNA Systems, Incorporated
Date	May 12
Person	Joe Smith
Person Position	CEO
Currency	23000000 USD and 9.84 USD
Measurement	32%, 47%
Country	China
Noun Group	proposed merger, analyst call, 1st quarter revenue weak sales, first trading day

Variant Identification and Grouping

Variant identification and grouping allow Inxight ThingFinder to accurately classify all relevant entities in a document, even one-word entities, and to provide true counts reflecting the number and location of ALL appearances of a given entity. For example, Inxight ThingFinder recognizes that the appearance of the word "Smith" in the example refers to the earlier identified person "Joe Smith."

Canonical Form	Variant Forms	No. of Appearances
Joe Smith	Smith, Mr. Smith, J. Smith	2
CNA Systems, Incorporated	CNA, CNA Systems, CNAI	5

Normalization

Normalization takes much of the guesswork out of metadata creation, search, data mining and link analysis processes by creating standard formats (e.g. ISO) for certain entity types such as dates or measurements.

Entity Found	Normalization
May 12	05/12
$23 million	23000000 USD
CNA Systems, Incorporated	CNA Systems, Inc.

Relevance Ranking

The entities extracted by Inxight ThingFinder are given relevance scores reflecting their importance to the document as a whole, making ThingFinder an essential part of any data categorization solution.

Customization

In addition to the out-of-the-box entity types included with the system, Inxight ThingFinder can be customized to extract other relevant items, such as WatchLists or project names. The optional Inxight ThingFinder Professional module further extends the power of ThingFinder by allowing developers to define custom entity and link types using regular expression patterns. For instance, developers can add chemical names or bank account numbers, and more.

About **Inxight**

Inxight Software, Inc. is the leading provider of enterprise software solutions for information discovery. Using Inxight solutions, companies can better discover, retrieve and connect with unstructured, semi-structured or structured text. Inxight is the only company that provides a complete, scalable solution to enable information discovery in all major languages. Customers include enterprise companies such as Air Products, DaimlerChrysler, Novartis, Purdue Pharma and Thomson, multiple U.S. and foreign government agencies, including the Department of Defense, Defense Intelligence Agency, Department of Homeland Security and Commonwealth Secretariat, and software OEMs such as SAP, SAS and IBM. The company has offices throughout the United States and Europe. For more information, visit www.inxight.com, or call 408-738-6200 or toll-free 1-888-414-4949 (in the U.S.) or +44 (0) 1252 761314 (in Europe).

Technical Specifications

Operating Systems

- Microsoft Windows 2000, XP and 2003 Server with MSVC 6, and 2003 Server with MSVC 7.0
- Sun Solaris 8.0 and 9.0 with GCC 3.2.3 and Solaris 9.0 with Forte 32 bit
- Red Hat Linux ES 2.1 with GCC 2.9.6, AS 3.0 with GCC 3.2.3, and AS 4.0 with GCC 3.4.3
- AIX 5.2 with Visual Age 5.0
- HPUX 11.11 with aCC A 03.37

Available Language Modules

Arabic, Chinese (Simplified), Chinese (Traditional), Dutch, English, Farsi (Persian), French, German, Italian, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Danish, Bokmål, Nynorsk, Finnish

Contact Inxight about the availability of other languages.

Pre-Defined Entity Categories

Entity categories can be customized and new categories may be added. Not all entity types are supported in all languages.

- Address
- City
- Company
- Country
- Currency
- Date
- Day
- Holiday
- Internet Address
- Measure
- Month
- Noun Group
- Organization
- Percent
- Person
  - Position
  - Given Name
  - Family Name
  - Suffix
  - Affiliation
- Phone Number
- Place
  - Regions
  - Political
  - Geographical Areas
- Product
- Social Security Number
- State
- Ticker Symbol
- Time
- Time Period
- Vehicle
  - Make
  - Model
  - Color
  - VIN
  - License Plate
- Year

Availability

Inxight ThingFinder is available as a software development kit or as an Inxight SmartDiscovery service.

*Contact Inxight Sales at **sales@inxight.com** for more information or a demo.*

www.inxight.com

Biological Entity Relationships Skill Cartridge™

Text Intelligence™

Skill Cartridge™ library

Biological Entity Relationships

Hunting for Relevant Information

Efficient information access is a major challenge in biological, chemical and clinical research today. As a result, we are observing a steadily growing demand to integrate information from various sources and across different disciplines in life sciences. However, a large portion of this information is only available from scientific articles and patent documents that are stored in free text format. The volume of this literature is growing exponentially and makes it almost impossible for researchers and scientists to retrieve all relevant information on a specific topic and keep up with current research.

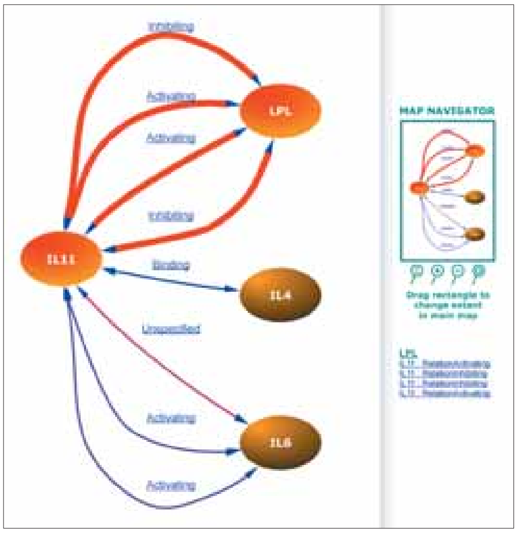
TEMIS Text Mining technology is a powerful and highly accurate solution for the transformation of large collections of literature into readily actionable and domain-specific knowledge.

Mining Biomedical Literature…

TEMIS Text Mining is a powerful technology to manage unstructured data with unparalleled accuracy. A search engine will return thousands of unworkable results from a query, but information extraction will focus on domain-specific and value-added content.

The ***Biological Entity Relationships* Skill Cartridge™** processes each biomedical article to identify and extract meaningful relationships (expression, regulation, activation, etc…) between 10 types of biological entities (i.e. genes, proteins, cells, process, etc.). However, these entities may be referred to very differently and the complexity of the corresponding expressions may range from ambiguous acronyms like "NMBR" to complex combinations of terms like "Corticotropin releasing factor receptor 2". Conventional indexers will not be able to recognize many cases of such expressions in the text. The names of biomedical entities are often composed of several words, furthermore they are not unique and may include several synonyms.

A single protein may be referred to as: Interleukin 6, HGF, HSF, BSF2, IL-6, IFNB2, B-cell stimulatory factor 2, BSF-2, Interferon beta-2, Hybridoma growth factor, IL 6, BSF 2, interleukin6, IFNB 2 or IFNB-2.



Biological Entity Relationships Skill Cartridge™

- A **Skill Cartridge™** is a set of customizable knowledge components describing the information relevant for extraction.
- A knowledge component can be a lexicon and/or an extraction rule.
- An extraction rule describes a sentence structure that characterizes a concept.

For the specific needs of Life Science Research, TEMIS offers the ***Biological Entity Relationships* Skill Cartridge™**. This Skill Cartridge combines the bio-computing expertise of Fraunhofer Institute for Algorithms and Scientific Computing (SCAI) for the recognition of protein and gene entities with the know-how of TEMIS Text Mining technology in a single package.

For more information:

Contact

TEMIS France

>> Headquarters

Tour Gamma B
193-197 rue de Bercy
75582 Paris Cedex 12
France
Tel.: +33 (0)1 40 04 46 70
info@temis-group.com

>> Grenoble

Tel.: +33 (0)4 56 38 24 00
info@temis-group.com

TEMIS Germany

>> Heidelberg

Tel.: +49 - 6221 13753-12
info.de@temis-group.com

TEMIS USA

>> Philadelphia

Tel.: +1 571 235 83 95
info.us@temis-group.com

TEMIS Italy

>> Modena

Tel.: +39 - 059 237 634
info.it@temis-group.com

www.temis.com

Text Intelligence™

The ***Biological Entity Relationships* Skill Cartridge™** performs an identification and extraction of the names of biomedical entities (including genes and proteins) together with their relationships from documents with unequaled relevance. Each organism-specific gene or protein has a unique identifier as well as a root form facilitating an easy linkage into biomedical databases (i.e. Swiss-Prot, Entre-GENE, etc.) and providing a defined way to integrate to existing systems.

The ***Biological Entity Relationships* Skill Cartridge™** is able to identify relationships between biological entities and to view them as interactive html reports or as graphical relational networks. This interactive vizualisation in the CYTOSCAPE tool helps scientists identify biological mechanisms and pathways, validate targets and define new therapeutic strategies for diseases or syndromes. Extraction results can be easily sorted by entity or by relationship type (gene expression, localization, activation, inhibition, regulation, binding, interaction) and highlighted in the context of the original document.

Why choose the Biological Entity Relationships Skill Cartridge™?

The ability to identify and extract biomedical entities and their relations helps to:

- Access and organize all the much needed information about Biological Entities, Mechanisms or Targets,
- Compare one's findings to others,
- Build knowledge from heterogeneous sources and innovative inputs,
- Gain new insights into the molecular foundations of diseases by combining gene/protein interactions with other relevant entities like Disorders, Process, Cells and Tissues.

Discover the TEMIS Product Range

The ***Biological Entity Relationships* Skill Cartridge™** is part of the **Skill Cartridge™ Library**. The collection includes general **Skill Cartridges™** (Analytics, Text Mining 360°) and Skill Cartridges™ that are specific to the Life Sciences (Biological Entity Relationships, Medical Entity Relationships, Chemical Entity Relationships, Competitive Intelligence Life Sciences Edition).

The ***Biological Entity Relationships* Skill Cartridge™** is loaded into **Insight Discoverer™ Extractor**, the TEMIS information extraction server dedicated to the analysis of text documents.
Its results can be used with others TEMIS solutions:

- **Online Miner™**, the Enterprise knowledge portal.
- **Insight Discoverer™ Clusterer**, the automated classification server that dynamically groups documents according to their semantic similarity.
- **Insight Discoverer™ Categorizer**, the document categorization server that automatically classifies unstructured documents into pre-defined categories, combining statistical and linguistic analysis rules.

For more information about TEMIS products please visit: www.temis.com

ABOUT FRAUNHOFER SCAI

The Department of Bioinformatics at Fraunhofer SCAI focuses on two major aspects of modern life science informatics: data management as well as data analysis for biomedical research. The Fraunhofer Institute SCAI is working on applied mathematics, numerical simulation, high performance parallel computing, and bioinformatics. The Fraunhofer Institute is Germany's largest organisation for applied research. It currently maintains 57 research institutes in Germany and other countries with about 13,000 employees.

Contact:
Juliane Fluck
Tel. +49 2241 14-2188
juliane.fluck@scai.fhg.de
www.scai.fraunhofer.de

SCAI
Fraunhofer Institut
Algorithmen und Wissenschaftliches Rechnen

Specifications

>> Operating systems:
- Windows NT, 2000, XP workstation or server versions
- Linux

>> API:
- Java (RMI - Remote Method Invocation)

>> Source languages:
English.

>> Formats :
over 50 input formats (including MS Word, PDF and HTML).

Biological Entity Relationships Skill Cartridge™

INTER PUBLI GRAPHIQUE 01 46 21 14 60

In print edition: Online translators go by the numbers

The Washington Post 2011-02-22 Health&Science P1

Google, Yahoo! BabelFish use math principles to translate documents online

By Konstantin Kakaes
Special to The Washington Post
Monday, February 21, 2011; 10:22 AM

Early one morning in 2007, Libby Casey was trying to do her laundry in a guesthouse in Reykjavik, Iceland. When she couldn't figure out how to use the washing machine, she opened up the instruction manual.

The guide was written in German, which Casey cannot read, so she typed bits of it into an Internet translation tool. "It occurs nobody endlschleudern, however, intercatapults" is one result she got. Stumped, she pressed some buttons and eventually managed to wash her clothes, in an elongated wash cycle that kept her pinned down for three hours.

Libby's quandary will come as no surprise to anyone who has tried to use a computer to translate things. For decades, machine t ranslation was mostly useful if you were trying to be funny. But in the last few years, as anyone using Google Translate, Babel Fish or many other translation Web sites can tell you, things have changed dramatically. And all because of an effort begun in the 1980s to remove humans from the equation.

As the late Frederick Jelinek, who pioneered work on speech recognition at IBM in the 1970s, is widely quoted as saying: "Every time I fire a linguist, my translation improves." (He later denied putting it so harshly.)

Up to that point, researchers working on machine translation used linguistic models. By getting a computer to understand how a sentence worked grammatically in one language, the thought was, it would be possible to create a sentence meaning the same thing in another language. But the differing rules in different languages made it difficult.

Jelinek and his group at IBM argued that by using statistics and probability theory, instead of language rules, a computer could do a better job of converting one language into another. Translation, they basically argued, was as much a mathematical problem as a linguistic one.

The computer wouldn't understand the

The Washington Post

Google, Yahoo! BabelFish use math principles to translate documents online

meaning of what it was translating, but by creating a huge database of words and sentences in different languages, the computer could be programmed to find the most common sentence constructions and alignment of words, and how these were likely to correspond between languages. (Warren Weaver, a mathematician at the Rockefeller Foundation, had first raised the idea of a statistical model for translation in a 1947 letter in which he wrote: "When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols.' ")

The IBM effort began with proceedings from the Canadian parliament, which were published in English and French. "A couple guys drove to Canada and left with two suitcases full of tapes that contained the proceedings," says Daniel Marcu, co-founder of Language Weaver, the first start-up to use the new statistical techniques in 2002.

Jelinek's group began by using a computer to automatically align sentences in the French and English versions of the parliamentary documents. It did this by pairing sentences from the same point in the proceedings that were of roughly equal lengths. If an opening sentence in English was 20 words long but the French opening was two sentences of about 10 words, the computer would pair the English sentence with the two French ones. The IBM researchers then used statistical methods and deductions to identify sentence structures and groups of words that were most common in the paired sentences.

As researchers got hold of more documents and translations of them in different languages, the database of common words and groups of words grew, providing increasing accuracy and nuance. This is the essence of the system today.

Although the IBM group's initiative began more than 20 years ago, it has taken time for computer scientists at IBM and elsewhere to refine those techniques, for computers to become powerful enough to manage the complexity of the many linguistic probabilities (such as multiword phrases and idioms) and for databases to grow large enough - billions of words in various languages - to provide translations nuanced enough to be usable. This is easier when

The Washington Post

Google, Yahoo! BabelFish use math principles to translate documents online

dealing with closely related languages, such as French and Spanish, and with languages that have lots of translated documents with which to build a database. European languages do well in computer translations in part because the workings of the European Union must be published in the 23 "official and working languages" of the EU; these documents can then be used as raw data for researchers.

A major step in computer translation occurred in 2007 - around the time that Libby Casey was struggling with those Reykjavik washer instructions - when Google introduced the first free, statistically based translation software. (Other Web-based translation programs were still using the older linguistic rule-based systems.)

"Suddenly we see enormous progress in this technology because of Google's push," says Dimitris Sabatakakis, chief executive of Systran, one of the oldest computer translation companies. (Systran powered Google Translate until 2007 and is still the engine behind the widely known Yahoo! Babel Fish computer translation service, which now uses a hybrid system combining both statistical and linguistic models for translation.)

All this means that someone such as Michael Cavendish, a lawyer based Jacksonville, Fla., can do human-rights work related to China. "Machine translation has been a godsend for someone like me who has trouble conversing in foreign languages, because I never got a chance to study them in depth," he said recently.

When Cavendish writes documents, e-mails or Twitter posts to communicate with dissidents and others in Chinese, he finds that a computer translation is pretty good - provided he keeps his English simple. So he doesn't go on about "ex post facto laws," he said, but simply says: "China arrested this man today for something that was legal yesterday."

After shunning linguistic system for many years, the statistical translation mainstream is now again embracing grammar and other language-specific rules to capture some nuances and improve accuracy.

The Washington Post

Google, Yahoo! BabelFish use math principles to translate documents online

Experts say that improvements in translation systems are only going to continue as the databases they use grow larger and as computer scientists are better able to incorporate linguistic information. Soon, researchers say, there will be more and better "speech to speech" software, which will allow simultaneous translation in meetings, for instance. The Pentagon is particularly interested in giving deployed soldiers the ability to communicate with locals: One project is focusing on translations between English and Pashto, which is spoken in Afghanistan and Pakistan.

Even as the field rapidly evolves, though, the kind of odd translations that Libby Casey encountered doing her laundry in Reykjavik are unlikely to vanish entirely - as Sandra Alboum recently found out. Alboum, who runs a translation company in Arlington, was perusing a manual for a half-million-dollar steel-manipulation machine that a client of hers had translated, using a computer, from German into English. "Do not step under floating burdens," it said.

She had to check the manual herself to figure out what was meant: "Do not stand under suspended loads."

Kakaes is a writer living in Washington.

View all comments that have been posted about this article.

http://www.washingtonpost.com/wp-dyn/content/article/2011/02/21/AR2011022102191_pf.html http://googlesystem.blogspot.com/2007/10/google-translate-switches-to-googles.html

Google Operating System

Unofficial news and tips about Google

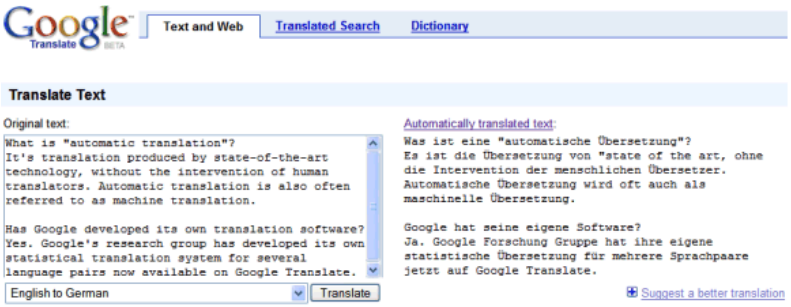
Monday, October 22, 2007

Google Switches to Its Own Translation System

Google switched the translation system from Systran to its own machine translation system for all the 25 language pairs available on the site. Until now, Google used its own system only for Arabic, Chinese, and Russian.

"Most state-of-the-art commercial machine translation systems in use today have been developed using a rules-based approach and require a lot of work by linguists to define vocabularies and grammars. Several research systems, including ours, take a different approach: we feed the computer with billions of words of text, both monolingual text in the target language, and aligned text consisting of examples of human translations between the languages. We then apply statistical learning techniques to build a translation model," explains Franz Och.

You can compare the new Google Translate with Babel Fish, a site that uses Systran to provide translations. The switch is a sign that Google's system has improved a lot and could soon be ready for expanding its coverage.



Translate Text

Original text:

```
What is "automatic translation"?
It's translation produced by state-of-the-art
technology, without the intervention of human
translators. Automatic translation is also often
referred to as machine translation.

Has Google developed its own translation software?
Yes. Google's research group has developed its own
statistical translation system for several
language pairs now available on Google Translate.
```

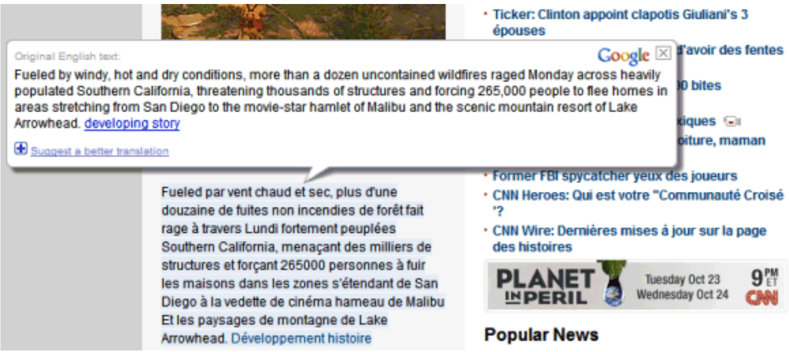
English to German | Translate

Automatically translated text:

```
Was ist eine "automatische Übersetzung"?
Es ist die Übersetzung von "state of the art, ohne
die Intervention der menschlichen Übersetzer.
Automatische Übersetzung wird oft auch als
maschinelle Übersetzung.

Google hat seine eigene Software?
Ja. Google Forschung Gruppe hat ihre eigene
statistische Übersetzung für mehrere Sprachpaare
jetzt auf Google Translate.
```

Suggest a better translation



Original English text:

Fueled by windy, hot and dry conditions, more than a dozen uncontained wildfires raged Monday across heavily populated Southern California, threatening thousands of structures and forcing 265,000 people to flee homes in areas stretching from San Diego to the movie-star hamlet of Malibu and the scenic mountain resort of Lake Arrowhead. developing story

Suggest a better translation

Fueled par vent chaud et sec, plus d'une douzaine de fuites non incendies de forêt fait rage à travers Lundi fortement peuplées Southern California, menaçant des milliers de structures et forçant 265000 personnes à fuir les maisons dans les zones s'étendant de San Diego à la vedette de cinéma hameau de Malibu Et les paysages de montagne de Lake Arrowhead. Développement histoire

- Ticker: Clinton appoint clapotis Giuliani's 3 épouses
- d'avoir des fentes
- 0 bites
- xiques
- oiture, maman
- Former FBI spycatcher yeux des joueurs
- CNN Heroes: Qui est votre "Communauté Croisé "?
- CNN Wire: Dernières mises à jour sur la page des histoires

Popular News

{ Thanks, Steve Rubel. }

Posted by Alex Chitu at 10/22/2007 04:56:00 PM
Labels: Google Translate

Lecture 6.2b Reading 3a (optional)

Reading 3a (optional) gives more detail. Only for those really interested in this topic

• Open source NLP software can help. A representative sampling of open source software in a nice NLP flowchart	45 - 50
• Transforming Unstructured Text into Actionable Data - slides, esp. p. 56 - 59	51 - 68
• ThingFinder Concepts Guide. A detailed description how ThingFinder works and a good introduction to natural language processing (NLP) in general. Start at p. 47 The XeLDA paper (next)may be even better.	69 - 106
• XeLDA: integrate a linguistic engine in your applications Another good introduction to natural language processing (NLP) with many illustrations. Perhaps even better than the ThingFinde paper	107 - 141
• Jaime Carbonell. Natural Language Processing 33 slides with very nice examples. Good overview	143 - 175
• Libby, Elizabeth. Natural Language Processing. Article in the Encyclopedia of Library and Information Science	177 - 190
• Inxight SmartDiscovery Awareness Server Interesting product that uses linguistic technology to post-process Google search results; for examples, extracts information from text into a relational database. Not sure this product still exists	191 - 194
• Going beyond Google. Another document on the same system	195 - 199
• **Several TEMIS products (from the Luxid or the earlier Insight suites) for illustration of applications**	
• Luxid in the medical domain. 18 slides	201 - 218
• Analyzing patent literature to gain competitive insight with Luxid (medical domain) 42 slides	219 - 260
• Insight Discoverer Extractor short	261 - 262
• Insight Discoverer Extractor long	263 - 266
• Insight Discoverer Categorizer short	267 - 268
• Insight Discoverer Categorizer long.	269 - 273

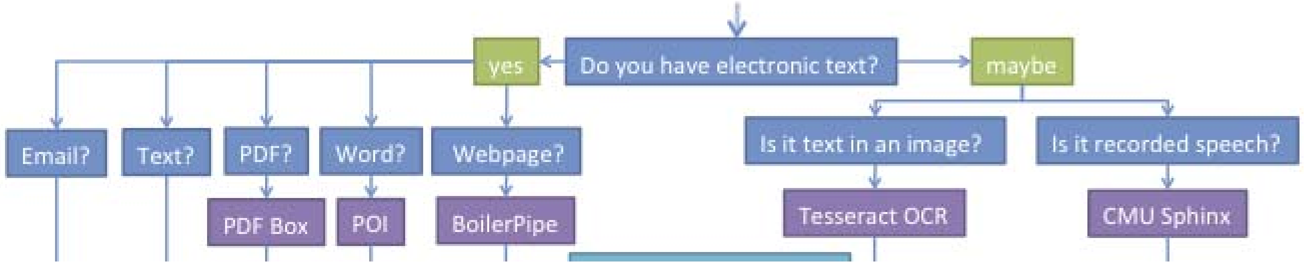
http://entopix.com/so-you-need-to-understand-language-data-open-source-nlp-software-can-help/

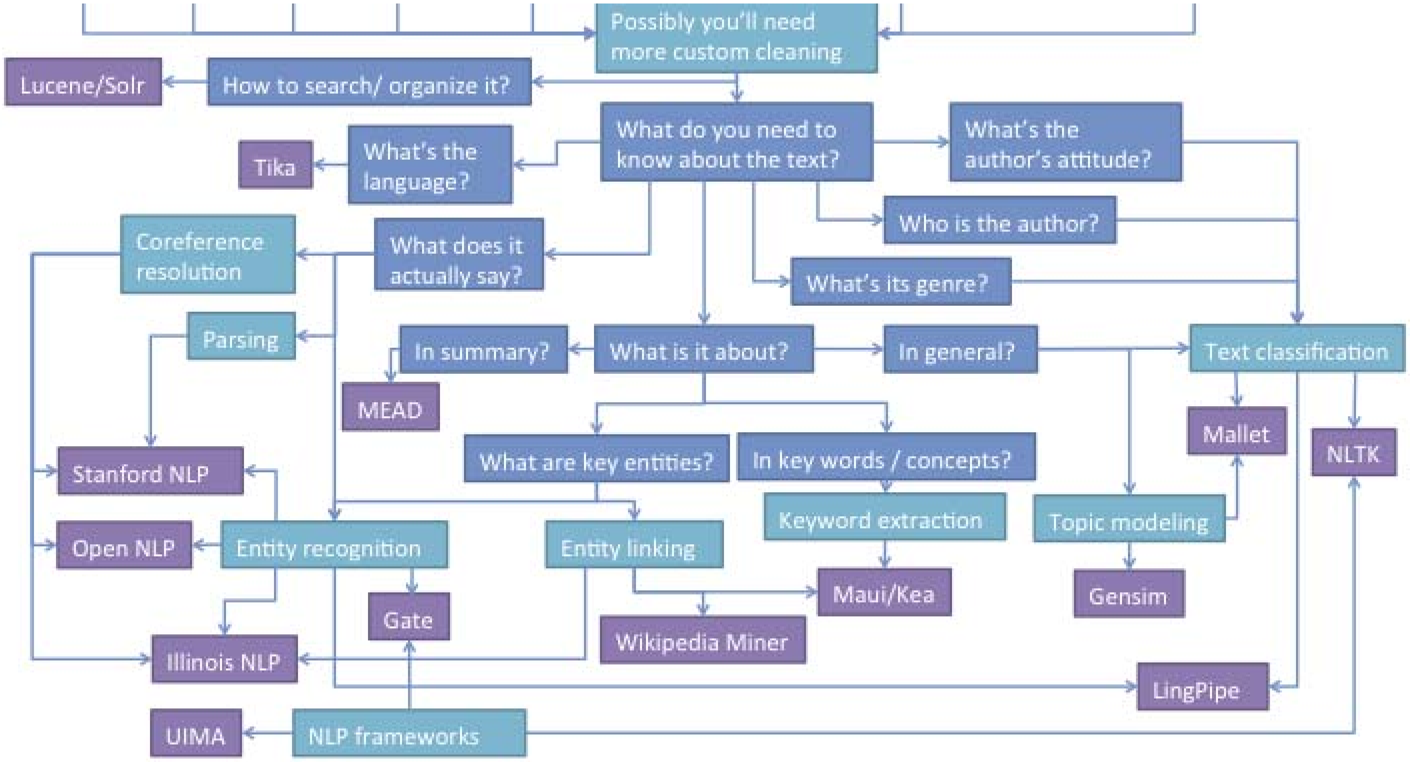
ABOUT SERVICES CASE STUDIES BLOG CONTACT

So, you need to understand language data? Open-source NLP software can help!

july 2, 2014 by zelandiya

Understanding language is not easy, even for us humans, but computers are slowly getting better at it. 50 years ago, the psychiatrist chat bot Elyza could successfully initiate a therapy session but very soon you understood that she was responding using simple pattern analysis. Now, the IBM's supercomputer Watson defeats human champions in a quiz show live on TV. The software pieces required to understand language, like the ones used by Watson, are complex. But **believe it or not, many of these pieces are actually available for free as open-source**. This post summarizes how open-source software can help you analyze language data using this flow chart as a guideline.





If your language data is already available as **text**, it is most likely to be stored in files. Apache libraries like POI and PDFBox extract text from the most common formats. Apache Tika is a toolkit that uses such libraries to extract text and other types of metadata from most types of documents. It has the additional feature of identifying the document language if needed. A great way of organizing and accessing text data is by using a search engine like Apache Solr, which also comes as open-source.

If your language data is on the web, most likely it is part of some other larger website. BoilerPipe helps discard irrelevant text like navigational menu or ads.

If your language data is stored in **images** (note that often PDFs are also simply images of text), you need to use OCR to extract the text. Tesseract can work well out of the box, or can be tuned to get commercial grade quality.

Finally, if your language data is stored as **audio**, you should try CMU Sphinx for converting speech to text. Depending on how diverse the speakers are and what they say, the results may be quite usable. I must add that the support CMU provides via email, forums and IRC is actually better than the support most commercial software vendors offer.

Independently on where the language data comes from, it most likely is what we NLP people call "dirty" and requires cleaning. It may have some idiosyncrasies that will prevent libraries trained on clean text to work well. Cleaning isn't complex work, but it can be time consuming. It's best to automate it with custom-written scripts. It's a bit like with painting a room: the more time you spend preparing, the better the end result.

Now, it all depends on what exactly you need to find out about your text, and what kind of data you already have. For example, methods that fall under **text classification** (also: text categorization) can be quite versatile and powerful. They can be used for detecting spam, guessing genre, estimating overall attitude or sentiment, and can even predict characteristics of the text author, like their age and gender. But these methods also rely on two assumptions: a) the classes are known in advance and b) training data available for each of these classes. The quality of a text classification approach always depends on the number of examples you are training on and how distinct are the classes. Toolkits like NLTK and LingPipe offer text classification as one of their features. But in fact, most machine learning libraries have features that makes them easily applicable to text, e.g. Weka. There is also LibShortText, useful when you analyze short text like tweets.

If you have a lot of data, although none is labeled with classes and in fact you have no idea what classes the data contains, try **topic modeling**, which clusters documents, while providing a series of possible labels (topics) for each cluster. Mallet offers both text classification and topic modelling. Gensim is a scalable topic modeling library in Python.

One step deeper in terms of what we can derive from text, is **keyword extraction** (also: tagging, keyphrase or subject indexing) Compared to text classification and topic modeling, keyword extraction assumes that the number of possible topics (or categories, or tags) can be large (thousands or more), that each document may have a different number of matching topics and that these topics are well-formed phrases. Keywords can be chosen from document text or from a predefined vocabulary, which ensures their consistency. Kea and Maui both extract keywords, but use somewhat different techniques.

Whereas text classification, topic modeling and keyword extraction all summarize the content of the document categorically, it is also possible to summarize by extracting the most relevant complete sentences from text. A **text summarization** platform MEAD allows one to try out different summarization algorithms and evaluate their performance using standard metrics.

Going beyond key categories and phrases, one could extract all those concepts and entities mentioned in text and identify their meaning. **Entity linking** performs this by disambiguating entities to their unique ids in a knowledge base. For example, a technique called wikification links phrases appearing in text to articles in Wikipedia. Wikipedia Miner and Illinois Wikifier both specialize in this task.

If you are interested specifically in names of people, organizations and locations, **entity recognition** (also: named entity tagging, NER) is the technique to use. Illinois Named Entity Tagger, OpenNLP and Stanford NLP all perform this task. If you are not working with documents similar to news articles, make sure to annotate some data by hand and train these tools first.

Going deeper we can determine what the text actually says, i.e. what are the facts expressed in its sentences. When people speak or write, they use pronouns and other ways to refer to the same entities in order to avoid repetition and sound nice. The computer first needs to understand which pronoun and expression refers to which entity. The technique to do this is called **coreference resolution** (also: anaphor resolution). The same three entity recognition libraries offer these capabilities.

The next step is to determine how the entities are connected semantically and syntactically. This is called **parsing**. Stanford NLP is known for its statistical parser that works in many languages, but other libraries, like MaltParser and OpenNLP also offer their versions.

The majority of NLP software is available as a set of Java or Python toolkits. If neither of these langauges is your thing, look out for ports to other languages, such as the Stanford.NLP.Net in F#.

One advantage of using toolkits is that they make it easy to pass the output from one NLP component to another. However, sometimes, you need to combine components from different libraries. UIMA and GATE both mitigate this problem by offering frameworks, which can combine components from different authors, some of which can be open-source and others commercial, into a single systems. Another way to do this, is to use the NLP Interchange Format (NIF) which connects NLP components using an RDF/OWL format for encoding their input/output.

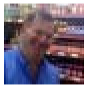
So, why is it possible for such complex software to be free? Researchers at universities are spending decades improving performance of many individual components and publish their results in conference papers. Releasing the software as open-source allows them to improve the state-of-the-art by collaborating on the same task over time and across universities, benchmark solutions against each other and using smaller components to build more complex NLP systems. It also gets researchers more citations. Everyone benefits.

| 5 comments

← Using NLP to deliver relevant content in Online Publishing

5 comments on "So, you need to understand language data? Open-source NLP software can help!"

Pingback: So, you need to understand language data? Open-...



David Medinets says:
July 4, 2014 at 2:27 pm

Very helpful. Thanks.

REPLY

Pingback: open source NPL solutions | Kanteron Secret Blog

Pingback: Procesamiento de Lenguaje Natural y software libre para ello | silta di*IT : thoughts on IT – reflexiones sobre IT

Pingback: So, you need to understand language data? Open-...

Leave a Reply

Your email address will not be published. Required fields are marked *

Name *

Email *

Website

□

□ × 6 = 36

Comment

□

Post Comment

Transforming Unstructured Text into Actionable Data with Business Objects – Inxight Software

Presented by:

Mike Morrow 408-309-7072

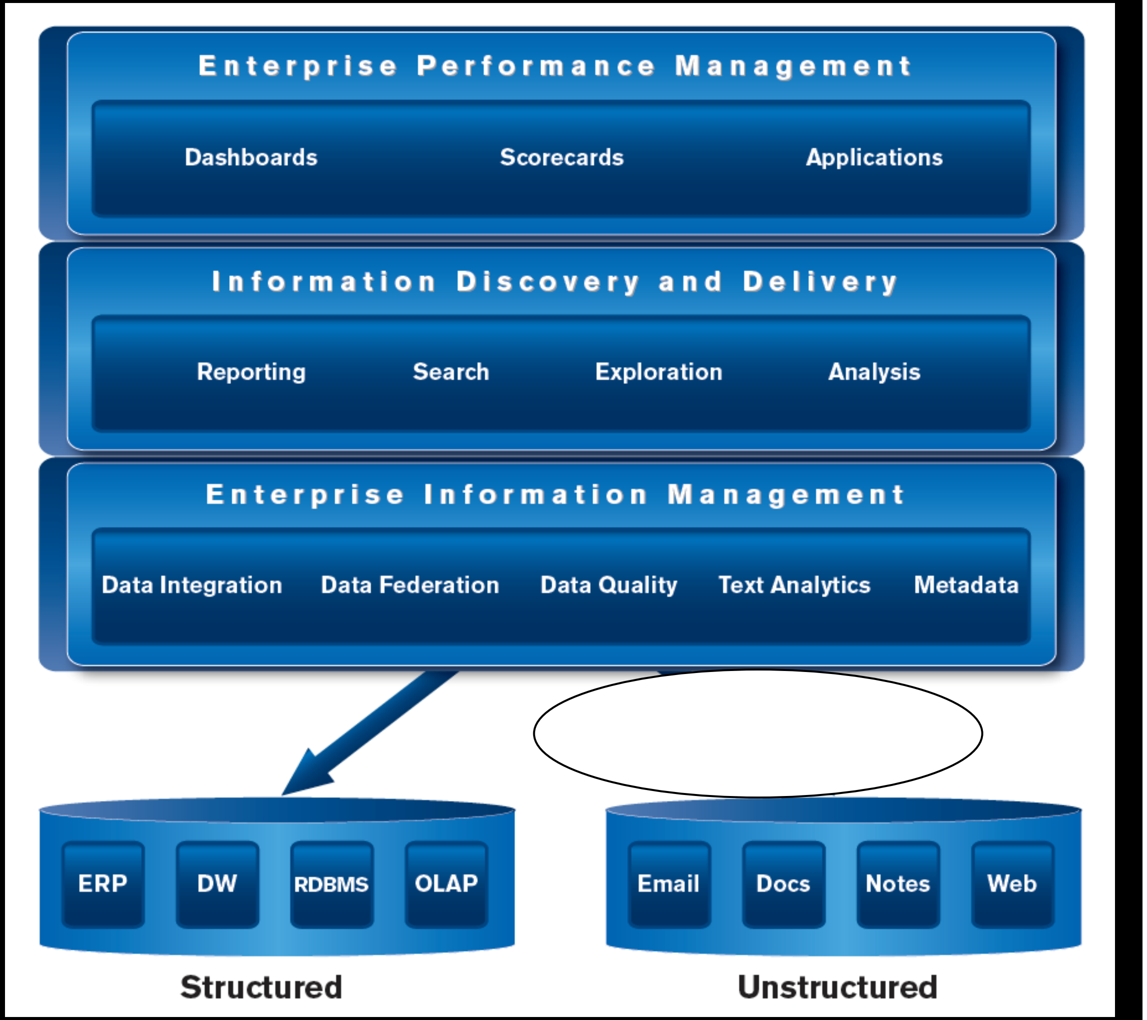
Mark Holly 303-530-0812

AGENDA

- Why Business Objects Acquired Inxight
- What we have - Overview of the Ten(10) Products
- Demo and Architecture Review.
- Questions & Answer Session.

BUSINESS OBJECTS + INXIGHT:

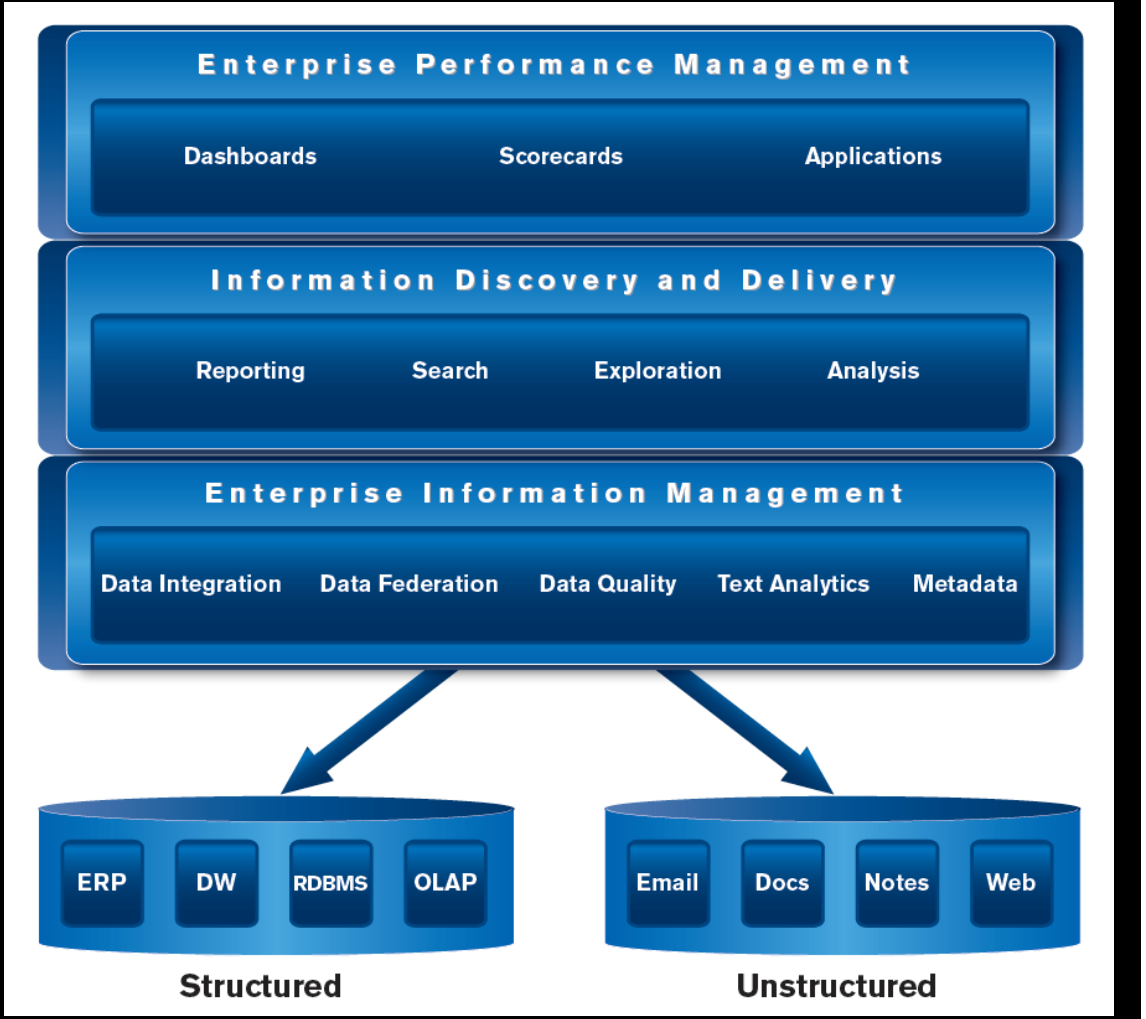
How can your Customers manage their business, when they can't measure it?



BUSINESS OBJECTS + INXIGHT:

How can your Customers manage their business, when they can't measure it?

Business Objects provides the tools to measure all your data!



Inxight's information intelligence and discovery solutions transform unstructured data into timely, actionable information.

Business Objects

At Xerox PARC: The science of Natural Language Processing was begun; The results of over 25 years of Research and Ten years on the market allows you the ability to add features so your customers can:

- Decrease Time spent gathering data
- Make Better Decisions → Faster
- Increase Top Line Revenue by Uncovering New Opportunities
- Reduce costly mistakes and litigation thru better data management

10 Products fall into 4 Categories

- **Text Analytics:**
  - Linguist X Platform
  - Extraction Products
  - Categorizer
  - Summarizer
- **Federated Search:**
  - Awareness Server
- **Data Cleansing:**
  - Text Analysis
  - Metadata Management System
- **Visualizations:**
  - StarTree, TableLens, and TimeWall

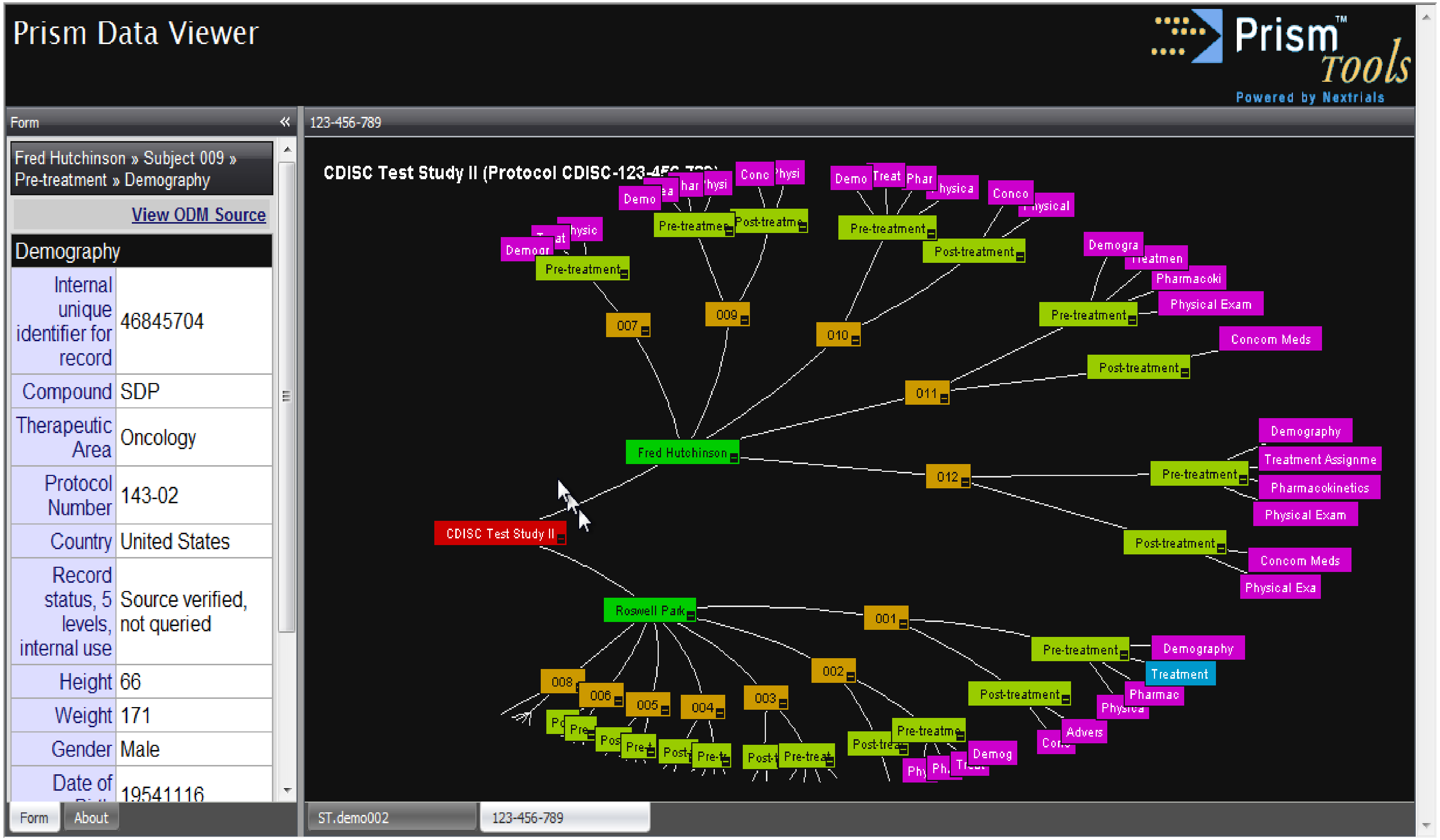
Benefits and Features of the 10 Products

Product	Benefit	Feature
LXP	Hi Accuracy Clean Understanding Mult. Langs. out of the BOX	32 languages POS Tokenization Stemming
Extraction	Reduced time and effort spent identifying Data and relevant links	Easy Rule Writing capability Efficient Name Catalog feature
Categorizer	Reduces manual interaction with large data Sets	Hybrid approach: Learn By Example Name Catalogs Easy Rule Extensibility
Summarizer	Reduces Review Time of large & Complex Documents	Accurately portrays document Highlights
Awareness Server (Federated Search)	Reduces Acquisition Time	Ability cast one single query against multiple sources
Text Analysis	Improved Decision making	Advanced business logic via links between unstructured and structured data

Benefits and Features of the 10 Products

Product	Benefit	Feature
MMS (Metadata Manag.System)	Reduces Errors in complex Unstructured Data	Supports Enterprise or Global Metadata Management – Cleansing of data
StarTree	Reduces time to Action	Reduces complex relationships to easy to understand trees of visually linked data
TimeWall	Decreases time to decision for complex time based business relationships	Depicts complex data on 3-demensional time based data in an easy to understand visual setting
TableLens	Speeds up complex / multidimensional decision making	can show 50,000 rows of complex Multi faceted data

Nextrials



Prism Data Viewer

Form

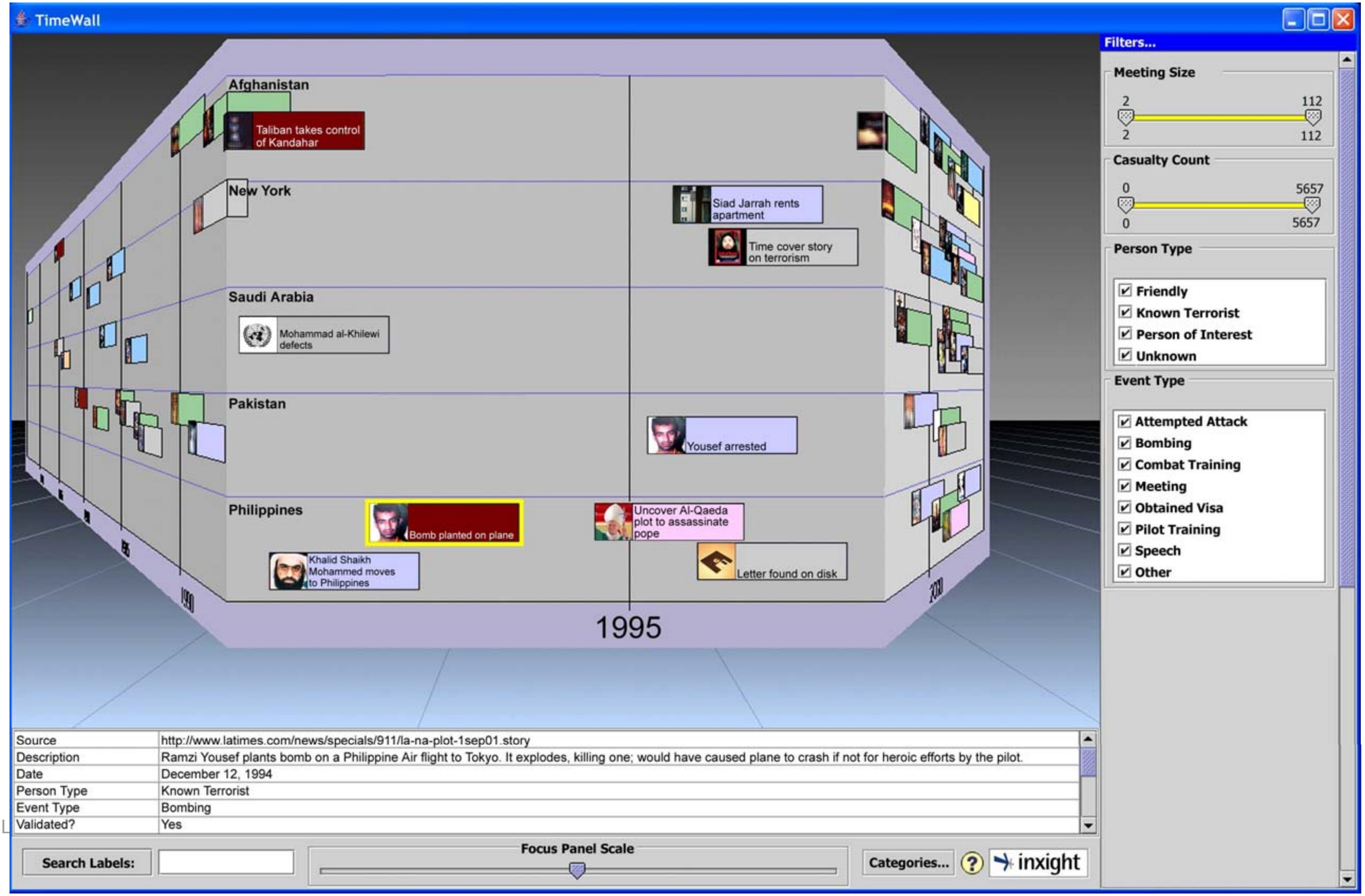
Fred Hutchinson » Subject 009 » Pre-treatment » Demography

View ODM Source

Demography

Internal unique identifier for record	46845704
Compound	SDP
Therapeutic Area	Oncology
Protocol Number	143-02
Country	United States
Record status, 5 levels, internal use	Source verified, not queried
Height	66
Weight	171
Gender	Male
Date of Birth	19541116

TimeWall SDK:
TIMELINES AND STORYBOARDS


TimeWall
Afghanistan
Taliban takes control of Kandahar
New York
Siad Jarrah rents apartment
Time cover story on terrorism
Saudi Arabia
Mohammad al-Khilewi defects
Pakistan
Yousef arrested
Philippines
Bomb planted on plane
Uncover Al-Qaeda plot to assassinate pope
Khalid Shaikh Mohammed moves to Philippines
Letter found on disk
1990
1995
2000
Search Labels:
Focus Panel Scale
Categories...
inxight

Filters...

Meeting Size

2 112

2 112

Casualty Count

0 5657

0 5657

Person Type

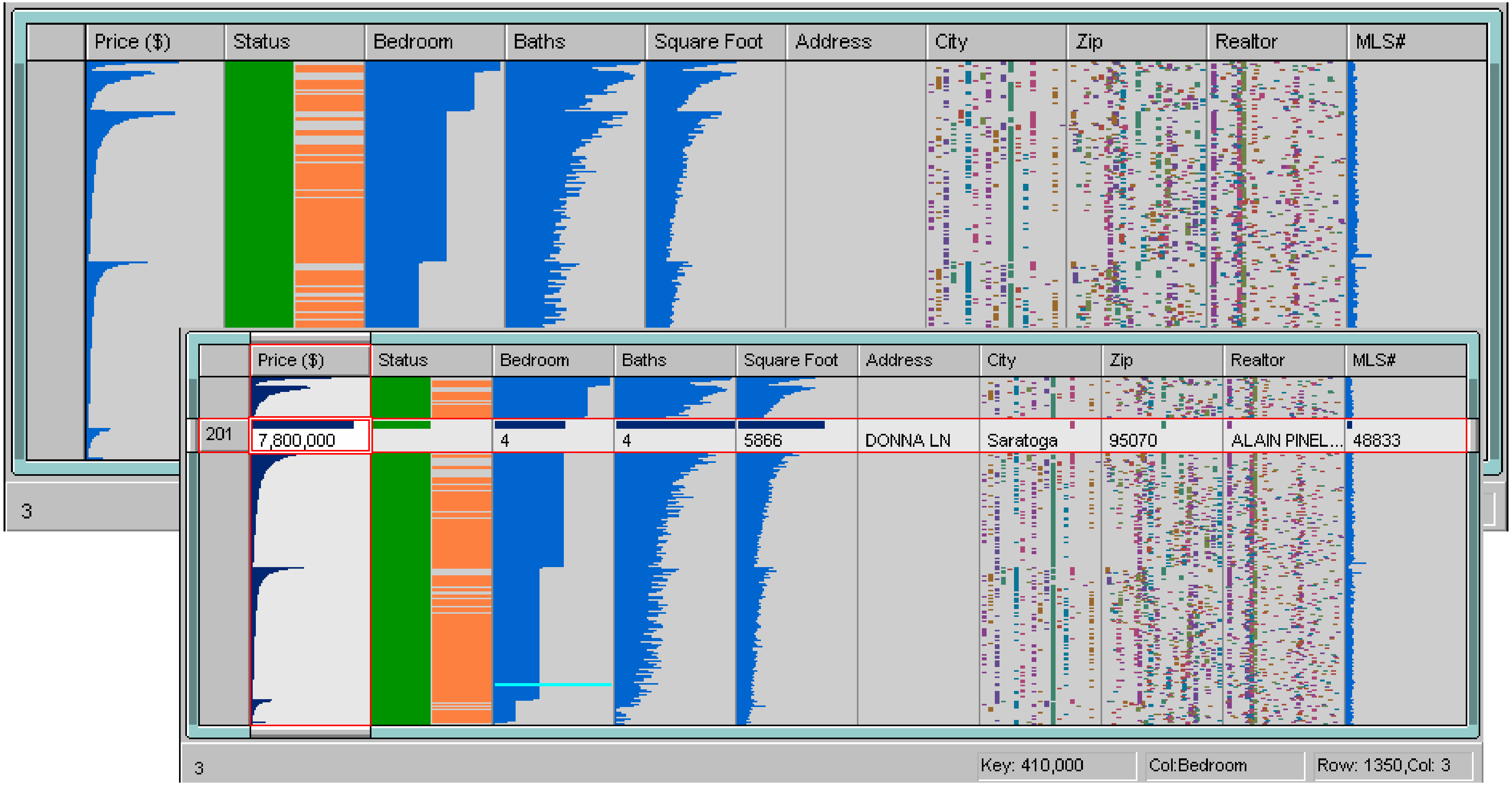
- ☑ Friendly
- ☑ Known Terrorist
- ☑ Person of Interest
- ☑ Unknown

Event Type

- ☑ Attempted Attack
- ☑ Bombing
- ☑ Combat Training
- ☑ Meeting
- ☑ Obtained Visa
- ☑ Pilot Training
- ☑ Speech
- ☑ Other

Source	http://www.latimes.com/news/specials/911/la-na-plot-1sep01.story
Description	Ramzi Yousef plants bomb on a Philippine Air flight to Tokyo. It explodes, killing one; would have caused plane to crash if not for heroic efforts by the pilot.
Date	December 12, 1994
Person Type	Known Terrorist
Event Type	Bombing
Validated?	Yes

TableLens SDK:
TRENDS AND CORRELATIONS



	Price ($)	Status	Bedroom	Baths	Square Foot	Address	City	Zip	Realtor	MLS#
201	7,800,000		4	4	5866	DONNA LN	Saratoga	95070	ALAIN PINEL...	48833

3

Key: 410,000 | Col:Bedroom | Row: 1350,Col: 3

Who can afford to lose 1.5Billion?

- Why didn't Microsoft see this coming?
- Where was the data showing the growing PROBLEM?
- What could have prevented the catastrophe?
  - Extraction from Business Objects

The Next Generation Now
Xbox 360™ is all about a new generation of extreme fun. Get in on the best games, the most powerful system, a worldwide community, and much more.
Learn More >>>
XBOX 360

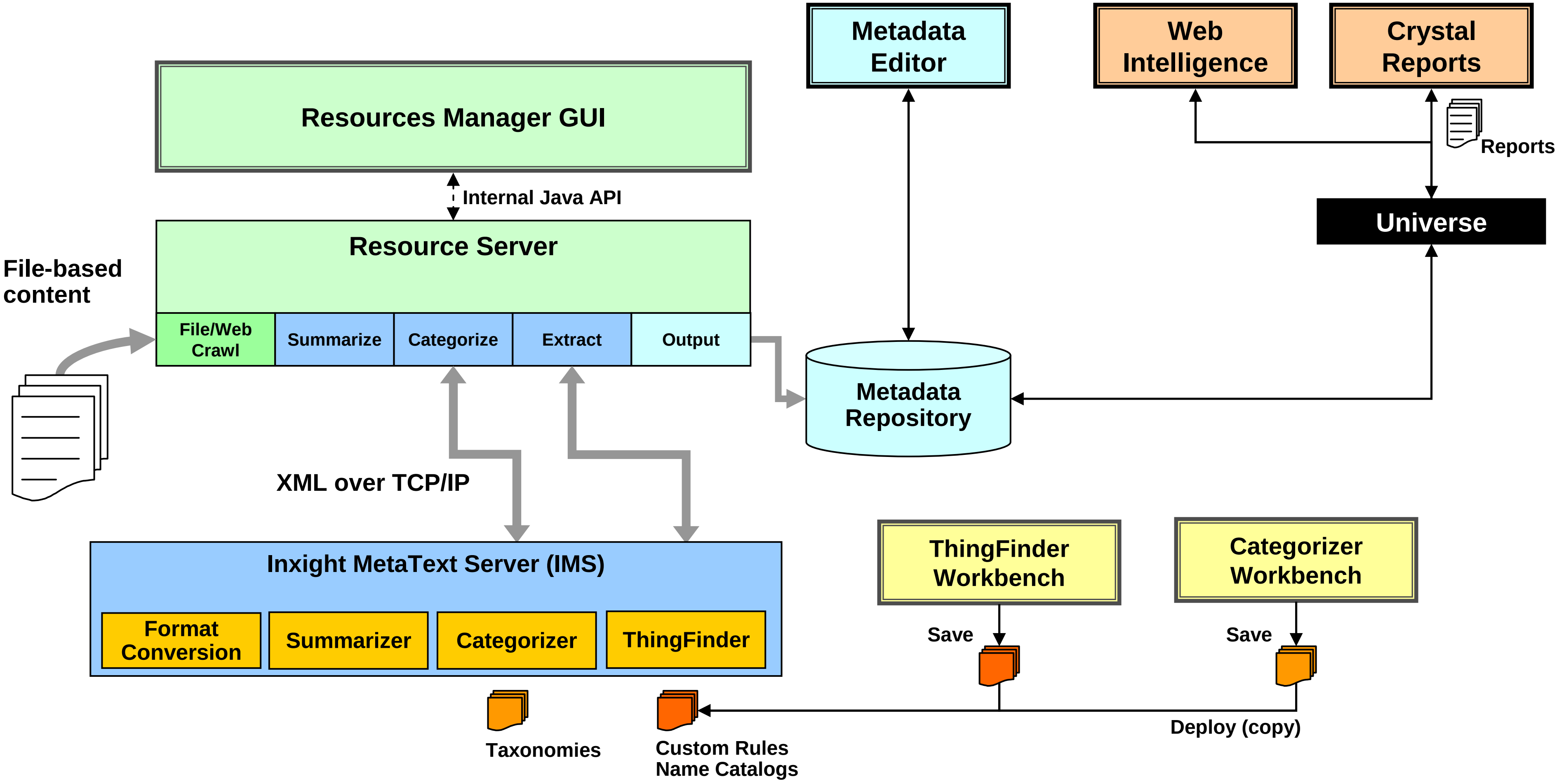
What an Email can cost you?

If your Halliburton and an email contains “copy” in context to a competitor, the judgment could be:

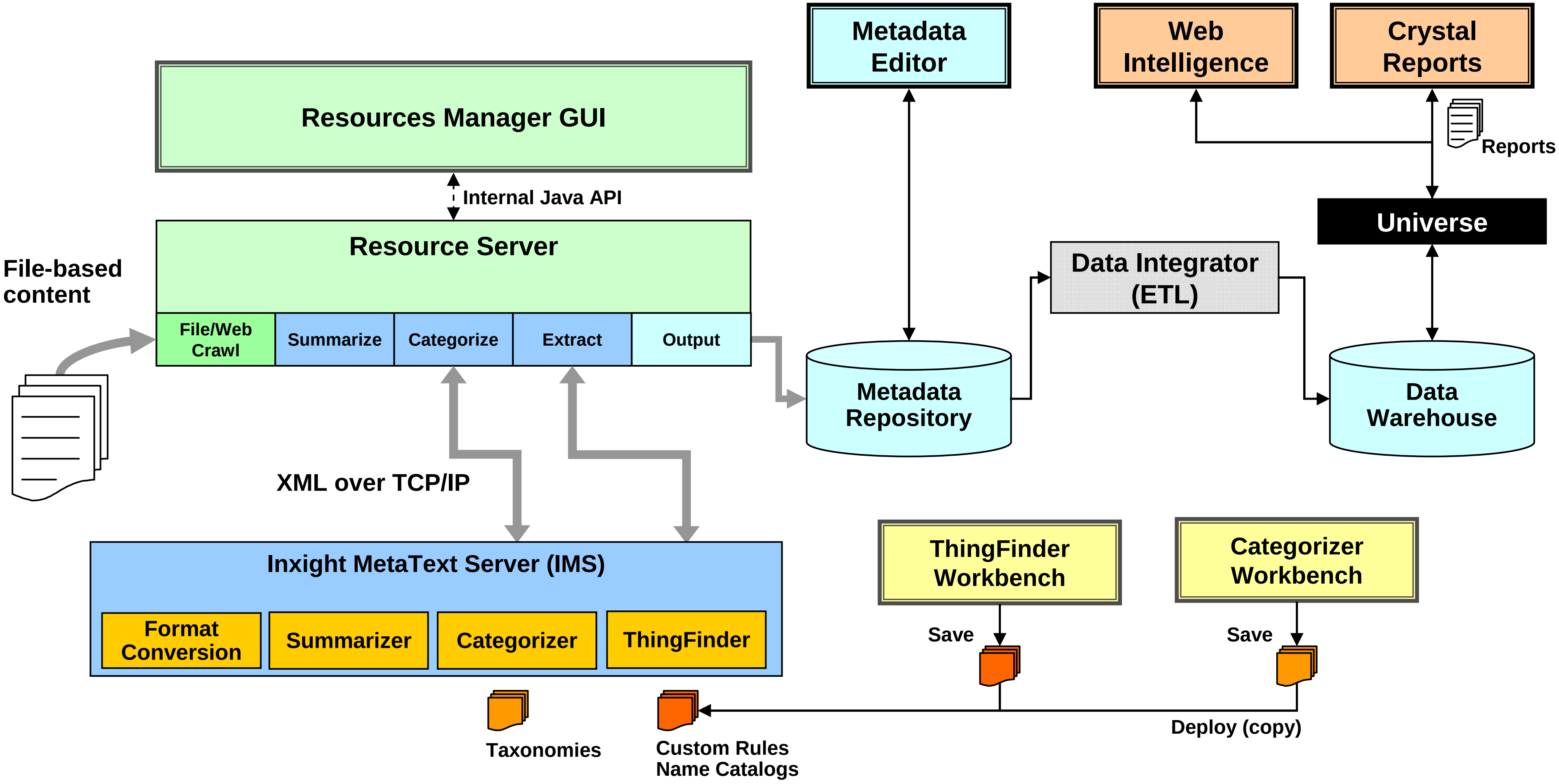
$100,000,000.00

- Obviously you’d want to find it and know your liability before it gets discovered by apposing council
- Components from Business Objects can now empower compliance applications, e-mail Control, and Content Management products

BusinessObjects™ INXIGHT Architecture (for BOE)



BusinessObjects™ INXIGHT Architecture (for BOE) + Data Integrator



APPLICATIONS WE CAN Facilitate

General

- Voice of the Customer (sentiment) analysis, voice of the employee analysis, competitive analysis

Publishing, Media & Entertainment

- Buzz tracking, automated tagging

Financial Services

- Regulatory Compliance, Fraud Detection, Insurance Claims Analysis

Manufacturing

- Warranty Analysis, Contract Analysis, Six Sigma Compliance

Healthcare

- Common Cause Diagnosis, Regional Diagnosis, Patient Sentiment

Education

- Application Analysis, Student Records Analysis, Administrative Records

State and Local

- Constituent Analysis, Claims/Bid Tracking and Analysis

Questions And Answer Session

We can help you get powerful NEW features to market in less time

Schedule a time for a next step discussion

Contact Info:

Mike Morrow 408-309-7072

mmorrow@businessobjects.com

Mark Holly 303-530-0812

mholly@businessobjects.com

BusinessObjects ThingFinder™ Concepts Guide

Downnloaded from http://help.sap.com/help.sap.com/
businessobject/product.../
xi3sp1_TA_thingfinder-4-3.pdf 2014-09-09 p. 53 - 90

Start on p. 8 in extract (61 in full document

BusinessObjects ThingFinder 4.3

Copyright

Third-party contributors

Document version 4.3 IDR-3

Contents

About This Guide

chapter 1

Business Objects ThingFinder™ is a powerful technology for enabling customized extraction applications. Business Objects ThingFinder analyzes text and automatically identifies and extracts more than 35 key entity types out of the box, including people, dates, places, companies or other things from any text data source, in multiple languages. The ThingFinder Professional module extends the power of extraction by enabling the detection and extraction of activities, events and relationships between entities and giving users a competitive edge with relevant information for their business needs.

This guide provides a conceptual framework for understanding ThingFinder, its functions, and its components.

This preface contains the following sections:

- Audience for this Guide
- Organization of This Guide
- Related Documentation
- Technical Support

Audience for this Guide

This guide is written for analysts and application developers working with Business Objects extraction products, including, but not limited to ThingFinder®, ThingFinder Workbench™, Business Objects™ Intelligent Search, and Business Objects™ Text Analysis Server.

Organization of This Guide

This guide contains the following chapters and appendices:

Chapter 2: Introducing ThingFinder—Surveys product features and provides a technical overview of ThingFinder.

Chapter 3: ThingFinder Processing—Describes the tasks ThingFinder performs to process documents.

Chapter 4: Glossary—A glossary of ThingFinder terminology.

Related Documentation

The following documentation contains information related and complementary to this guide:

- *ThingFinder Customization Guide*—Describes how to create and use name catalogs and custom extraction rules to create your own extraction patterns.
- *ThingFinder Language Guide and Reference*—Describes how to configure the ThingFinder language modules and provides reference information for each module.
- *ThingFinder SDK Getting Started Guide*—Describes ThingFinder SDK requirements, installation and guidelines to follow when developing applications based on your operation system platform.
- *ThingFinder SDK Programmer's Guide and Reference*—Describes how to use the ThingFinder C++ API within your programs and provides the API reference.
- *ThingFinder SDK Java API Getting Started Guide*—Describes the installation and configuration process for ThingFinder's Java API.

Technical Support

For all technical support issues, please visit our Customer Support Web site at `http://technicalsupport.businessobject.com`.

Introducing ThingFinder

chapter 2

Business Objects products that perform extraction tasks use Business Objects ThingFinder™ technology. Business Objects ThingFinder™ is a multilingual suite of tools for building high-level applications with entity extraction and information retrieval components. Business Objects ThingFinder analyzes text and automatically identifies and extracts more than 35 key entity types out of the box, including people, dates, places, companies an so on, from any text data source, in multiple languages. In addition, the ThingFinder Professional module extends the power of extraction by enabling the detection and extraction of activities, events and relationships between entities.

Extracting entities from a document tells us what the document is about—the people, organizations, places and other parties described in the document. Extraction involves processing and analyzing text documents, finding entities of interest, assigning them to the appropriate type, and presenting this metadata in a standard format. Extraction applications are as diverse as your information needs. Some examples of relationships and events that can be extracted with ThingFinder Professional include:

- Co-occurrence and associations of brand names, company names, people, supplies, and more
- Competitive and market intelligence such as competitors' activities, merger and acquisition events, press releases, contact information, and so on
- A person's associations, activities, or role in a particular event
- Customer claim information, defect reports or patient information such as adverse drug effects
- Various alphanumeric patterns such as ID numbers, contract dates, profits, and so on

ThingFinder technology goes beyond conventional character matching tools for information retrieval, which can only seek exact matches for specific strings. Thanks to its strong linguistic foundation in Business Objects LinguistX Platform™, ThingFinder uses a deep understanding of the semantics of words. In addition to known entity matching, ThingFinder performs the complementary function of new entity discovery.

To customize entity extraction, ThingFinder enables you to specify your own list of entities in a *name catalog*. Name catalogs enable you to store entities and manage name variation. Known entity names can be standardized using the name catalog. ThingFinder also performs normalization of certain numeric expressions, such as dates.

In addition, ThingFinder Professional adds tools for building custom rules that enable the detection and extraction of activities, events, and relationships between entities. ThingFinder Professional extends the power of extraction offering by allowing users to:

- Define custom relationship and event types that are unique to your data sets and requirements
- Discover entities and concepts based on patterns rather than lists of known entities
- Disambiguate entities using contextual information.
- Leverage ThingFinder's predefined entities and/or customized vocabulary files in your rules

ThingFinder Professional provides a simple and powerful rule writing language to define patterns for discovery and extraction, including regular expression operators, linguistic operators (word stems, part-of-speech tags, phrase and clause boundaries), list matching, input matching filters, case insensitive matching, and much more. Leveraging Business Objects’s deep linguistic analysis, ThingFinder Professional performs pronoun resolution and grammatical function assignment, so custom extraction rules can be targeted for high-precision results or generalized to catch loose associations.

This chapter provides an overview of ThingFinder capabilities, covering the following topics:

- What is Entity Extraction?
- The ThingFinder Solution
- ThingFinder Architecture

What is Entity Extraction?

Entity extraction is the process by which ThingFinder identifies entities in input documents, classifies them according to type, and where possible, normalizes them to a standard format. ThingFinder can extract entities using lists of specific named entities, and it can also discover new entities using linguistic models. Named entities are often proper names, such as the names of specific and unique people, companies, or places. Other specified entity types include currency amounts and dates, among others. ThingFinder classifies each extracted entity by type and presents this metadata in a standardized format along with the entity’s character offset into the document, length, and other attributes.

For ThingFinder, an *entity* is defined as a pairing of a specific name and its type. For example, several entities are given here with their type:

Canada/COUNTRY

Pope John Paul/PERSON

General Motors Corporation/ORGANIZATION

These *entity types* play a crucial role in the definition of an entity. Entity types are used to classify entities extracted from documents and entities stored in a name catalog. ThingFinder contains an extensive set of predefined entity types, including but not limited to:

- ADDRESS
- ADDRESS_ INTERNET
- CITY
- ORGANIZATION
- COUNTRY
- CURRENCY
- DATE
- HOLIDAY
- LANGUAGE
- MONTH
- ORGANIZATION
- PERCENT
- PERSON
- PHONE
- PRODUCT
- PUBLICATION
- SSN
- STATE
- VEHICLE
- YEAR

ThingFinder Professional custom extraction rules can use predefined and custom entity types to specify and extract more complex information based on the relationship between entities, along with the linguistic attributes of the entities (such as stem, part-of-speech, syntactic function, and so on).

The set of supported entity types differs by language; for more information, refer to the *ThingFinder Language Guide and Reference*. You can extend upon existing entity types by customizing a name catalog; refer to the *ThingFinder Programmer's Guide and Reference* for details.

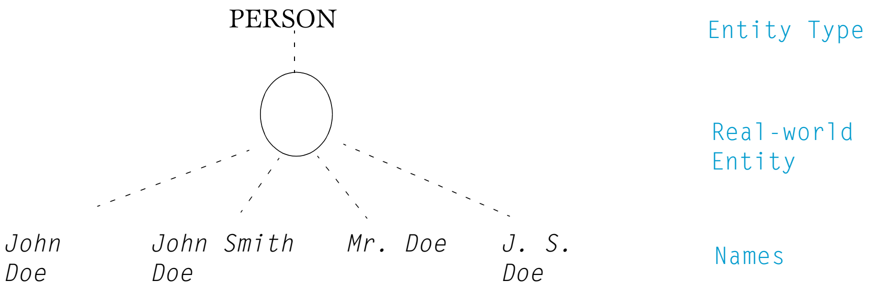
Challenges in Entity Extraction

Knowledge management and information retrieval applications must handle the problems of Name Variation and Ambiguity.

Name Variation

A given entity can be referred to in more than one way. For example, *John Doe* in the sentence *John Doe reported the incident* can be replaced by all the variations shown in Figure 2-1, and potentially others. Similarly, *The United States of America*, *United States*, *America,* and *USA* are various ways to refer to the same country.

Figure 2-1 :Name variation: more than one way to refer to the same entity

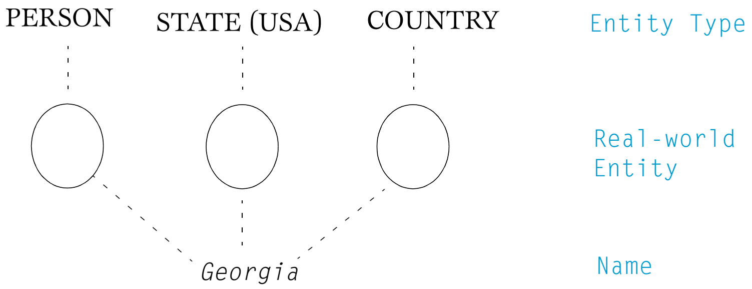


Knowledge management applications must know that these different forms are referring to the same individual.

Ambiguity

A particular name may refer to more than one entity. For example, in the sentence, *Georgia's budget is balanced*, the identity of *Georgia* is ambiguous. Depending on interpretation, the sentence is talking about very different things, as shown in Figure 2-2.

Figure 2-2 :Ambiguity: one name can refer to more than one entity



Knowledge management applications must know which of the possible interpretations is the correct one for the current document.

To summarize, the relationship between real-world entities and their names as found in text documents is many to many. ThingFinder provides tools for recognizing variant names and for distinguishing the different interpretations of ambiguous examples.

The ThingFinder Solution

ThingFinder combines the following tools for recognizing entities, events, and relationships from text documents:

- A set of sophisticated *language modules* to discover entities based on ThingFinder’s inherent knowledge of the semantics of words and the linguistic context in which these words occur.

  The language modules perform more sophisticated linguistic processing than string matching because they combine knowledge about sequences of word tags with intelligence about word semantics.
- Reference to a *name catalog*—a compiled database of named entities, their canonical forms and their common variations.
- *Custom extraction rules* created with ThingFinder Professional—the ability to create custom patterns to extract entities and facts that are specific to your needs.

ThingFinder can use these techniques individually or in combination.

In addition, there are several supplementary operations that ThingFinder uses to enhance its results. For instance, ThingFinder can determine that certain names are aliases of each other and refer to the same entity.

Linguistic Processing

ThingFinder performs linguistic processing by using tools that include semantic and syntactic knowledge of words. In general, linguistic processing identifies paragraphs, sentences, and clauses, and then identifies semantic and syntactical information within the text. Presently, there are two modes for linguistic processing: *standard* and *advanced*.

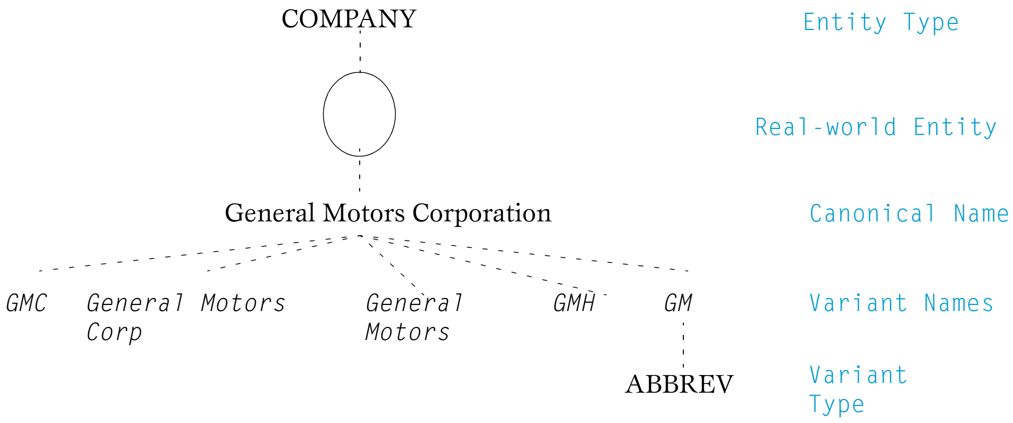
- *Standard linguistic processing*—is available for all supported languages, and is the default behavior for all supported languages.
- *Advanced parsing*—is available for English custom rule-based extraction. Advanced parsing offers richer noun phrase structure, noun phrase coordination, syntactic function attributes, and pronominal resolution. Advanced parsing is used only when custom extraction rules are used.

Name Catalog

A ThingFinder name catalog is an easy to use customization tool that specifies a list of entities that ThingFinder should always extract while processing text. You can use a name catalog to store name variations in a structured way that is accessible through the ThingFinder API. The name catalog structure can help standardize references to an entity.

Name catalogs distinguish between *canonical* and *variant* names. The canonical name is the most standard, complete or precise form for a given entity. For example, *United Parcel Service of America* is the canonical name for that company. A canonical name may have one or more variant names embedded under it. A *variant name* is less standard or complete than a canonical name. For example, United Parcel Service and UPS are both variant names for the same company. While each canonical must have a type, variants can optionally have their own type; for example, you might define a variant type ABBREV to include abbreviations. Figure 2-3 shows the structure of a name catalog entry:

Figure 2-3 :Name Catalog hierarchy



For more information, refer to the *ThingFinder Programmer's Guide and Reference*.

Custom Extraction Rules

ThingFinder Professional extraction rules are patterns written using regular expressions and linguistic attributes that define patterns for the entities, events, and relations you need to find. These rules are written using the

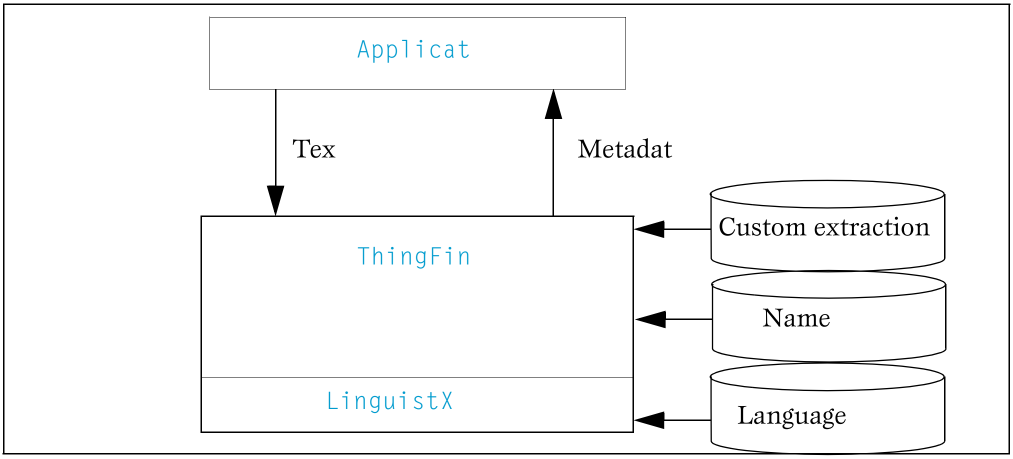
ThingFinder Workbench or CGUL, they are compiled and then they are used by the ThingFinder extraction engine to identify and extract matching patterns from text. You can define and extract the following types of information by writing rules of varying complexity:

- Entities—A pairing of a specific name and its type. For example, *Canada*/ COUNTRY.
- Events—Two or more entities whose relationship indicates an occurrence or a change of state. For example, *John Smith* landed in *Plymouth* in *1675*.
- Relations—Two or more entities that have a specific relationship. For example, *John Smith* met *Pocahontas* in 1675.

ThingFinder Architecture

This section surveys some essential technical aspects of ThingFinder.

Figure 2-4 :ThingFinder architecture



Note: It is not necessary to license Business Objects LinguistX Platform separately for proper ThingFinder operation, as ThingFinder embeds all required components.

The workflow shown in Figure 2-4 is described here:

1. The application sends to ThingFinder the text on which to perform extraction.

2. Using the language modules, ThingFinder performs extraction on the text, normalizes the format of entities where possible, compares these results with the name catalog, applies custom extraction rules, and finally generates a list of the entities to be returned. Optionally, ThingFinder refines its results by performing further ambiguity resolution and grouping aliases together.
3. ThingFinder results are returned to the application as lists of metadata.

ThingFinder Operations

The standard ThingFinder processing algorithm includes the operations summarized below. You can perform some of the operations independently, or you can skip some.

Known Entity Matching

Known entity matching is accomplished by using the following methods:

- *Language modules*—ThingFinder language modules contain known entities, their syntactic and semantic information, and the entity type to which they can belong. You can use a language module alone to perform known entity matching and extraction, or you can use it in conjunction with a name catalog or custom extraction rules, or both.
- *Name catalogs*—Name catalogs contain a given form of an entity, including its canonical name, and its variants and their types. You can use a name catalog alone or you can use it in conjunction with a language module or custom extraction rules, or both.
- *Custom Extraction rules*—ThingFinder Professional custom extraction rules contain patterns that include regular expressions and syntactic information to match specific entities. You must use custom extraction rules in conjunction with a language module. You can also use it in conjunction with a name catalog, which, in addition, can be used within extraction rules.

New Entity Discovery

ThingFinder uses specialized language modules to recognize entities in running text. For example, a noun phrase following *Mr.* is often a PERSON, and a noun phrase preceding *Inc.* is often a COMPANY.

ThingFinder Professional enables you to define your own extraction rules to recognize entities, events, and relations that are specific to your needs.

Entity Construction

ThingFinder compares the results of entity extraction from various language modules, with the records found in the name catalog to determine the entities and assign types to them. In its results, ThingFinder returns information about how it determined the entity and its type, for example, whether the entity was found in the name catalog or some other way. Entity names are standardized where possible.

When ThingFinder cannot determine the type of an entity, it might return more than one type so that the calling application can perform entity disambiguation.

Normalization

Numeric expressions in the categories DATE, CURRENCY, PERCENT, and YEAR are normalized according to the norms of the International Standards Organization, or ISO.

Post-processing

Following the standard processing, there are several optional operations that can refine your results. *Conjecture* determines the type of an ambiguous entity by comparing it with identified entities from the same document. *Aliasing* groups together references to the same entity. *Relevance ranking* provides a score of how relevant an entity is to the overall themes in its containing document.

Precision and Recall

Here we'll briefly review the accuracy measures most often used to describe the correctness and completeness of information retrieval systems, including entity extraction systems—*precision* and *recall*.

Precision and recall statistics assume that, for any question, a system will produce a variable number of answers, some correct and some not. The question posed to an entity extraction system is: *What entities are contained in a given document, and what is the type of each?* We can define precision and recall in these terms as follows:

`Precision` The number of correct answers as a percentage of all answers a system produces.

`Recall` The number of correct answers *actually* produced as a percentage of the total number of correct answers that *can* be produced.

ThingFinder combines processes aimed at increasing both recall and precision. Recall is increased during the generation of candidate entities, while precision is the focus during entity selection.

Output

ThingFinder produces a collection of entities, packaged as an ordered list. The information associated with each extracted entity includes its name, entity type, and its offset, or position, in the text, so that its original text form can easily be fetched. These structures are in memory and are accessible through the API, for use by the calling application. The character encoding of output structures is identical to that of the input. Applications are responsible for processing this metadata and presenting it to users.

ThingFinder Processing

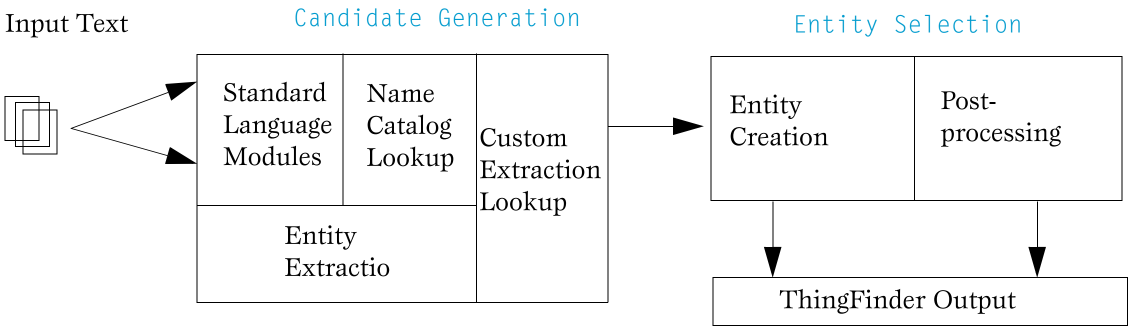
3

chapter

This chapter describes, in greater detail, the process of entity extraction. When performing entity extraction, certain steps are required while others are optional.

Figure 3-5 illustrates the ThingFinder workflow:

Figure 3-5 :ThingFinder workflow



This chapter contains the following topics:

- Candidate Generation
- Entity Selection

Candidate Generation

ThingFinder's first task is to generate candidate entities. As part of this process, ThingFinder processes all input with its language modules. Candidate generation involves the following phases: *pre-processing*, *grouping*, and *name catalog* lookup. If you use ThingFinder Professional, candidate generation also includes *custom extraction lookup*.

ThingFinder *discovers* entities using a process called *Grouping*. This process maximizes recall by finding all candidate entities. The ThingFinder language modules take many factors into consideration during grouping, including the linguistic distribution of word tags, semantic knowledge about words, capitalization, punctuation, and the presence of designator words, such as a title preceding a proper noun.

The *name catalog lookup* maximizes precision if you have created a name catalog containing the entities you want to extract. It can also generate entity candidates when it finds matching entities.

ThingFinder Professional *custom extraction lookup* expands extraction capabilities to also include events and relations, if you have created rules that express the patterns you want to extract. It also generates extraction candidates when it finds matches.

Pre-processing

The first phase in the ThingFinder process is to segment the input text into linguistic chunks that can be analyzed and manipulated. The operations described in this section are part of ThingFinder pre-processing, and this information is intended to help you understand how ThingFinder works. Your application won't do anything specifically related to these operations. Nor can a ThingFinder application directly access these operations or their objects. The output of these operations is available only for internal use by ThingFinder.

Pre-processing consists of the following operations:

- **Segment Generation**—The input text is broken into segments, which are chunks of text, normally containing one or more paragraphs, that are used for further processing. Their size is controlled by the `<maximum-segment-size>` parameter, as described in the configuration chapter of the *ThingFinder Language Guide*.
- **Language and Encoding Identification**—To process text, ThingFinder must know the *language*, *format*, and *character encoding* of the input text. ThingFinder can automatically identify these properties, or you can supply the values, depending on your application.

  For example, when using the ThingFinder SDK, you can specify the values in the parameters of the `TF_Finder` constructor, or you can specify "auto" to instruct ThingFinder to identify the properties automatically.
- **Word Segmentation**—ThingFinder identifies words in the input buffer, including words numbers, abbreviations, multi-word tokens, and punctuation. At the same time, a determination is made for whether punctuation marks should be considered part of a word or separate. For example, the period in *Ms.* is a part of the word, while a period ending a sentence is a separate unit.

Grouping

During grouping, ThingFinder processes the input text using a series of language modules that specialize in the recognition of various tag sequences. These tags are assigned to each token in the text based on contextual co-occurrence, and they form the basis for the grouping operation. As a result, ThingFinder identifies phrases like *big old house*, *crucial question*, *New York City*, *Microsoft Corporation*, and so on.

ThingFinder assigns one or more entity type labels to each discovered phrase. Some labels are very specific, such as PERSON, DATE and VEHICLE_LIC (license plate), while others are generic, labeling an entity as a miscellaneous proper noun (PROP_MISC). Entities with generic types can be identified during a post-processing operation or through the use of custom extraction rules.

If you are using ThingFinder Professional, you can also define custom rules to help you identify extraction patterns that are specific to your requirements. In a second pass after grouping, ThingFinder uses your compiled custom rules, looking for matches. These are output in parallel with the results from the standard processing.

The grouping operation is optional. You can also choose to extract entities using only a name catalog.

Note: The output of custom rules matching can overlap with the output from standard processing and with name catalog only extraction.

Name Catalog Lookup

In this operation, ThingFinder searches the name catalog for matching entries. In addition to the entity candidates generated by the language modules, the name catalog lookup generates entity candidates, and the lookups contribute to the precision of the entity generation. You can use a name catalog to identify lists of known entities, and to match variant names in a standard output format. Because name catalogs are language-independent, this lookup is also independent of the document language. The name catalog lookup is optional.

The name catalog lookup is divided into two types, determined by the presence of wildcards in the name catalog entries. For more information, refer to the *ThingFinder Customization Guide*.

- **Entries with Wildcards**—The name catalog looks up sentences to match entries with wildcards. Sentences form the candidate set for matching with wildcard name catalog entries. Matching entries are returned with the name catalog entity type.
- **Entries without Wildcards**—The name catalog looks up the entire text segment used for processing, without regard to sentence and paragraph boundaries. Matching entries are extracted from the text and returned with the name catalog entity type.

Entity Selection

Entity selection is the process of narrowing down the possible entity candidates to one entity if possible. This process combines the discovery and the lookup results. First, the output of the grouping operation is compared with that of the name catalog lookup; then, these are combined to form non-overlapping entities. A selection is made according to several principles that maximize precision, and entities are returned with their entity types and the method by which the type was assigned.

Note: Entity selection diminishes overlapping entities if possible. However, there are items that return overlapping entities, including custom extraction rules (these do not participate in the selection process at all), some name catalog entities (once name catalog entity can overlap another), and common mentions.

This section also describes three optional post-processing steps that help precision: *aliasing*, *conjecture* and *relevance ranking*.

Selection Principles

This section discusses the principles used in selecting an entity and its type, and how they interact in some possible scenarios. ThingFinder selects entity types according to the following principles:

- Length The longest matching group takes precedence.
- Occurrence in the Name Catalog Occurrence in the name catalog ensures that the entity will be found.
- Linguistic Confidence Some entity types are more likely than other types, and miscellaneous proper nouns are least likely.
- Entity Type Weighting You can assign more weight to specific entity types.
- Filters If you specify filters, then only requested entity types are returned.

The following sections also describe the related topics of Conjecturing and Ambiguity.

Length

As a general rule, ThingFinder gives precedence to the group with the largest scope, i.e. the longest matching group is selected. As an example, consider the scenario where a group is identified along with enclosed *sub-groups*, as in the sentence *He lives at 1600 Pennsylvania Avenue*, which contains the following groups:

[1600 Pennsylvania Avenue]ADDRESS

[1600]ADDRESS_STR_NUM

[1600]YEAR

[Pennsylvania Avenue]ADDRESS_STR

[Pennsylvania Avenue]PLACE_OTHER

[Pennsylvania Avenue]PROP_MISC

[Pennsylvania]STATE

ThingFinder selects the ADDRESS group as the longest group. Additionally, the ADDRESS sub-entities are accessible because their semantics match the containing group:

[1600 Pennsylvania Avenue]ADDRESS

[1600]ADDRESS_STR_NUM

[Pennsylvania Avenue]ADDRESS_STR

Other smaller groups are not relevant in this context and are discarded.

Occurrence in the Name Catalog

After length, the next principle of entity selection in a name catalog is *occurrence*. As a general rule, if you list an entity in your name catalog, it is found during entity extraction. An entity that is found by the language modules but is also found in the name catalog is evaluated as follows:

- If the entity type is `PROP_MISC` and the entity is the same length as the name catalog entry, then the name catalog entry takes precedence.
- If the entity contains smaller entities listed in the name catalog, the entity found by the language modules take precedence over the name catalog entities.

  For example, in the *He lives at 1600 Pennsylvania Avenue* discussed above, [Pennsylvania]STATE is not returned even if it’s in your name catalog. To avoid this, you can use the name catalog alone for extraction, or you can use the category filters to exclude the larger entity type.
- If the entity is ambiguous, then both types are returned.

Linguistic Confidence

For each entity, ThingFinder determines the most likely type. Linguistic confidence comes into play when the language modules detect more than one possible type for an entity with the same length. For example, *Houston* may be a PERSON_GIV, but it is most commonly a CITY.

During entity selection, ThingFinder assigns a confidence score between 1 and 30 to each found entity. This score indicates ThingFinder's degree of confidence in its identification of the current entity. A score of 30 indicates a custom rule match, 1 indicates that the entity is ambiguous, and 2-29 indicate a combination of factors, including language modules, category, proximity of similar entities in the text, and so on.

The following describes the linguistic confidence values and their meanings:

Value	Description
30	Custom rule match
25	Name Catalog entry and ThingFinder match the same type
20	Name Catalog entry only match
15	Name Catalog entry and ThingFinder entity match different entity types
10	ThingFinder entity type only match
9	Strong conjecture to custom entity
8	Strong conjecture to ThingFinder entity
6	Weak conjecture
4	Ambiguous custom entity output
3	Ambiguous ThingFinder entity output
1	PROP_MISC

Entity Type Weighting

ThingFinder lets you assign relative weights to different entity types. When an entity is ambiguous between two different types, the entity type weighting is used to select the final interpretation. For instance, if you know that the current set of documents consists largely of market reports, you weigh towards the COMPANY interpretation.

Note: Entity type weighting only affects the output of the language modules and does not resolve ambiguities involving multiple entries in the name catalog.

Filters

You can use a filter while extracting entities to specify the entity types that you wish to extract. If ambiguous entities are found, they are resolved if possible, and the results are output.

Method and Confidence

ThingFinder indicates the *method* used in entity selection, as one of the following:

Method	Description
Unique	The entity name and category are unique according to the language modules.
Conjectured	The entity name or category were ambiguous, and ThingFinder resolved the ambiguities by comparing several entities.
Name Catalog	The entity name and category were found in the specified name catalog.
Ambiguous	The entity name or category were ambiguous, and ThingFinder was unable to resolve the ambiguity.
Custom Grouper Entity	The entity name and category derived from custom extraction rules.
Custom Grouper Fact	The fact name and category derived from a custom extraction rules.

During entity selection, ThingFinder assigns a confidence score between 1 and 30 to each found entity. This score indicates ThingFinder's degree of confidence in its identification of the current entity. A score of 30 indicates a custom rule match, 25 indicates unique identification, and 1 indicates that the entity is ambiguous, and 2-29 indicate a combination of factors, including category, proximity of similar entities in the text, and so on.

Ambiguity

An entity is ambiguous when its type cannot be identified uniquely. This happens when there is either too little or too much information about the entity.

- There is too little information when an entity is identified only as a proper noun and no type can be assigned. For example, in *The XYZ is funny*, *XYZ* is a noun according to the context, but it can't be further identified. In such cases, ThingFinder classifies the entity in question as PROP_MISC, indicating a miscellaneous proper noun.
- There is more than one possible entity type:
  - If an entity matches more than one entity in the name catalog. The matching entities may or may not belong to the same type. For example, in analyzing *United announced Q3 results*, if the name catalog lists *United* as a variant of more than one entity, e.g. of

United Airlines and *United Steel*, then it will find several entities sharing the name variant *United*. In this case, multiple entities are returned.

- When different entities have the same confidence, then they are returned with an ambiguous reading. The ambiguity can be resolved later via conjecture or by category weighting in the `<language>-tf.config` file. An example of this is the sentence *Georgia's budget is balanced*, in which *Georgia* is ambiguous between a PERSON, STATE, and COUNTRY.

Post-Processing

There are three optional post-processing steps that ThingFinder performs:

- Conjecturing
- Aliasing
- Relevance Ranking

Note: These steps can improve precision or recall. However they are resource-intensive, and therefore they should be turned off, if they are not already.

Conjecturing

ThingFinder *conjecturing* is informed guesswork to resolve the type of an unknown or ambiguous entity, based on comparison with similar entities in the input buffer. Conjecture also searches the name catalog to match unknown or ambiguous entities with the canonical forms of entities stored there. An entity is ambiguous either because it has been labeled PROP_MISC or because more than one entity type has been assigned to it with the same confidence.

During conjecturing, ThingFinder searches the entire input buffer or document, both forward and backward, for similar, co-occurring entities, starting with the longest entities. Word properties such as case, punctuation, titles, etc. are ignored during conjecturing. If the document contains more than one candidate for conjecture, then ThingFinder uses distance to select an entity type. That is, an unknown entity is conjectured to the closest, matching, known entity. When conjecturing an ambiguous entity, ThingFinder only seeks to resolve the ambiguity within one of the types already assigned.

The distinction between strong and weak conjecture reflects the expected accuracy of the applied conjecturing. The more similar two entities are, the more likely it is that they have the same referent.

Consider the PROP_MISC example below and the ThingFinder output for it:

Example Foobar was campaigning for her second term in office. Ms. Foobar, known for her appetite for sushi, gathered her supporters for a benefit dinner.

Output

```
[Foobar]PROP_MISC
[Ms. Foobar]PERSON
```

ThingFinder examines [Foobar]PROP_MISC and finds [Ms. Foobar]PERSON in the same input buffer. ThingFinder then conjectures that *Foobar* is also of type PERSON. [*Foobar*]PERSON is returned with its method marked as *tf_conjectured*.

Consider the following ambiguous example, which assumes that *nation's capital* is listed in a name catalog as a variant name for *Washington DC*:

Example Washington is buzzing with excitement. The nation's capital has a new leader.

Output

```
[Washington]CITY/STATE
[nation's capital]CITY
```

Washington is ambiguous between CITY and STATE, and conjecture seeks to resolve the ambiguity. During conjecture, ThingFinder will compare [Washington]CITY/STATE with the canonical form of the [nation's capital]CITY, which is *Washington DC*, and conjecture that it is a CITY rather than a STATE. This same process may yield incorrect results in certain cases, e.g. the following, where *Washington* should be typed as a STATE:

Example Washington is buzzing with excitement. For the first time, the northwestern state is sending a president to the nation's capital.

Aliasing

An alias is an alternative name for an entity, and aliasing is the process of grouping together text references to the same entity, regardless of their variation in form. For instance, if *William Jefferson Clinton* and *Mr. Clinton* are found in the same input buffer, and they're listed as variant names in the name catalog, then they'll be grouped as aliases of the same entity.

The ThingFinder aliasing capability handles the following variations:

Feature	Description	Aliased Examples
Case	Case is ignored.	REUTERS INC == Reuters Inc
Punctuation	Punctuation is ignored.	Reuters, Inc. == Reuters Inc.
Abbreviations	Abbreviations are matched.	Reuters Incorporated == Reuters Inc
Acronyms	An acronym is formed from the initial letter(s) of each successive word in a named entity and usually refers to the same object as the full form.	I.B.M. == International Business Machines NASA == National Aeronautics and Space Administration
Titles	Titles such as *Col*. and *Dr*. are ignored during aliasing. **Note:** No distinction is made for gender-specific titles like *Mr*. and *Mrs*.	William Clinton == Mr. Clinton == Mrs. Clinton
Name catalog	Any entity discovered in the name catalog is aliased by matching the canonical form. Aliasing groups entities found by the language modules with those in a (user-defined) name catalog. However, the entity type name is case-sensitive.	Big Blue == I.B.M.
Multi-word tokens	An entity containing one or more multi-word tokens is aliased to the same string of characters segmented differently.	Mike Tyson (single multi-word token) == Mike Tyson (two tokens)

When enabled, ThingFinder collects aliases into an alias group. Some alias groups contain only a single reference, indicating that the relevant entity was referred to only once. The reference is nevertheless considered an “alias”, so that each entity is available without the duplication inherent in a full list of entity references.

You can configure how aliasing is performed with respect to titles and abbreviations by modifying the `tf.aliasing-config` file. In that file, you can specify titles and abbreviations that should be ignored for aliasing. For more information about configuring aliasing, refer to the configuration chapter in the *ThingFinder Language Guide and Reference*.

Note: Aliasing doesn’t cover cases such as William==Bill and Robert==Bob.

Relevance Ranking

Relevance ranking measures the relevance of a given entity to the core themes of its document, providing a way to quickly identify those entities that reflect the subject of the document. Relevance ranking requires that entity aliasing has been performed.

The relevance score for a given entity is based on the following factors:

- `Count of coreferential entities` Co-referential entities are counted, and aliased and conjectured entities are assigned the same relevance score.
- `Entity type weight` The entity type weight, if set in the `tf.relevance-config` file.
- `Sentence weight` This takes into account that some sentences are more important than others, such as titles and headings.

You can configure relevance ranking by modifying the `tf.relevance-config` file. Among other things, this file lets you control the weight of different entity types and the weight given to entities occurring in titles and headings. For more information about configuring relevance ranking, refer to the configuration chapter in the *ThingFinder Language Guide and Reference*.

Relevance scores are in the range of -1 and 100. A value of -1 indicates that relevance ranking wasn’t performed because it wasn’t requested or because aliasing wasn’t performed. A value of -1 is also returned for sub-entities, which are not ranked for relevance.

Note: The relevance score and confidence score calculations are independent of each other.

Glossary

4

chapter

Term	Definition
alias	A variant name, or another name for the same entity.
alias group	When several text references refer to the same entity, they are aliases of each other and can be collected into an alias group.
alias list	An alias list contains zero or more alias groups.
aliasing	The process of determining which entities refer to the same objects and are therefore aliases of each other. ThingFinder uses co-occurrence analysis and a name catalog to determine that entities in an input buffer are aliases of one other. When enabled, ThingFinder collects aliases into an alias group. Some alias groups contain only a single reference, indicating that the relevant entity was referred to only once.
ambiguity	Ambiguity occurs when a reference to an entity can be identified with more than one entity type or when there's not enough information to uniquely identify the entity type. For example, a name like *Houston* can be either a CITY or a PERSON_FAM.
canonical name	A canonical name is the standard form for an entity; generally it is the longest, most precise or official name of the object. The canonical name is listed in a name catalog.
CGUL	Custom Grouper User Language. CGUL is a token-based language that enables you to perform pattern matching using character or token-based regular expressions combined with linguistic attributes to define custom rules.
conjecture	The process of determining the type of an ambiguous entity. An entity may be ambiguous either when its type can't be identified or when it is assigned more than one type. ThingFinder performs conjecture, if possible, by comparing the ambiguous entity with known entities that occur in the same document. For example, if a text contains *[Mr. Dixon*]PERSON and later contains [*Dixon*]PROP_MISC, ThingFinder conjectures the second reference as a PERSON.
co-reference	Mention of a given entity in an input buffer, by any of its possible names. *Esther Dyson*, *Dyson*, *Ms. Dyson*, and *Esther* are all possible references to an entity [Esther Dyson]PERSON. A reference may be the canonical name or variant name from a name catalog. All references to the same entity can be collected in an alias group.
custom entity type	Entity types you create by writing rules that define each entity type. Custom entity types enable you to perform specialized extraction, customized to your specific needs.

Term	Definition
disambiguation	The process of resolving ambiguity. Ambiguity arises when there is more than one possible type for an entity or the entity's type is not known.
entity	An entity denotes the names of people, places, and things that can be extracted from text. Each ThingFinder entity is defined as a pairing of a name and its type. For example, [Winston Churchill]PERSON is an entity in which *Winston Churchill* is the entity name and PERSON is the entity type.
entity type	The category an entity falls into. For example, [Janet Reno]PERSON is an entity in which *Janet Reno* is the entity name and PERSON is the entity type. You can define your own entity types by using the name catalog and by writing CGUL rules.
entity sub-type	A hierarchical specification of an entity type that enables the distinction between different varieties of the same entity type. For example, to distinguish Land Vehicles from Air Vehicles.
enumeration	The ThingFinder feature that automatically generates predictable variant names for entities listed in a name catalog. An enumeration tag must be listed in the name catalog entry for the given entity.
event	An event denotes an activity, event, or action that can be extracted from text.
language module	A set of files containing knowledge about a given natural language.
method	The method ThingFinder used to identify the entity and its type.
name catalog	A user-defined repository of information about entities—their canonical name and variant names, their entity types, etc.
name catalog compiler	A tool for compiling files into the name catalog format that lets you add and remove entities, override entity types, add name variants, add and remove entity types, etc.
normalization	The process of normalizing the form of a name to a predefined standard. For instance, ThingFinder normalizes numeric entities in the entity types DATE, CURRENCY, PERCENT, and YEAR only into standard formats defined by the International Standards Organization.
noun group	A NOUN_GROUP is any common noun sequence consisting of two or more related nouns and not identified as a name, measure, or identifier. The entity type NOUN_GROUP is supported in all the ThingFinder language modules.
relationship	A relationship denotes two or more entities that have a specific connection, or a connection between events and their participants, that can be extracted from text.
relevance ranking	Measures the relevance of a given entity to the central themes of the document in which it occurs.

Term	Definition
segment	An unprocessed piece of text from the input document. A segment holds text and metadata for one or more complete paragraphs of the text being processed.
sub-entity	An embedded entity of the same semantic type as the containing entity. The sub-entity has a prefix that matches that of the larger, containing entity. For example, [1600]ADDRESS_NUM is a sub-entity of [1600 Pennsylvania Ave.]ADDRESS, but [1600]YEAR is not, because it doesn't have a matching ADDRESS prefix.
variant name	An alternative name for the same entity. An entity can have zero or more variant names associated with its canonical name, all of which are variants. For instance, *United Parcel Service of America, Inc.*, *United Parcel Service*, and *UPS* are all variant names of the same entity.
variant type	A variant name in a name catalog can optionally be typed. A variant type indicates what sort of name the variant is. For example, to identify a given variant as an abbreviation, you might define a type ABBREV in your name catalog.
word segmentation	The process of breaking input text into its component parts—words, multi-word tokens, numbers, and punctuation marks. This is also known as *tokenization*.

Index

XeLDA®: integrate a linguistic engine in your applications

Modeling language...

Understanding written language is a formidable challenge for information science: the wealth of language requires sophisticated automated processing. Inter and intra-company communication, by definition multilingual, and the increasing volume of documents available in electronic format require high-performance content analysis and management, data/text mining, indexing, authoring/translation support and information retrieval tools.

XeLDA®, the linguistic engine

XeLDA® is a multilingual linguistic engine. It models and standardizes unstructured documents in order to automatically exploit their content. Based on a technology developed through 20 years of research and development, XeLDA® supplies advanced solutions to the issues of processing written information, providing expertise, command of natural language and multilingualism.

Who is XeLDA® designed for?

Are you a solutions developer, software publisher or systems integrator?
XeLDA® enables you to embed the most advanced natural language processing services in your applications. With XeLDA®, you bring true added value to your customers by offering them new products and services, based on accurate understanding of language.

Why choose XeLDA®?

XeLDA® and its range or services have been designed to optimize third-party applications based on a series of advantages:

- ➜ Renowned XFST linguistic technology from the Xerox® Laboratories (Xerox® Finite State Transducers)
- ➜ Client/server or standalone architecture
- ➜ Open and modular architecture
- ➜ Fast processing
- ➜ Unicode-compliant
- ➜ Robust
- ➜ Easy to integrate (API)
- ➜ Results generated in XML format
- ➜ Available in 16 languages

It's a long way **round (adverb)**.

How do you **round (verb)** a number to 2 decimal places with this spreadsheet?

We began a new **round (noun)** of negotiations.

Text Intelligence™

TEMIS

For more information:

Contact

TEMIS France

>> Headquarters

Tour Gamma B
193-197 rue de Bercy
75582 Paris Cedex 12
France
Tel. : +33 (0)1 40 04 46 70
Fax : +33 (0)1 40 04 46 80
info@temis-group.com

>> Grenoble

Tel. : +33 (0)4 56 38 24 00
Fax : +33 (0)4 56 38 24 01
info@temis-group.com

TEMIS Germany

>> Heidelberg

Tel. : +49 - 6221 13753-12
Fax : +49 - 6221 13753-14
info.de@temis-group.com

TEMIS Italy

>> Modena

Tel. : +39 - 059 237 634
Fax : +39 - 059 220 093
info.it@temis-group.com

TEMIS United Kingdom

>> Bristol

Tel. : + 44 (0)117 949 8646
info.uk@temis-group.com

TEMIS USA

>> Washington

Tel. : +1 703 823 57 45
info.us@temis-group.com

www.temis-group.com

Text Intelligence™

How XeLDA® works

XeLDA® offers a scalable range of services based on natural language processing components that you can integrate in your business applications:

>> **Language identification:** automatically recognizes the language used by each document

>> **Segmentation:** divides a text into sentences

>> **Tokenization:** splits a text into basic lexical units

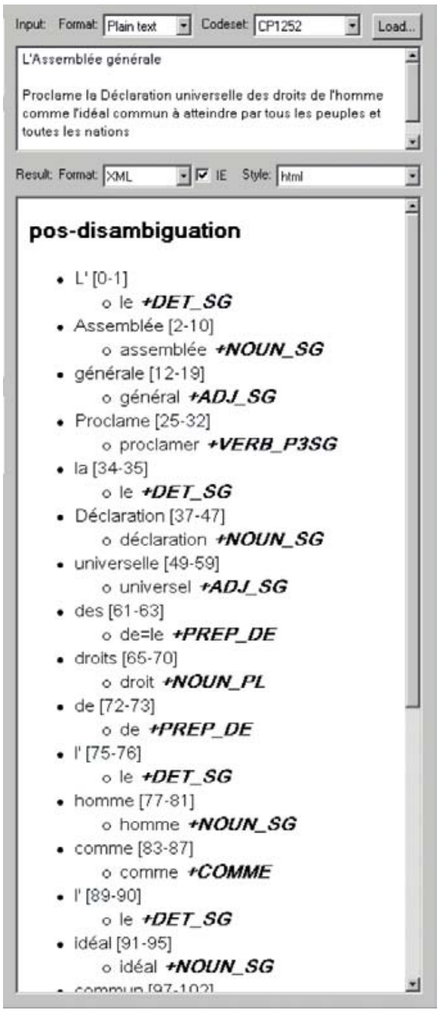
>> **Morphological analysis:** returns the normalized form (the lemma) and the potential grammatical categories for all the words identified during the tokenization stage

>> **Morpho-syntactic disambiguation:** determines the exact grammatical category of a word according to its context

>> **Extraction of noun phrases:** extracts sequences of words that form noun phrases

>> **Dictionary lookup:** identifies the context of a word to find the corresponding dictionary entry

>> **Recognition of idiomatic expressions:** recognizes the expressions found in a text.



Input: Format: Plain text Codeset: CP1252

L'Assemblée générale

Proclame la Déclaration universelle des droits de l'homme comme l'idéal commun à atteindre par tous les peuples et toutes les nations

Result: Format: XML ☑ IE Style: html

pos-disambiguation

- L' [0-1]
  - le ***+DET_SG***
- Assemblée [2-10]
  - assemblée ***+NOUN_SG***
- générale [12-19]
  - général ***+ADJ_SG***
- Proclame [25-32]
  - proclamer ***+VERB_P3SG***
- la [34-35]
  - le ***+DET_SG***
- Déclaration [37-47]
  - déclaration ***+NOUN_SG***
- universelle [49-59]
  - universel ***+ADJ_SG***
- des [61-63]
  - de=le ***+PREP_DE***
- droits [65-70]
  - droit ***+NOUN_PL***
- de [72-73]
  - de ***+PREP_DE***
- l' [75-76]
  - le ***+DET_SG***
- homme [77-81]
  - homme ***+NOUN_SG***
- comme [83-87]
  - comme ***+COMME***
- l' [89-90]
  - le ***+DET_SG***
- idéal [91-95]
  - idéal ***+NOUN_SG***
- commun [97-102]

Discover the TEMIS product range

XeLDA® is the technology used at the core of Insight Discoverer™ Extractor, the information extraction engine, and XTS, the corporate terminology suite.
Insight Discoverer™ Extractor offers solutions tailored to your company's different business areas using Skill Cartridges™.

Specifications

>> Operating systems and compilers supported:

- Windows NT, 2000, XP with Microsoft Visual C++ 6
- Solaris 9 with Sun Compiler Forte 7
- Linux RedHat 8 with gcc 3.2
- Java Runtime Environment 1.3 and higher

>> API :

- C++, Java,
- documentation.

>> Source languages:

Czech, Dutch, English, French, German, Greek, Hungarian, Italian, Polish, Portuguese, Russian, Spanish.

New ▶ Nordic Language Pack: Danish, Finnish, Swedish, Norwegian (Bokmal).

Other languages are being developed.

>> Bilingual dictionaries:

Dutch, English, French, German, Italian, Spanish.

XeLDA®

For more information:

Contact us
- By e-mail:
info@temis-group.com
- Or visit our Web site :
www.temis-group.com

TEMIS France
>> Headquarters
Tour Gamma B
193-197 rue de Bercy
75582 Paris Cedex 12
France
Tel. : +33 (0)1 40 04 46 70
Fax : +33 (0)1 40 04 46 80
info@temis-group.com

TEMIS France
>> Grenoble
Tel. : +33 (0)4 76 61 51 83
Fax : +33 (0)4 76 61 41 21
info@temis-group.com

TEMIS Germany
>> Heidelberg
Tel. : +49 - 6221 13753-12
Mob. : +49 - 174 6595 713
Fax : +49 - 6221 13753-14
info.de@temis-group.com

TEMIS Italy
>> Modena
Tel. : +39 - 059 237 634
Mob. : +39 - 349 4131 718
Fax : +39 - 059 220 093
info.it@temis-group.com

TEMIS UK
>> Bristol
Tel. : + 44 (0)117 949 8646
info.uk@temis-group.com

Text Intelligence™

TEMIS

XeLDA®

white paper

Contents

About This Overview

The Xerox Linguistic Development Architecture (XeLDA) is a toolkit for developing custom linguistic applications. The XeLDA engine can be used to transform, normalize, and extract information from text.

This overview introduces the XeLDA linguistic services and the modes in which XeLDA applications can operate.

Audience

This overview is designed for researchers and developers in the fields of computing and linguistics who understand linguistic terminology.

Conventions Used

This overview uses the following style conventions:

- `Monospaced font`: this typeface is used for any text that appears on the computer screen or text that you should type. It is also used for file names, functions, and examples.
- *`Monospaced italic font`*: this typeface is used for any text that serves as a placeholder for a variable. For example, *`dataType`* indicates that the word *`dataType`* is a placeholder for an actual data type, such as `CString`.
- Internal cross references: this format is used to indicate cross-references within the manual. If you are working on an electronic copy of the manual, click the cross-reference to go directly to the section it references.

What This Manual Contains

This overview contains the following sections:

- Introduction to XeLDA
  Provides a brief introduction to XeLDA's linguistic services, the languages supported by XeLDA, and the environments in which XeLDA runs.
- Linguistic Services
  Provides an overview of each of XeLDA's linguistic services, including examples of each.
- Overview of XeLDA Architecture
  Describes the two modes in which XeLDA can operate.
- Developing Application With XeLDA
  Introduces the XeLDA toolkit for developing natural language processing applications.
- Glossary of Terms
  Contains definitions of the terms used in this document.

Other XeLDA Documentation

In addition to this overview, the documentation set for XeLDA includes the following manuals:

- *XeLDA Installation Guide:* provides procedures for installing XeLDA.
- *XeLDA C++ API Programmer's Guide*: provides information about creating custom natural language processing applications using the XeLDA C++ API.
- *XeLDA Java API Programmer's Guide*: provides information about creating custom Java applications using the XeLDA Java API.
- *XeLDA C++API Reference Manual*: provides a complete reference to all of XeLDA's public C++ classes.
- *XeLDA Java API Reference Manual*: provides a complete reference to all of XeLDA's public Java classes.
- *XeLDA Tagsets*: for each language supported by XeLDA, a file describing the Part-Of-Speech tags is available in the directory `docs` of the `xelda` installation directory.
- *XeLDA Customization Guide*: provides detailed information about building custom XeLDA services.
- *XeLDA Server User's Guide*: describes the server shipped with XeLDA.
- *XeLDA Command-Line Client User's Guide*: describes the client shipped with XeLDA.

Introduction to XeLDA

XeLDA is an engine for transforming, normalizing, and extracting information from text.

It is a comprehensive set of tools, which can be incorporated into applications written in Java or C++ to provide text processing in a number of different natural languages by applying linguistics theory and science.

The Advanced Technology and Systems (ATS) group at XRCE in Grenoble, France, created XeLDA from research by the Multilingual Theory and Technology (MLTT) group. XeLDA stands for Xerox Linguistic Development Architecture.

XeLDA was initially supported by the Multilingual and Knowledge Management Solutions (MKMS) division of the Xerox Innovation Group (XIG).

Following the acquisition by the TEMIS group of the entire business activity of XeLDA (product development, marketing and sales), the whole product is now fully supported by TEMIS, starting from July 2nd, 2003.

Facilities

The facilities provided by XeLDA, called *linguistic services*, include:

- **Language identification**: recognizes the language used by a selected text.
- **Tokenization**: divides the selected text into words.
- **Morphological analysis**: provides the normalized form and all potential part of speech categories for each word identified during tokenization.
- **Part of speech disambiguation**: finds the correct grammatical category of a word according to its context within a text.
- **Noun phrase extraction**: identifies a sequence of words that behaves together as a noun.
- **Dictionary lookup**: retrieves a word’s context and uses this context to find the correct entry in the dictionary.
- **Idiom recognition**: recognizes idiomatic expressions in a text.
- **Relational morphology**: groups words according to their derivational family.

Each linguistic service is described in detail later in this document.

Languages

The XeLDA architecture can handle most written languages.

Most XeLDA services support West-European languages and several East-European languages.

Middle East and more East-European languages are in active development.

Environment

XeLDA runs in the Intel/Linux, Sparc/Solaris and Windows environments and can be tailored for others.

XeLDA is written in C++ but can also connect to applications written in Java.

To develop a Windows application that uses the XeLDA SDK, you must work on Windows 2000 and compile using the Microsoft Visual C++ Compiler version 6.0.

For Solaris applications, use Sun Solaris 8 and compile using the Sun WorkShop 6 update 2 C++ Compiler.

For Linux applications, use RedHat 8 and compile using the GNU gcc 3.2 C++ Compiler.

Architecture

XeLDA acts as a *server* to a client application and is based on an open architecture that lends itself to the integration of future Xerox proprietary Finite State research results as well as third party linguistics components.

XeLDA has two operating modes: client-server and stand-alone (or monoprocess) mode. In client-server mode, all linguistic processing takes place on the server machine. In stand-alone mode, the client and server are merged into a single executable.

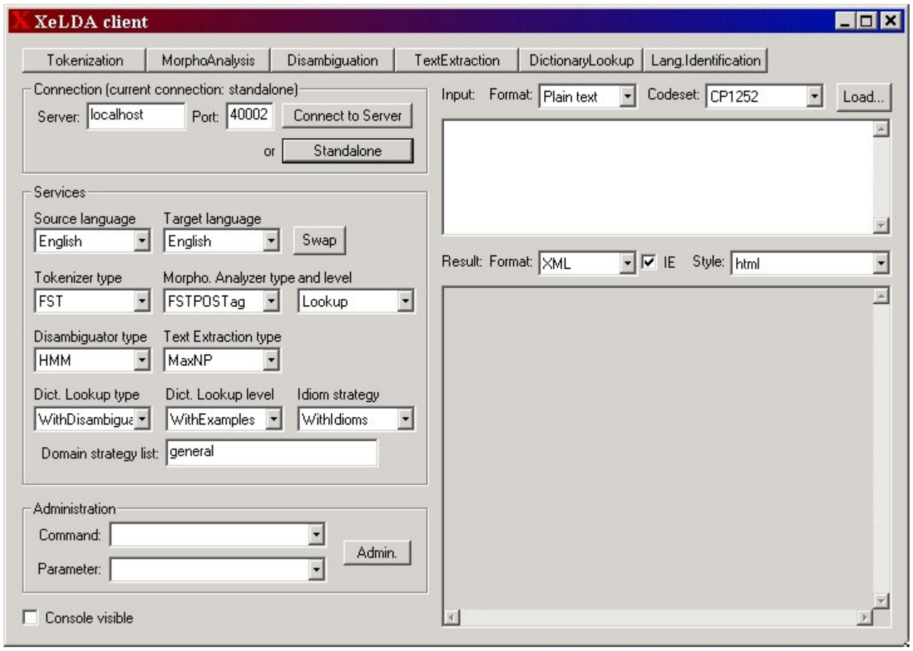
XeLDA embraces the open architecture of a package running in a UNIX or Windows environment. As such, it is designed to provide a framework into which developers and researchers can seamlessly integrate further linguistics services and resources.

Potential applications of the XeLDA engine include comprehension aids and translation or syntax checking, terminology extraction, and other general authoring tools.

Using XeLDA

XeLDA is a toolkit that provides linguistic processing to client applications.

A sample Windows application, wxeldac, is bundled with the XeLDA toolkit and demonstrates XeLDA’s linguistic services.



Connection (current connection: standalone)

Server: localhost Port: 40002

Source language: English Target language: English

Tokenizer type: FST Morpho. Analyzer type and level: FSTPOSTag Lookup

Disambiguator type: HMM Text Extraction type: MaxNP

Dict. Lookup type: WithDisambigua Dict. Lookup level: WithExamples Idiom strategy: WithIdioms

Domain strategy list: general

Command:

Parameter:

□ Console visible

Input: Format: Plain text Codeset: CP1252

Result: Format: XML ☑ IE Style: html

Figure 1 - The wxeldac application main screen, which demonstrates XeLDA functions

Some features are available only through programming. See the *XeLDA C++ Programmer’s Guide* for more information.

For more information

- For C++ or Java examples, go to the `samples` folder in the XeLDA install tree. A description, a brief explanation of the technique used, the available languages, an example, and reference data are given for each service.

About XeLDA's Linguistic Services

This section describes the linguistic services provided by XeLDA in detail and provides examples of each. XeLDA provides the following linguistic services:

- Language Identification
- Tokenization
- Morphological Analysis
- Part of Speech Disambiguation
- Text Extraction
- Dictionary Lookup
- Idiom Recognition
- Relational Morphology

Language Identification

Description

The language identification service recognizes the language used to write a text, even if the text is written without accented characters.

It is called either explicitly or transparently when the user does not specify the language for a particular service.

The language identification service is extremely fast and reliable, working optimally with five words or more.

How it works

The language identification service uses the Trigram and Short Words method, also known as TRISHORT.

The language identifier service has been trained on many texts in various languages so it has a statistical background. Using this statistical information, it determines the following:

- The frequency of the three-character sequences in every language.
- The frequency of common short words (five characters or less) such as "the", "of", and "is".

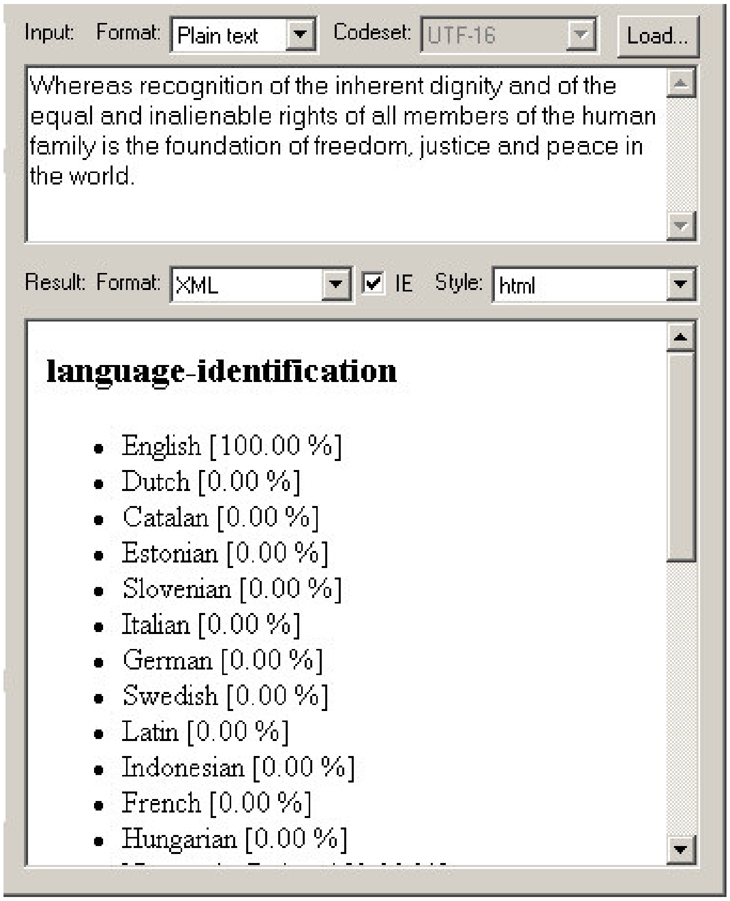
Languages supported

38 languages are supported.

For more information

- For programming details, see the *XeLDA C++ API Programmer's Guide* and the *XeLDA Java Programmer's Guide*.
- For the C++ example, see the "`LanguageIdentification`" folder.

Example



Input: Format: Plain text Codeset: UTF-16

Whereas recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world.

Result: Format: XML ☑ IE Style: html

language-identification

- English [100.00 %]
- Dutch [0.00 %]
- Catalan [0.00 %]
- Estonian [0.00 %]
- Slovenian [0.00 %]
- Italian [0.00 %]
- German [0.00 %]
- Swedish [0.00 %]
- Latin [0.00 %]
- Indonesian [0.00 %]
- French [0.00 %]
- Hungarian [0.00 %]

Tokenization

Description

The tokenization service performs the basic function of cutting a text into words.

It is generally harder that it sounds. Punctuation, abbreviations, units of measure, clitics, multiword expressions, and special cases must be considered. Some examples follow:

- In French, "parce que" and "aujourd'hui" are single words, while "l'ami" and "j'ai" are composed of two words.
- In English, "e.g." and "i.e." are single words, and "don't" is composed of two words. The possessive "'s" as in "St Paul's Cathedral" must be kept with the previous word.
- The dash may represent the hyphenation of a single word that did not fit on the line.
- A character may belong to two words. For example, in French, "donne-le" should be tokenized as "donne-" + "-le".
- In German and agglutinative languages, a single word may be composed of several parts. For example, "Bundesfinanzminister" is made of three nouns. The tokenization service keeps the word together, and it is the morphological analyser's job to extract the individual parts.

How it works

For all supported languages, XeLDA tokenizes texts with a finite-state transducer (FST).

For unsupported languages, there is a simple algorithm that uses spaces and punctuation for basic tokenization.

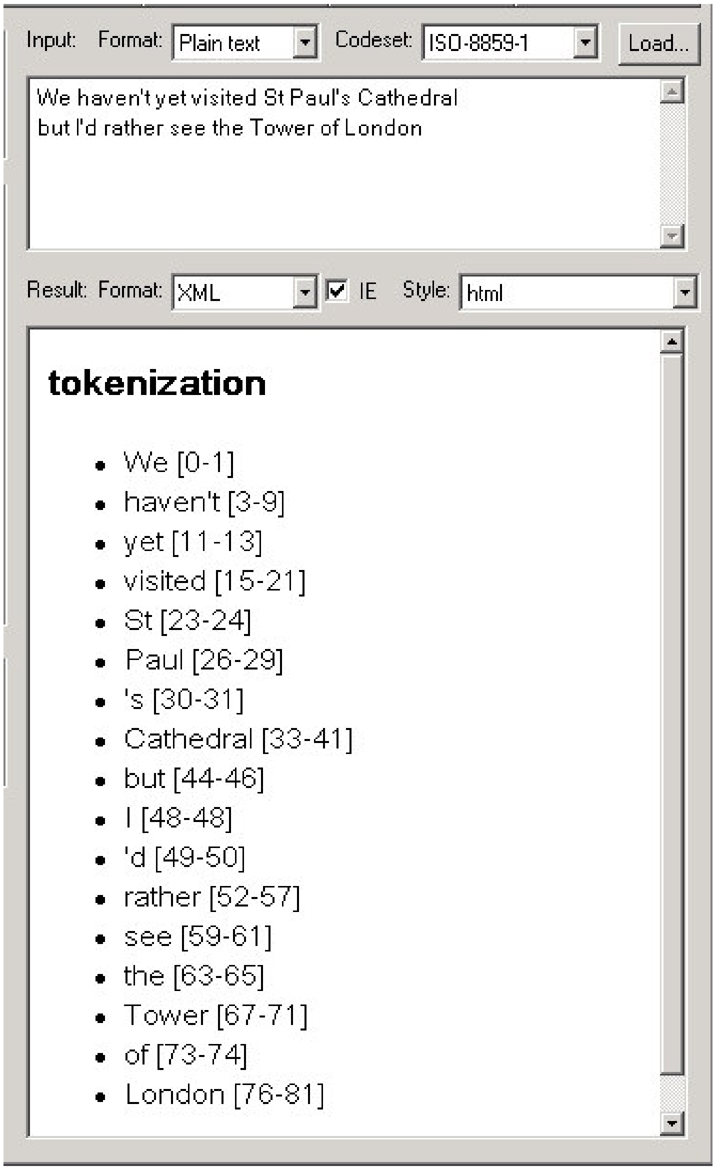
Languages Supported

Czech, Dutch, English, French, German, Greek, Hungarian, Italian, Polish, Portuguese, Russian, and Spanish. Other languages are in development.

For more information

- For programming details, see the *XeLDA C++ API Programmer's Guide* and the *XeLDA Java Programmer's Guide*.
- For the C++ example, see the "`Tokenization`" folder.

Example



We haven't yet visited St Paul's Cathedral
but I'd rather see the Tower of London

tokenization

- We [0-1]
- haven't [3-9]
- yet [11-13]
- visited [15-21]
- St [23-24]
- Paul [26-29]
- 's [30-31]
- Cathedral [33-41]
- but [44-46]
- I [48-48]
- 'd [49-50]
- rather [52-57]
- see [59-61]
- the [63-65]
- Tower [67-71]
- of [73-74]
- London [76-81]

Morphological Analysis

Description

The morphological analysis service finds all the possible combinations of *base form* and *part of speech* corresponding to a given inflected word.

For example, the word "levels" will be analyzed as either "level", a plural noun, or as "to level", a verb in third person singular present tense.

The base form of a word is the form used as headword in a dictionary. This is language-dependent. For example, in French and English the base form for a verb is the infinitive, while in Latin the base form is the first person of the present tense.

Various morphological phenomena (such as gender, number, case, contraction and elisions, and vowel harmony) are taken into account, depending on the language being analyzed.

The part of speech is coded as one or a series of *tags*. Each tag represents a grammatical category. However, the tags do not adhere to all of the usual grammatical categories, sometimes being combined to optimise processing and improve efficiency. Examples of tags are "+Noun", "+Coord" (coordinating pronoun), "+SubjP" (Present subjunctive).

There is a different set of tags for each language. The tag set is optimized for the next step, disambiguation. The disambiguation service decides which one of the different possibilities is the right one in the given context.

The tags list for each supported language is found in the `docs` directory of the `xelda` installation directory.

How it works

The morphological analysis service uses finite state transducers (FST). A FST transforms one string into another in an extremely speed- and space-efficient manner. FSTs are combined in *strategies*, a series of transducers, the output of each being the input of the next. If a strategy fails to find a match, then the next one is used.

Using a series of transducers is necessary for several reasons. A word may not be properly accentuated or capitalized, in which case one of the transducers will normalize it to the "standard" form before passing it to the next. Also, a specialized transducer may be created with domain-specific terms.

Finally, neologisms and unknown words are taken into account by a "guesser", which will try to find the part of speech of an unknown word. For English this is based on rules like:

- -----y = adverb
- ---ed = verb/adj

- ---ing = verb, noun, adj
- digit = number

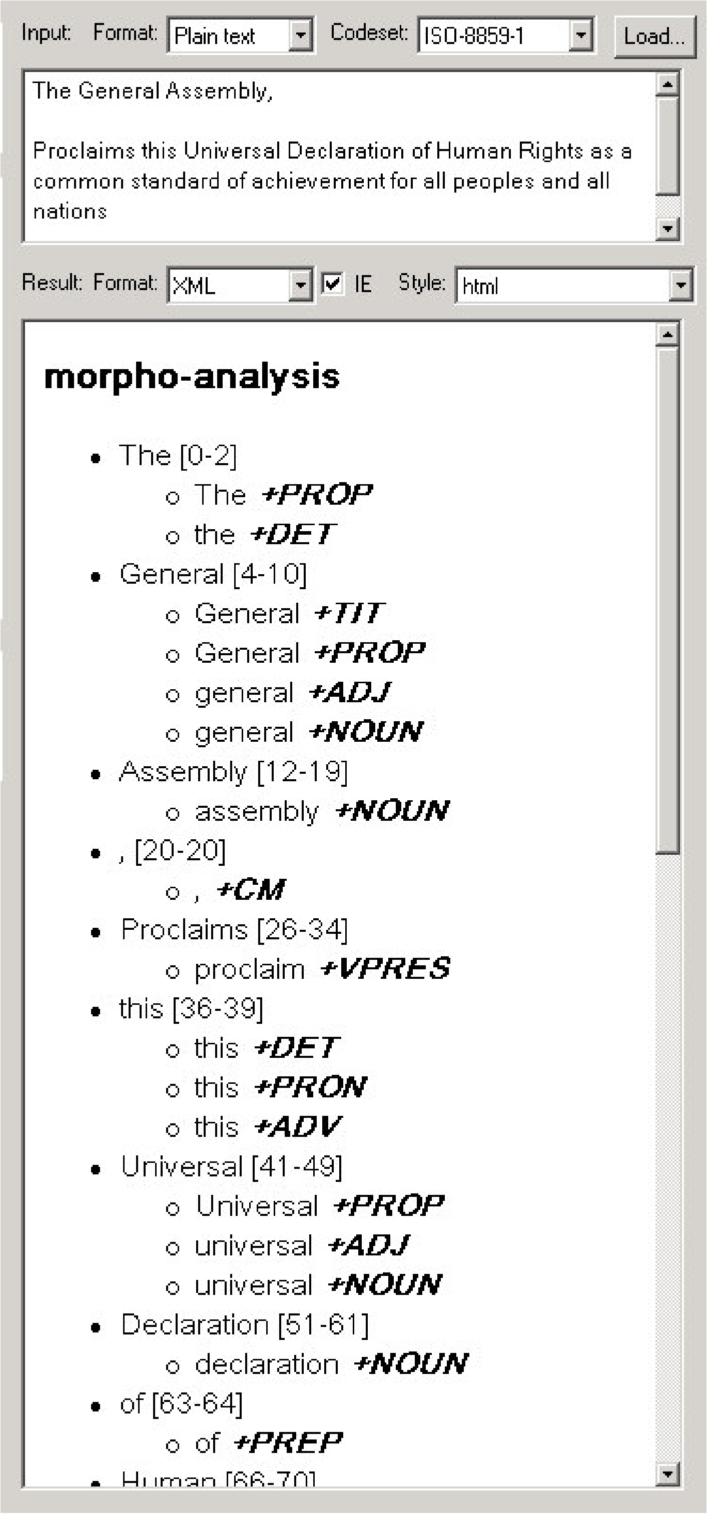
Languages supported

Czech, Dutch, English, French, German, Greek, Hungarian, Italian, Polish, Portuguese, Russian and Spanish. Other languages are in development.

For more information

- See the *XeLDA Tagsets* files in `xelda/docs` for the tags associated with the language being analysed.
- For programming details, see the *XeLDA C++ API Programmer's Guide* and the *XeLDA Java Programmer's Guide*.
- For the C++ example, see the "`MorphoAnalysis`" folder.

Example



The General Assembly,

Proclaims this Universal Declaration of Human Rights as a common standard of achievement for all peoples and all nations

morpho-analysis

- The [0-2]
  - The *+PROP*
  - the *+DET*
- General [4-10]
  - General *+TIT*
  - General *+PROP*
  - general *+ADJ*
  - general *+NOUN*
- Assembly [12-19]
  - assembly *+NOUN*
- , [20-20]
  - , *+CM*
- Proclaims [26-34]
  - proclaim *+VPRES*
- this [36-39]
  - this *+DET*
  - this *+PRON*
  - this *+ADV*
- Universal [41-49]
  - Universal *+PROP*
  - universal *+ADJ*
  - universal *+NOUN*
- Declaration [51-61]
  - declaration *+NOUN*
- of [63-64]
  - of *+PREP*
- Human [66-70]

Part of Speech Disambiguation

Description

The part of speech disambiguation service follows the morphological analysis. It removes the ambiguity by choosing one base form plus tag for each word from the list proposed by the morphological analysis service.

To make the choice, the part of speech disambiguation service chooses the most probable tag in the word context. Many other modules depend on the results produced by the disambiguation service, making its accuracy important. Depending on the language, the part of speech disambiguation service finds the grammatical category of a word with an accuracy of more than 95%.

The disambiguation service is a basic component for many more sophisticated services.

How it works

The disambiguation service relies on a Hidden Markov Model engine. This engine has been trained on carefully chosen, manually tagged corpus for each language and the result of this training is used by XeLDA to disambiguate new texts.

The disambiguation service takes the results from the morphological analysis service and retains, for each word, the most probable tag in the word context.

The output of the disambiguation service is usually a single base form + tag for each input word. However, some semantic ambiguities may remain after part of speech disambiguation. For example, in French the word "MODELE" is disambiguated to either `modèle+NOUN_SG` or `modelé+NOUN_SG`. The part of speech ambiguity has been removed, but not the semantic ambiguity.

As a convenience of language, the phrase "Part of Speech Tagging" is often used to describe the sequence of tokenization, morphological analysis, and disambiguation.

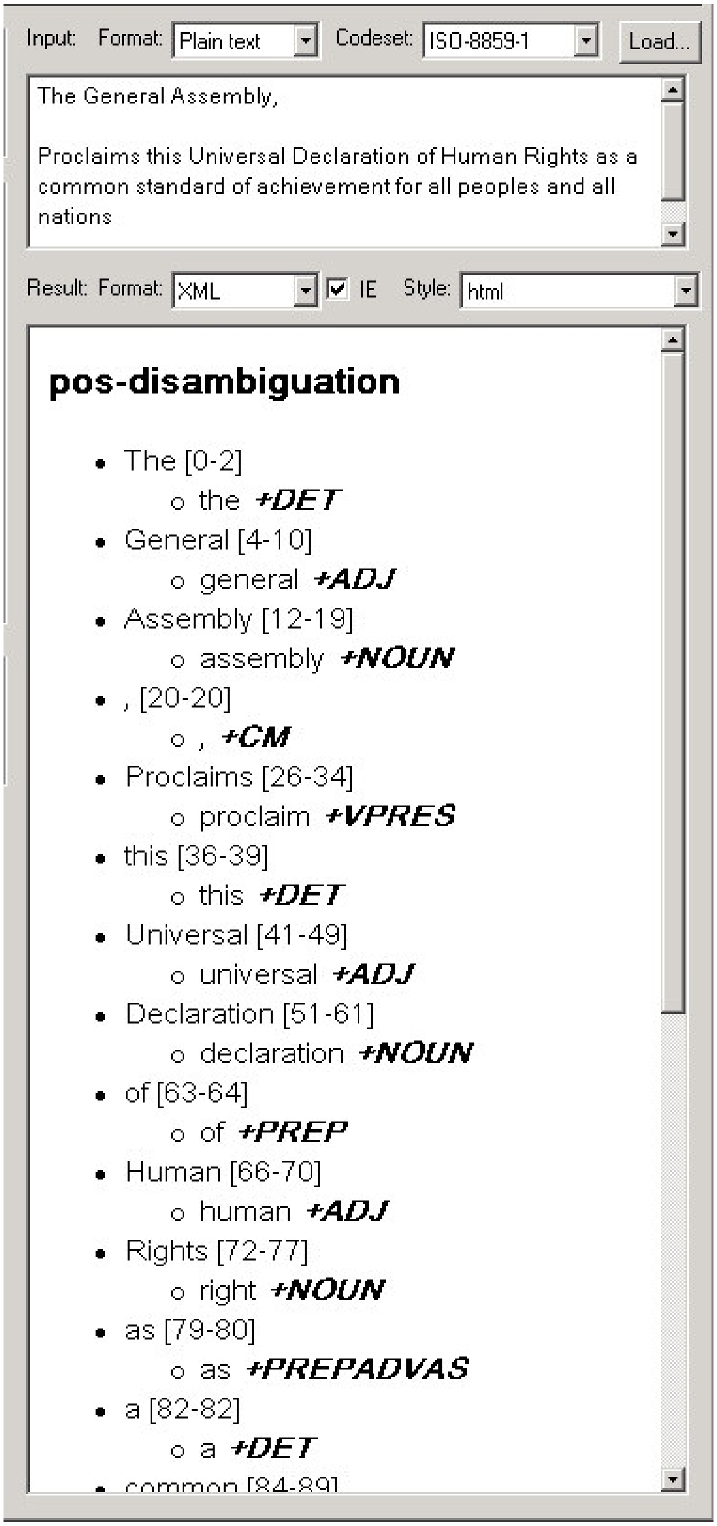
Languages supported

Czech, Dutch, English, French, German, Greek, Hungarian, Italian, Polish, Portuguese, Russian and Spanish. Other languages are in development.

For more information

- For programming details, see the *XeLDA C++ API Programmer's Guide* and the *XeLDA Java Programmer's Guide*.
- For the C++ example, see the "`Disambiguation`" folder.

Example



Input: Format: Plain text Codeset: ISO-8859-1

The General Assembly,

Proclaims this Universal Declaration of Human Rights as a common standard of achievement for all peoples and all nations

Result: Format: XML ☑ IE Style: html

pos-disambiguation

- The [0-2]
  - the +DET
- General [4-10]
  - general +ADJ
- Assembly [12-19]
  - assembly +NOUN
- , [20-20]
  - , +CM
- Proclaims [26-34]
  - proclaim +VPRES
- this [36-39]
  - this +DET
- Universal [41-49]
  - universal +ADJ
- Declaration [51-61]
  - declaration +NOUN
- of [63-64]
  - of +PREP
- Human [66-70]
  - human +ADJ
- Rights [72-77]
  - right +NOUN
- as [79-80]
  - as +PREPADVAS
- a [82-82]
  - a +DET
- common [84-89]

Text Extraction

Description

A noun phrase is a sequence of terms that behave together as a noun.

"Noun phrase" and "sequence of terms" are themselves noun phrases, as are "red power cord" and "geostationary communications satellite".

The noun phrase extraction service is a powerful tool to get candidate terms for a glossary, or as a front-end for other linguistic processing.

How it works

The noun phrase extraction service uses the XeLDA tokenization, morphological analysis, and disambiguation services.

It extracts noun phrases according to language-dependent patterns, resulting from Xerox linguistic research.

The individual components of the noun phrases are also returned, with their disambiguated base form and part of speech.

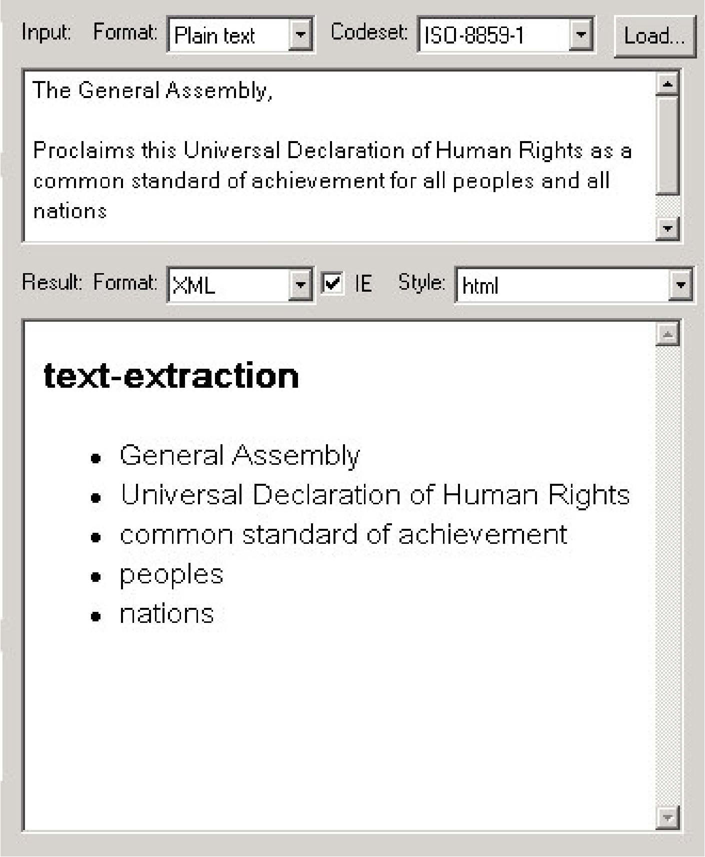
Languages supported

Czech, Dutch, English, French, German, Greek, Hungarian, Italian, Polish, Portuguese, Russian and Spanish. Other languages are in development.

For more information

- For programming details, see the *XeLDA C++ API Programmer's Guide* and the *XeLDA Java Programmer's Guide*.
- For the C++ example, see the "`TextExtraction`" folder.

Example



The General Assembly,

Proclaims this Universal Declaration of Human Rights as a common standard of achievement for all peoples and all nations

text-extraction

- General Assembly
- Universal Declaration of Human Rights
- common standard of achievement
- peoples
- nations

Dictionary Lookup

Description

A word can have several meanings, depending on the context in which it is used. The dictionary lookup service retrieves the word context and can use this context to find the correct entry, the *target*, in a dictionary.

Several types of dictionaries are supported, including monolingual, synonym, and bilingual.

The dictionary lookup service may be contextual or not, can look for idiomatic expressions, and can return examples from the dictionary.

How it works

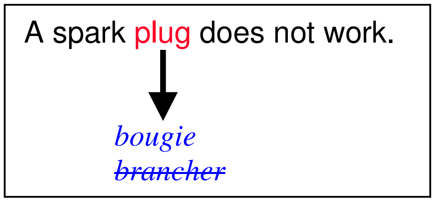
The dictionaries used are encoded in XML or SGML and compiled into a XeLDA dictionary.

XeLDA first does a disambiguation on the word, in context, to get the proper base form and part of speech.

Idiomatic expression recognition is attempted. If there is a match, this expression is shown.

The dictionary is then searched and the result (translation for a bilingual dictionary) is shown.

For example, a text would be analysed as follows:



Examples or idiomatic expressions or both can be shown, depending on the options set by the user.

Languages supported

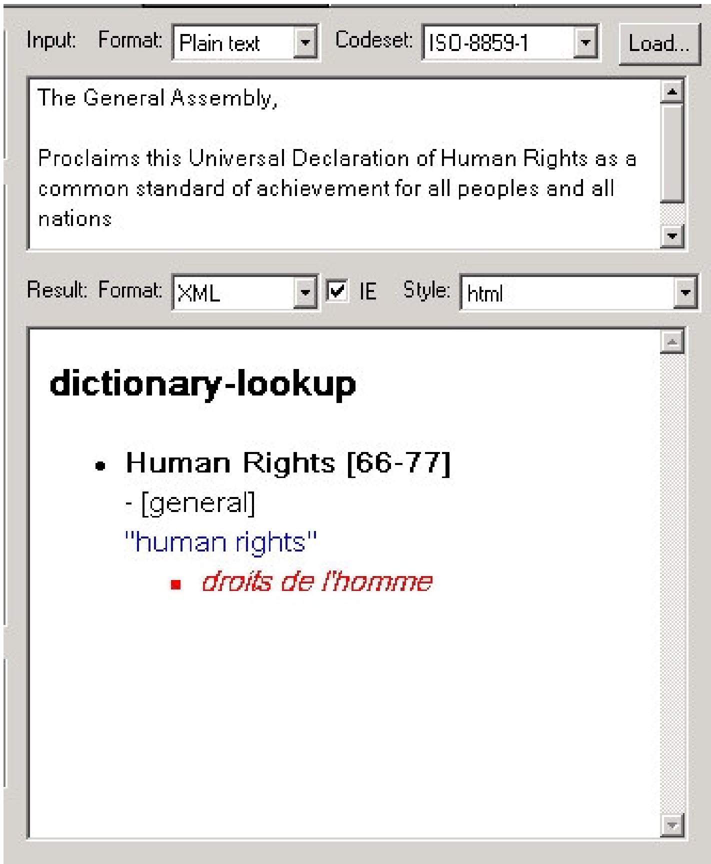
English, French, German, Italian and Spanish as source languages, any language as the target language.

For more information

- For programming details, see the *XeLDA C++ API Programmer's Guide* and the *XeLDA Java Programmer's Guide*.
- For the C++ example, see the "`DictionaryLookup`" folder.

Example

The contextual dictionary lookup of the word "Rights" using an English-French bilingual dictionary occurs as follows in the wxeldac sample application:



The General Assembly,

Proclaims this Universal Declaration of Human Rights as a common standard of achievement for all peoples and all nations

dictionary-lookup

- **Human Rights [66-77]**
  - [general]
  "human rights"
    - *droits de l'homme*

Idiom Recognition

DescriptionAn idiom is a group of words that have a meaning not deducible from the meaning of the individual words.

Idioms may have some variability, for example verbs may be conjugated, nouns may be inflected, and adverbs or adjectives may be inserted. Some words have a particular meaning when they disambiguate in a certain way.

Some idiomatic expressions may be used for domain-specific expressions or may point to particular entities. For example, a name followed by "Inc." or its variants points to a particular kind of company.

Some examples of idiomatic expressions follow:

- "To take the bull by the horns", which may be interpreted as "he took quickly and firmly the bull by the horns."
- "<noun>", followed by variants or abbreviations of "Chief Executive Officer", pointing both to a proper name and rank in a company.
- "spark plug", "operating system", "gasket joint", and "reverse" can be disambiguated as nouns depending upon the context.

XeLDA provides an idiomatic regular expression language (IDAREX) to describe idiomatic expressions. With IDAREX, you can define:

- Invariants (exact spelling).
- Words that can appear optionally. For example, an adjective can appear in a phrase but is not required.
- Repetition.
- Specific words when they disambiguate in a specific category, such as adjective, determiner, noun, verb. For example, you could specify that the word "level" be counted only when used as an adjective.
- Any word of a specific category: adjectives, adverbs, negatives, and clitic pronouns.
- Macros helping to define particular constructs, such as VPRON(verb): Verb + variable clitic. In Spanish, this macro is interpreted as follows: VPRON(dar) = darle, le he dado, me das, nos daba, les dará, and so on.

How it works

XeLDA provides an idiomatic expression compiler. The user may define custom expressions, using the categories and macros provided with the language library, and incorporate the compilation result in the XeLDA configuration files.

The idiom recognition service is available through the dictionary lookup service.

For more information

- For programming details, see the *XeLDA C++ API Programmer's Guide* and the *XeLDA Java Programmer's Guide*.
- For the C++ example, see the "`DictionaryLookup`" folder.

Example

Input text:

..., U.S. District judge Thomas Penfield Jackson wrote in a 50-page opinion ...
... James Barksdale, the former chief executive of Netscape Communications Corp. ...

IDAREX expression:

NP [judge N: | (DET) (ADJ) :chief :executive (:of)] NP

Result:

..., U.S. District judge Thomas Penfield Jackson wrote in a 50-page opinion ...

... James Barksdale, the former chief executive of Netscape Communications Corp., ...

Languages supported

English, French, German, Italian and Spanish have predefined word categories and macros. Others can be added on request.

Relational Morphology

Description

Often, you need to know the family to which a word belongs. For example, the word "Presidential" is related to the word "President".

The relational morphology service allows the grouping of words that belong to the same derivational family. Each family is characterized by a representative word. There are two relational morphology services integrated into XeLDA:

- **Relational morphology service**: this service returns a single word representative of the derivational family.
- **Reverse relational morphology service**: this service returns all the words belonging to the derivational family.

These services are used for the following:

- Categorizing words and documents.
- Expanding queries by finding all related words and expressions.

How it works

The relational morphology service is a special type of morphological analyser that defines the relational family based on the prefix of a word.

The relational morphology service first requests the disambiguation service on the whole text. Then, for each word in the text, it:

- Gets the base form of a word.
- Performs a relational morphological analysis on the base form.

Languages supported

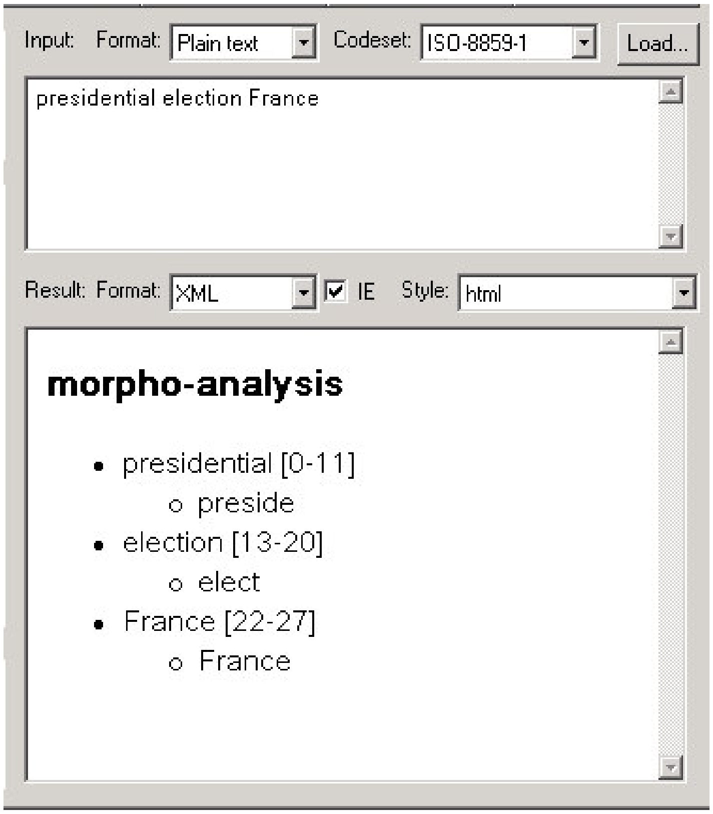
English, French, German, Italian and Spanish.

For more information

- For programming details, see the *XeLDA C++ API Programmer's Guide* and the *XeLDA Java Programmer's Guide*.
- For the C++ example, see the "`MorphoAnalysis`" folder.

Examples

The relational morphology of the text "presidential election France" occurs as follows in the wxeldac sample application:



Format: Plain text

Codeset: ISO-8859-1

presidential election France

Result: Format: XML

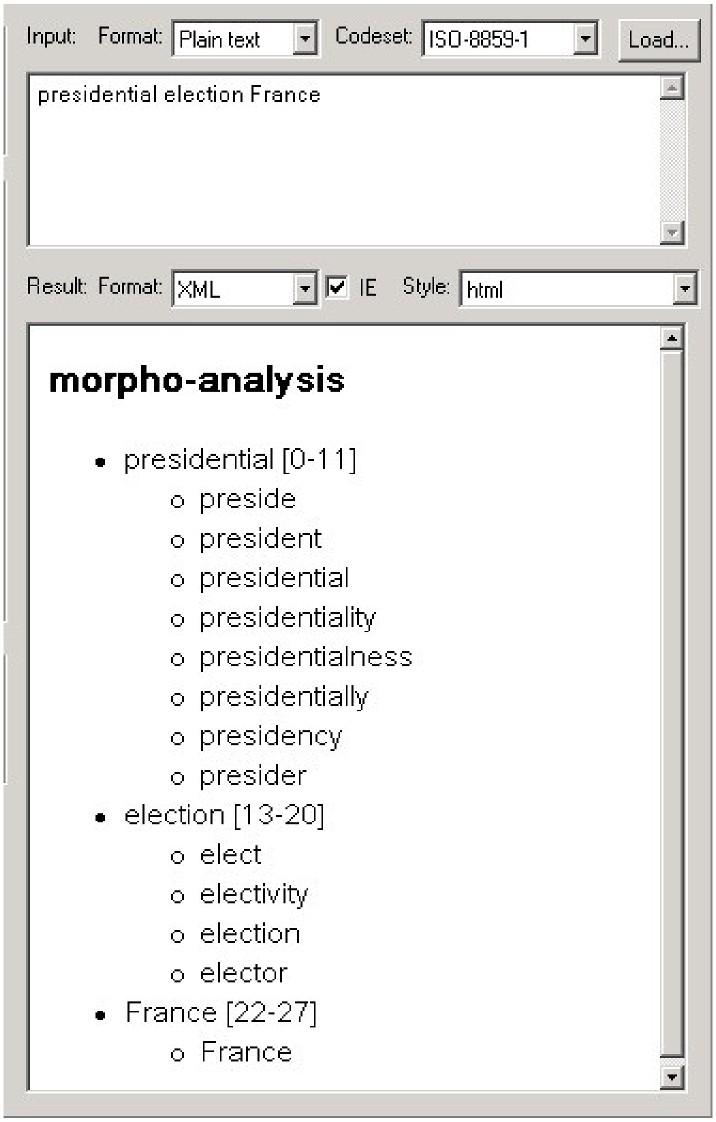
☑ IE

Style: html

morpho-analysis

- presidential [0-11]
  - preside
- election [13-20]
  - elect
- France [22-27]
  - France

The reverse relational morphology of the text "presidential election France" occurs as follows in the wxeldac sample application:



Input: Format: Plain text Codeset: ISO-8859-1

presidential election France

Result: Format: XML ☑ IE Style: html

morpho-analysis

- presidential [0-11]
  - preside
  - president
  - presidential
  - presidentiality
  - presidentialness
  - presidentially
  - presidency
  - presider
- election [13-20]
  - elect
  - electivity
  - election
  - elector
- France [22-27]
  - France

Overview of XeLDA Architecture

This section describes the two modes in which XeLDA can operate, and explains how XeLDA processes requests and returns results.

XeLDA modes

XeLDA has two main operating modes: client-server or stand-alone, also called "monoprocess".

The operating modes are transparent to the user and the client application developer.

Client-server mode

In client-server mode, all linguistic processing takes place on a server machine. The client process on the local computer sends requests and retrieves the results through the network.

There may be many servers on the network. The communication takes place through the usual TCP/IP mechanisms. For example, www.xelda.com and www.xelda.com:40002 are both potentially valid addresses for a XeLDA server.

Because processing occurs over the network, the server may be as near as the next office or as far as the next continent. It is even possible to have the server and the client on the same machine, although in this case, the stand-alone mode may be more efficient.

The client-server mode is designed so that a few powerful servers are installed to serve several lightweight clients.

See Figure 2 on the following page for a diagram of communication between the client and server when XeLDA operates in client-server mode.

Stand-alone mode

Use the stand-alone mode when there is only one machine involved and the server and client are merged in a single executable. In stand-alone mode, the network component disappears and all processing is done locally.

See Figure 3 on the following page for a diagram of an application operating in stand-alone mode.

Requests and results

XeLDA functions the same way in either mode. A client application creates a request object describing the linguistic service it needs.

This request is sent to the server via the communication link (which can be a direct connection in the case of a stand-alone application). This request is executed on the server by the XeLDA kernel, which calls the appropriate linguistic service.

The result is generated either in raw format or formatted according to what the client specifies. Raw format is easier to manipulate by the client whereas a formatted result is easier to display or print. After the result is generated, it is sent back to the client.

Client-server mode

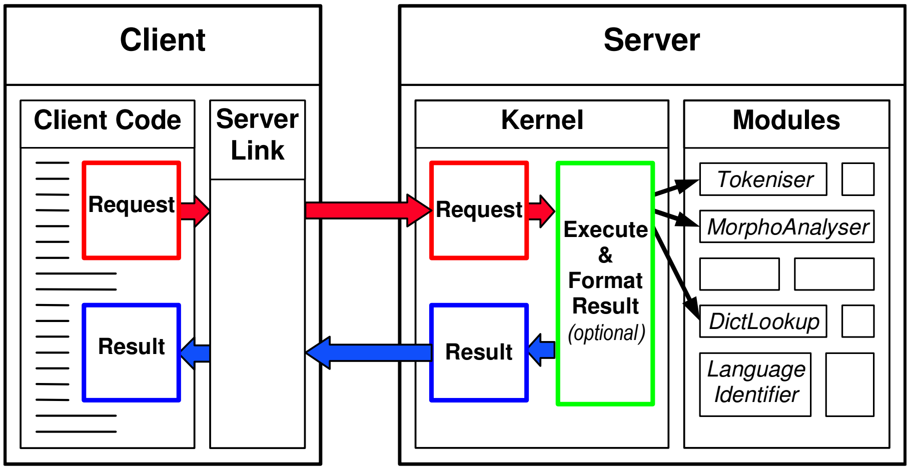


Figure 2 – Client-Server Mode

Stand-alone mode

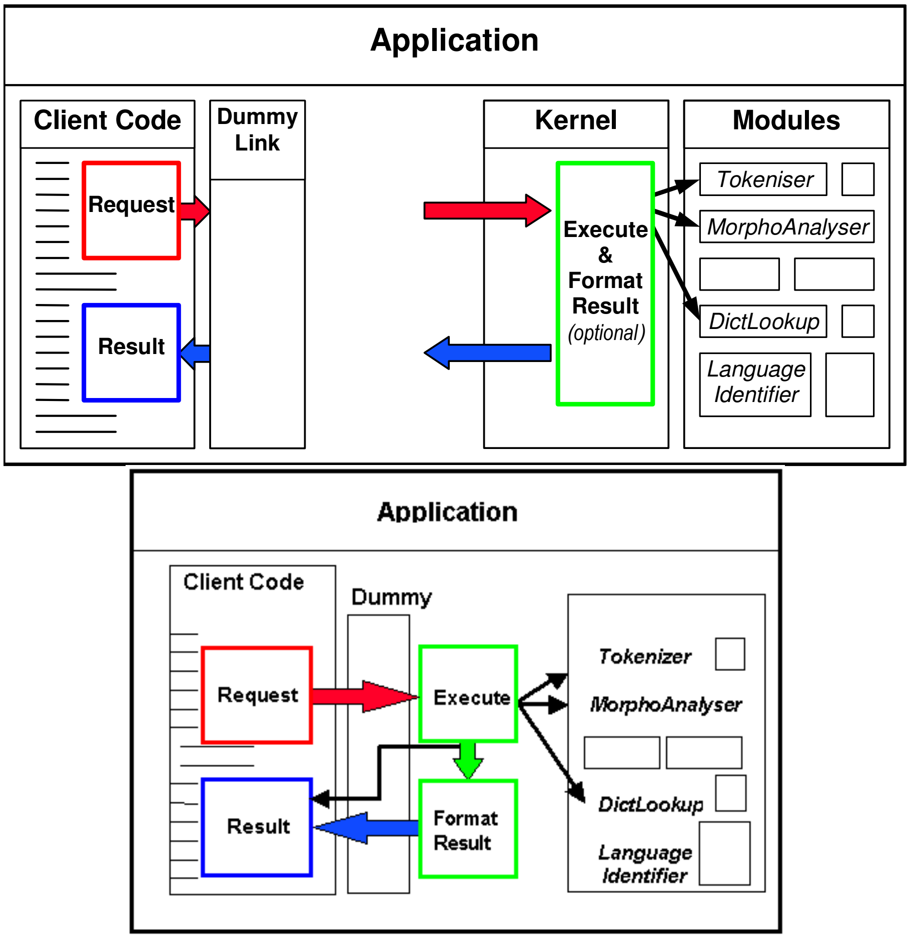


Figure 3 – Stand-Alone Mode

Developing Applications With XeLDA

XeLDA as a toolkit

The XeLDA services are available in the XeLDA toolkit for creating custom language processors, either using XeLDA services "out of the box" or combining them to get the results you need.

XeLDA can be customized, meaning customer lexicons and expressions can be integrated into the XeLDA toolkit.

XeLDA is an evolving product. It constantly incorporates new modules and languages from Xerox or third-party linguistic researchers. One of XeLDA's design goals is the ability to quickly integrate cutting-edge results into mainstream processing.

For more information

For more information, refer to the following documents:

- The *XeLDA C++ API Programmer's Guide* for information about using the XeLDA C++ SDK to create custom language processing applications.
- The *XeLDA Java API Programmer's Guide* for information about creating custom Java applications using the XeLDA Java API.
- The *XeLDA Reference Manuals* for information about all of XeLDA's public classes.

Contact

For more information on XeLDA, please contact:
Cyril Chantrier, Temis
6, chemin de Maupertuis F-38240 MEYLAN
Email: xelda-info@temis-group.com
Tel: +33 [0]4 76 61 51 83
Fax: +33 [0]4 76 61 51 89

Glossary of Terms

ANSI: American National Standards Institute. For more information, go to http://www.ansi.org/.

Character: the smallest component of written language with semantic value.

Character Set: a complete group of characters for one or more writing systems.

Clix: A C++ library that provides a collection of general functionalities, grouped by packages.

Coded Character Set: a mapping from a set of abstract characters to a set of integers.

Dictionary Lookup: a linguistic service that retrieves a word's context and uses this context to find the correct entry in the dictionary.

Finite State Technology: see *FST*.

FST: Finite State Technology. An FST is a network of states and transitions that work as an abstract machine to perform dedicated linguistic tasks. For example, the XeLDA morphological analyzer tokenization resources are encoded as FSTs.

IANA: Internet Assigned Numbers Authority, the central coordinator for the assignment of unique parameter values for Internet protocols. For more information, go to http://www.iana.org/.

IDAREX: Idiomatic Regular Expression language.

Idiom: a group of words having a meaning not deducible from each of the individual words.

Idiom Recognition: a linguistic service that recognizes idiomatic expressions in a text.

IETF: Internet Engineering Task Force, an internal community of network service providers concerned with Internet architecture and operation. For more information go to http://www.ietf.org/.

ISO: International Organization for Standardization, a worldwide federation of national standards bodies. For more information, go to http://www.iso.ch/.

Language Guesser: see *Language Identification.*

Language Identification: a linguistic service that recognizes the language and character set used to write a document.

Lexeme: a word or token.

Module: basic component of XeLDA, such as a service, a Clix component, the kernel, or an application.

Morphological Analyzer: a linguistic service that provides the normalized form and all potential part of speech categories for each token identified during tokenization.

Noun Phrase Extraction: a linguistic service that identifies a sequence of words that behaves together as a noun. It is the continuation of a chain of linguistic services including the tokenization service, the morphological analysis service and the part of speech disambiguation service.

Part of Speech (POS) Disambiguation: a linguistic service that finds the correct grammatical category of a word according to its context.

Relational Morphology: a linguistic service that returns the representative of the derivational family of a word.

Reverse Relational Morphology: a linguistic service that returns the representative of a word's derivational family and all the words belonging to this family.

Senses: the list of base forms and part of speech information for a word.

Service: the client portion of the communication link between the client and the server. It contains general information about what server to contact and supports descriptions specific to the client environment, such as user name, machine name, operating system, etc.

Tans: Translation Aid Network Service. This project was originally a technology transfer project and is the base for the XeLDA framework.

Target Entry: the correct entry for a word in a dictionary. The dictionary lookup service retrieves the target entry depending on the context of a word.

Tokenization Service: a linguistic service that divides a sequence of input characters into words or tokens.

TRISHORT: stands for the trigram and shortword methods. XeLDA supports one type of language identifier, TRISHORT. It is based on a combination of the trigram and shortword methods.

Unicode Standard: the Unicode Standard is the work of a private consortium. For more information, go to http://www.unicode.org/.

15-381 Artificial Intelligence

- www.cs.cmu.edu/~15381/Lectures/NLP.ppt

Natural Language Processing

Jaime Carbonell

13-February-2003

OUTLINE

- Overview of NLP Tasks
- Parsing: Augmented Transition Networks
- Parsing: Case Frame Instantiation
- Intro to Machine Translation

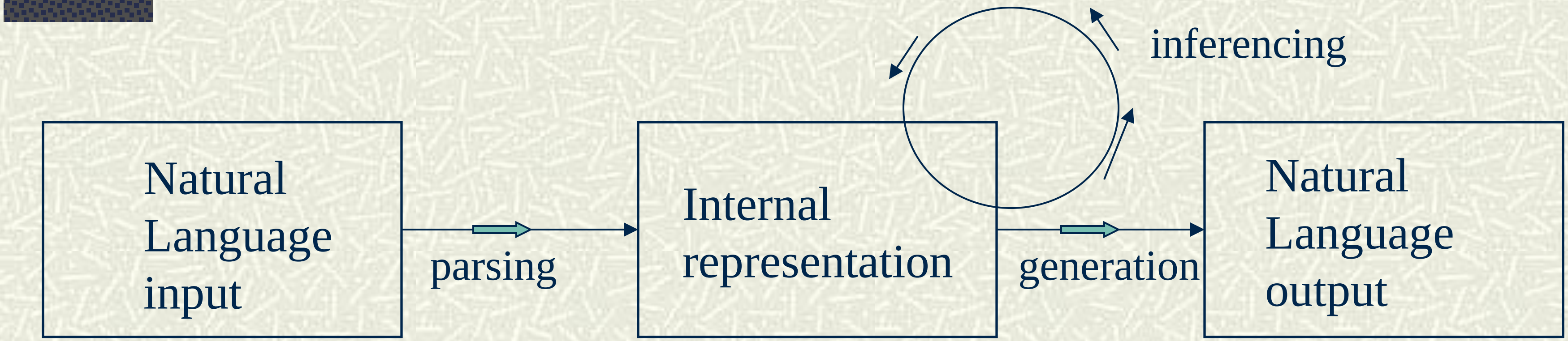
NLP in a Nutshell

- Objectives:
  - To study the nature of language (Linguistics)
  - As a window into cognition (Psychology)
  - As a human-interface technology (HCI)
  - As a technology for text translation (MT)
  - As a technology for information management (IR)

Component Technologies

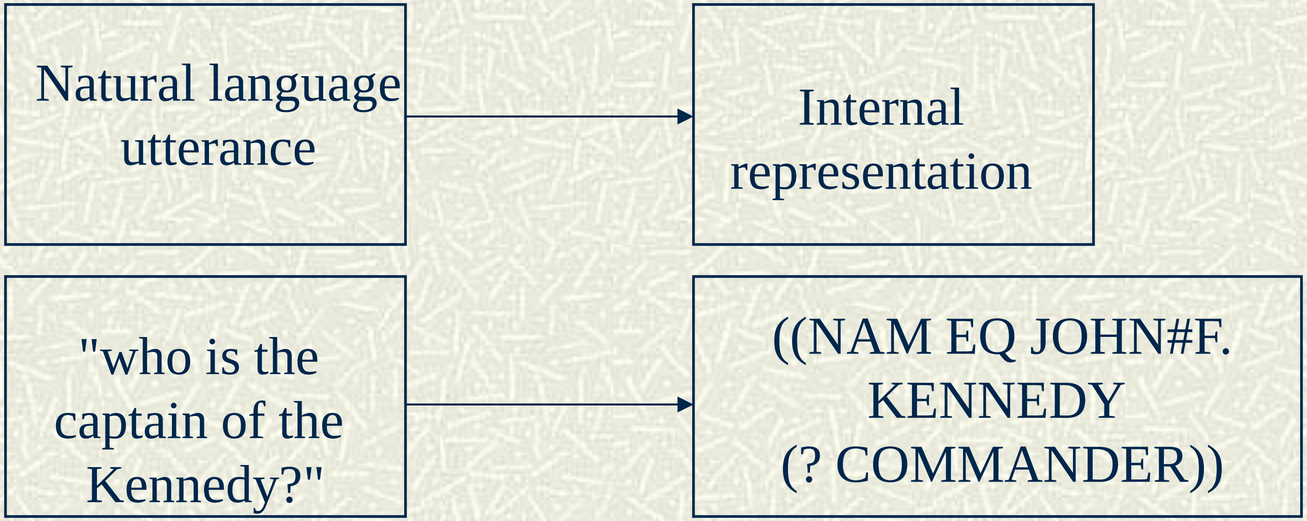
- Text NLP
  - Parsing: text → internal representation such as parse trees, frames, FOL,…
  - Generation: representation → text
  - Inference: representation → fuller representation
  - Filter: huge volumes text → relevant-only text
  - Summarize: clustering, extraction, presentation
- Speech NLP
  - Speech recognition: acoustics → text
  - Speech synthesis: text → acoustics
  - Language modeling: text → p(text | context)
  - …and all the text-NLP components

Outline of an NLP System



- Natural language processing involves translation of input into an unambiguous internal representation before any further inferences can be made or any response given.
- In applied natural language processing:
  - Little additional inference is necessary after initial translation
  - Canned text templates can often provide adequate natural language output
  - So translation into internal representation is central problem

Translation into Internal Representation



Natural language utterance

Internal representation

"who is the captain of the Kennedy?"

((NAM EQ JOHN#F. KENNEDY (? COMMANDER))

- Examples of representations:
  - DB query language (for DB access)
  - Parse trees with word sense terminal nodes (for machine translation)
  - Case frame instantiations (for a variety of applications)
  - Conceptual dependency (for story understanding)

Ambiguity Makes NLP Hard

- Syntactic

  I saw the Grand Canyon flying to New York.
  Time flies like an arrow.

- Word Sense

  The man went to the **bank** *to get some cash.*
  and jumped in.

- Case

  He ran the mile **in** *four minutes.*
  the Olympics.

- Referential

  I took the cake from the table and washed **it**.
  ate **it.**

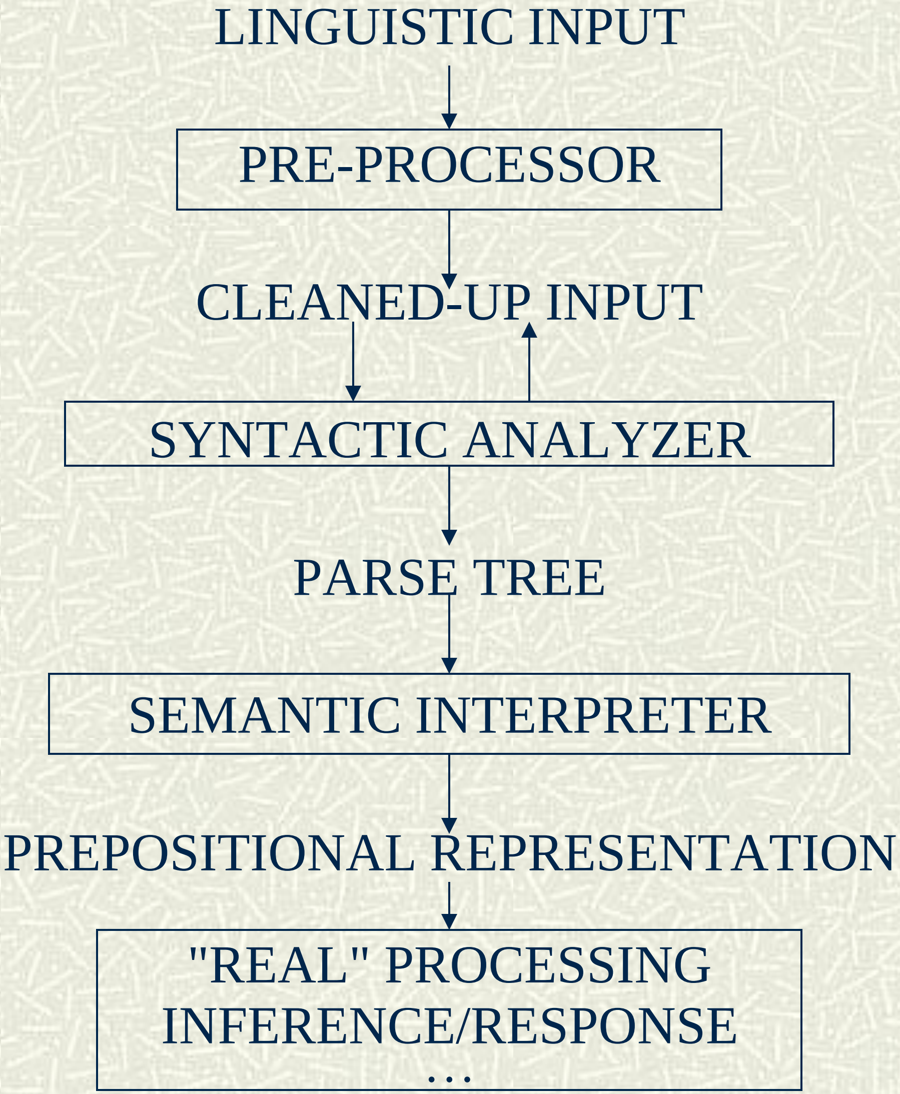
- Indirect Speech Acts

  Can you open the window? I need some air.

Parsing in NLP

- Parsing Technologies
  - Parsing by template matching (e.g. ELIZA)
  - Parsing by direct grammar application (e.g. LR, CF)
  - Parsing with Augmented Transition Networks (ATNs)
  - Parsing with Case Frames (e.g. DYPAR)
  - Unification-Based parsing methods (e.g. GLR/LFG)
  - Robust parsing methods (e.g. GLR*)
- Parsing Complexity
  - Unambiguous Context-Free → $O(n^2)$ (e.g. LR)
  - General CF → $O(n^3)$ (e.g. Early, GLR, CYK)
  - Context-Sensitive → $O(2^n)$
  - NLP is "mostly" Context Free
  - Semantic constraints reduce average case complexity
    In practice: $O(n^2) < O(NLP) < O(n^3)$

Classical Period



LINGUISTIC INPUT

PRE-PROCESSOR

CLEANED-UP INPUT

SYNTACTIC ANALYZER

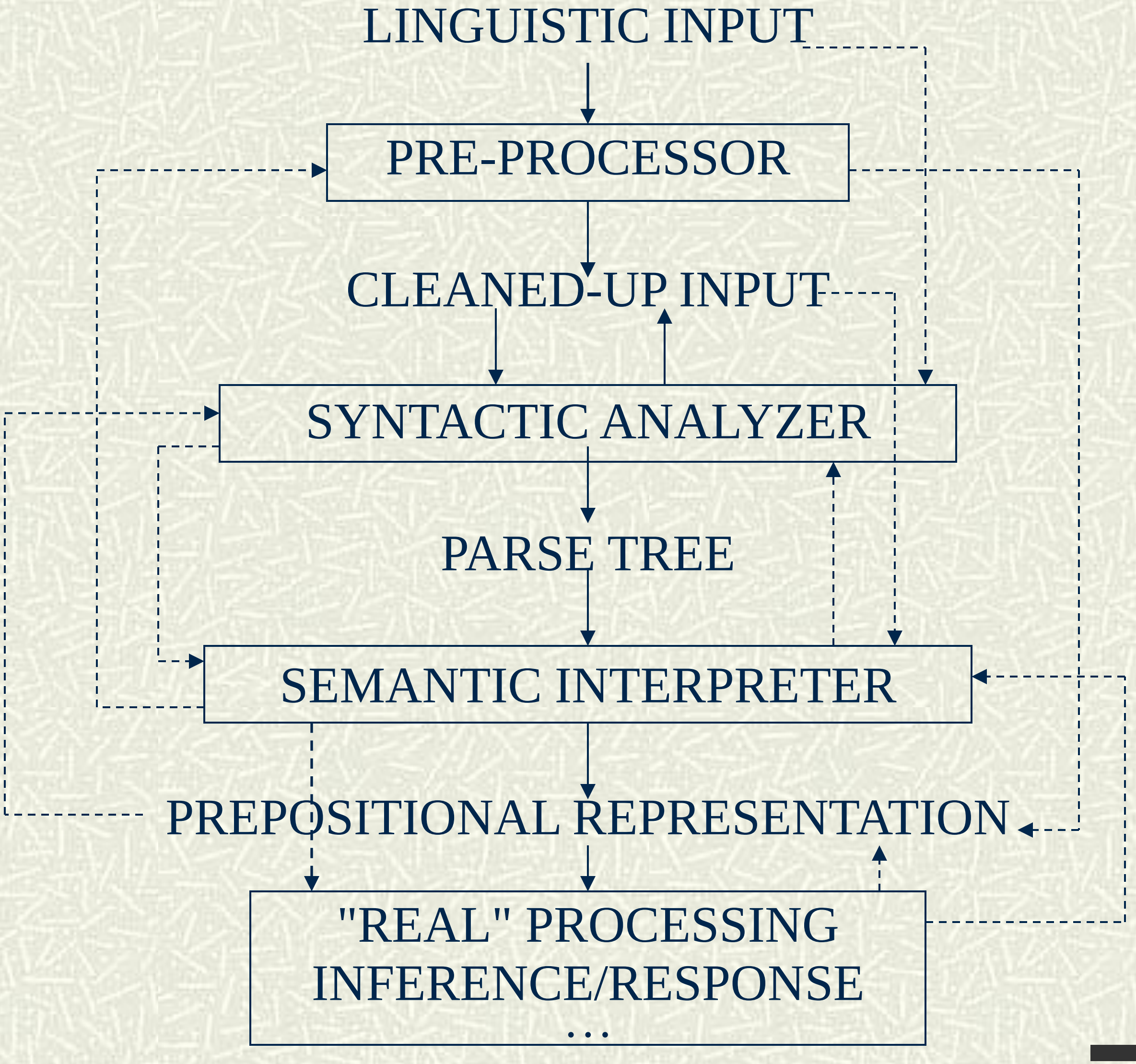
PARSE TREE

SEMANTIC INTERPRETER

PREPOSITIONAL REPRESENTATION

"REAL" PROCESSING
INFERENCE/RESPONSE
. . .

Baroque Period



LINGUISTIC INPUT

PRE-PROCESSOR

CLEANED-UP INPUT

SYNTACTIC ANALYZER

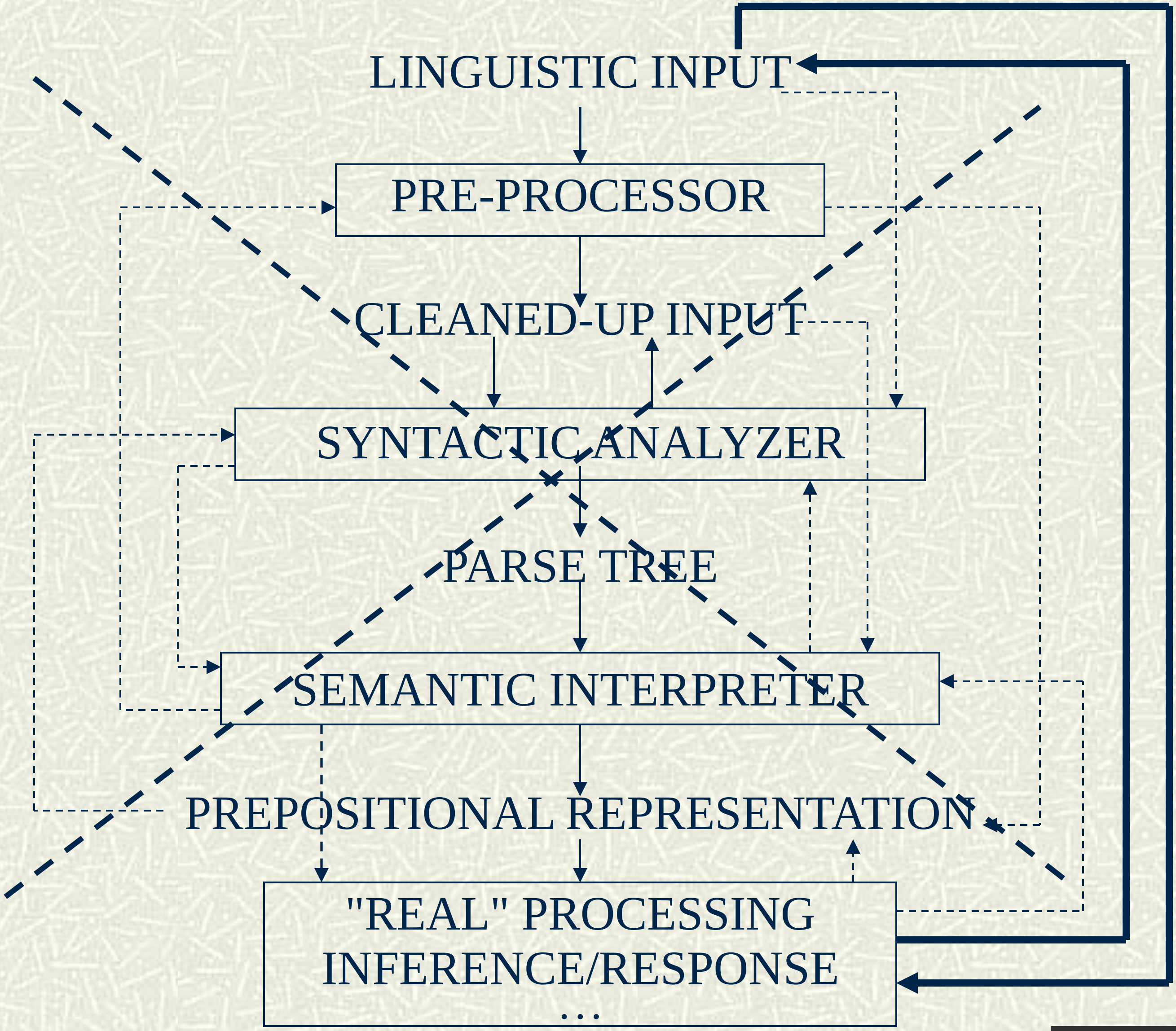
PARSE TREE

SEMANTIC INTERPRETER

PREPOSITIONAL REPRESENTATION

"REAL" PROCESSING
INFERENCE/RESPONSE
. . .

Renaissance



LINGUISTIC INPUT

PRE-PROCESSOR

CLEANED-UP INPUT

SYNTACTIC ANALYZER

PARSE TREE

SEMANTIC INTERPRETER

PREPOSITIONAL REPRESENTATION

"REAL" PROCESSING
INFERENCE/RESPONSE
...

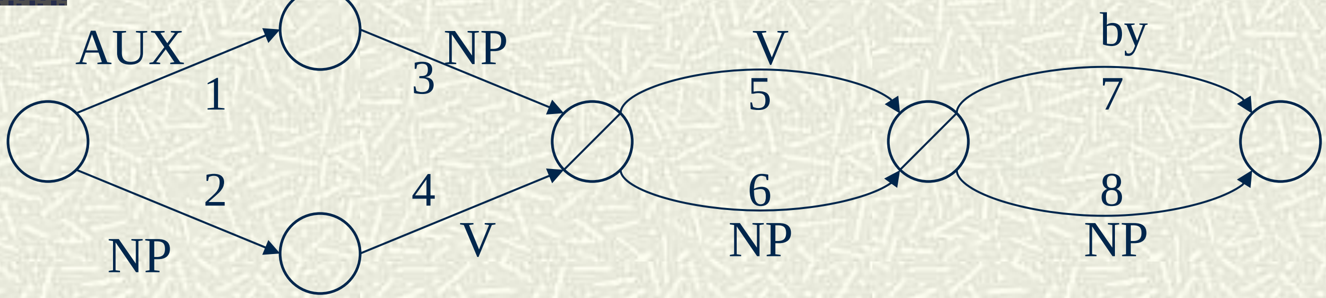
Context-Free Grammars

- Example:

  S → NP VP
  VP → V NP
  ADJ → big | green
  V→ nibbled | nibbles | nibble
  NP → DET N | DET ADJ N
  DET → the | a | am
  N → rabbit | rabbit | carrot

- Advantages:
  - Simple to define
  - Efficient parsing algorithms
- Disadvantages:
  - Can't enforce agreements in a concise way
  - Can't capture relationships between similar utterances (e.g. passive and active)
  - No semantic checks (as in all syntactic approaches)

Example ATN



1	T	(SETR V *) (SETR TYPE “QUESTION”)
2	T	(SETR SUBJ *) (SETR TYPE “DECLARATIVE”)
3	(agrees * V)	(SETR SUBJ*)
4	(agrees SUBJ *)	(SETR V *)
5	(AND (GETF PPRT) (= V “BE”))	(SETR OBJ SUBJ) (SETR V*) (SETR AGFLAG T) (SETR SUBJ “SOMEONE”)
6	(TRANSV)	(SETR OBJ *)
7	AGFLAG	(SETR AGFLAG FALSE)
8	T	(SETR SUBJ *)

Lifer Semantic Grammars

- Example domain—access to DB of US Navy ships

  S → <present> the <attribute> of <ship>

  <present> → what is | [can you] tell me

  <attribute> → length | beam | class

  <ship> → the <shipname>

  <shipname> → kennedy | enterprise

  <ship> → <classname> class ship

  <classname> →kitty hawk | lafayette

- Example inputs recognized by above grammar:

  what is the length of the Kennedy

  can you tell me the class of the Enterprise

  what is the length of Kitty Hawk class ships

  - Not all categories are "true" syntactic categories
  - Words are recognized by their context rather than category (e.g. *class*)
  - Recognition is strongly directed
  - Strong direction useful for spelling correction

Semantic Grammars Summary

- Advantages:
  - Efficient recognition of limited domain input
  - Absence of overall grammar allows pattern-matching possibilities for idioms, etc.
  - No separate interpretation phase
  - Strength of top-down constraints allow powerful ellipsis mechanisms

    What is the length of the Kennedy? The Kittyhawk?
- Disadvantages:
  - Different grammar required for each new domain
  - Lack of overall syntax can lead to "spotty" grammar coverage (e.g. fronting possessive in "<attribute> of <ship>") doesn't imply fronting in "<rank> of <officer>")
  - Difficult to develop grammars
  - Suffers from same fragility as ATNs

Case Frames

Case frames were introduced by Fillmore (a linguist) to account for essential equivalence of sentences like:

- "John broke the window with a hammer"
- "The window was broken by John with a hammer"
- "Using a hammer, John broke the window"

[*head*: BREAK
agent: JOHN
object: WINDOW
instrument: HAMMER
]

Case Frames

Fillmore postulated finite set of cases applicable to all actions:

[*head:* <the action>
agent: <the active causal agent agent instigating the action>
object: <the object upon which the action is done>
instrument: <an instrument used to assist in the action>
recipient: <the receiver of an action-often the I-OBJ>
directive:<the target of an (usually physical) action>
locative: <the location where the action takes place>
benefactive: <the entity for whom the action is taken>
source: <where the object acted upon comes from>
temporal <when the action takes place>
co-agent: <a secondary or assistant active agent>]

Case Frame Examples

- “John broke the window with a hammer on Elm Street for Billy on Tuesday”
- “John broke the window with Sally”
- “Sally threw the ball at Billy”
- “Billy gave Sally the baseball bat”
- “Billy took the bat from his house to the playground”

Uninstantiated Case Frame

```
[CASE-F:  [HEADER [NAME: “move”]  [PATTERN: <move>]]
     [OBJECT:  [VALUE:   ________ ]
               [POSITION:  DO]
               [SEM-FILLER:  <file> |  <directory>]]
     [DESTINATION:  [VALUE: __________   ]
                    [MARKER:  <dest> ]
                    [SEM-FILLER:  <directory> | <O-port> ]]
     [SOURCE:  [VALUE:  __________   ]
                [MARKER:  <source> ]
                [SEM-FILLER:  <directory> | <I-port> ]]]
```

Case-Frame Grammar Fragments

HEADER PATTERN determines which case frame to instantiate

<move> → "move" | "transfer" | …

<delete> → "delete" | "erase" | "flush" | …

LEXICAL MARKERS are prepositions++ that assign NPs to cases

<dest> → "to" | "into" | "onto" | …

<source> → "from" | "in" | "that's in" | …

POSITIONAL INDICATORS also assign NPs to cases

DO means "direct object position" (unmarked NP right of V)

SUBJ means "subject position" (unmarked NP left of V)

Case Frame Instantiation Process

- Select which case-frame(s) match input string
  - Match header-patterns against input
- Set up constraint-satisfaction problem
  - SEM-FILLER, POSITION, MARKER → constraints
  - At-most one value per case → constraint
  - Any required case must be filled → constraint
  - At-most one case per input-substring → constraint
- Solve constraint-satisfaction problem
  - Use least-commitment, or satisfiability algorithm

Instantiated Case Frame

S1: “Please transfer foo.c from the diskette to my notes directory”

[CASE-F: [HEADER [NAME: “move”] [VALUE: S1]]

[OBJECT: [VALUE: “foo.c”]]

[DESTINATION: [VALUE: “notes directory”]]

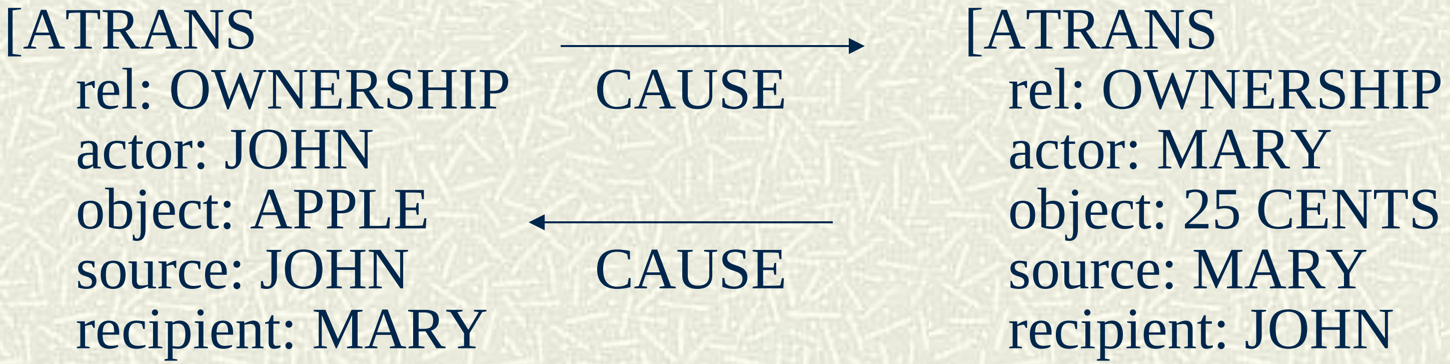
[SOURCE: [VALUE: “diskette”]]]

Conceptual Dependency

- Canonical representation of NL developed by Schank
- Computational motivation—organization of inferences

[ATRANS
rel: POSSESSION
actor: JOHN
object: BALL
source: JOHN
recipient: MARY]
"John gave Mary a ball"

[ATRANS
rel: POSSESSION
actor: MARY
object: BALL
source: JOHN
recipient: MARY]
"Mary took the ball from John"



[ATRANS
rel: OWNERSHIP
actor: JOHN
object: APPLE
source: JOHN
recipient: MARY

CAUSE
CAUSE

[ATRANS
rel: OWNERSHIP
actor: MARY
object: 25 CENTS
source: MARY
recipient: JOHN

"John sold an apple to Mary for 25 Cents."

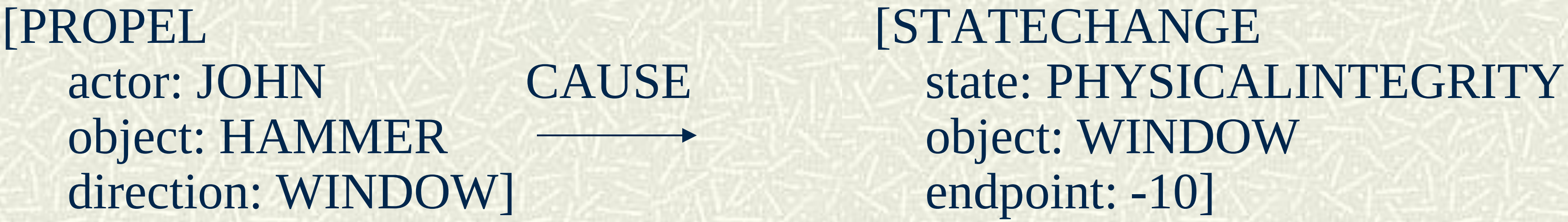
Conceptual Dependency

Other conceptual dependency primitive actions include:

- PTRANS--Physical transfer of location
- MTRANS--Mental transfer of information
- MBUILD--Create a new idea/conclusion from other info
- INGEST--Bring any substance into the body
- PROPEL--Apply a force to an object

States and *causal* relations are also part of the representation:

ENABLE	(State *enables* an action)
RESULT	(An action *results* in a state change)
INITIATE	(State or action *initiates* mental state)
REASON	(Mental state is the internal *reason* for an action)



[PROPEL
actor: JOHN
object: HAMMER
direction: WINDOW]

CAUSE →

[STATECHANGE
state: PHYSICALINTEGRITY
object: WINDOW
endpoint: -10]

"John broke the window with a hammer"

Robust Parsing

Spontaneously generated input will contain errors and items outside an interface's grammar

- Spelling errors

  tarnsfer *Jim Smith from* **Econoics** *237* **too** *Mathematics 156*

- Novel words

  transfer Smith out of Economics 237 to **Basketwork** *100*

- Spurious phrases

  please *enroll Smith* **if that's possible** *in* **I think** *Economics 237*

- Ellipsis or other fragmentary utterances

  also Physics 314

- Unusual word order

  In Economics 237 Jim Smith enroll

- Missing words

  enroll Smith Economics 237

What Makes MT Hard?

- Word Sense
  “Comer” [Spanish] → eat, capture, overlook
  “Banco” [Spanish] → bank, bench

- Specificity
  “Reach” (up) → “atteindre” [French]
  “Reach” (down) → “baisser” [French]
  14 words for “snow” in Inupiac

- Lexical holes
  “Shadenfreuder” [German] → happiness in the misery of others, no such English word

- Syntactic Ambiguity (as discussed earlier)

Bar Hillel's Argument

1. Text must be (minimally) understood before translation can proceed effectively.
2. Computer understanding of text is too difficult.
3. Therefore, Machine Translation is infeasible.

- *Bar Hillel (1960)*

Premise 1 is accurate

Premise 2 *was* accurate in 1960

Some forms of text comprehension are becoming possible with present AI technology, but we have a long way to go. Hence, Bar Hillel's conclusion is losing its validity, but only gradually.

What Makes MT Hard?

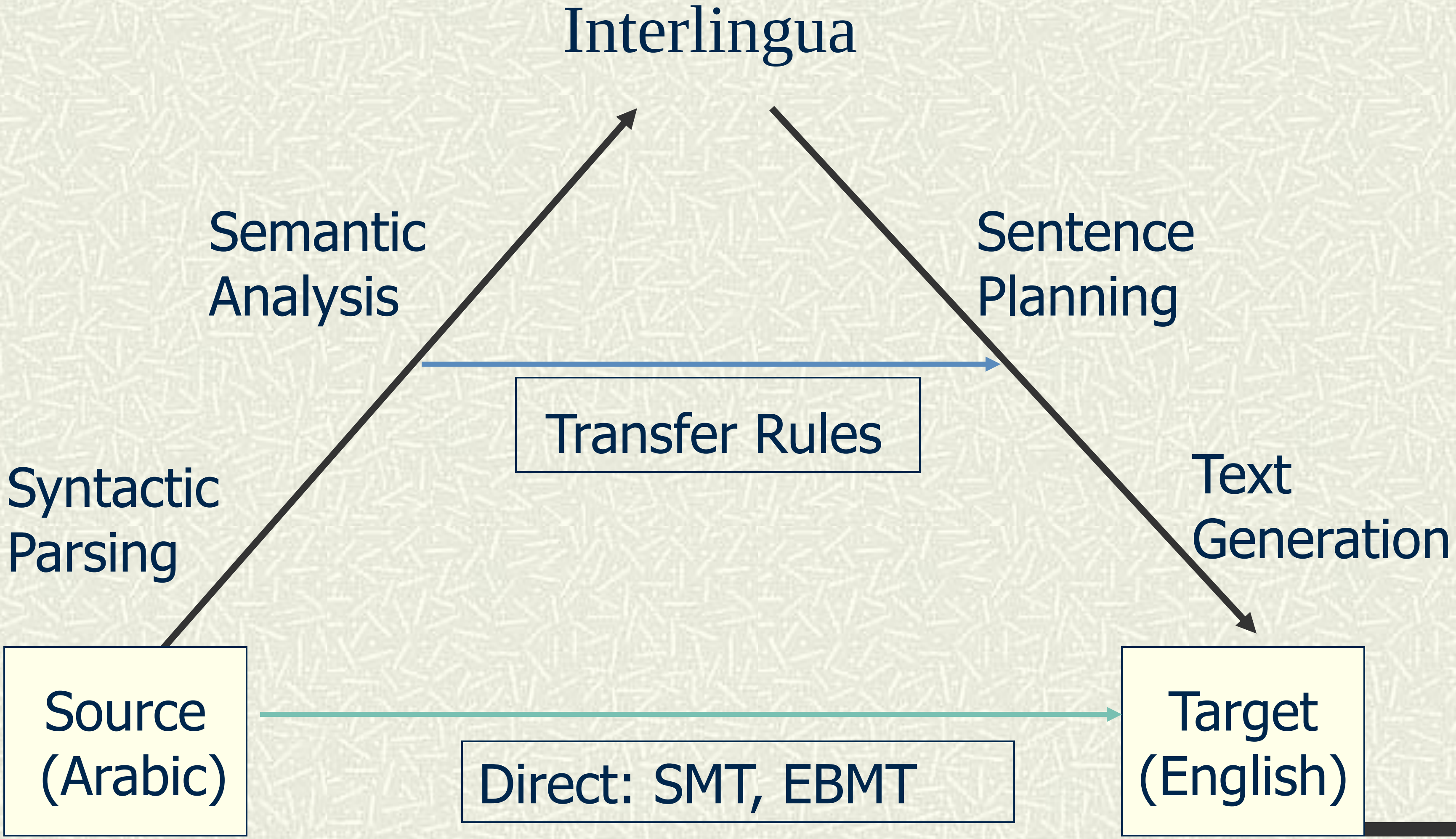
- Word Sense
  “Comer” [Spanish] → eat, capture, overlook
  “Banco” [Spanish] → bank, bench

- Specificity
  “Reach” (up) → “atteindre” [French]
  “Reach” (down) → “baisser” [French]
  14 words for “snow” in Inupiac

- Lexical holes
  “Shadenfreuder” [German] → happiness in the misery of others, no such English word

- Syntactic Ambiguity (as discussed earlier)

Types of Machine Translation



Interlingua

Semantic Analysis

Sentence Planning

Transfer Rules

Syntactic Parsing

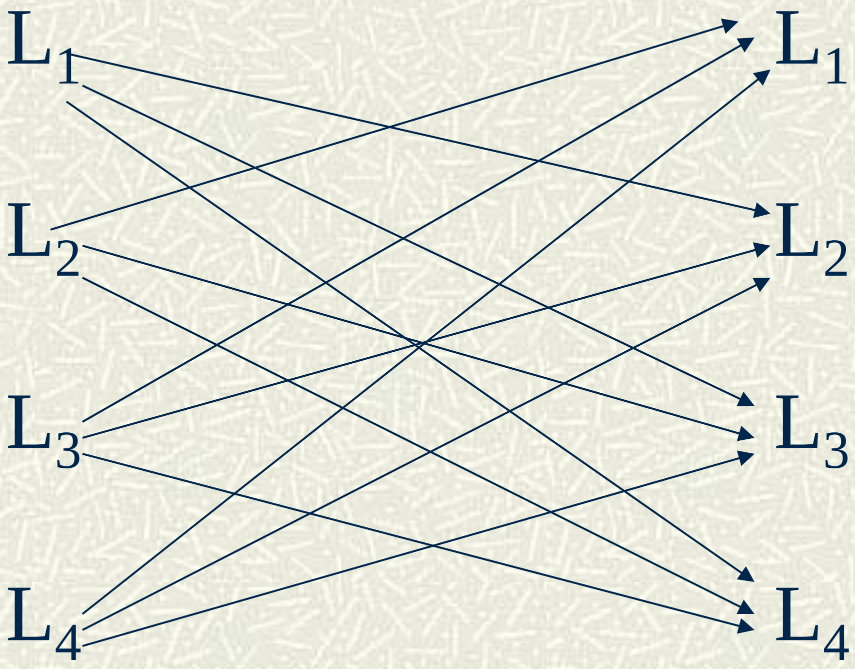
Text Generation

Source (Arabic)

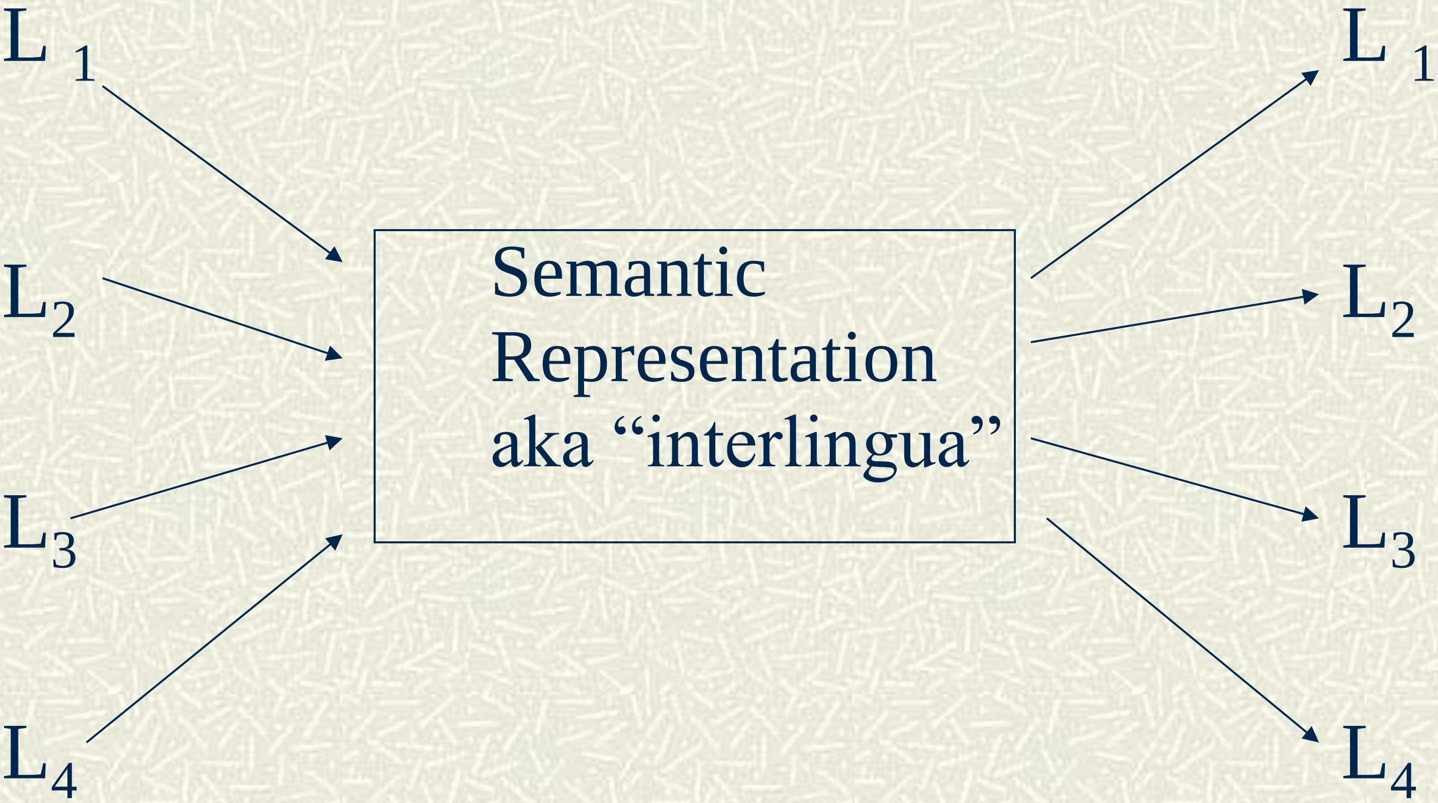
Direct: SMT, EBMT

Target (English)

Transfer Grammars: N(N-1)



Interlingua Paradigm for MT (2N)



For N = 72, T/G → 5112 grammars, Interlingua → 144

Beyond Parsing, Generation and MT

- Anaphora and Ellipsis Resolution
  - "Mary got a nice present from Cindy. *It* was *her* birthday."
  - "John likes oranges and Mary apples."
- Dialog Processing
  - "Speech Acts" (literal → intended message)
  - Social Role context →s peech act selection
  - "General" context sometimes needed
- Example

10-year old: "I want a juicy Hamburger!"
Mother: "Not today, perhaps tomorrow…"

General: "I want a juicy Hamburger."
Aide: "Yes, sir!!"

Prisoner 1: "I want a juicy Hamburger."
Prisoner 2: "Wouldn't that be nice for once."

Social Role Determines Interpretation

10-year old: “I want a juicy Hamburger!”
Mother: “Not today, perhaps tomorrow…”

General: “I want a juicy Hamburger!”
Aide: “Yes, sir!!”

Prisoner 1: “I want a juicy Hamburger!”
Prisoner 2: “Wouldn't that be nice for once!”

Merit Cigarette Advertisement

Merit
Smashes
Taste
Barrier.

-National Smoker Study

Majority of smokers confirm 'Enriched Flavor' cigarette matches taste of leading high tar brands.

Why do we intepret barrier-smashing as good?
[Metaphors, Metonomy, … other hard stuff]

UBLIS 571 Soergel. A supplemental reading for Lecture 6.2b, Applications on p. 13. This is more about how NLP is done, the Feldman Jabberwocky reading more about applications in searching

Natural Language Processing[1]. Elizabeth Liddy

INTRODUCTION

Natural Language Processing (NLP) is the computerized approach to analyzing text that is based on both a set of theories and a set of technologies. And, being a very active area of research and development, there is not a single agreed-upon definition that would satisfy everyone, but there are some aspects, which would be part of any knowledgeable person's definition. The definition I offer is:

> **Definition:** Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.

Several elements of this definition can be further detailed. Firstly the imprecise notion of *'range of computational techniques'* is necessary because there are multiple methods or techniques from which to choose to accomplish a particular type of language analysis.

'Naturally occurring texts' can be of any language, mode, genre, etc. The texts can be oral or written. The only requirement is that they be in a language used by humans to communicate to one another. Also, the text being analyzed should not be specifically constructed for the purpose of the analysis, but rather that the text be gathered from actual usage.

The notion of *'levels of linguistic analysis'* (to be further explained in Section 2) refers to the fact that there are multiple types of language processing known to be at work when humans produce or comprehend language. It is thought that humans normally utilize all of these levels since each level conveys different types of meaning. But various NLP systems utilize different levels, or combinations of levels of linguistic analysis, and this is seen in the differences amongst various NLP applications. This also leads to much confusion on the part of non-specialists as to what NLP really is, because a system that uses any subset of these levels of analysis can be said to be an NLP-based system. The difference between them, therefore, may actually be whether the system uses 'weak' NLP or 'strong' NLP.

'Human-like language processing' reveals that NLP is considered a discipline within Artificial Intelligence (AI). And while the full lineage of NLP does depend on a number of other disciplines, since NLP strives for human-like performance, it is appropriate to consider it an AI discipline.

'For a range of tasks or applications' points out that NLP is not usually considered a goal in and of itself, except perhaps for AI researchers. For others, NLP is the means for

[1] Liddy, E. D. In Encyclopedia of Library and Information Science, 2nd Ed. Marcel Decker, Inc.

accomplishing a particular task. Therefore, you have Information Retrieval (IR) systems that utilize NLP, as well as Machine Translation (MT), Question-Answering, etc.

Goal

The goal of NLP as stated above is "*to accomplish human-like language processing*". The choice of the word 'processing' is very deliberate, and should not be replaced with 'understanding'. For although the field of NLP was originally referred to as Natural Language Understanding (NLU) in the early days of AI, it is well agreed today that while the goal of NLP is true NLU, that goal has not yet been accomplished. A full NLU System would be able to:

1. Paraphrase an input text
2. Translate the text into another language
3. Answer questions about the contents of the text
4. Draw inferences from the text

While NLP has made serious inroads into accomplishing goals 1 to 3, the fact that NLP systems cannot, of themselves, draw inferences from text, NLU still remains the goal of NLP.

There are more practical goals for NLP, many related to the particular application for which it is being utilized. For example, an NLP-based IR system has the goal of providing more precise, complete information in response to a user's real information need. The goal of the NLP system here is to represent the true meaning and intent of the user's query, which can be expressed as naturally in everyday language as if they were speaking to a reference librarian. Also, the contents of the documents that are being searched will be represented at all their levels of meaning so that a true match between need and response can be found, no matter how either are expressed in their surface form.

Origins

As most modern disciplines, the lineage of NLP is indeed mixed, and still today has strong emphases by different groups whose backgrounds are more influenced by one or another of the disciplines. Key among the contributors to the discipline and practice of NLP are: *Linguistics* - focuses on formal, structural models of language and the discovery of language universals - in fact the field of NLP was originally referred to as Computational Linguistics; Computer *Science* - is concerned with developing internal representations of data and efficient processing of these structures, and; Cognitive *Psychology* - looks at language usage as a window into human cognitive processes, and has the goal of modeling the use of language in a psychologically plausible way.

Divisions

While the entire field is referred to as Natural Language Processing, there are in fact two distinct focuses – language processing and language generation. The first of these refers to the analysis of language for the purpose of producing a meaningful representation, while the latter refers to the production of language from a representation. The task of Natural Language Processing is equivalent to the role of reader/listener, while the task of Natural Language Generation is that of the writer/speaker. While much of the theory and technology are shared by these two divisions, Natural Language Generation also requires a planning capability. That is, the generation system requires a plan or model of the goal of the interaction in order to decide what the system should generate at each point in an interaction. We will focus on the task of natural language analysis, as this is most relevant to Library and Information Science.

Another distinction is traditionally made between language understanding and speech understanding. Speech understanding starts with, and speech generation ends with, oral language and therefore rely on the additional fields of acoustics and phonology. Speech understanding focuses on how the 'sounds' of language as picked up by the system in the form of acoustical waves are transcribed into recognizable morphemes and words. Once in this form, the same levels of processing which are utilized on written text are utilized. All of these levels, including the phonology level, will be covered in Section 2; however, the emphasis throughout will be on language in the written form.

BRIEF HISTORY OF NATURAL LANGUAGE PROCESSING

Research in natural language processing has been going on for several decades dating back to the late 1940s. Machine translation (MT) was the first computer-based application related to natural language. While Weaver and Booth (1); (2) started one of the earliest MT projects in 1946 on computer translation based on expertise in breaking enemy codes during World War II, it was generally agreed that it was Weaver's memorandum of 1949 that brought the idea of MT to general notice and inspired many projects (3). He suggested using ideas from cryptography and information theory for language translation. Research began at various research institutions in the United States within a few years.

Early work in MT took the simplistic view that the only differences between languages resided in their vocabularies and the permitted word orders. Systems developed from this perspective simply used dictionary-lookup for appropriate words for translation and reordered the words after translation to fit the word-order rules of the target language, without taking into account the lexical ambiguity inherent in natural language. This produced poor results. The apparent failure made researchers realize that the task was a lot harder than anticipated, and they needed a more adequate theory of language. However, it was not until 1957 when Chomsky (4) published *Syntactic Structures* introducing the idea of generative grammar, did the field gain better insight into whether or how mainstream linguistics could help MT.

During this period, other NLP application areas began to emerge, such as speech recognition. The language processing community and the speech community then was split into two camps with the language processing community dominated by the theoretical perspective of generative grammar and hostile to statistical methods, and the speech community dominated by statistical information theory (5) and hostile to theoretical linguistics (6).

Due to the developments of the syntactic theory of language and parsing algorithms, there was over-enthusiasm in the 1950s that people believed that fully automatic high quality translation systems (2) would be able to produce results indistinguishable from those of human translators, and such systems should be in operation within a few years. It was not only unrealistic given the then-available linguistic knowledge and computer systems, but also impossible in principle (3).

The inadequacies of then-existing systems, and perhaps accompanied by the over-enthusiasm, led to the ALPAC (Automatic Language Processing Advisory Committee of the National Academy of Science - National Research Council) report of 1966. (7) The report concluded that MT was not immediately achievable and recommended it not be funded. This had the effect of halting MT and most work in other applications of NLP at least within the United States.

Although there was a substantial decrease in NLP work during the years after the ALPAC report, there were some significant developments, both in theoretical issues and in construction of prototype systems. Theoretical work in the late 1960's and early 1970's focused on the issue of how to represent meaning and developing computationally tractable solutions that the then-existing theories of grammar were not able to produce. In 1965, Chomsky (8) introduced the transformational model of linguistic competence. However, the transformational generative grammars were too syntactically oriented to allow for semantic concerns. They also did not lend themselves easily to computational implementation. As a reaction to Chomsky's theories and the work of other transformational generativists, case grammar of Fillmore, (9), semantic networks of Quillian, (10), and conceptual dependency theory of Schank, (11) were developed to explain syntactic anomalies, and provide semantic representations. Augmented transition networks of Woods, (12) extended the power of phrase-structure grammar by incorporating mechanisms from programming languages such as LISP. Other representation formalisms included Wilks' preference semantics (13), and Kay's functional grammar (14).

Alongside theoretical development, many prototype systems were developed to demonstrate the effectiveness of particular principles. Weizenbaum's ELIZA (15) was built to replicate the conversation between a psychologist and a patient, simply by permuting or echoing the user input. Winograd's SHRDLU (16) simulated a robot that manipulated blocks on a tabletop. Despite its limitations, it showed that natural language understanding was indeed possible for the computer (17). PARRY (18) attempted to embody a theory of paranoia in a system. Instead of single keywords, it used groups of keywords, and used synonyms if keywords were not found. LUNAR was developed by Woods (19) as an interface system to a database that consisted of information about lunar rock samples using augmented transition network and procedural semantics (20).

In the late 1970's, attention shifted to semantic issues, discourse phenomena, and communicative goals and plans (21). Grosz (22) analyzed task-oriented dialogues and proposed a theory to partition the discourse into units based on her findings about the relation between the structure of a task and the structure of the task-oriented dialogue. Mann and Thompson (23) developed Rhetorical Structure Theory, attributing hierarchical structure to discourse. Other researchers have also made significant contributions, including Hobbs and Rosenschein (24), Polanyi and Scha (25), and Reichman (26).

This period also saw considerable work on natural language generation. McKeown's discourse planner TEXT (27) and McDonald's response generator MUMMBLE (28) used rhetorical predicates to produce declarative descriptions in the form of short texts, usually paragraphs. TEXT's ability to generate coherent responses online was considered a major achievement.

In the early 1980s, motivated by the availability of critical computational resources, the growing awareness within each community of the limitations of isolated solutions to NLP problems (21), and a general push toward applications that worked with language in a broad, real-world context (6), researchers started re-examining non-symbolic approaches that had lost popularity in early days. By the end of 1980s, symbolic approaches had been used to address many significant problems in NLP and statistical approaches were shown to be complementary in many respects to symbolic approaches (21).

In the last ten years of the millennium, the field was growing rapidly. This can be attributed to: a) increased availability of large amounts of electronic text; b) availability of computers with increased speed and memory; and c) the advent of the Internet. Statistical approaches succeeded in dealing with many generic problems in computational linguistics such as part-of-speech identification, word sense disambiguation, etc., and have become standard throughout NLP (29). NLP researchers are now developing next generation NLP systems that deal reasonably well with general text and account for a good portion of the variability and ambiguity of language.

LEVELS OF NATURAL LANGUAGE PROCESSING

The most explanatory method for presenting what actually happens within a Natural Language Processing system is by means of the 'levels of language' approach. This is also referred to as the synchronic model of language and is distinguished from the earlier sequential model, which hypothesizes that the levels of human language processing follow one another in a strictly sequential manner. Psycholinguistic research suggests that language processing is much more dynamic, as the levels can interact in a variety of orders. Introspection reveals that we frequently use information we gain from what is typically thought of as a higher level of processing to assist in a lower level of analysis. For example, the pragmatic knowledge that the document you are reading is about biology will be used when a particular word that has several possible senses (or meanings) is encountered, and the word will be interpreted as having the biology sense.

Of necessity, the following description of levels will be presented sequentially. The key point here is that meaning is conveyed by each and every level of language and that since humans have been shown to use all levels of language to gain understanding, the more capable an NLP system is, the more levels of language it will utilize.

(Figure 1: Synchronized Model of Language Processing)

Phonology

This level deals with the interpretation of speech sounds within and across words. There are, in fact, three types of rules used in phonological analysis: 1) phonetic rules – for sounds within words; 2) phonemic rules – for variations of pronunciation when words are spoken together, and; 3) prosodic rules – for fluctuation in stress and intonation across a sentence. In an NLP system that accepts spoken input, the sound waves are analyzed and encoded into a digitized signal for interpretation by various rules or by comparison to the particular language model being utilized.

Morphology

This level deals with the componential nature of words, which are composed of morphemes – the smallest units of meaning. For example, the word *preregistration* can be morphologically analyzed into three separate morphemes: the prefix *pre*, the root *registra,* and the suffix *tion*. Since the meaning of each morpheme remains the same across words, humans can break down an unknown word into its constituent morphemes in order to understand its meaning. Similarly, an NLP system can recognize the meaning conveyed by each morpheme in order to gain and represent meaning. For example, adding the suffix *–ed* to a verb, conveys that the action of the verb took place in the past. This is a key piece of meaning, and in fact, is frequently only evidenced in a text by the use of the *-ed* morpheme.

Lexical

At this level, humans, as well as NLP systems, interpret the meaning of individual words. Several types of processing contribute to word-level understanding – the first of these being assignment of a single part-of-speech tag to each word. In this processing, words that can function as more than one part-of-speech are assigned the most probable part-of-speech tag based on the context in which they occur.

Additionally at the lexical level, those words that have only one possible sense or meaning can be replaced by a semantic representation of that meaning. The nature of the representation varies according to the semantic theory utilized in the NLP system. The following representation of the meaning of the word *launch* is in the form of logical predicates. As can be observed, a single lexical unit is decomposed into its more basic properties. Given that there is a set of semantic primitives used across all words, these simplified lexical representations make it possible to unify meaning across words and to produce complex interpretations, much the same as humans do.

```
launch (a large boat used for carrying people on rivers, lakes harbors, etc.)
((CLASS BOAT) (PROPERTIES (LARGE)
(PURPOSE  (PREDICATION (CLASS CARRY) (OBJECT PEOPLE))))
```

The lexical level may require a lexicon, and the particular approach taken by an NLP system will determine whether a lexicon will be utilized, as well as the nature and extent of information that is encoded in the lexicon. Lexicons may be quite simple, with only the words and their part(s)-of-speech, or may be increasingly complex and contain information on the semantic class of the word, what arguments it takes, and the semantic limitations on these arguments, definitions of the sense(s) in the semantic representation utilized in the particular system, and even the semantic field in which each sense of a polysemous word is used.

Syntactic

This level focuses on analyzing the words in a sentence so as to uncover the grammatical structure of the sentence. This requires both a grammar and a parser. The output of this level of processing is a (possibly delinearized) representation of the sentence that reveals the structural dependency relationships between the words. There are various grammars that can be utilized, and which will, in turn, impact the choice of a parser. Not all NLP applications require a full parse of sentences, therefore the remaining challenges in parsing of prepositional phrase attachment and conjunction scoping no longer stymie those applications for which phrasal and clausal dependencies are sufficient. Syntax conveys meaning in most languages because order and dependency contribute to meaning. For example the two sentences: *'The dog chased the cat.'* and *'The cat chased the dog.'* differ only in terms of syntax, yet convey quite different meanings.

Semantic

This is the level at which most people think meaning is determined, however, as we can see in the above defining of the levels, it is all the levels that contribute to meaning. Semantic processing determines the possible meanings of a sentence by focusing on the interactions among word-level meanings in the sentence. This level of processing can include the semantic disambiguation of words with multiple senses; in an analogous way to how syntactic disambiguation of words that can function as multiple parts-of-speech is accomplished at the syntactic level. Semantic disambiguation permits one and only one sense of polysemous words to be selected and included in the semantic representation of the sentence. For example, amongst other meanings, *'file'* as a noun can mean either a folder for storing papers, or a tool to shape one's fingernails, or a line of individuals in a queue. If information from the rest of the sentence were required for the disambiguation, the semantic, not the lexical level, would do the disambiguation. A wide range of methods can be implemented to accomplish the disambiguation, some which require information as to the frequency with which each sense occurs in a particular corpus of interest, or in general usage, some which require consideration of the local context, and others which utilize pragmatic knowledge of the domain of the document.

Discourse

While syntax and semantics work with sentence-length units, the discourse level of NLP works with units of text longer than a sentence. That is, it does not interpret multi-sentence texts as just concatenated sentences, each of which can be interpreted singly. Rather, discourse focuses on the properties of the text as a whole that convey meaning by making connections between component sentences. Several types of discourse processing can occur at this level, two of the most common being anaphora resolution and discourse/text structure recognition. Anaphora resolution is the replacing of words such as pronouns, which are semantically vacant, with the appropriate entity to which they refer (30). Discourse/text structure recognition determines the functions of sentences in the text, which, in turn, adds to the meaningful representation of the text. For example, newspaper articles can be deconstructed into discourse components such as: Lead, Main Story, Previous Events, Evaluation, Attributed Quotes, and Expectation (31).

Pragmatic

This level is concerned with the purposeful use of language in situations and utilizes context over and above the contents of the text for understanding The goal is to explain how extra meaning is *read into* texts without actually being encoded in them. This requires much world knowledge, including the understanding of intentions, plans, and goals. Some NLP applications may utilize knowledge bases and inferencing modules. For example, the following two sentences require resolution of the anaphoric term ‘they’, but this resolution requires pragmatic or world knowledge.

> *The city councilors refused the demonstrators a permit because **they** feared violence.*
>
> *The city councilors refused the demonstrators a permit because **they** advocated revolution.*

Summary of Levels

Current NLP systems tend to implement modules to accomplish mainly the lower levels of processing. This is for several reasons. First, the application may not require interpretation at the higher levels. Secondly, the lower levels have been more thoroughly researched and implemented. Thirdly, the lower levels deal with smaller units of analysis, e.g. morphemes, words, and sentences, which are rule-governed, versus the higher levels of language processing which deal with texts and world knowledge, and which are only regularity-governed. As will be seen in the following section on Approaches, the statistical approaches have, to date, been validated on the lower levels of analysis, while the symbolic approaches have dealt with all levels, although there are still few working systems which incorporate the higher levels.

APPROACHES TO NATURAL LANGUAGE PROCESSING

Natural language processing approaches fall roughly into four categories: symbolic, statistical, connectionist, and hybrid. Symbolic and statistical approaches have coexisted since the early days of this field. Connectionist NLP work first appeared in the 1960’s. For a long time, symbolic approaches dominated the field. In the 1980’s, statistical approaches regained popularity as a result of the availability of critical computational resources and the need to deal with broad, real-world contexts. Connectionist approaches also recovered from earlier criticism by demonstrating the utility of neural networks in NLP. This section examines each of these approaches in terms of their foundations, typical techniques, differences in processing and system aspects, and their robustness, flexibility, and suitability for various tasks.

Symbolic Approach

Symbolic approaches perform deep analysis of linguistic phenomena and are based on explicit representation of facts about language through well-understood knowledge representation schemes and associated algorithms (21). In fact, the description of the levels of language analysis in the preceding section is given from a symbolic perspective. The primary source of evidence in symbolic systems comes from human-developed rules and lexicons.

A good example of symbolic approaches is seen in logic or rule-based systems. In logic-based systems, the symbolic structure is usually in the form of logic propositions. Manipulations of such structures are defined by inference procedures that are generally truth preserving. Rule-based systems usually consist of a set of rules, an inference engine, and a workspace or working memory. Knowledge is represented as facts or rules in the rule-base. The inference engine repeatedly selects a rule whose condition is satisfied and executes the rule.

Another example of symbolic approaches is semantic networks. First proposed by Quillian (10) to model associative memory in psychology, semantic networks represent knowledge through a set of nodes that represent objects or concepts and the labeled links that represent relations between nodes. The pattern of connectivity reflects semantic organization, that is; highly associated concepts are directly linked whereas moderately or weakly related concepts are linked through intervening concepts. Semantic networks are widely used to represent structured knowledge and have the most connectionist flavor of the symbolic models (32).

Symbolic approaches have been used for a few decades in a variety of research areas and applications such as information extraction, text categorization, ambiguity resolution, and lexical acquisition. Typical techniques include: explanation-based learning, rule-based learning, inductive logic programming, decision trees, conceptual clustering, and K nearest neighbor algorithms (6; 33).

Statistical Approach

Statistical approaches employ various mathematical techniques and often use large text corpora to develop approximate generalized models of linguistic phenomena based on actual examples of these phenomena provided by the text corpora without adding significant linguistic or world knowledge. In contrast to symbolic approaches, statistical approaches use observable data as the primary source of evidence.

A frequently used statistical model is the Hidden Markov Model (HMM) inherited from the speech community. HMM is a finite state automaton that has a set of states with probabilities attached to transitions between states (34). Although outputs are visible, states themselves are not directly observable, thus “hidden” from external observations. Each state produces one of the observable outputs with a certain probability.

Statistical approaches have typically been used in tasks such as speech recognition, lexical acquisition, parsing, part-of-speech tagging, collocations, statistical machine translation, statistical grammar learning, and so on.

Connectionist Approach

Similar to the statistical approaches, connectionist approaches also develop generalized models from examples of linguistic phenomena. What separates connectionism from other statistical methods is that connectionist models combine statistical learning with various theories of representation - thus the connectionist representations allow transformation, inference, and manipulation of logic formulae (33). In addition, in connectionist systems, linguistic models are harder to observe due to the fact that connectionist architectures are less constrained than statistical ones (35); (21).

Generally speaking, a connectionist model is a network of interconnected simple processing units with knowledge stored in the weights of the connections between units (32). Local interactions among units can result in dynamic global behavior, which, in turn, leads to computation.

Some connectionist models are called localist models, assuming that each unit represents a particular concept. For example, one unit might represent the concept “mammal” while another unit might represent the concept “whale”. Relations between concepts are encoded by the weights of connections between those concepts. Knowledge in such models is spread across the network, and the connectivity between units reflects their structural relationship. Localist models are quite similar to semantic networks, but the links between units are not usually labeled as they are in semantic nets. They perform well at tasks such as word-sense disambiguation, language generation, and limited inference (36).

Other connectionist models are called distributed models. Unlike that in localist models, a concept in distributed models is represented as a function of simultaneous activation of multiple units. An individual unit only participates in a concept representation. These models are well suited for natural language processing tasks such as syntactic parsing, limited domain translation tasks, and associative retrieval.

Comparison Among Approaches

From the above section, we have seen that similarities and differences exist between approaches in terms of their assumptions, philosophical foundations, and source of evidence. In addition to that, the similarities and differences can also be reflected in the processes each approach follows, as well as in system aspects, robustness, flexibility, and suitable tasks.

Process: Research using these different approaches follows a general set of steps, namely, data collection, data analysis/model building, rule/data construction, and application of rules/data in system. The data collection stage is critical to all three approaches although statistical and connectionist approaches typically require much more data than symbolic approaches. In the data analysis/model building stage, symbolic approaches rely on human analysis of the data in order to form a theory while statistical approaches manually define a statistical model that is an approximate generalization of the collected data. Connectionist approaches build a connectionist model from the data. In the rule / data construction stage, manual efforts are typical for symbolic approaches and the theory formed in the previous step may evolve when new cases are encountered. In contrast, statistical and connectionist approaches use the statistical or connectionist model as guidance and build rules or data items automatically, usually in relatively large quantity. After building rules or data items, all approaches then automatically apply them to specific tasks in the system. For instance, connectionist approaches may apply the rules to train the weights of links between units.

System aspects: By system aspects, we mean source of data, theory or model formed from data analysis, rules, and basis for evaluation.

- *Data:* As mentioned earlier, symbolic approaches use human introspective data, which are usually not directly observable. Statistical and connectionist approaches are built on the basis of machine observable facets of data, usually from text corpora.

- *Theory or model based on data analysis:* As the outcome of data analysis, a theory is formed for symbolic approaches whereas a parametric model is formed for statistical approaches and a connectionist model is formed for connectionist approaches.

- *Rules:* For symbolic approaches, the rule construction stage usually results in rules with detailed criteria of rule application. For statistical approaches, the criteria of rule application are usually at the surface level or under-specified. For connectionist approaches, individual rules typically cannot be recognized.

- *Basis for Evaluation:* Evaluation of symbolic systems is typically based on intuitive judgments of unaffiliated subjects and may use system-internal measures of growth such as the number of new rules. In contrast, the basis for evaluation of statistical and connectionist systems are usually in the form of scores computed from some evaluation function. However, if all approaches are utilized for the same task, then the results of the task can be evaluated both quantitatively and qualitatively and compared.

Robustness: Symbolic systems may be fragile when presented with unusual, or noisy input. To deal with anomalies, they can anticipate them by making the grammar more general to accommodate them. Compared to symbolic systems, statistical systems may be more robust in the face of unexpected input provided that training data is sufficient, which may be difficult to be assured of. Connectionist systems may also be robust and fault tolerant because knowledge in such systems is stored across the network. When presented with noisy input, they degrade gradually.

Flexibility: Since symbolic models are built by human analysis of well-formulated examples, symbolic systems may lack the flexibility to adapt dynamically to experience. In contrast, statistical systems allow broad coverage, and may be better able to deal with unrestricted text (21) for more effective handling of the task at hand. Connectionist systems exhibit flexibility by dynamically acquiring appropriate behavior based on the given input. For example, the weights of a connectionist network can be adapted in real-time to improve performance. However, such systems may have difficulty with the representation of structures needed to handle complex conceptual relationships, thus limiting their abilities to handle high-level NLP (36).

Suitable tasks: Symbolic approaches seem to be suited for phenomena that exhibit identifiable linguistic behavior. They can be used to model phenomena at all the various linguistic levels described in earlier sections. Statistical approaches have proven to be effective in modeling language phenomena based on frequent use of language as reflected in text corpora. Linguistic phenomena that are not well understood or do not exhibit clear regularity are candidates for statistical approaches. Similar to statistical approaches, connectionist approaches can also deal with linguistic phenomena that are not well understood. They are useful for low-level NLP tasks that are usually subtasks in a larger problem.

To summarize, symbolic, statistical, and connectionist approaches have exhibited different characteristics, thus some problems may be better tackled with one approach while other problems by another. In some cases, for some specific tasks, one approach may prove adequate, while in other cases, the tasks can get so complex that it might not be possible to choose a single best approach. In addition, as Klavans and Resnik (6) pointed out, there is no such thing as a “purely statistical” method. Every use of statistics is based upon a symbolic model and statistics alone is not adequate for NLP. Toward this end, statistical approaches are not at odds with symbolic approaches. In fact, they are rather complementary. As a result, researchers have begun developing hybrid techniques that utilize the strengths of each approach in an attempt to address NLP problems more effectively and in a more flexible manner.

NATURAL LANGUAGE PROCESSING APPLICATIONS

Natural language processing provides both theory and implementations for a range of applications. In fact, any application that utilizes text is a candidate for NLP. The most frequent applications utilizing NLP include the following:

- Information Retrieval – given the significant presence of text in this application, it is surprising that so few implementations utilize NLP. Recently, statistical approaches for accomplishing NLP have seen more utilization, but few systems other than those by Liddy (37) and Strzalkowski (38) have developed significant systems based on NLP

.

- Information Extraction (IE) – a more recent application area, IE focuses on the recognition, tagging, and extraction into a structured representation, certain key elements of information, e.g. persons, companies, locations, organizations, from large collections of text. These extractions can then be utilized for a range of applications including question-answering, visualization, and data mining.

- Question-Answering – in contrast to Information Retrieval, which provides a list of potentially relevant documents in response to a user's query, question-answering provides the user with either just the text of the answer itself or answer-providing passages.

- Summarization – the higher levels of NLP, particularly the discourse level, can empower an implementation that reduces a larger text into a shorter, yet richly-constituted abbreviated narrative representation of the original document.

- Machine Translation – perhaps the oldest of all NLP applications, various levels of NLP have been utilized in MT systems, ranging from the 'word-based' approach to applications that include higher levels of analysis.

- Dialogue Systems – perhaps the omnipresent application of the future, in the systems envisioned by large providers of end-user applications. Dialogue systems, which usually focus on a narrowly defined application (e.g. your refrigerator or home sound system), currently utilize the phonetic and lexical levels of language. It is believed that utilization of all the levels of language processing explained above offer the potential for truly habitable dialogue systems.

CONCLUSIONS

While NLP is a relatively recent area of research and application, as compared to other information technology approaches, there have been sufficient successes to date that suggest that NLP-based information access technologies will continue to be a major area of research and development in information systems now and far into the future.

Acknowledgement

Grateful appreciation to Xiaoyong Liu who contributed to this entry while she was a Ph.D. student and a Research Assistant in the Center for Natural Language Processing in the School of Information Studies at Syracuse University.

Inxight Software, Inc.
U.S. | 500 Macara Avenue, Sunnyvale, CA 94085
Phone: 408.738.6299 | Fax: 408.738.6311
www.inxight.com | sales@inxight.com

Inxight Federal Systems
11951 Freedom Drive, Suite 1300, Reston, VA 20190
Phone: 703.251.4429 | Fax: 703.251.4440
www.inxightfedsys.com | sales@inxightfedsys.com

Inxight SmartDiscovery™ Awareness Server

Uncover New Revenue Opportunities and Speed Time-to-Information through the Power of Inxight Extraction

Inxight SmartDiscovery Awareness Server is a proven federated search and alert solution that helps users derive insight and intelligence from hundreds of high-value information sources through a single interface. Alerts and page tracking automatically inform users of new and updated information impacting their business.

The average knowledge worker spends up to four hours a day searching for information. Having access to timely and complete information is critical for informed decisions. Whether you are a pharmaceutical researcher, product manager, government analyst or legal compliance expert, you need to be able to **search for and be alerted to information residing in both internal and external sources – with only one query.**

Inxight SmartDiscovery Awareness Server is a powerful **federated search** solution that can bring together a variety of information sources, including the open Internet, Deep Web, and subscription content, as well as your internal document, data and records repositories, such as Google Search Appliance and Oracle Secure Enterprise Search installations.

Then, Inxight's sophisticated extraction technology places the results in **intelligent clusters** by extracting and analyzing the most **relevant people, places and organizations.** This **helps you pinpoint documents quickly** and see non-obvious relationships that would otherwise be overlooked.

Inxight helps you stay informed in real-time by giving you the ability **to set up personalized email and mobile alerts** based on changes to a specific document or web page, or the appearance of new information of interest in Deep Web, Internet, news, internal systems or other sources. The Awareness Server is a relentless assistant, relieving you of the drudgery of research and **giving you a jump on the competition.**

In short, SmartDiscovery Awareness Server **automates the process of information discovery.** This means that **instead of spending two to four hours a day looking for information, you can spend those hours *using* the information you find.**

Using Awareness Server, you can:

- Automatically pull together analyst reports, news articles and SEC filings for complete competitive analysis.
- Monitor your brand globally.
- Access information on a given topic, regardless of source, for compliance.
- Monitor chatter from a variety of classified sources, blogs and other open sources.
- Be alerted when competitors' Websites change.

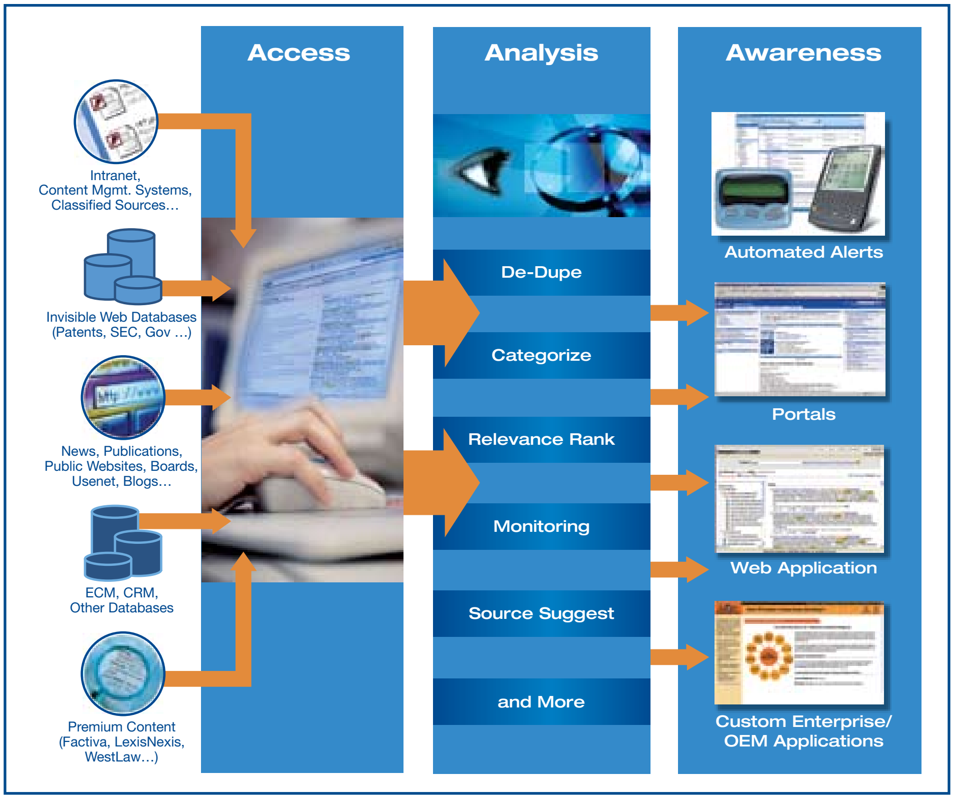
Deliver relevant information to each user's desktop or mobile device based on their individual information needs

Inxight SmartDiscovery Awareness Server greatly reduces the time users spend staying on top of the important changes in their business.

The Alert module in SmartDiscovery Awareness Server includes the ability to set up email and landing page alerts based on changes to a specific document or Web page or the appearance of newly available information of interest in Deep Web, Internet, news, internal systems, or other sources.

Page tracking keeps users aware of changes to a single Web page or document of interest – price changes on a competitive product, patent filing status changes from a patent database, new press releases or news coverage listed on a competitor's Website – whatever information is important to you and your job.

Saved search alerts provide users with email or mobile alerts on the latest newly available information that might impact business decisions – new product coverage, new Usenet postings, new patents and more. The system knows what users have already seen and what is new to them. Users need ask only once to stay on top of daily changes.



Access - Analysis - Awareness with Inxight SmartDiscovery Awareness Server

Inxight SmartDiscovery Awareness Server is a proven federated search and alert solution that adds helps users derive insight and intelligence from hundreds of high-value information sources through a single interface. Alerts and page tracking automatically inform users of new and updated information impacting their business.

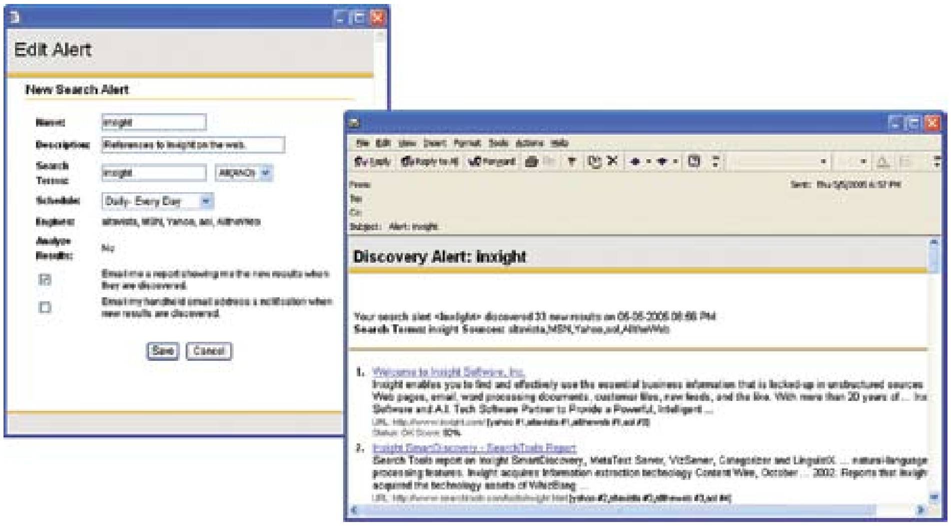
Filtering and Preview Options

Numerous filtering and preview options allow users to **quickly understand and locate relevant information from billions of potential hits** spread around the world in hundreds of internal and external systems.

For example, **relevance ranking** across documents from multiple content sources enables users to **see the most relevant documents first, regardless of source.**

SmartDiscovery Awareness Server quickly categorizes each matching result record and provides a **View by Concept** display that **organizes search results into a visual tree structure of matching categories, enabling rapid navigation by subject.**

Dynamic summarization enables users to browse quickly though volumes of information and quickly find the most relevant documents, **reducing unnecessary click-throughs and saving precious time and effort.**



Edit Alert

New Search Alert

Name: inxight

Description: References to Inxight on the web.

Search Terms: inxight ALL(AND)

Schedule: Daily- Every Day

Engines: altavista, MSN, Yahoo, aol, AlltheWeb

Analyze Results: No

☑ Email me a report showing me the new results when they are discovered.

☐ Email my handheld email address a notification when new results are discovered.

Save
Cancel
File Edit View Insert Format Tools Actions Help
Reply Reply to All Forward

From:
To:
Cc:
Subject: Alert: inxight

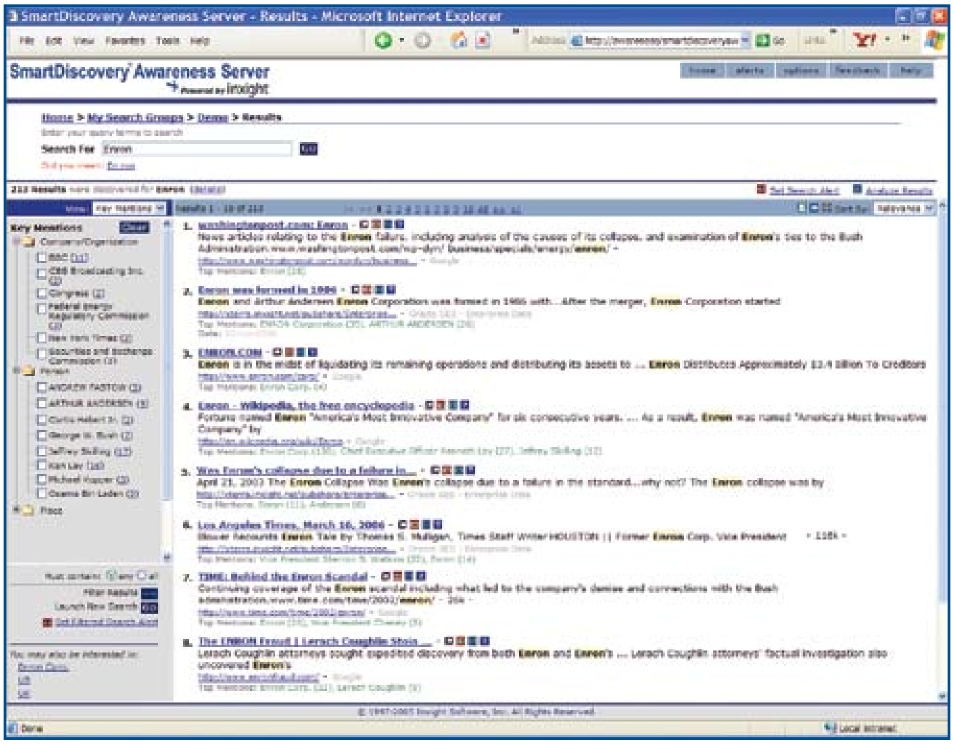
Sent: Thu 5/5/2005 6:57 PM

Discovery Alert: inxight

Your search alert <inxight> discovered 33 new results on 05-05-2005 06:56 PM
Search Terms: inxight **Sources:** altavista,MSN,Yahoo,aol,AlltheWeb

1. Welcome to Inxight Software, Inc.
Inxight enables you to find and effectively use the essential business information that is locked-up in unstructured sources Web pages, email, word processing documents, customer files, new feeds, and the like. With more than 20 years of ... Inx Software and A.I. Tech Software Partner to Provide a Powerful, Intelligent ...
URL: http://www.inxight.com/ [yahoo #1,altavista #1,alltheweb #1,aol #3]
Status: OK Score: 80%
2. Inxight SmartDiscovery - SearchTools Report
Search Tools report on Inxight SmartDiscovery, MetaText Server, VizServer, Categorizer and LinguistX ... natural-language processing features. Inxight acquires information extraction technology Content Wise, October ... 2002: Reports that Inxight acquired the technology assets of VizBang ...
URL: http://www.searchtools.com/tools/inxight.html [yahoo #2,altavista #3,alltheweb #3,aol #4]

Search alerts keep you on top of the latest developments related to your search queries.



SmartDiscovery Awareness Server - Results - Microsoft Internet Explorer

File Edit View Favorites Tools Help
Address
Go
Links
SmartDiscovery Awareness Server
Powered by inxight
home
alerts
options
feedback
help

Home > My Search Groups > Demo > Results

Enter your query terms to search

Search For: Enron

Did you mean: En ron

212 Results were discovered for **Enron** (details)

Key Mentions

- Company/Organization
  - ☐ BBC (11)
  - ☐ CBS Broadcasting Inc. (2)
  - ☐ Congress (2)
  - ☐ Federal Energy Regulatory Commission (2)
  - ☐ New York Times (2)
  - ☐ Securities and Exchange Commission (2)
- Person
  - ☐ ANDREW FASTOW (2)
  - ☐ ARTHUR ANDERSEN (3)
  - ☐ Curtis Hebert Jr. (2)
  - ☐ George W. Bush (2)
  - ☐ Jeffrey Skilling (13)
  - ☐ Ken Lay (10)
  - ☐ Michael Kopper (2)
  - ☐ Osama Bin Laden (2)
- Place

Must contain: ○ any ○ all

You may also be interested in:
Enron Corp.
US
UK

1. washingtonpost.com: Enron
News articles relating to the **Enron** failure, including analysis of the causes of its collapse, and examination of **Enron**'s ties to the Bush Administration.www.washingtonpost.com/wp-dyn/ business/specials/energy/**enron**/ -
http://www.washingtonpost.com/wp-dyn/business... - Google
Top Mentions: Enron (28)
2. Enron was formed in 1986
Enron and Arthur Andersen **Enron** Corporation was formed in 1986 with... After the merger, **Enron** Corporation started
http://... - Oracle [illegible] - Enterprise Data
Top Mentions: ENRON Corporation (35), ARTHUR ANDERSEN (26)
Date: [illegible]
3. ENRON.COM
Enron is in the midst of liquidating its remaining operations and distributing its assets to ... **Enron** Distributes Approximately $1.9 Billion To Creditors
http://www.enron.com/corp/ - Google
Top Mentions: Enron Corp. (4)
4. Enron - Wikipedia, the free encyclopedia
Fortune named **Enron** "America's Most Innovative Company" for six consecutive years. ... As a result, **Enron** was named "America's Most Innovative Company" by
http://en.wikipedia.org/wiki/Enron - Google
Top Mentions: Enron Corp. (130), Chief Executive Officer Kenneth Lay (27), Jeffrey Skilling (12)
5. Was Enron's collapse due to a failure in...
April 21, 2003 The **Enron** Collapse Was **Enron**'s collapse due to a failure in the standard...why not? The **Enron** collapse was by
http://... - Oracle [illegible] - Enterprise Data
Top Mentions: Enron (11), Andersen (8)
6. Los Angeles Times, March 16, 2006
Blower Recounts **Enron** Tale By Thomas S. Mulligan, Times Staff Writer HOUSTON | | Former **Enron** Corp. Vice President - 118k -
http://... - Oracle [illegible] - Enterprise Data
Top Mentions: Vice President Sherron S. Watkins (21), Enron (14)
7. TIME: Behind the Enron Scandal
Continuing coverage of the **Enron** scandal including what led to the company's demise and connections with the Bush administration.www.time.com/time/2002/**enron**/ - 38k -
http://www.time.com/time/2002/enron/ - Google
Top Mentions: Enron (15), Vice President Cheney (3)
8. The ENRON Fraud | Lerach Coughlin Stoia ...
Lerach Coughlin attorneys sought expedited discovery from both **Enron** and **Enron**'s ... Lerach Coughlin attorneys' factual investigation also uncovered **Enron**'s
http://www.enronfraud.com/ - Google
Top Mentions: Enron Corp. (11), Lerach Coughlin (3)

© 1997-2005 Inxight Software, Inc. All Rights Reserved.

Search the public Web, Deep Web, news, subscription content and internal information – all from one screen.

inxight

Inxight SmartDiscovery Awareness Server Features:

Search and Access

- **Global Search:** Users can select from more than 130 pre-configured content sources to perform searches in real time.
- **Rich Boolean, Proximity and Field-based Operators:** The system will automatically translate and transform the universal query provided by the user (using either basic or advanced search) to the form appropriate to each source.
- **Single Sign-On and Authentication:** The system leverages existing SSO and authentication schemes to provide secure access to subscription and other password-protected data sources.

Knowledge and Understanding

- **Eliminates Duplicate Results:** Deletes duplicate result records from multiple sources.
- **Relevance Ranking:** Allows you to quickly see documents from a variety of sources based on their relevance to your information needs.
- **Advanced Sorting and Filtering Options:** Results can be sorted and filtered by relevance, date, source, category or other criteria.
- **"More Like This" Searching:** Users can highlight an individual result and the system will automatically find other similar results.
- **Source Suggest:** Awareness Server will suggest new sources to add to your personal search universe based on your queries.

Awareness and Alerts

- **Collaborative Results Sharing:** Results can be shared with others via email, or formatted for printing and reading offline.
- **Search Alerts:** Personalized search alerts keep you aware of the latest developments related to your search queries.
- **Page Alerts:** Page alerts can be set to monitor individual pages for changes.
- **Background Searches:** Federated searches can be configured to run in the background. The system will notify the user with the results when they are ready.

About **Inxight**

Inxight Software, Inc. is the leading provider of enterprise software solutions for information discovery. Using Inxight solutions, organizations can access and analyze unstructured, semi-structured and structured text to extract key information to enable business intelligence. Inxight is the only company that provides a complete, scalable solution enabling information discovery in more than 30 languages. Customers include enterprise companies such as Air Products, AOL, Merrill Lynch, Morgan Stanley, Novartis and Thomson, multiple U.S. and foreign government agencies, including the Department of Defense, Defense Intelligence Agency, Department of Homeland Security and Commonwealth Secretariat, and software OEMs such as SAP, SAS, Oracle and IBM. The company has offices throughout the United States and Europe. For more information, visit www.inxight.com or call 1-408-738-6200 or +44 (0) 1252 761314.

Technical Specifications

- Open interfaces (APIs), including XML, SOAP and Web Services, allow for easy integration with existing or new enterprise applications.
- Provides search results from Microsoft Office documents, PDFs, XML, HTML and text.
- Works with standards-based and proprietary security mechanisms for access and authentication.
- The product is scalable to support large numbers of users, delivering results to users' requests in seconds.
- Enables a wide range of sources and options to be customized rapidly to meet the needs of the organization, individual departments or even individual users.

Operating Systems

- Windows 2003, and Windows XP on Intel 32-bit processors
- Solaris 8 and Solaris 9 on SPARC processors
- Red Hat Linux AS 4.0 and ES 4.0

Application Servers

- Apache Tomcat 5.0 or 5.5
- BEA WebLogic 9.1
- IBM WebSphere 6.0

Database Servers

- Oracle 9i
- MySQL 4.1
- Microsoft SQL Server 2000

Java Platforms

- Java 2 Standard Edition 1.4.2 and Java Standard Edition 5.0

Browsers

- Microsoft Internet Explorer 5.5 and 6.0

® inxight

ENTERPRISE
PROFESSIONAL
Google

Inxight Software, Inc.
500 Macara Avenue, Sunnyvale, CA 94085
Phone: 408.738.6200 | Fax: 408.738.6311
www.inxight.com | sales@inxight.com

Going Beyond Google with Inxight SmartDiscovery®

Google.com puts the Web's information at your fingertips. It's one of the world's best search indexers. The Google Search Appliance does the same thing for your Website or corporate intranet.

However, the challenge is what to do with all of this information. The key to "organizing the world's information to make it universally accessible and useful" is not only to *index* the world's information, but to be able to electronically *"read"* that information and structure it – pulling out key entities (people, places, companies…), concepts, relations, and events trapped in unstructured text – and then be able to visualize that information in meaningful ways.

Inxight picks up where Google and other search indexers leave off.

Inxight SmartDiscovery Awareness Server

Inxight SmartDiscovery Awareness Server is a powerful federated search, clustering and alert solution that makes discovering information easier than ever before.

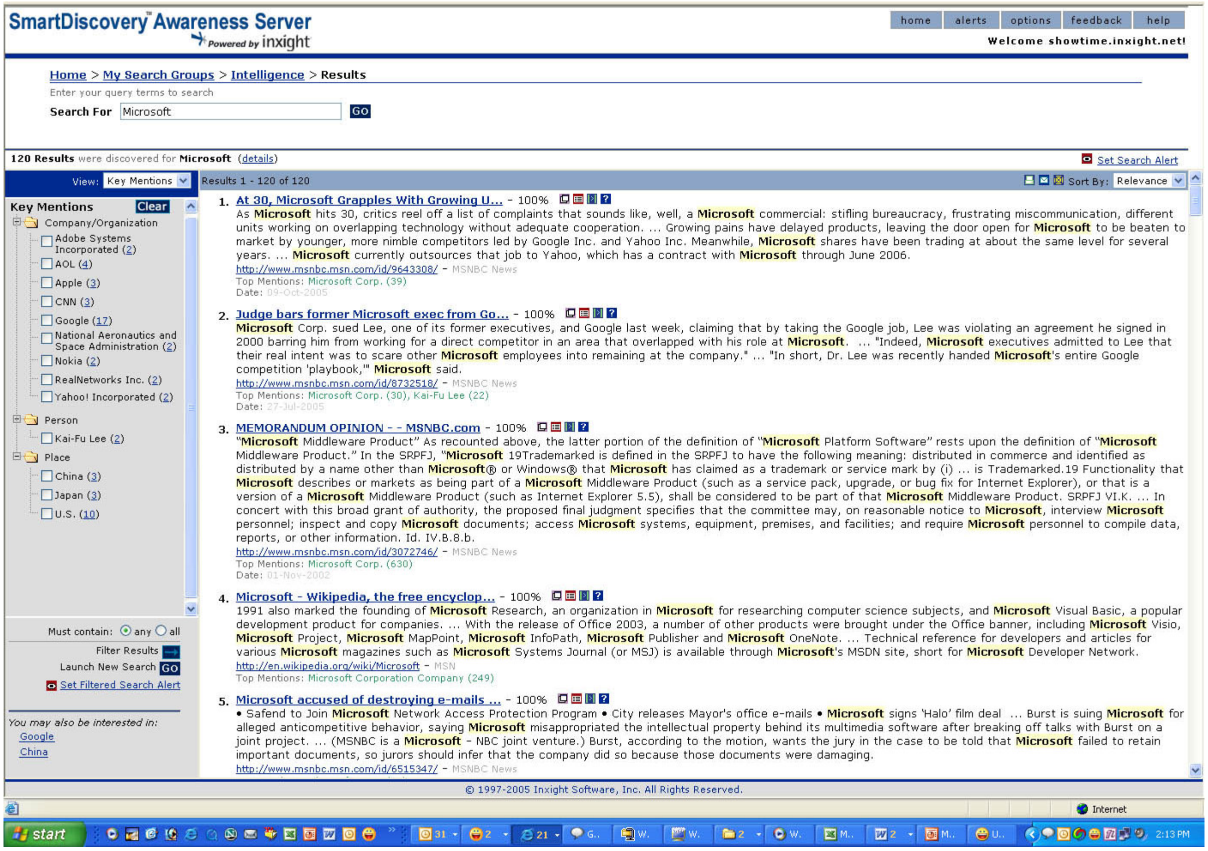
Using SmartDiscovery Awareness Server, you can:

- Search multiple Google Search Appliances that are geographically or departmentally dispersed within your company with only one query.
- Search disparate search indexes and systems within your company, such as Verity indexes, Lotus Domino installations, and Google Search Appliance installations, again with only one query.
- Search Google and other sources (patent databases, analyst subscriptions, etc.) where you don't have permission to index.
- Cluster and filter search results by entity, concept, category or source for faster retrieval of documents of interest.
- Automatically pull together analyst reports, news articles, patent filings, and SEC filings for complete competitive analysis.
- Monitor brand chatter from a variety of blogs, news articles, analyst reports and other sources.
- Be alerted to when competitors' websites change.
- Find the right document - quickly - when you need it.

How can you do this? With a single query on SmartDiscovery Awareness Server, you can securely access Deep Web sources, such as patent databases, SEC filings, and industry publications, Public Web sources, such as blogs, competitor's websites and online news, subscription-only sources, such as PR Newswire and LexisNexis, enterprise sources, such as

Lotus Domino, Google Search Appliance indexes, intranets and Documentum, and classified sources. Results are displayed as a relevance-ranked list, or search queries can be saved as alerts, and results can be emailed to you.

In seconds, you can cluster and filter your federated search results by the most relevant people, companies, places, concepts, weapons, vehicles and other entities mentioned in them. More than 35 entities are available out-of-the-box, and you can customize the system to track new list-based or pattern-based topics such as products, competitors, terrorist names, part numbers, gene sequences, etc. You can also cluster and filter results by pre-defined taxonomies or by source.


SmartDiscovery™ Awareness Server
Powered by inxight
home
alerts
options
feedback
help
Welcome showtime.inxight.net!
Home > My Search Groups > Intelligence > Results
Enter your query terms to search
GO
View: Key Mentions
Key Mentions
Clear
Sort By: Relevance
Must contain: ● any ○ all
Filter Results
Launch New Search GO
Set Filtered Search Alert
Set Search Alert
start
Internet
2:13 PM

Search For: Microsoft

120 Results were discovered for **Microsoft** (details)

Company/Organization
- □ Adobe Systems Incorporated (2)
- □ AOL (4)
- □ Apple (3)
- □ CNN (3)
- □ Google (17)
- □ National Aeronautics and Space Administration (2)
- □ Nokia (2)
- □ RealNetworks Inc. (2)
- □ Yahoo! Incorporated (2)

Person
- □ Kai-Fu Lee (2)

Place
- □ China (3)
- □ Japan (3)
- □ U.S. (10)

You may also be interested in:
Google
China

Results 1 - 120 of 120

1. At 30, Microsoft Grapples With Growing U... - 100%
As **Microsoft** hits 30, critics reel off a list of complaints that sounds like, well, a **Microsoft** commercial: stifling bureaucracy, frustrating miscommunication, different units working on overlapping technology without adequate cooperation. ... Growing pains have delayed products, leaving the door open for **Microsoft** to be beaten to market by younger, more nimble competitors led by Google Inc. and Yahoo Inc. Meanwhile, **Microsoft** shares have been trading at about the same level for several years. ... **Microsoft** currently outsources that job to Yahoo, which has a contract with **Microsoft** through June 2006.
http://www.msnbc.msn.com/id/9643308/ - MSNBC News
Top Mentions: Microsoft Corp. (39)
Date: 09-Oct-2005

2. Judge bars former Microsoft exec from Go... - 100%
Microsoft Corp. sued Lee, one of its former executives, and Google last week, claiming that by taking the Google job, Lee was violating an agreement he signed in 2000 barring him from working for a direct competitor in an area that overlapped with his role at **Microsoft**. ... "Indeed, **Microsoft** executives admitted to Lee that their real intent was to scare other **Microsoft** employees into remaining at the company." ... "In short, Dr. Lee was recently handed **Microsoft**'s entire Google competition 'playbook,'" **Microsoft** said.
http://www.msnbc.msn.com/id/8732518/ - MSNBC News
Top Mentions: Microsoft Corp. (30), Kai-Fu Lee (22)
Date: 27-Jul-2005

3. MEMORANDUM OPINION - - MSNBC.com - 100%
"**Microsoft** Middleware Product" As recounted above, the latter portion of the definition of "**Microsoft** Platform Software" rests upon the definition of "**Microsoft** Middleware Product." In the SRPFJ, "**Microsoft** 19Trademarked is defined in the SRPFJ to have the following meaning: distributed in commerce and identified as distributed by a name other than **Microsoft**® or Windows® that **Microsoft** has claimed as a trademark or service mark by (i) ... is Trademarked.19 Functionality that **Microsoft** describes or markets as being part of a **Microsoft** Middleware Product (such as a service pack, upgrade, or bug fix for Internet Explorer), or that is a version of a **Microsoft** Middleware Product (such as Internet Explorer 5.5), shall be considered to be part of that **Microsoft** Middleware Product. SRPFJ VI.K. ... In concert with this broad grant of authority, the proposed final judgment specifies that the committee may, on reasonable notice to **Microsoft**, interview **Microsoft** personnel; inspect and copy **Microsoft** documents; access **Microsoft** systems, equipment, premises, and facilities; and require **Microsoft** personnel to compile data, reports, or other information. Id. IV.B.8.b.
http://www.msnbc.msn.com/id/3072746/ - MSNBC News
Top Mentions: Microsoft Corp. (630)
Date: 01-Nov-2002

4. Microsoft - Wikipedia, the free encyclop... - 100%
1991 also marked the founding of **Microsoft** Research, an organization in **Microsoft** for researching computer science subjects, and **Microsoft** Visual Basic, a popular development product for companies. ... With the release of Office 2003, a number of other products were brought under the Office banner, including **Microsoft** Visio, **Microsoft** Project, **Microsoft** MapPoint, **Microsoft** InfoPath, **Microsoft** Publisher and **Microsoft** OneNote. ... Technical reference for developers and articles for various **Microsoft** magazines such as **Microsoft** Systems Journal (or MSJ) is available through **Microsoft**'s MSDN site, short for **Microsoft** Developer Network.
http://en.wikipedia.org/wiki/Microsoft - MSN
Top Mentions: Microsoft Corporation Company (249)

5. Microsoft accused of destroying e-mails ... - 100%
• Safend to Join **Microsoft** Network Access Protection Program • City releases Mayor's office e-mails • **Microsoft** signs 'Halo' film deal ... Burst is suing **Microsoft** for alleged anticompetitive behavior, saying **Microsoft** misappropriated the intellectual property behind its multimedia software after breaking off talks with Burst on a joint project. ... (MSNBC is a **Microsoft** - NBC joint venture.) Burst, according to the motion, wants the jury in the case to be told that **Microsoft** failed to retain important documents, so jurors should infer that the company did so because those documents were damaging.
http://www.msnbc.msn.com/id/6515347/ - MSNBC News

© 1997-2005 Inxight Software, Inc. All Rights Reserved.

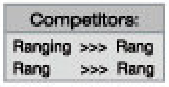
These capabilities are powered by Inxight's deep understanding of language.

Language Understanding

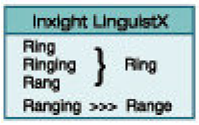
Inxight's language understanding offers intelligent analysis of more than 30 languages. This analysis includes:

- Identification of paragraphs, sentences, clauses, and phrases within text.
- Accurate identification and management of capitalization and case normalization.
- **Word Segmentation (Tokenization)** - Identifies meaningful units of text at a granular level, including:
    - Individual words and word particles
    - Abbreviations and contractions (i.e. "don't")
    - Punctuation (periods, commas, exclamation marks, etc.)

- **Stemming** - Identifies true stems (base forms) for each surface form token; normalizes words to the most basic form for more efficient indexing and better recall in search.



Competitors:		
Ranging	>>>	Rang
Rang	>>>	Rang

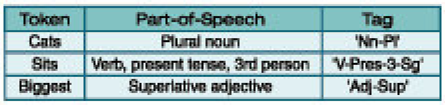


Inxight LinguistX

Ring
Ringing } Ring
Rang
Ranging >>> Range

- **De-Compounding** - Splits compound words into distinct elements -- particularly important in languages such as German and Dutch where words are freely joined together.

- **Part-of-Speech Tagging** - Identifies and labels the part-of-speech of each word in context, including grammatical category (noun, verb, etc.) and sub-class attributes (singular vs. plural nouns, present vs. past tense verbs, etc.).
-



Token	Part-of-Speech	Tag
Cats	Plural noun	'Nn-Pl'
Sits	Verb, present tense, 3rd person	'V-Pres-3-Sg'
Biggest	Superlative adjective	'Adj-Sup'

Entity and Concept Extraction

Inxight then builds on these underlying language components to out-of-the-box extract 35 different key entities (people, places, organizations, weapons, vehicles, dates, etc.) and concepts from electronic text. The Inxight system can also be easily extended to identify and extract custom entities, such as project names, lists of “persons of interest,” date/timestamps, chemical compound names or formulae, serial or part numbers, and patent numbers.

Relation and Event Extraction

Again, Inxight builds on the underlying language components to allow the system to detect relations and events, such as personal associations, merger and acquisition activities, or travel events. It can be extended to find custom relations and events such as company contact information, brand co-occurrence or medication adverse effects.

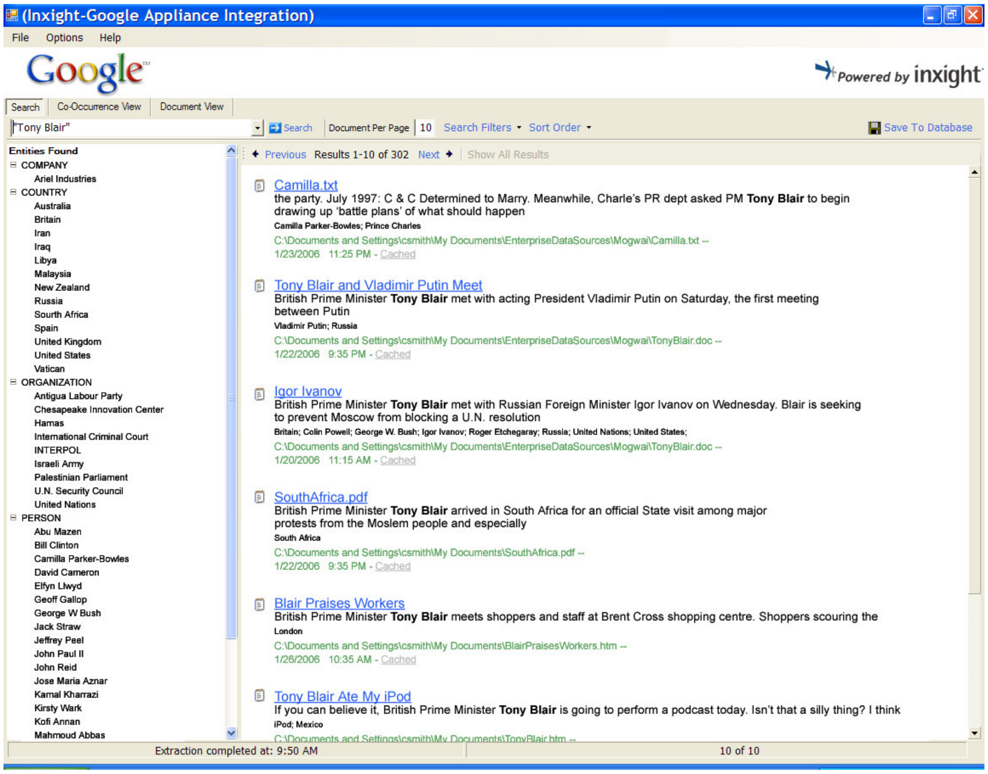
Variant Identification and Normalization

Variant identification and grouping allow Inxight to accurately classify all relevant entities in a document, even one-word entities, and to provide true counts reflecting the number and location of ALL appearances of a given entity. For example, Inxight recognizes that the appearance of the word "Smith" in the example refers to the earlier identified person "Joe Smith." Normalization takes much of the guesswork out of metadata creation, search, data mining and link analysis processes by creating standard formats (e.g., ISO) for certain entity categories such as dates or measurements.

Enriched Applications

In addition to powering more effective information retrieval through Inxight SmartDiscovery Awareness Server, this deep understanding of language also allows documents indexed by Google to be enriched with information, powering applications in routing, data mining, and relationship and trend visualization.

For example, in the custom application below, the user first queries their Google Search Appliance for information on "Tony Blair." Relevant entities are extracted from the result set and this information is saved to a MySQL metadata repository.



"Tony Blair"

Document Per Page: 10

Results 1-10 of 302

Entities Found
- COMPANY
  - Ariel Industries
- COUNTRY
  - Australia
  - Britain
  - Iran
  - Iraq
  - Libya
  - Malaysia
  - New Zealand
  - Russia
  - Sourth Africa
  - Spain
  - United Kingdom
  - United States
  - Vatican
- ORGANIZATION
  - Antigua Labour Party
  - Chesapeake Innovation Center
  - Hamas
  - International Criminal Court
  - INTERPOL
  - Israeli Army
  - Palestinian Parliament
  - U.N. Security Council
  - United Nations
- PERSON
  - Abu Mazen
  - Bill Clinton
  - Camilla Parker-Bowles
  - David Cameron
  - Elfyn Llwyd
  - Geoff Gallop
  - George W Bush
  - Jack Straw
  - Jeffrey Peel
  - John Paul II
  - John Reid
  - Jose Maria Aznar
  - Kamal Kharrazi
  - Kirsty Wark
  - Kofi Annan
  - Mahmoud Abbas

Camilla.txt
the party. July 1997: C & C Determined to Marry. Meanwhile, Charle's PR dept asked PM **Tony Blair** to begin drawing up 'battle plans' of what should happen
Camilla Parker-Bowles; Prince Charles
C:\Documents and Settings\csmith\My Documents\EnterpriseDataSources\Mogwai\Camilla.txt --
1/23/2006 11:25 PM - Cached

Tony Blair and Vladimir Putin Meet
British Prime Minister **Tony Blair** met with acting President Vladimir Putin on Saturday, the first meeting between Putin
Vladimir Putin; Russia
C:\Documents and Settings\csmith\My Documents\EnterpriseDataSources\Mogwai\TonyBlair.doc --
1/22/2006 9:35 PM - Cached

Igor Ivanov
British Prime Minister **Tony Blair** met with Russian Foreign Minister Igor Ivanov on Wednesday. Blair is seeking to prevent Moscow from blocking a U.N. resolution
Britain; Colin Powell; George W. Bush; Igor Ivanov; Roger Etchegaray; Russia; United Nations; United States;
C:\Documents and Settings\csmith\My Documents\EnterpriseDataSources\Mogwai\TonyBlair.doc --
1/20/2006 11:15 AM - Cached

SouthAfrica.pdf
British Prime Minister **Tony Blair** arrived in South Africa for an official State visit among major protests from the Moslem people and especially
South Africa
C:\Documents and Settings\csmith\My Documents\SouthAfrica.pdf --
1/22/2006 9:35 PM - Cached

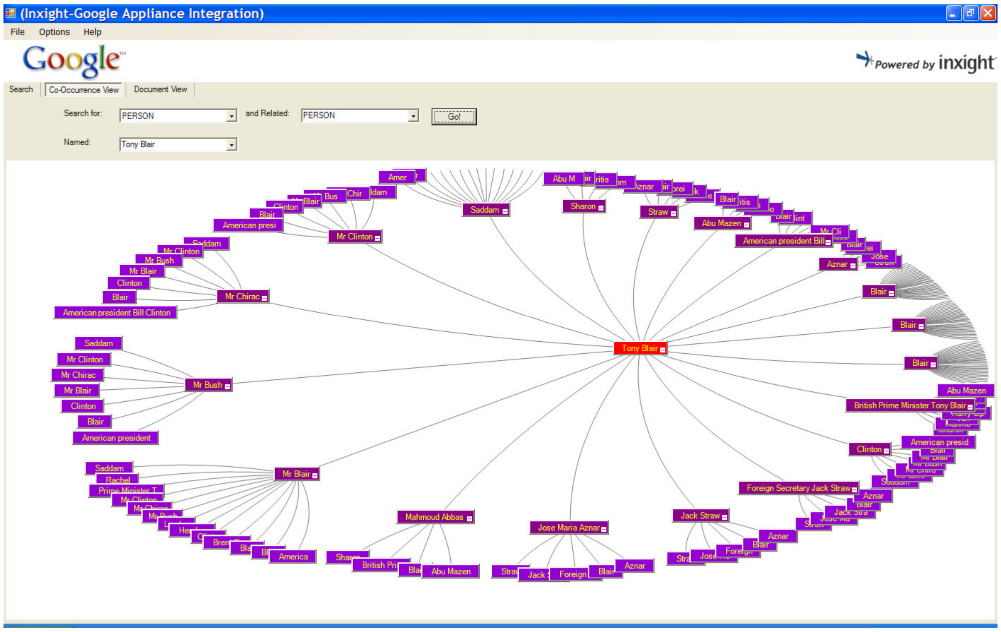
Blair Praises Workers
British Prime Minister **Tony Blair** meets shoppers and staff at Brent Cross shopping centre. Shoppers scouring the
London
C:\Documents and Settings\csmith\My Documents\BlairPraisesWorkers.htm --
1/26/2006 10:35 AM - Cached

Tony Blair Ate My iPod
If you can believe it, British Prime Minister **Tony Blair** is going to perform a podcast today. Isn't that a silly thing? I think
iPod; Mexico
C:\Documents and Settings\csmith\My Documents\TonyBlair.htm --

Extraction completed at: 9:50 AM

10 of 10

The user can then see which entities co-occurred within a sentence in the document result set. They can also explore different events, such as seeing what people sent emails to each other, or what people had meetings with each other. This capability is powered by an Inxight StarTree® visualization:



Capitalize on Your Google Investment with Inxight SmartDiscovery

Google is one of the world's best search indexers. It puts the Web's information at your fingertips, while the Google Search Appliance does the same thing for your website or corporate intranet.

Inxight, a Google Enterprise Partner, takes this information and helps you harness it in more meaningful ways. Inxight SmartDiscovery Awareness Server augments it with sources that cannot (because of accessibility or copyright regulations) be indexed by the public Web or by internal systems.

Inxight SmartDiscovery Analysis Server electronically "reads" that information and structures it – pulling out key entities (people, places, companies…), concepts, relations, and events trapped in unstructured text. Inxight visualizations such as Inxight StarTree then provide the ability to visualize that information in meaningful ways.

For more information or a demonstration, contact Inxight Software at www.inxight.com or email sales@inxight.com.

About Inxight

Inxight Software, Inc. is the leading provider of enterprise software solutions for information discovery. Using Inxight solutions, organizations can access and analyze unstructured, semistructured and structured text to extract key information to enable business intelligence. Inxight is the only company that provides a complete, scalable solution enabling information discovery in more than 30 languages. Customers include enterprise companies such as Air Products, Novartis, Procter & Gamble and Thomson, multiple U.S. and foreign government agencies, including the Department of Defense, Defense Intelligence Agency, Department of Homeland Security and Commonwealth Secretariat, and software OEMs such as SAP, SAS, Oracle and IBM. The company has offices throughout the United States and Europe. For more information, visit www.inxight.com or call 1-408-738-6200 or +44 (0) 1252 761314.

Welcome to TEMIS Webinar "Luxid® for Life Sciences"

Agenda

1 **Welcome & Introduction**

Guillaume Mazieres – Vice President Sales & Marketing – 5 ′

2 **Luxid® for Life Sciences – Presentation**

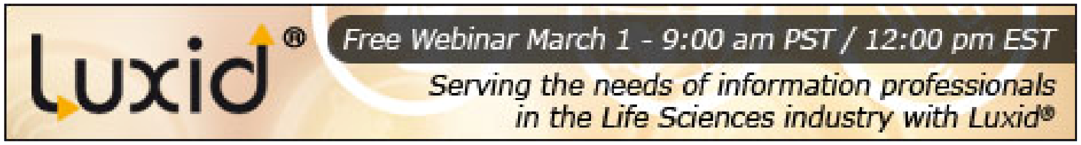
Chris Beguel – Business Development Manager – 10 ′

3 **Luxid® for Life Sciences – Demonstration**

Mathieu Plantefol – Product Manager Life Sciences – 25 ′

4 **Q&A session**

All – 5 ′



luxid®

Free Webinar March 1 - 9:00 am PST / 12:00 pm EST

Serving the needs of information professionals in the Life Sciences industry with Luxid®

Agenda

1 **Welcome & Introduction**

Guillaume Mazieres – Vice President Sales & Marketing – 5 ′

2 **Luxid® for Life Sciences – Presentation**

Chris Beguel – Business Development Manager – 10 ′

3 **Luxid® for Life Sciences – Demonstration**

Mathieu Plantefol – Product Manager Life Sciences – 25 ′

4 **Q&A session**

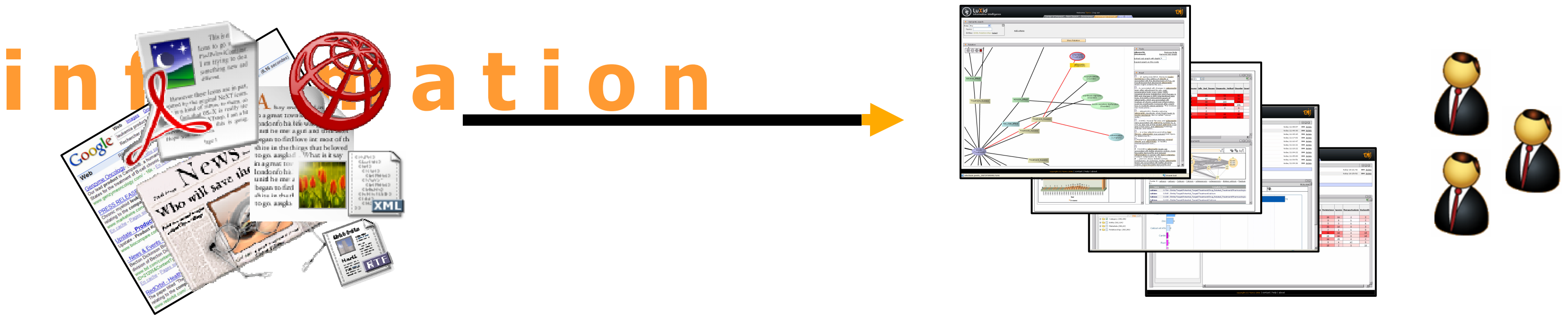
All – 5 ′

luxid®

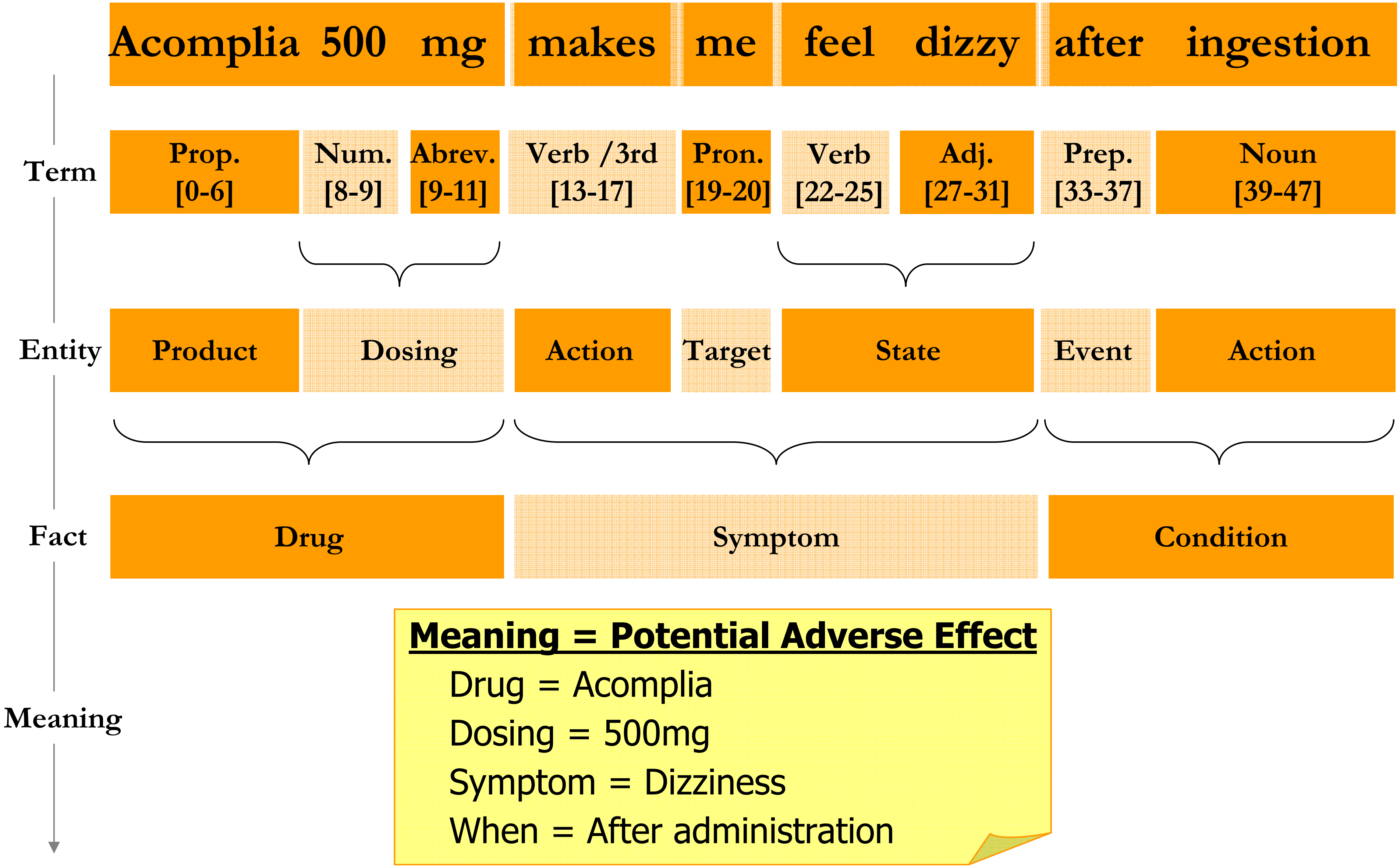
Free Webinar March 1 - 9:00 am PST / 12:00 pm EST

*Serving the needs of information professionals
in the Life Sciences industry with Luxid®*

Luxid® - Introduction



Innovation? *From Text to Meaning!*



Acomplia 500 mg makes me feel dizzy after ingestion

Term: Prop. [0-6] | Num. [8-9] | Abrev. [9-11] | Verb /3rd [13-17] | Pron. [19-20] | Verb [22-25] | Adj. [27-31] | Prep. [33-37] | Noun [39-47]

Entity: Product | Dosing | Action | Target | State | Event | Action

Fact: Drug | Symptom | Condition

Meaning

Meaning = Potential Adverse Effect

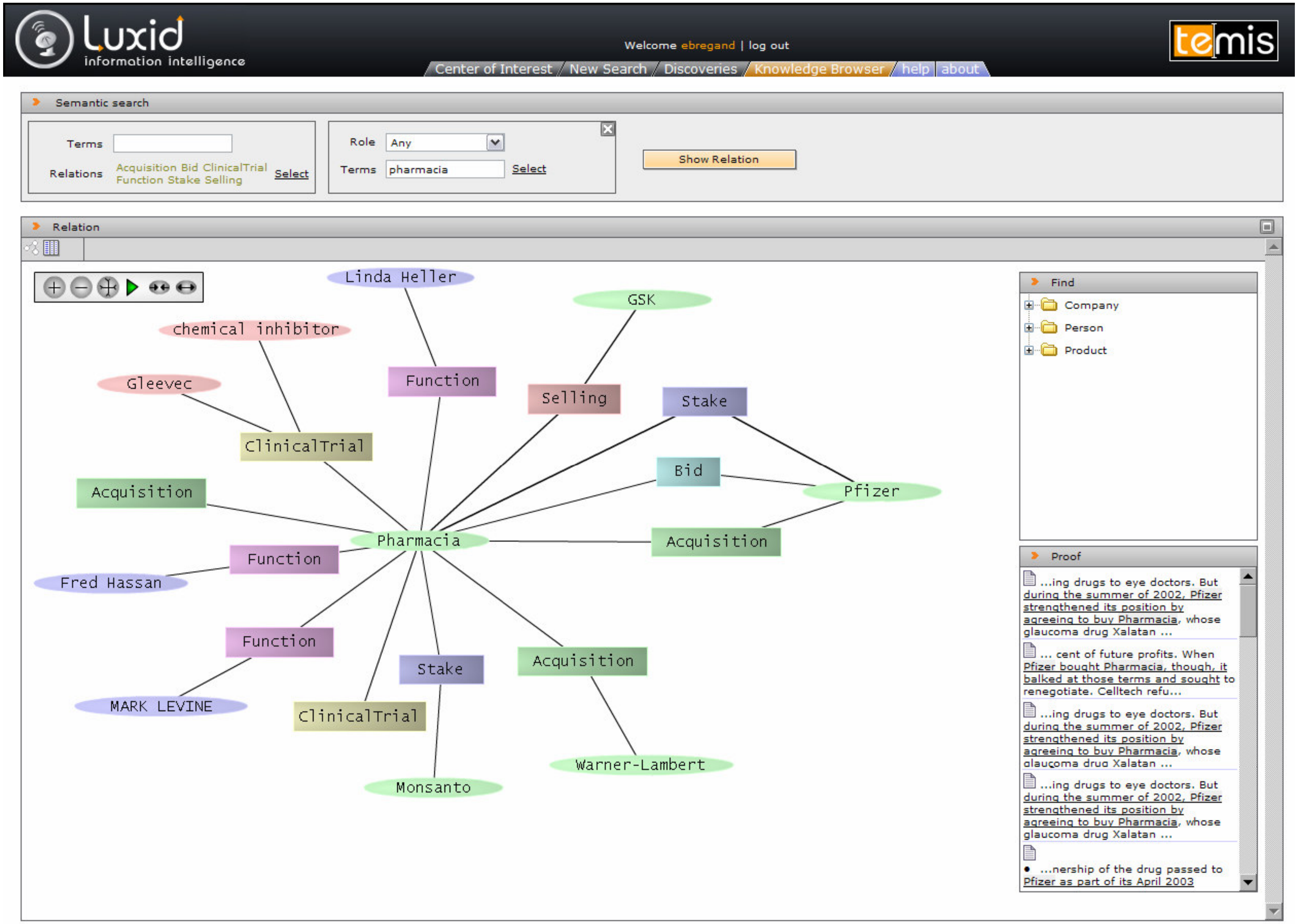
Drug = Acomplia

Dosing = 500mg

Symptom = Dizziness

When = After administration

And… *From Meaning to Knowledge!*



Welcome ebregand | log out

Semantic search

Terms

Relations Acquisition Bid ClinicalTrial Function Stake Selling Select

Role Any

Terms pharmacia Select

Show Relation

…ing drugs to eye doctors. But during the summer of 2002, Pfizer strengthened its position by agreeing to buy Pharmacia, whose glaucoma drug Xalatan …

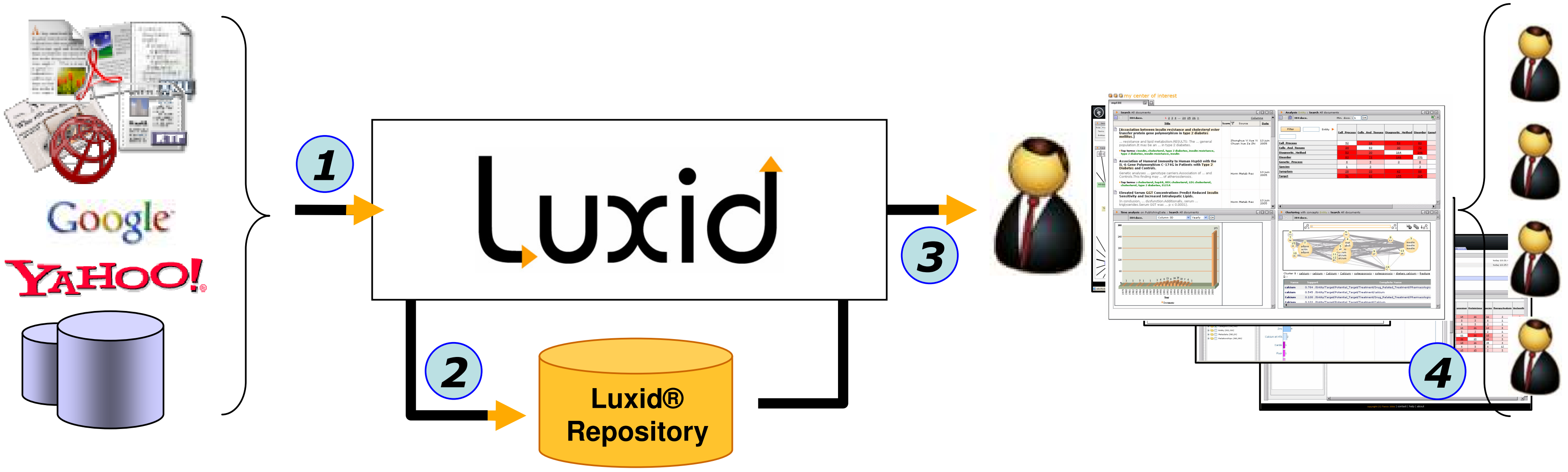
… cent of future profits. When Pfizer bought Pharmacia, though, it balked at those terms and sought to renegotiate. Celltech refu…

…ing drugs to eye doctors. But during the summer of 2002, Pfizer strengthened its position by agreeing to buy Pharmacia, whose glaucoma drug Xalatan …

…ing drugs to eye doctors. But during the summer of 2002, Pfizer strengthened its position by agreeing to buy Pharmacia, whose glaucoma drug Xalatan …

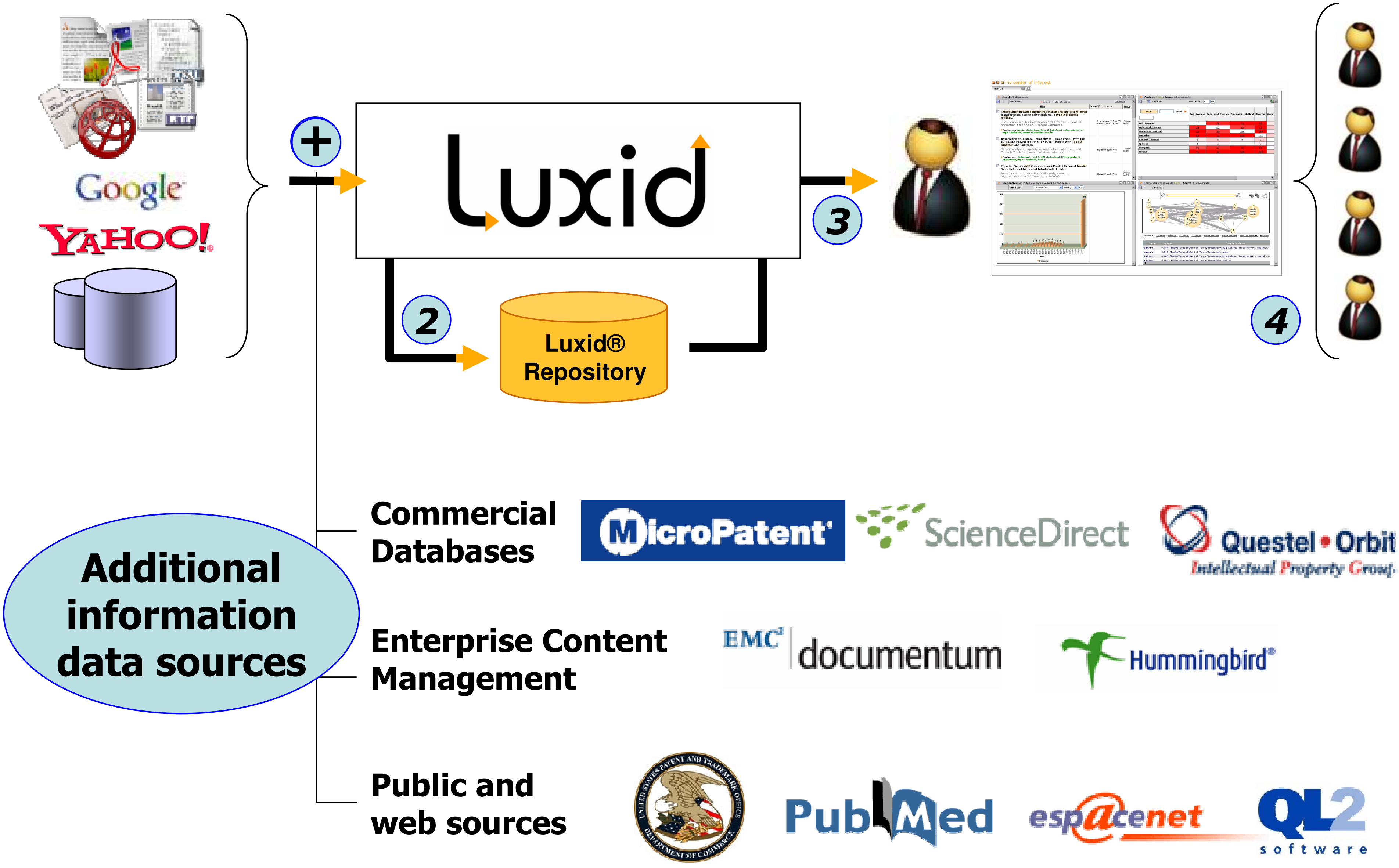
- …nership of the drug passed to Pfizer as part of its April 2003

Luxid® - Complete Corporate Solution



1. Gather textual data from any sources
2. Perform domain-specific & detailed content analysis
3. Discover and understand information
4. Share information and alerts on event

Luxid® *For Life Sciences*



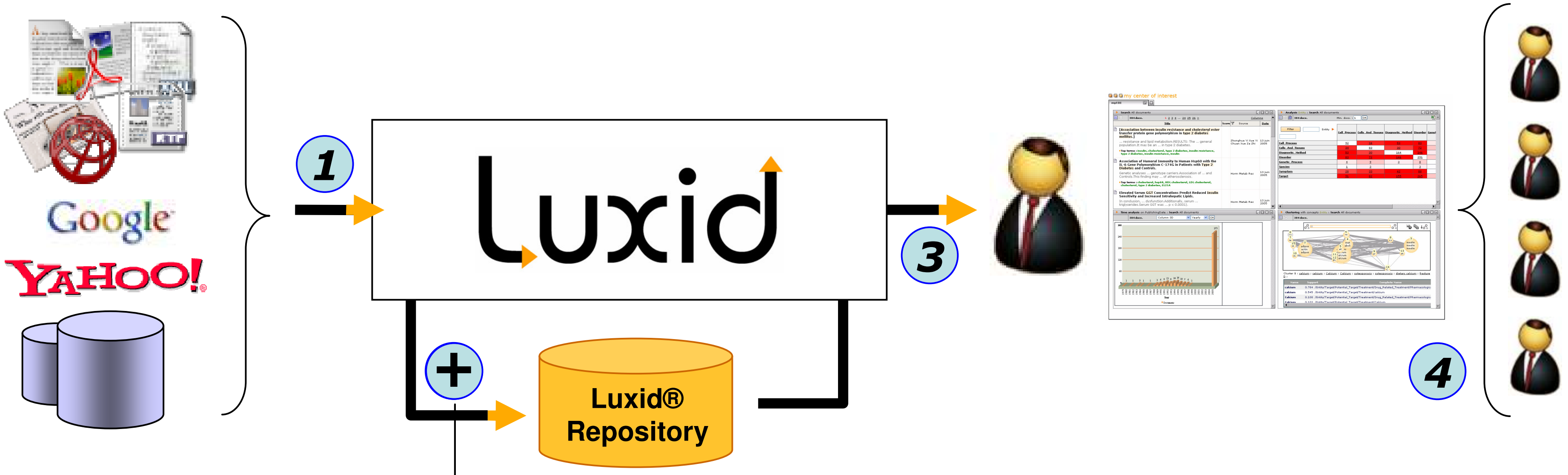
Additional information data sources

Commercial Databases

Enterprise Content Management

Public and web sources

Luxid® *For Life Sciences*



Additional domain-specific annotators

Biological Entities Relationships

Extraction of genes + proteins & interactions

Fraunhofer Institute Algorithms and Scientific Computing SCAI

Chemical Entities Relationships

Extraction of chemical names and structures

ELSEVIER MDL

Medical Entities Relationships

Extraction of diseases, disorders and adverse events

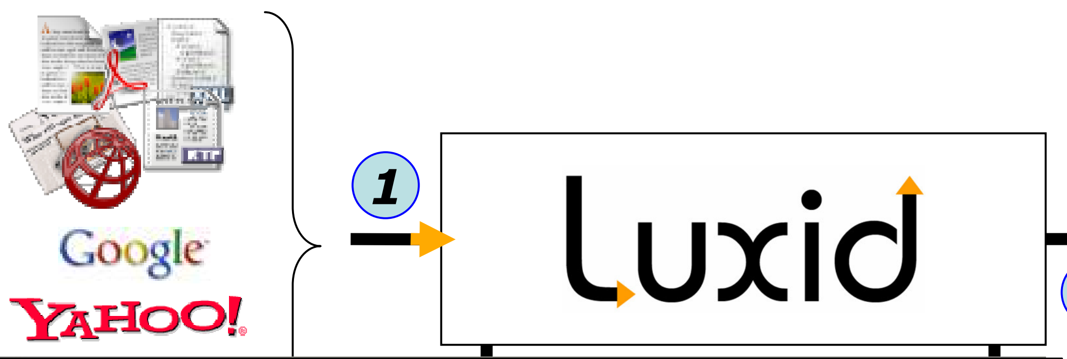
MeSH

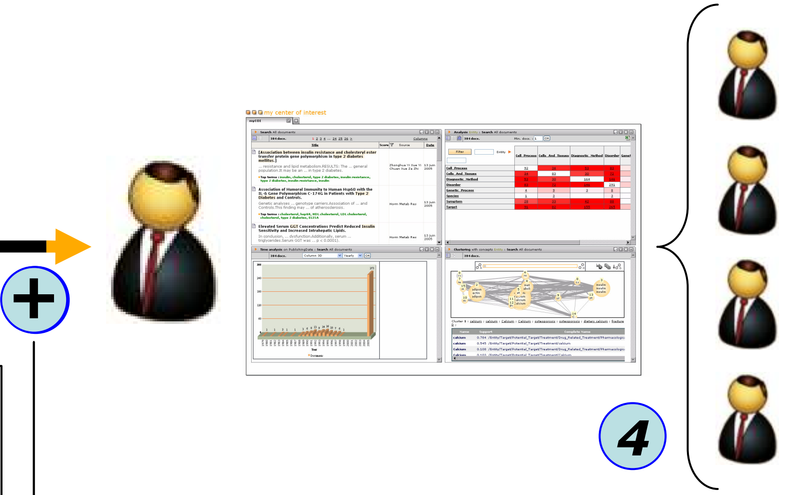
Identification of Knowledge Shift

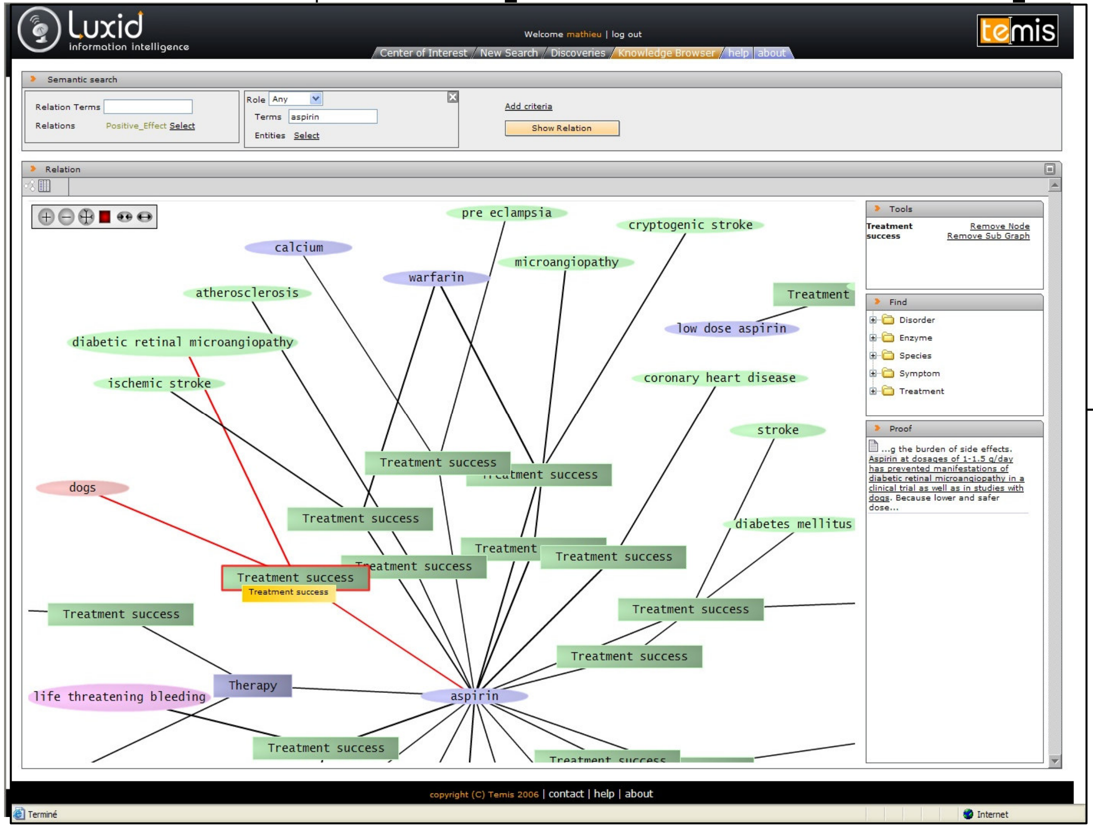
Extraction of paradigm change

GENEBIO Geneva Bioinformatics SA

Luxid® *For Life Sciences*







Welcome mathieu | log out

Semantic search

Relation Terms

Relations Positive_Effect Select

Role Any

Terms aspirin

Entities Select

Add criteria

Relation

Tools

Treatment success Remove Node Remove Sub Graph

Find

- Disorder
- Enzyme
- Species
- Symptom
- Treatment

Proof

...g the burden of side effects. Aspirin at dosages of 1-1.5 g/day has prevented manifestations of diabetic retinal microangiopathy in a clinical trial as well as in studies with dogs. Because lower and safer dose...

copyright (C) Temis 2006 | contact | help | about

Terminé

Internet

Additional domain-specific features

Sub-Structure Search

Biological Pathways

Key Business Benefits

1. Improve Research Effectiveness

- Gain productivity by automating scientific literature review and return interpretation time to scientists.
- Capture innovation at any stage by unveiling cross-domain knowledge.
- Minimize TCO through one single entry-point to information.
- Stimulate collaborative work by streamlining information sharing and optimizing content generation process.

Scientists
Scientific Information Managers
Business Unit/Therapeutic Areas Managers
Vice-President Discovery, Vice-President R&D

Key Business Benefits

1. Improve Research Effectiveness
2. Manage Intellectual Property Assets
   - Minimize infringement risks with a reliable FTO assessment
   - Cut down on legal costs by providing attorneys with timely and undisputable evidence for defending product rights
   - Optimize product portfolio by identifying market opportunities and qualifying business leads to sustain corporate development strategies
   - Gain a competitive edge through accurate and real-time monitoring of competitors & lead inventors patent activity

IP analysts
IP attorneys, Legal departments
Corporate Development groups, Licensing groups
Information Management Groups

Key Business Benefits

1. Improve Research Effectiveness
2. Manage Intellectual Property Assets
3. Reduce Adverse Event Related Risks
   - Reduce company exposure to risks with early and reliable detection of potential adverse events
   - Improve R&D effectiveness by automatically enriching research knowledge bases with all detected adverse effects
   - Ensure global compliance by adjusting detection rules and escalation procedures to regulation changes and country specific policies.

Drug Safety Specialists
Business Unit Managers
VP Marketing, VP Research and Development

Agenda

1 **Welcome & Introduction**

Guillaume Mazieres – Vice President Sales & Marketing – 5 ’

2 **Luxid® for Life Sciences – Presentation**

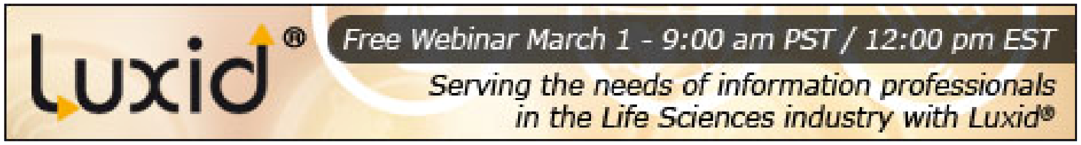
Chris Beguel – Business Development Manager – 10 ’

3 **Luxid® for Life Sciences – Demonstration**

Mathieu Plantefol – Product Manager Life Sciences – 25 ’

4 **Q&A session**

All – 5 ’



Free Webinar March 1 - 9:00 am PST / 12:00 pm EST

Serving the needs of information professionals
in the Life Sciences industry with Luxid®

Scenario 1 Target Validation

How to improve R&D research effectiveness using Luxid®?

- Use Case : target validation using Luxid®
- Scenario : find information about potential targets or biomarkers for atherosclerosis in 10 minutes inside a corpus of 22 000 Medline abstracts
- Benefits : combine faceted search, multiple analysis and semantic navigation

Demo

Scenario 2 : FTO analysis

How to manage IP assets using Luxid®?

- Use Case : FTO evaluation around chemical structures
- Scenario : make a first analysis in 10 minutes of hundreds of recent patents to find claims around known chemical structures and atherosclerosis
- Benefits : search directly documents using chemical structure, perform analysis on rich metadata content, drill up/down within MeSH ontology

Demo

Agenda

1 **Welcome & Introduction**

Guillaume Mazieres – Vice President Sales & Marketing – 5 ′

2 **Luxid® for Life Sciences – Presentation**

Chris Beguel – Business Development Manager – 10 ′

3 **Luxid® for Life Sciences – Demonstration**

Mathieu Plantefol – Product Manager Life Sciences – 25 ′

4 **Q&A session**

All – 5 ′

luxid®

Free Webinar March 1 - 9:00 am PST / 12:00 pm EST

Serving the needs of information professionals in the Life Sciences industry with Luxid®

Thank You For Your Time!

TEMIS unveils
Luxid® for Life Sciences

www.temis.com

Analyzing Patent Literature to Gain Competitive Insight with Luxid®

Sorrento, Italy, March 8 2007

Stefan Geißler - Managing Director, TEMIS Deutschland GmbH

Guillaume Mazieres - Vice President Sales & Marketing, TEMIS SA

Agenda

1 **Welcome & Introduction**

Stefan Geißler - Guillaume Mazieres

2 **Luxid® for Life Sciences – Presentation**

Guillaume Mazieres - 5 ′

3 **Patent Analysis Scenarios with Luxid® for Life Sciences**

A. Patent landscape around a class of chemical substance
B. Explore potential targets around a therapeutic area

Stefan Geißler – 40 ′

4 **Q&A session**

All – 5 ′

Agenda

temis

1 **Welcome & Introduction**

Stefan Geißler - Guillaume Mazieres

2 **Luxid® for Life Sciences – Presentation**

Guillaume Mazieres - 5 ′

3 **Patent Analysis Scenarios with Luxid® for Life Sciences**

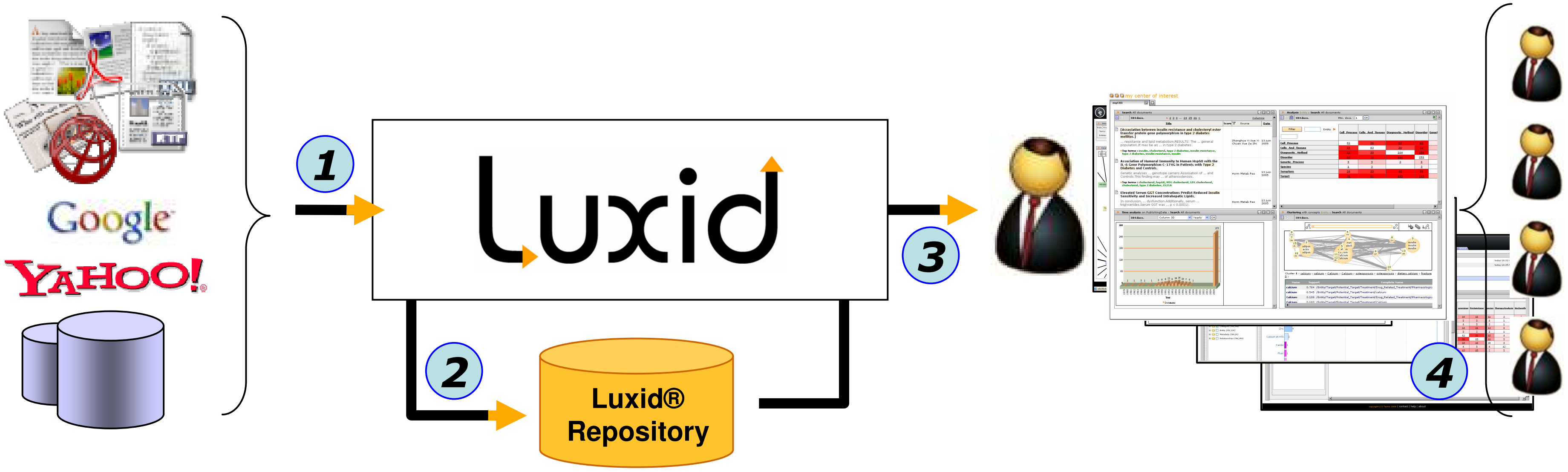
A. Patent landscape around a class of chemical substance
B. Explore potential targets around a therapeutic area

Stefan Geißler – 40 ′

4 **Q&A session**

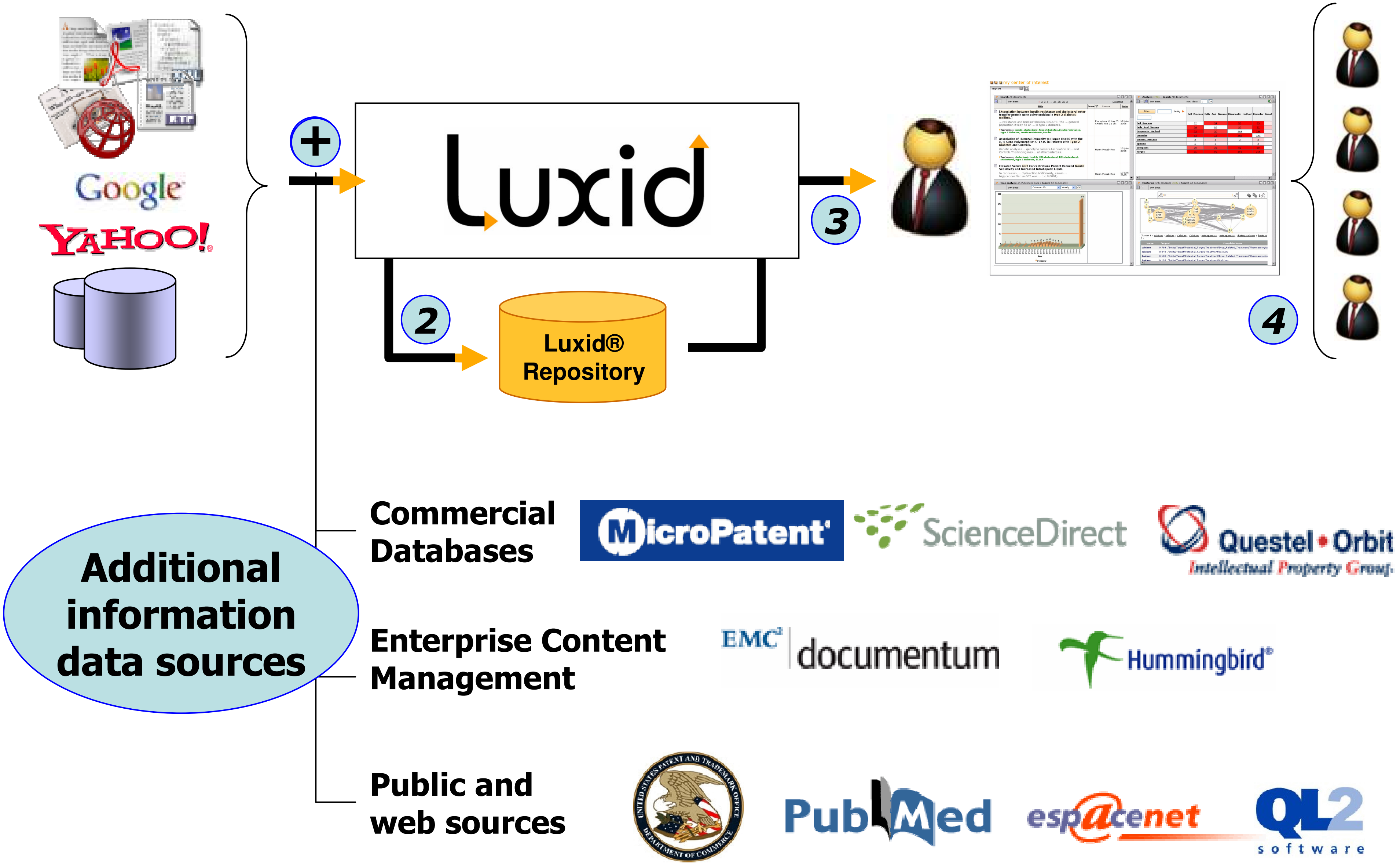
All – 5 ′

Luxid® - Complete Corporate Solution



1. Gather textual data from any sources
2. Perform domain-specific & detailed content analysis
3. Discover and understand information
4. Share information and alerts

Luxid® *For Life Sciences*



Additional information data sources

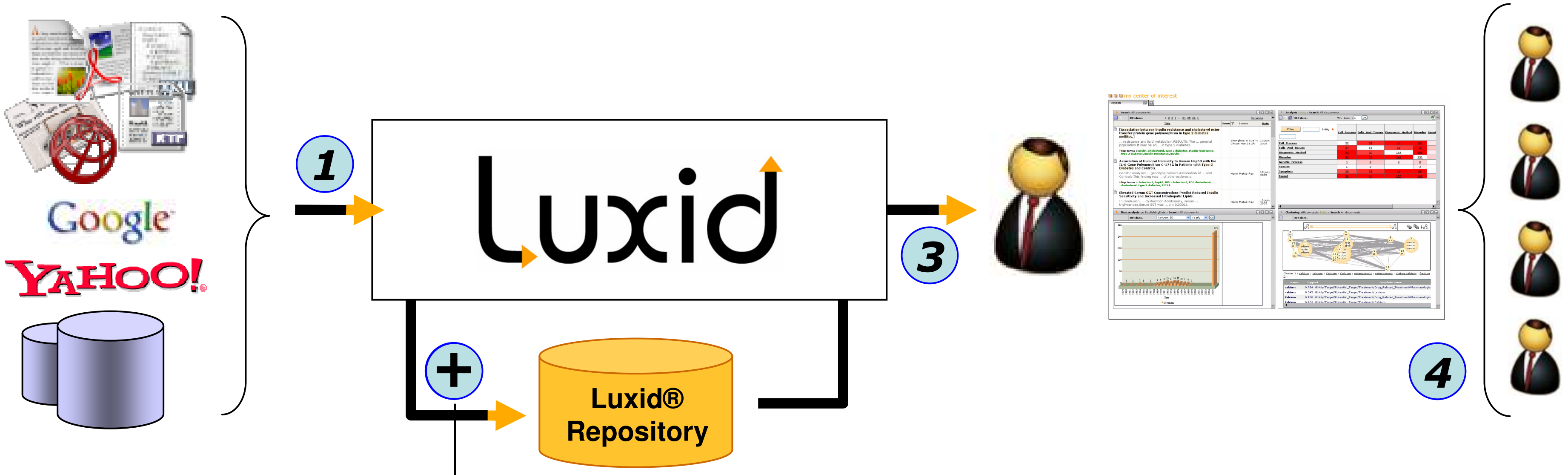
Commercial Databases

Enterprise Content Management

Public and web sources

Luxid® *For Life Sciences*

temis



Additional domain-specific annotators

Biological Entities Relationships
Extraction of genes + proteins & interactions

Fraunhofer Institute Algorithms and Scientific Computing SCAI

Chemical Entities Relationships
Extraction of chemical names and structures

ELSEVIER MDL

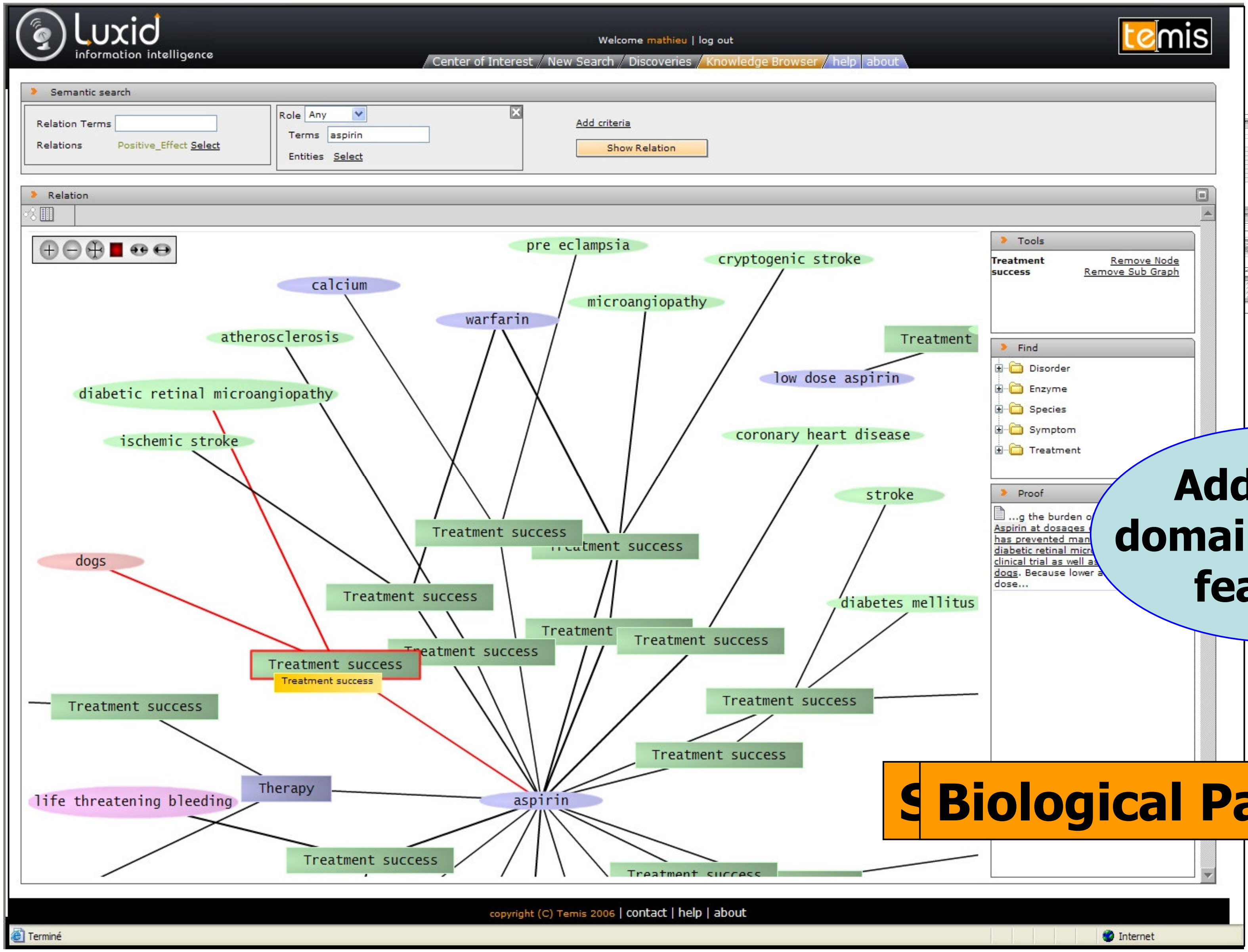
Medical Entities Relationships
Extraction of diseases, disorders and adverse events

MeSH

Identification of Knowledge Shift
Extraction of paradigm change

GENEBIO Geneva Bioinformatics SA

Luxid® *For Life Sciences*



Welcome mathieu | log out

Semantic search

Relation Terms
Relations Positive_Effect Select

Role Any
Terms aspirin
Entities Select

Add criteria

Show Relation

Relation

Tools

Treatment success Remove Node Remove Sub Graph

Find

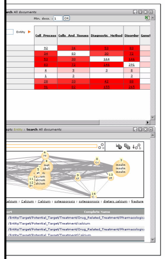
- Disorder
- Enzyme
- Species
- Symptom
- Treatment

Proof

...g the burden o
Aspirin at dosages
has prevented man
diabetic retinal micr
clinical trial as well a
dogs. Because lower a
dose...

copyright (C) Temis 2006 | contact | help | about

Terminé Internet



4

Additional domain-specific features

S Biological Pathways

Agenda

1 **Welcome & Introduction**

Stefan Geißler - Guillaume Mazieres

2 **Luxid® for Life Sciences – Presentation**

Guillaume Mazieres - 5 ′

3 **Patent Analysis Scenarios with Luxid® for Life Sciences**

A. Patent landscape around a class of chemical substance
B. Explore potential targets around a therapeutic area

Stefan Geißler – 40 ′

4 **Q&A session**

All – 5 ′

Scenarios

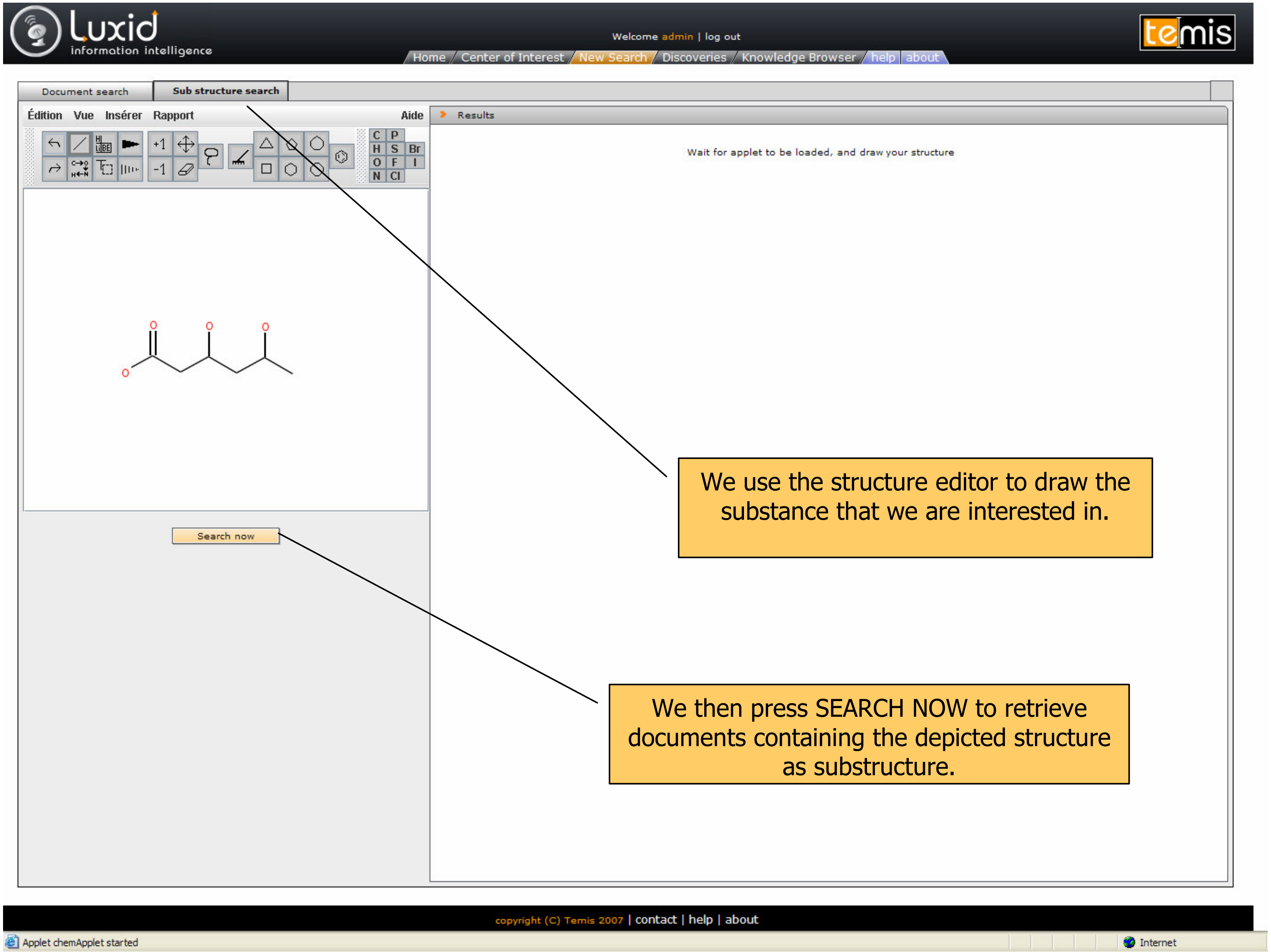
- Use TEMIS Luxid® to explore freedom to operate around a class of substances
- Document base: several hundred sample patents from the EPO
  - Analyzed by the Luxid® Annotation Factory with respect to Chemical, Biological and Medical terminology
  - Loaded into the Luxid® DB and made available through the Luxid web interface
- Questions:
  - Which patents are there that mention a given class of chemical substances?
  - Which applicant are active in which therapeutic area and when?

Scenario #1

Patent landscape around

a class of chemical substances

Draw the statin pharmacophore



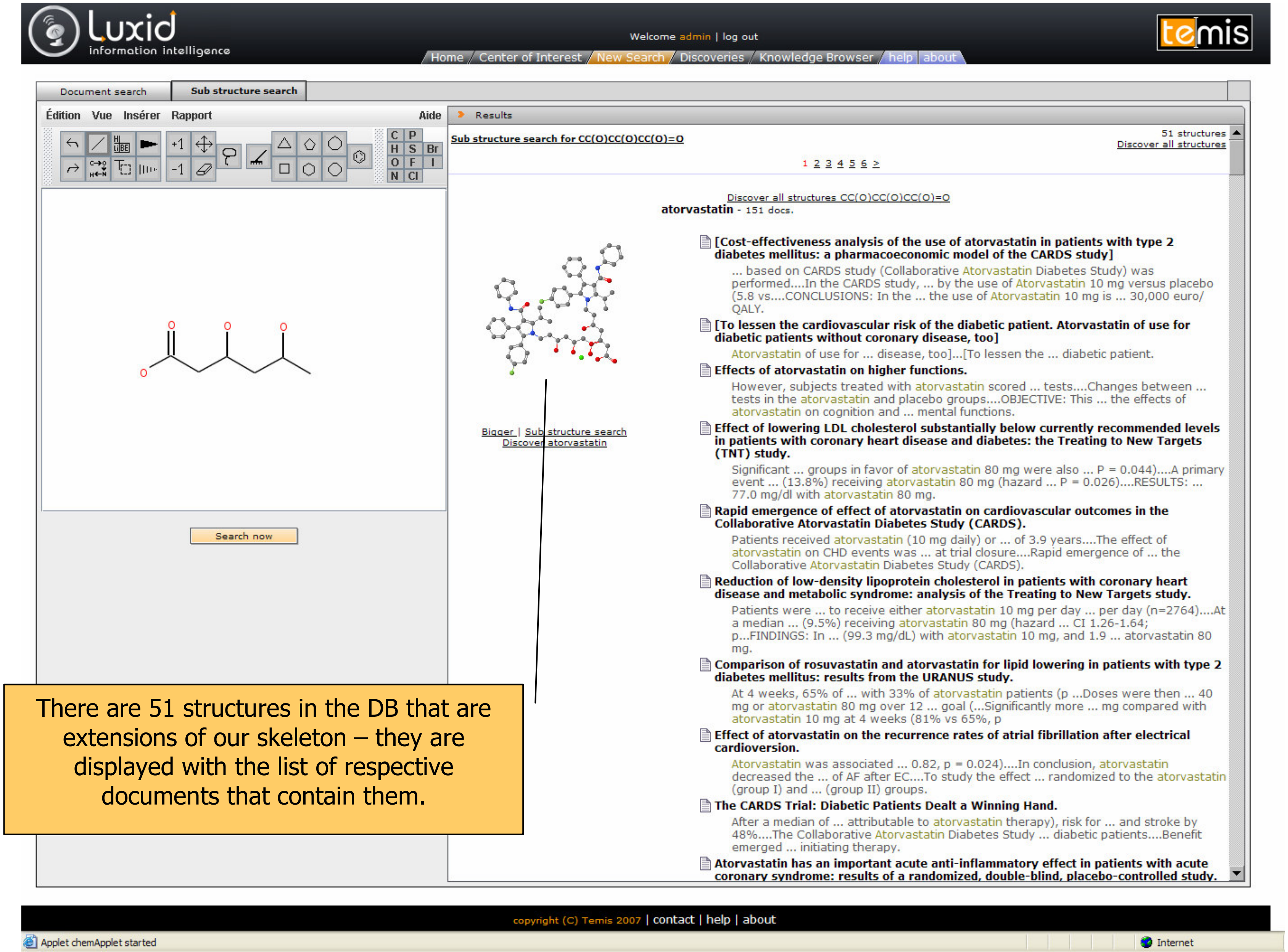
Wait for applet to be loaded, and draw your structure

Applet chemApplet started

We use the structure editor to draw the substance that we are interested in.

We then press SEARCH NOW to retrieve documents containing the depicted structure as substructure.

Retrieve molecules and documents



Sub structure search for CC(O)CC(O)CC(O)=O

51 structures
Discover all structures

1 2 3 4 5 6 >

Discover all structures CC(O)CC(O)CC(O)=O

atorvastatin - 151 docs.

[Cost-effectiveness analysis of the use of atorvastatin in patients with type 2 diabetes mellitus: a pharmacoeconomic model of the CARDS study]

... based on CARDS study (Collaborative Atorvastatin Diabetes Study) was performed....In the CARDS study, ... by the use of Atorvastatin 10 mg versus placebo (5.8 vs....CONCLUSIONS: In the ... the use of Atorvastatin 10 mg is ... 30,000 euro/QALY.

[To lessen the cardiovascular risk of the diabetic patient. Atorvastatin of use for diabetic patients without coronary disease, too]

Atorvastatin of use for ... disease, too]...[To lessen the ... diabetic patient.

Effects of atorvastatin on higher functions.

However, subjects treated with atorvastatin scored ... tests....Changes between ... tests in the atorvastatin and placebo groups....OBJECTIVE: This ... the effects of atorvastatin on cognition and ... mental functions.

Effect of lowering LDL cholesterol substantially below currently recommended levels in patients with coronary heart disease and diabetes: the Treating to New Targets (TNT) study.

Significant ... groups in favor of atorvastatin 80 mg were also ... P = 0.044)....A primary event ... (13.8%) receiving atorvastatin 80 mg (hazard ... P = 0.026)....RESULTS: ... 77.0 mg/dl with atorvastatin 80 mg.

Rapid emergence of effect of atorvastatin on cardiovascular outcomes in the Collaborative Atorvastatin Diabetes Study (CARDS).

Patients received atorvastatin (10 mg daily) or ... of 3.9 years....The effect of atorvastatin on CHD events was ... at trial closure....Rapid emergence of ... the Collaborative Atorvastatin Diabetes Study (CARDS).

Reduction of low-density lipoprotein cholesterol in patients with coronary heart disease and metabolic syndrome: analysis of the Treating to New Targets study.

Patients were ... to receive either atorvastatin 10 mg per day ... per day (n=2764)....At a median ... (9.5%) receiving atorvastatin 80 mg (hazard ... CI 1.26-1.64; p...FINDINGS: In ... (99.3 mg/dL) with atorvastatin 10 mg, and 1.9 ... atorvastatin 80 mg.

Comparison of rosuvastatin and atorvastatin for lipid lowering in patients with type 2 diabetes mellitus: results from the URANUS study.

At 4 weeks, 65% of ... with 33% of atorvastatin patients (p ...Doses were then ... 40 mg or atorvastatin 80 mg over 12 ... goal (...Significantly more ... mg compared with atorvastatin 10 mg at 4 weeks (81% vs 65%, p

Effect of atorvastatin on the recurrence rates of atrial fibrillation after electrical cardioversion.

Atorvastatin was associated ... 0.82, p = 0.024)....In conclusion, atorvastatin decreased the ... of AF after EC....To study the effect ... randomized to the atorvastatin (group I) and ... (group II) groups.

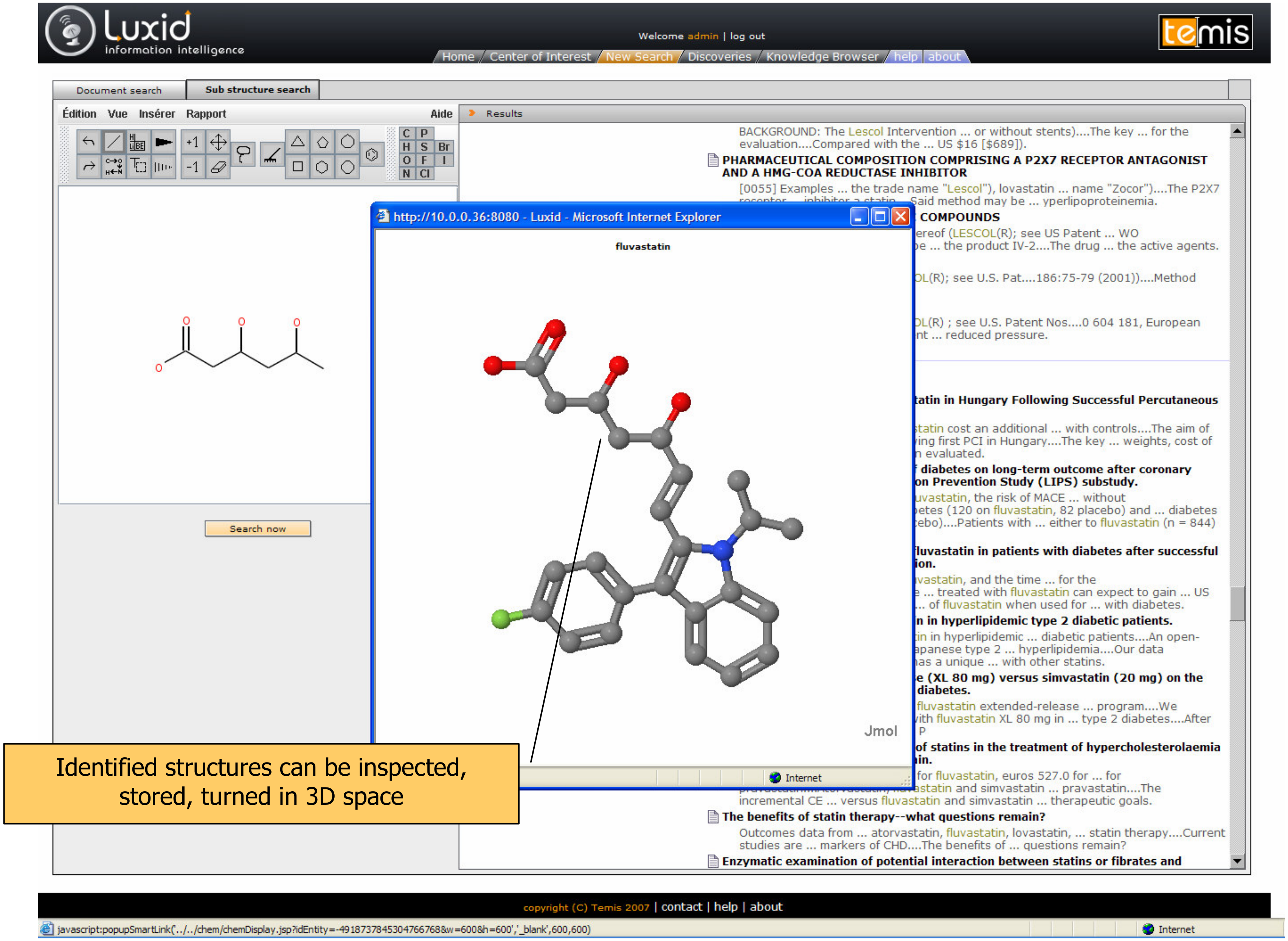
The CARDS Trial: Diabetic Patients Dealt a Winning Hand.

After a median of ... attributable to atorvastatin therapy), risk for ... and stroke by 48%....The Collaborative Atorvastatin Diabetes Study ... diabetic patients....Benefit emerged ... initiating therapy.

Atorvastatin has an important acute anti-inflammatory effect in patients with acute coronary syndrome: results of a randomized, double-blind, placebo-controlled study.

There are 51 structures in the DB that are extensions of our skeleton – they are displayed with the list of respective documents that contain them.

Show one statin



BACKGROUND: The Lescol Intervention ... or without stents)....The key ... for the evaluation....Compared with the ... US $16 [$689]).

PHARMACEUTICAL COMPOSITION COMPRISING A P2X7 RECEPTOR ANTAGONIST AND A HMG-COA REDUCTASE INHIBITOR

[0055] Examples ... the trade name "Lescol"), lovastatin ... name "Zocor")....The P2X7 receptor ... inhibitor a statin ... Said method may be ... yperlipoproteinemia.

COMPOUNDS

ereof (LESCOL(R); see US Patent ... WO
be ... the product IV-2....The drug ... the active agents.

OL(R); see U.S. Pat....186:75-79 (2001))....Method

OL(R) ; see U.S. Patent Nos....0 604 181, European
nt ... reduced pressure.

tatin in Hungary Following Successful Percutaneous

statin cost an additional ... with controls....The aim of
ving first PCI in Hungary....The key ... weights, cost of
n evaluated.

f diabetes on long-term outcome after coronary
on Prevention Study (LIPS) substudy.

uvastatin, the risk of MACE ... without
betes (120 on fluvastatin, 82 placebo) and ... diabetes
cebo)....Patients with ... either to fluvastatin (n = 844)

fluvastatin in patients with diabetes after successful
ion.

vastatin, and the time ... for the
e ... treated with fluvastatin can expect to gain ... US
... of fluvastatin when used for ... with diabetes.

n in hyperlipidemic type 2 diabetic patients.

tin in hyperlipidemic ... diabetic patients....An open-
apanese type 2 ... hyperlipidemia....Our data
has a unique ... with other statins.

e (XL 80 mg) versus simvastatin (20 mg) on the
diabetes.

fluvastatin extended-release ... program....We
vith fluvastatin XL 80 mg in ... type 2 diabetes....After
P

of statins in the treatment of hypercholesterolaemia
in.

for fluvastatin, euros 527.0 for ... for
astatin and simvastatin ... pravastatin....The
incremental CE ... versus fluvastatin and simvastatin ... therapeutic goals.

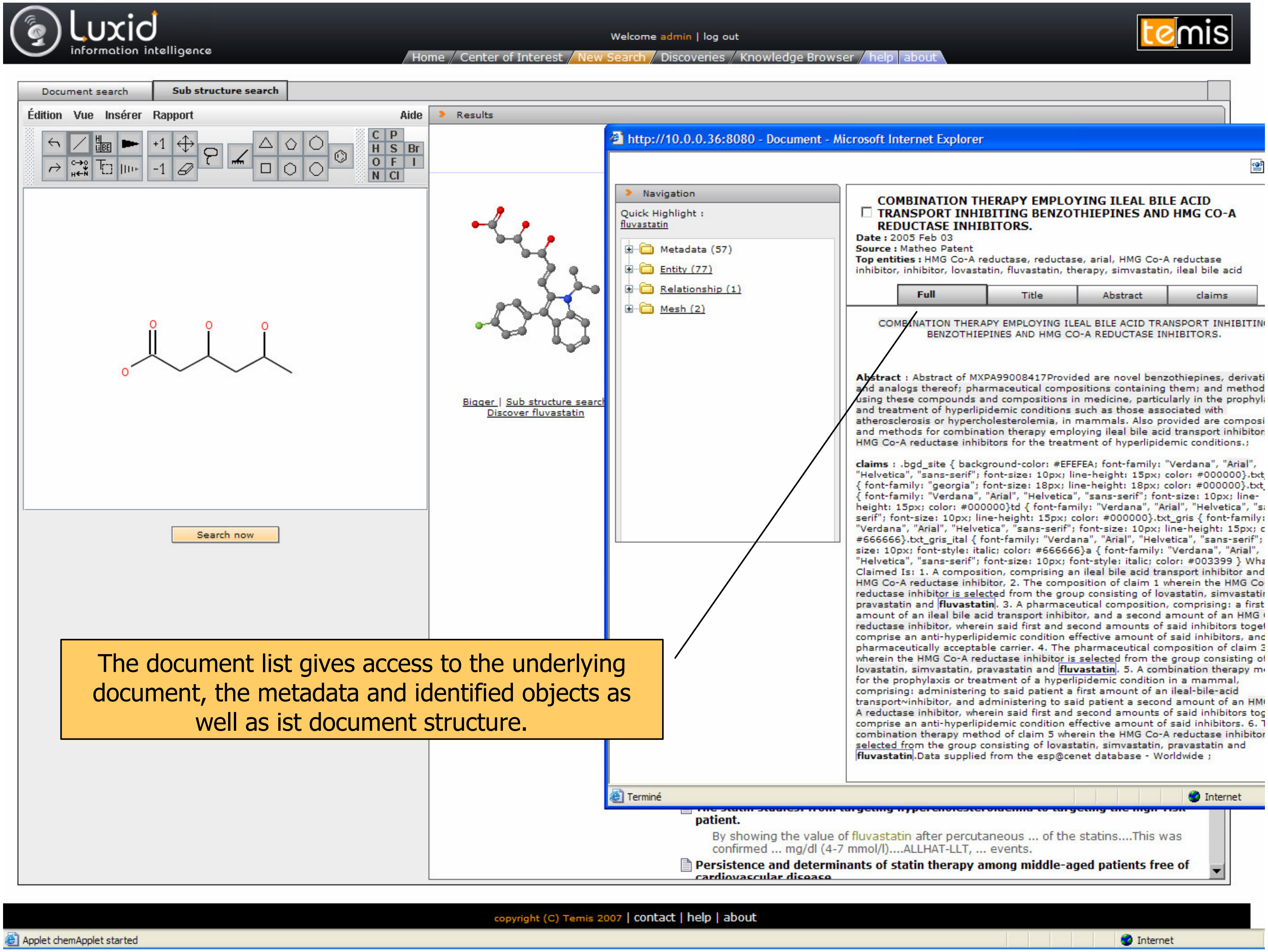
The benefits of statin therapy--what questions remain?

Outcomes data from ... atorvastatin, fluvastatin, lovastatin, ... statin therapy....Current studies are ... markers of CHD....The benefits of ... questions remain?

Enzymatic examination of potential interaction between statins or fibrates and

Identified structures can be inspected, stored, turned in 3D space

Show a document



□ **COMBINATION THERAPY EMPLOYING ILEAL BILE ACID TRANSPORT INHIBITING BENZOTHIEPINES AND HMG CO-A REDUCTASE INHIBITORS.**

Date : 2005 Feb 03
Source : Matheo Patent
Top entities : HMG Co-A reductase, reductase, arial, HMG Co-A reductase inhibitor, inhibitor, lovastatin, fluvastatin, therapy, simvastatin, ileal bile acid

COMBINATION THERAPY EMPLOYING ILEAL BILE ACID TRANSPORT INHIBITIN
BENZOTHIEPINES AND HMG CO-A REDUCTASE INHIBITORS.

Abstract : Abstract of MXPA99008417Provided are novel benzothiepines, derivati
and analogs thereof; pharmaceutical compositions containing them; and method
using these compounds and compositions in medicine, particularly in the prophyl
and treatment of hyperlipidemic conditions such as those associated with
atherosclerosis or hypercholesterolemia, in mammals. Also provided are composi
and methods for combination therapy employing ileal bile acid transport inhibitor
HMG Co-A reductase inhibitors for the treatment of hyperlipidemic conditions.;

claims : .bgd_site { background-color: #EFEFEA; font-family: "Verdana", "Arial",
"Helvetica", "sans-serif"; font-size: 10px; line-height: 15px; color: #000000}.txt
{ font-family: "georgia"; font-size: 18px; line-height: 18px; color: #000000}.txt
{ font-family: "Verdana", "Arial", "Helvetica", "sans-serif"; font-size: 10px; line-
height: 15px; color: #000000}td { font-family: "Verdana", "Arial", "Helvetica", "s
serif"; font-size: 10px; line-height: 15px; color: #000000}.txt_gris { font-family:
"Verdana", "Arial", "Helvetica", "sans-serif"; font-size: 10px; line-height: 15px; c
#666666}.txt_gris_ital { font-family: "Verdana", "Arial", "Helvetica", "sans-serif";
size: 10px; font-style: italic; color: #666666}a { font-family: "Verdana", "Arial",
"Helvetica", "sans-serif"; font-size: 10px; font-style: italic; color: #003399 } Wha
Claimed Is: 1. A composition, comprising an ileal bile acid transport inhibitor and
HMG Co-A reductase inhibitor, 2. The composition of claim 1 wherein the HMG Co
reductase inhibitor is selected from the group consisting of lovastatin, simvastati
pravastatin and **fluvastatin**. 3. A pharmaceutical composition, comprising: a first
amount of an ileal bile acid transport inhibitor, and a second amount of an HMG
reductase inhibitor, wherein said first and second amounts of said inhibitors toge
comprise an anti-hyperlipidemic condition effective amount of said inhibitors, and
pharmaceutically acceptable carrier. 4. The pharmaceutical composition of claim 3
wherein the HMG Co-A reductase inhibitor is selected from the group consisting o
lovastatin, simvastatin, pravastatin and **fluvastatin**. 5. A combination therapy m
for the prophylaxis or treatment of a hyperlipidemic condition in a mammal,
comprising: administering to said patient a first amount of an ileal-bile-acid
transport~inhibitor, and administering to said patient a second amount of an HM
A reductase inhibitor, wherein said first and second amounts of said inhibitors tog
comprise an anti-hyperlipidemic condition effective amount of said inhibitors. 6. T
combination therapy method of claim 5 wherein the HMG Co-A reductase inhibitor
selected from the group consisting of lovastatin, simvastatin, pravastatin and
fluvastatin.Data supplied from the esp@cenet database - Worldwide ;

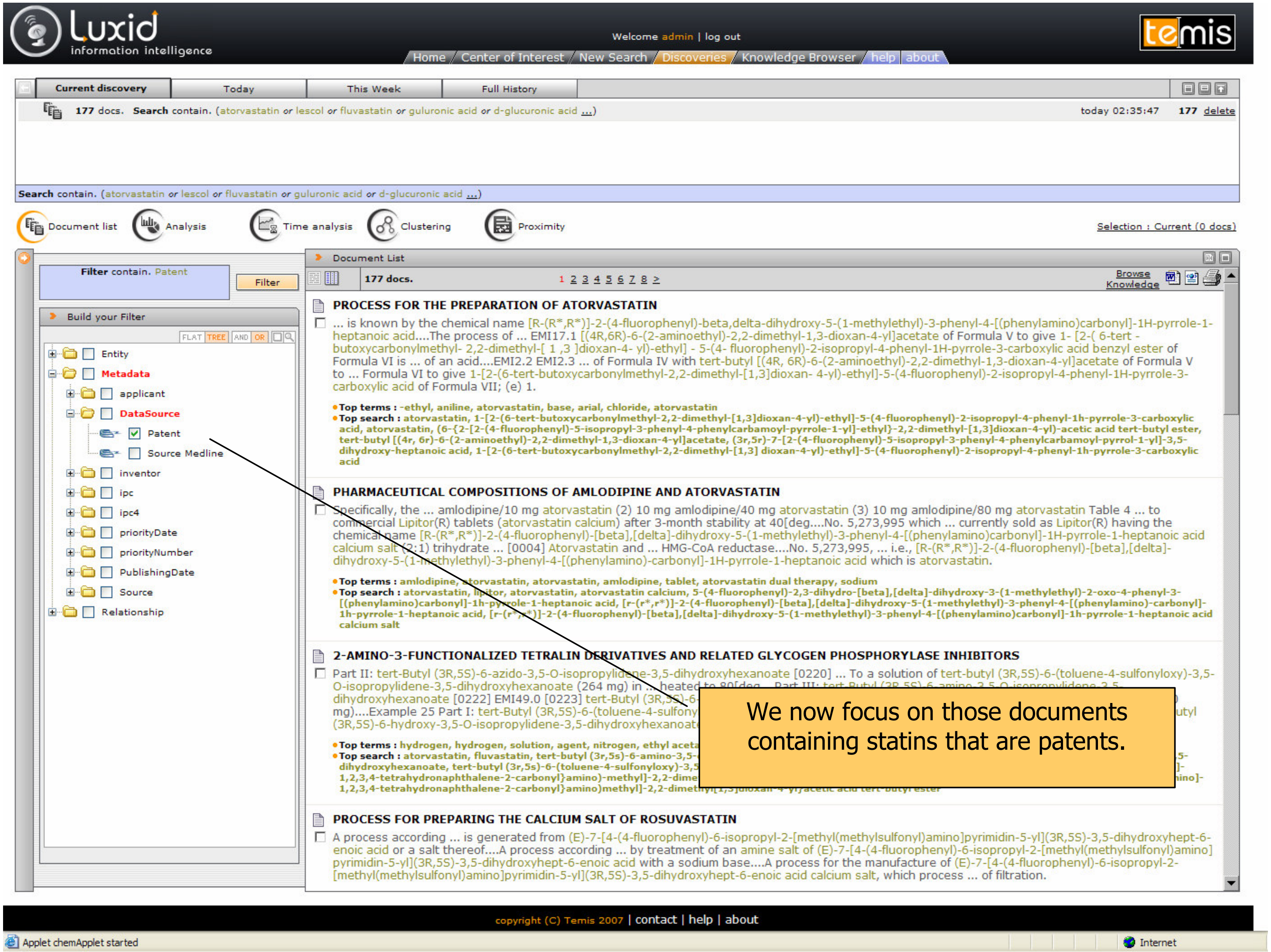
patient.

By showing the value of fluvastatin after percutaneous ... of the statins....This was confirmed ... mg/dl (4-7 mmol/l)....ALLHAT-LLT, ... events.

Persistence and determinants of statin therapy among middle-aged patients free of cardiovascular disease

The document list gives access to the underlying document, the metadata and identified objects as well as ist document structure.

Discover all structures



177 docs. **Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...) today 02:35:47 177 delete

Search contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...)

Filter contain. Patent

- ⊞ □ Entity
- ⊟ □ **Metadata**
  - ⊞ □ applicant
  - ⊟ □ **DataSource**
    - ☑ Patent
    - □ Source Medline
  - ⊞ □ inventor
  - ⊞ □ ipc
  - ⊞ □ ipc4
  - ⊞ □ priorityDate
  - ⊞ □ priorityNumber
  - ⊞ □ PublishingDate
  - ⊞ □ Source
- ⊞ □ Relationship

177 docs. 1 2 3 4 5 6 7 8 >

PROCESS FOR THE PREPARATION OF ATORVASTATIN

□ ... is known by the chemical name [R-(R*,R*)]-2-(4-fluorophenyl)-beta,delta-dihydroxy-5-(1-methylethyl)-3-phenyl-4-[(phenylamino)carbonyl]-1H-pyrrole-1-heptanoic acid....The process of ... EMI17.1 [(4R,6R)-6-(2-aminoethyl)-2,2-dimethyl-1,3-dioxan-4-yl]acetate of Formula V to give 1- [2-(6-tert -butoxycarbonylmethyl- 2,2-dimethyl-[1 ,3]dioxan-4- yl)-ethyl] - 5-(4- fluorophenyl)-2-isopropyl-4-phenyl-1H-pyrrole-3-carboxylic acid benzyl ester of Formula VI is ... of an acid....EMI2.2 EMI2.3 ... of Formula IV with tert-butyl [(4R, 6R)-6-(2-aminoethyl)-2,2-dimethyl-1,3-dioxan-4-yl]acetate of Formula V to ... Formula VI to give 1-[2-(6-tert-butoxycarbonylmethyl-2,2-dimethyl-[1,3]dioxan- 4-yl)-ethyl]-5-(4-fluorophenyl)-2-isopropyl-4-phenyl-1H-pyrrole-3-carboxylic acid of Formula VII; (e) 1.

- **Top terms :** -ethyl, aniline, atorvastatin, base, arial, chloride, atorvastatin
- **Top search :** atorvastatin, 1-[2-(6-tert-butoxycarbonylmethyl-2,2-dimethyl-[1,3]dioxan-4-yl)-ethyl]-5-(4-fluorophenyl)-2-isopropyl-4-phenyl-1h-pyrrole-3-carboxylic acid, atorvastatin, (6-{2-[2-(4-fluorophenyl)-5-isopropyl-3-phenyl-4-phenylcarbamoyl-pyrrole-1-yl]-ethyl}-2,2-dimethyl-[1,3]dioxan-4-yl)-acetic acid tert-butyl ester, tert-butyl [(4r, 6r)-6-(2-aminoethyl)-2,2-dimethyl-1,3-dioxan-4-yl]acetate, (3r,5r)-7-[2-(4-fluorophenyl)-5-isopropyl-3-phenyl-4-phenylcarbamoyl-pyrrol-1-yl]-3,5-dihydroxy-heptanoic acid, 1-[2-(6-tert-butoxycarbonylmethyl-2,2-dimethyl-[1,3] dioxan-4-yl)-ethyl]-5-(4-fluorophenyl)-2-isopropyl-4-phenyl-1h-pyrrole-3-carboxylic acid

PHARMACEUTICAL COMPOSITIONS OF AMLODIPINE AND ATORVASTATIN

□ Specifically, the ... amlodipine/10 mg atorvastatin (2) 10 mg amlodipine/40 mg atorvastatin (3) 10 mg amlodipine/80 mg atorvastatin Table 4 ... to commercial Lipitor(R) tablets (atorvastatin calcium) after 3-month stability at 40[deg....No. 5,273,995 which ... currently sold as Lipitor(R) having the chemical name [R-(R*,R*)]-2-(4-fluorophenyl)-[beta],[delta]-dihydroxy-5-(1-methylethyl)-3-phenyl-4-[(phenylamino)carbonyl]-1H-pyrrole-1-heptanoic acid calcium salt (2:1) trihydrate ... [0004] Atorvastatin and ... HMG-CoA reductase....No. 5,273,995, ... i.e., [R-(R*,R*)]-2-(4-fluorophenyl)-[beta],[delta]-dihydroxy-5-(1-methylethyl)-3-phenyl-4-[(phenylamino)-carbonyl]-1H-pyrrole-1-heptanoic acid which is atorvastatin.

- **Top terms :** amlodipine, atorvastatin, atorvastatin, amlodipine, tablet, atorvastatin dual therapy, sodium
- **Top search :** atorvastatin, lipitor, atorvastatin, atorvastatin calcium, 5-(4-fluorophenyl)-2,3-dihydro-[beta],[delta]-dihydroxy-3-(1-methylethyl)-2-oxo-4-phenyl-3-[(phenylamino)carbonyl]-1h-pyrrole-1-heptanoic acid, [r-(r*,r*)]-2-(4-fluorophenyl)-[beta],[delta]-dihydroxy-5-(1-methylethyl)-3-phenyl-4-[(phenylamino)-carbonyl]-1h-pyrrole-1-heptanoic acid, [r-(r*,r*)]-2-(4-fluorophenyl)-[beta],[delta]-dihydroxy-5-(1-methylethyl)-3-phenyl-4-[(phenylamino)carbonyl]-1h-pyrrole-1-heptanoic acid calcium salt

2-AMINO-3-FUNCTIONALIZED TETRALIN DERIVATIVES AND RELATED GLYCOGEN PHOSPHORYLASE INHIBITORS

□ Part II: tert-Butyl (3R,5S)-6-azido-3,5-O-isopropylidene-3,5-dihydroxyhexanoate [0220] ... To a solution of tert-butyl (3R,5S)-6-(toluene-4-sulfonyloxy)-3,5-O-isopropylidene-3,5-dihydroxyhexanoate (264 mg) in ... heated to 80[deg....Part III: tert-Butyl (3R,5S)-6-amino-3,5-O-isopropylidene-3,5-dihydroxyhexanoate [0222] EMI49.0 [0223] tert-Butyl (3R,5S)-6-) mg)....Example 25 Part I: tert-Butyl (3R,5S)-6-(toluene-4-sulfony utyl (3R,5S)-6-hydroxy-3,5-O-isopropylidene-3,5-dihydroxyhexanoat

- **Top terms :** hydrogen, hydrogen, solution, agent, nitrogen, ethyl aceta
- **Top search :** atorvastatin, fluvastatin, tert-butyl (3r,5s)-6-amino-3,5- 5-dihydroxyhexanoate, tert-butyl (3r,5s)-6-(toluene-4-sulfonyloxy)-3,5]-1,2,3,4-tetrahydronaphthalene-2-carbonyl}amino)-methyl]-2,2-dime ino]-1,2,3,4-tetrahydronaphthalene-2-carbonyl}amino)methyl]-2,2-dimethyl[1,3]dioxan-4-yl)acetic acid tert-butyl ester

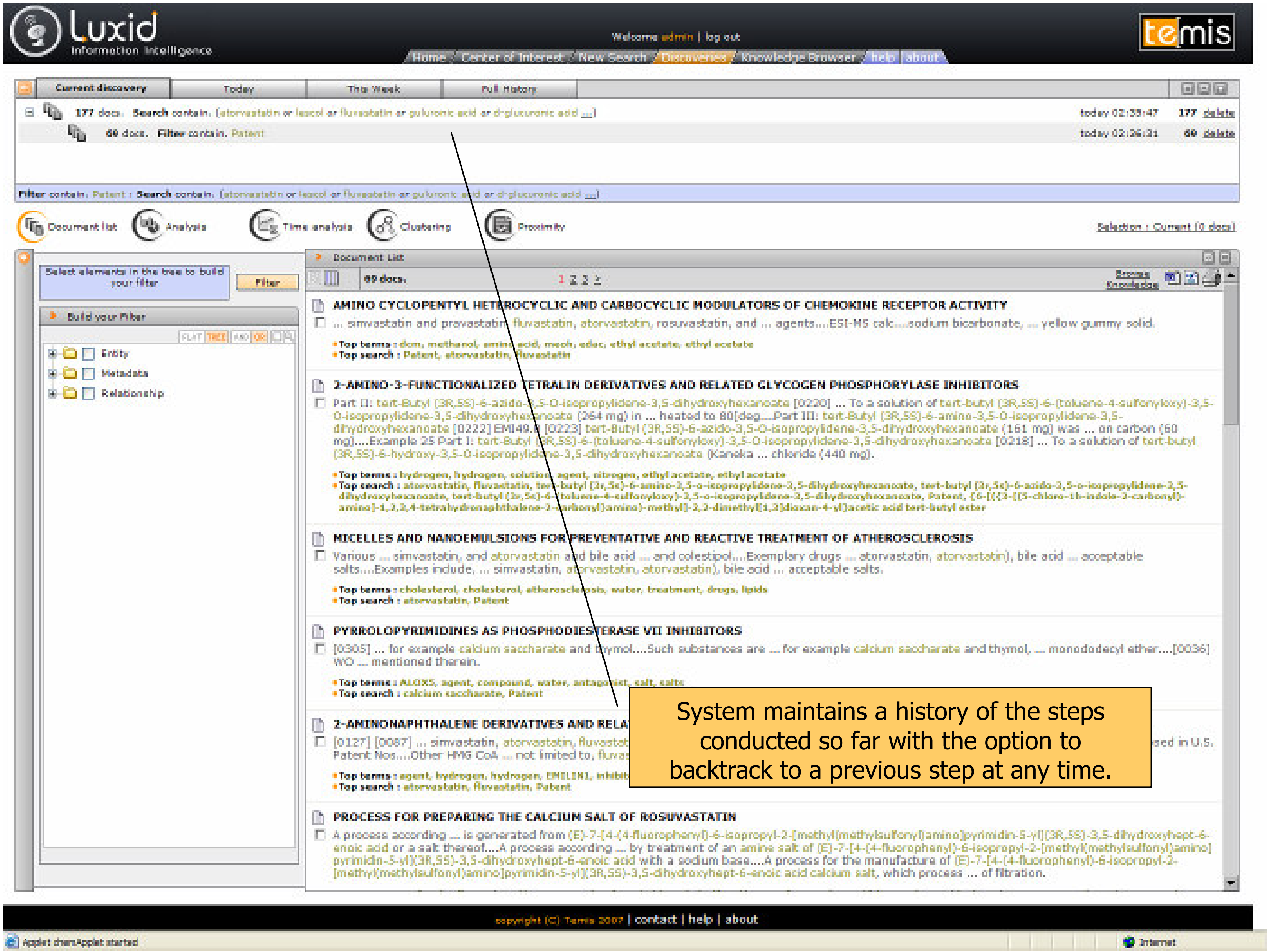
We now focus on those documents containing statins that are patents.

PROCESS FOR PREPARING THE CALCIUM SALT OF ROSUVASTATIN

□ A process according ... is generated from (E)-7-[4-(4-fluorophenyl)-6-isopropyl-2-[methyl(methylsulfonyl)amino]pyrimidin-5-yl](3R,5S)-3,5-dihydroxyhept-6-enoic acid or a salt thereof....A process according ... by treatment of an amine salt of (E)-7-[4-(4-fluorophenyl)-6-isopropyl-2-[methyl(methylsulfonyl)amino] pyrimidin-5-yl](3R,5S)-3,5-dihydroxyhept-6-enoic acid with a sodium base....A process for the manufacture of (E)-7-[4-(4-fluorophenyl)-6-isopropyl-2-[methyl(methylsulfonyl)amino]pyrimidin-5-yl](3R,5S)-3,5-dihydroxyhept-6-enoic acid calcium salt, which process ... of filtration.

copyright (C) Temis 2007 | contact | help | about
Applet chemApplet started
Internet

Filter on patent



177 docs. Search contain. (atorvastatin or lescol or fluvastatin or guluronic acid or d-glucuronic acid ...) today 02:33:47 177 delete

69 docs. Filter contain. Patent today 02:36:31 69 delete

Filter contain. Patent : Search contain. (atorvastatin or lescol or fluvastatin or guluronic acid or d-glucuronic acid ...)

69 docs.

AMINO CYCLOPENTYL HETEROCYCLIC AND CARBOCYCLIC MODULATORS OF CHEMOKINE RECEPTOR ACTIVITY

□ ... simvastatin and pravastatin fluvastatin, atorvastatin, rosuvastatin, and ... agents....ESI-MS calc....sodium bicarbonate, ... yellow gummy solid.

- **Top terms :** dcm, methanol, amino acid, meoh, edac, ethyl acetate, ethyl acetate
- **Top search :** Patent, atorvastatin, fluvastatin

2-AMINO-3-FUNCTIONALIZED TETRALIN DERIVATIVES AND RELATED GLYCOGEN PHOSPHORYLASE INHIBITORS

□ Part II: tert-Butyl (3R,5S)-6-azido-3,5-O-isopropylidene-3,5-dihydroxyhexanoate [0220] ... To a solution of tert-butyl (3R,5S)-6-(toluene-4-sulfonyloxy)-3,5-O-isopropylidene-3,5-dihydroxyhexanoate (264 mg) in ... heated to 80[deg....Part III: tert-Butyl (3R,5S)-6-amino-3,5-O-isopropylidene-3,5-dihydroxyhexanoate [0222] EMI49.1 [0223] tert-Butyl (3R,5S)-6-azido-3,5-O-isopropylidene-3,5-dihydroxyhexanoate (161 mg) was ... on carbon (60 mg)....Example 25 Part I: tert-Butyl (3R,5S)-6-(toluene-4-sulfonyloxy)-3,5-O-isopropylidene-3,5-dihydroxyhexanoate [0218] ... To a solution of tert-butyl (3R,5S)-6-hydroxy-3,5-O-isopropylidene-3,5-dihydroxyhexanoate (Kaneka ... chloride (440 mg).

- **Top terms :** hydrogen, hydrogen, solution, agent, nitrogen, ethyl acetate, ethyl acetate
- **Top search :** atorvastatin, fluvastatin, tert-butyl (3r,5s)-6-amino-3,5-o-isopropylidene-3,5-dihydroxyhexanoate, tert-butyl (3r,5s)-6-azido-3,5-o-isopropylidene-3,5-dihydroxyhexanoate, tert-butyl (3r,5s)-6-(toluene-4-sulfonyloxy)-3,5-o-isopropylidene-3,5-dihydroxyhexanoate, Patent, {6-[({3-[(5-chloro-1h-indole-2-carbonyl)-amino]-1,2,3,4-tetrahydronaphthalene-2-carbonyl}amino)-methyl]-2,2-dimethyl[1,3]dioxan-4-yl}acetic acid tert-butyl ester

MICELLES AND NANOEMULSIONS FOR PREVENTATIVE AND REACTIVE TREATMENT OF ATHEROSCLEROSIS

□ Various ... simvastatin, and atorvastatin and bile acid ... and colestipol....Exemplary drugs ... atorvastatin, atorvastatin), bile acid ... acceptable salts....Examples include, ... simvastatin, atorvastatin, atorvastatin), bile acid ... acceptable salts.

- **Top terms :** cholesterol, cholesterol, atherosclerosis, water, treatment, drugs, lipids
- **Top search :** atorvastatin, Patent

PYRROLOPYRIMIDINES AS PHOSPHODIESTERASE VII INHIBITORS

□ [0305] ... for example calcium saccharate and thymol....Such substances are ... for example calcium saccharate and thymol, ... monododecyl ether....[0036] WO ... mentioned therein.

- **Top terms :** ALOX5, agent, compound, water, antagonist, salt, salts
- **Top search :** calcium saccharate, Patent

2-AMINONAPHTHALENE DERIVATIVES AND RELA

□ [0127] [0087] ... simvastatin, atorvastatin, fluvastat ...sed in U.S. Patent Nos....Other HMG CoA ... not limited to, fluvas

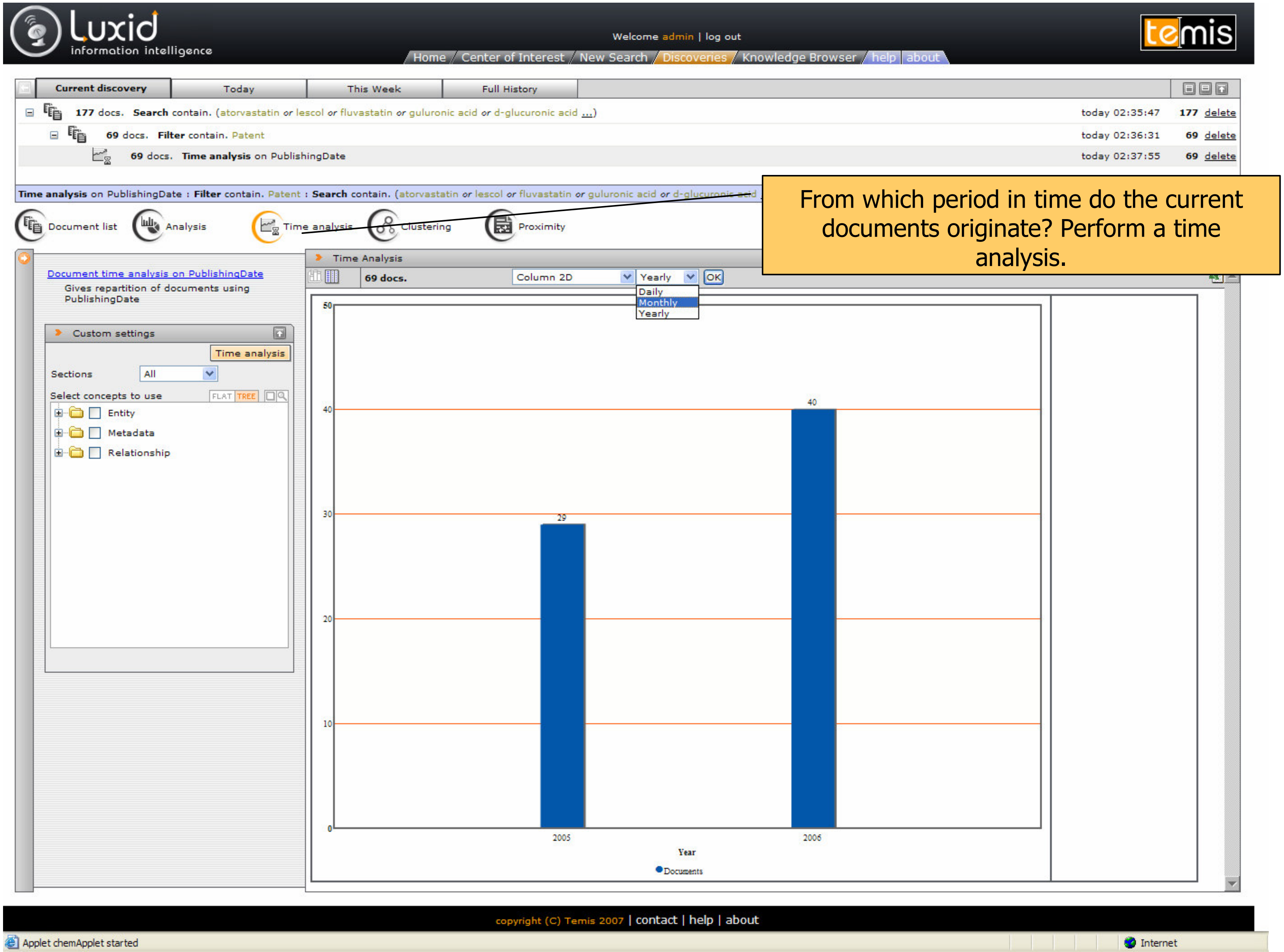
- **Top terms :** agent, hydrogen, hydrogen, EMILIN1, inhibit
- **Top search :** atorvastatin, fluvastatin, Patent

PROCESS FOR PREPARING THE CALCIUM SALT OF ROSUVASTATIN

□ A process according ... is generated from (E)-7-[4-(4-fluorophenyl)-6-isopropyl-2-[methyl(methylsulfonyl)amino]pyrimidin-5-yl](3R,5S)-3,5-dihydroxyhept-6-enoic acid or a salt thereof....A process according ... by treatment of an amine salt of (E)-7-[4-(4-fluorophenyl)-6-isopropyl-2-[methyl(methylsulfonyl)amino]pyrimidin-5-yl](3R,5S)-3,5-dihydroxyhept-6-enoic acid with a sodium base....A process for the manufacture of (E)-7-[4-(4-fluorophenyl)-6-isopropyl-2-[methyl(methylsulfonyl)amino]pyrimidin-5-yl](3R,5S)-3,5-dihydroxyhept-6-enoic acid calcium salt, which process ... of filtration.

System maintains a history of the steps conducted so far with the option to backtrack to a previous step at any time.

Switch to Time Analysis



Welcome admin | log out

Luxid information intelligence
temis
Home / Center of Interest / New Search / Discoveries / Knowledge Browser / help / about
Current discovery
Today
This Week
Full History

⊟ 177 docs. **Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...) today 02:35:47 **177** delete

⊟ 69 docs. **Filter** contain. Patent today 02:36:31 **69** delete

69 docs. **Time analysis** on PublishingDate today 02:37:55 **69** delete

Time analysis on PublishingDate : **Filter** contain. Patent : **Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid

Document time analysis on PublishingDate

Gives repartition of documents using PublishingDate

Sections: All

Select concepts to use

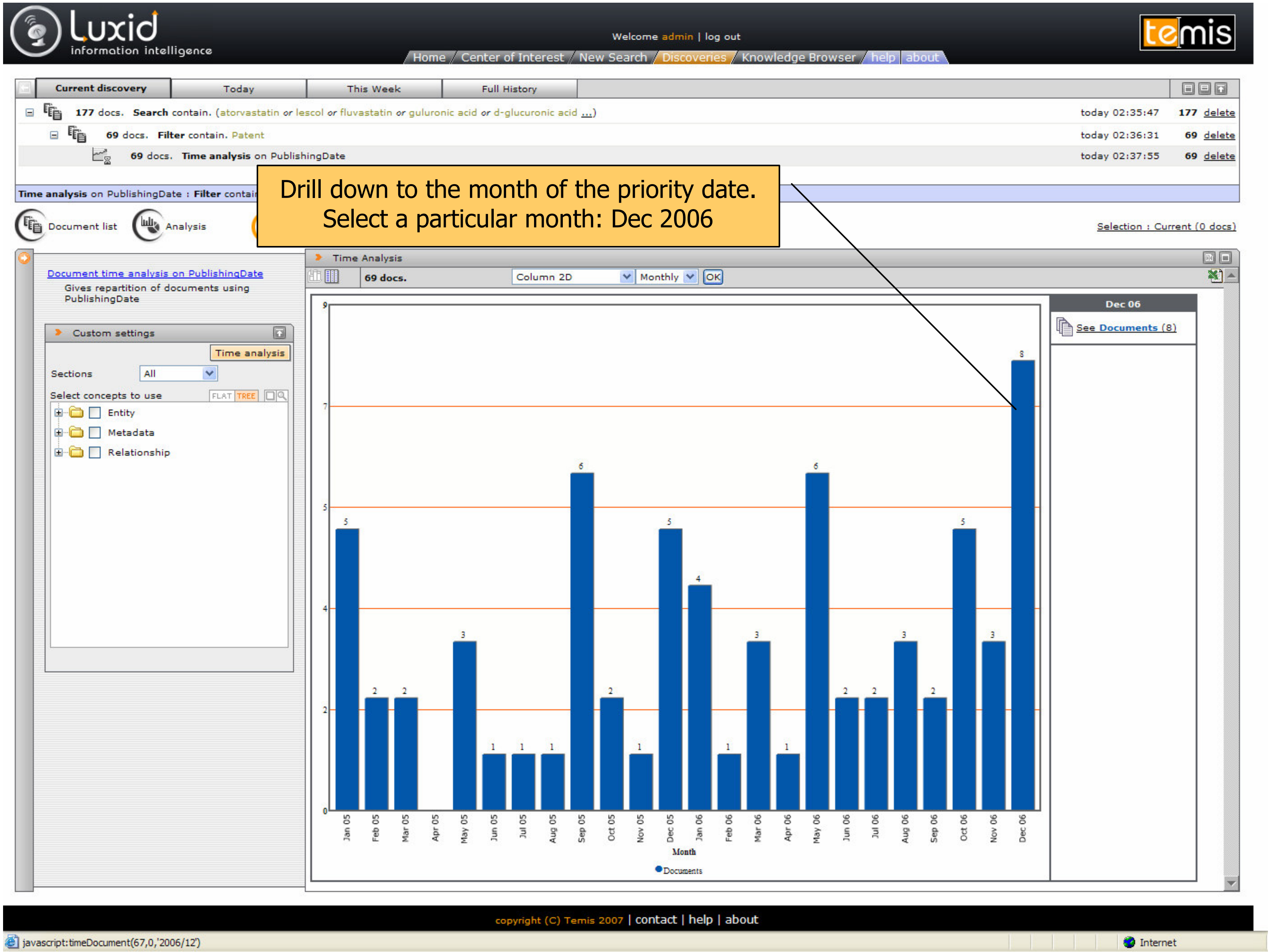
- ⊞ □ Entity
- ⊞ □ Metadata
- ⊞ □ Relationship

69 docs. Column 2D Yearly OK

copyright (C) Temis 2007 | contact | help | about

From which period in time do the current documents originate? Perform a time analysis.

Access monthly publication trends



177 docs. **Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...) today 02:35:47 177 delete

69 docs. **Filter** contain. Patent today 02:36:31 69 delete

69 docs. **Time analysis** on PublishingDate today 02:37:55 69 delete

Time analysis on PublishingDate : **Filter** contai

Drill down to the month of the priority date.
Select a particular month: Dec 2006

Document time analysis on PublishingDate

Gives repartition of documents using PublishingDate

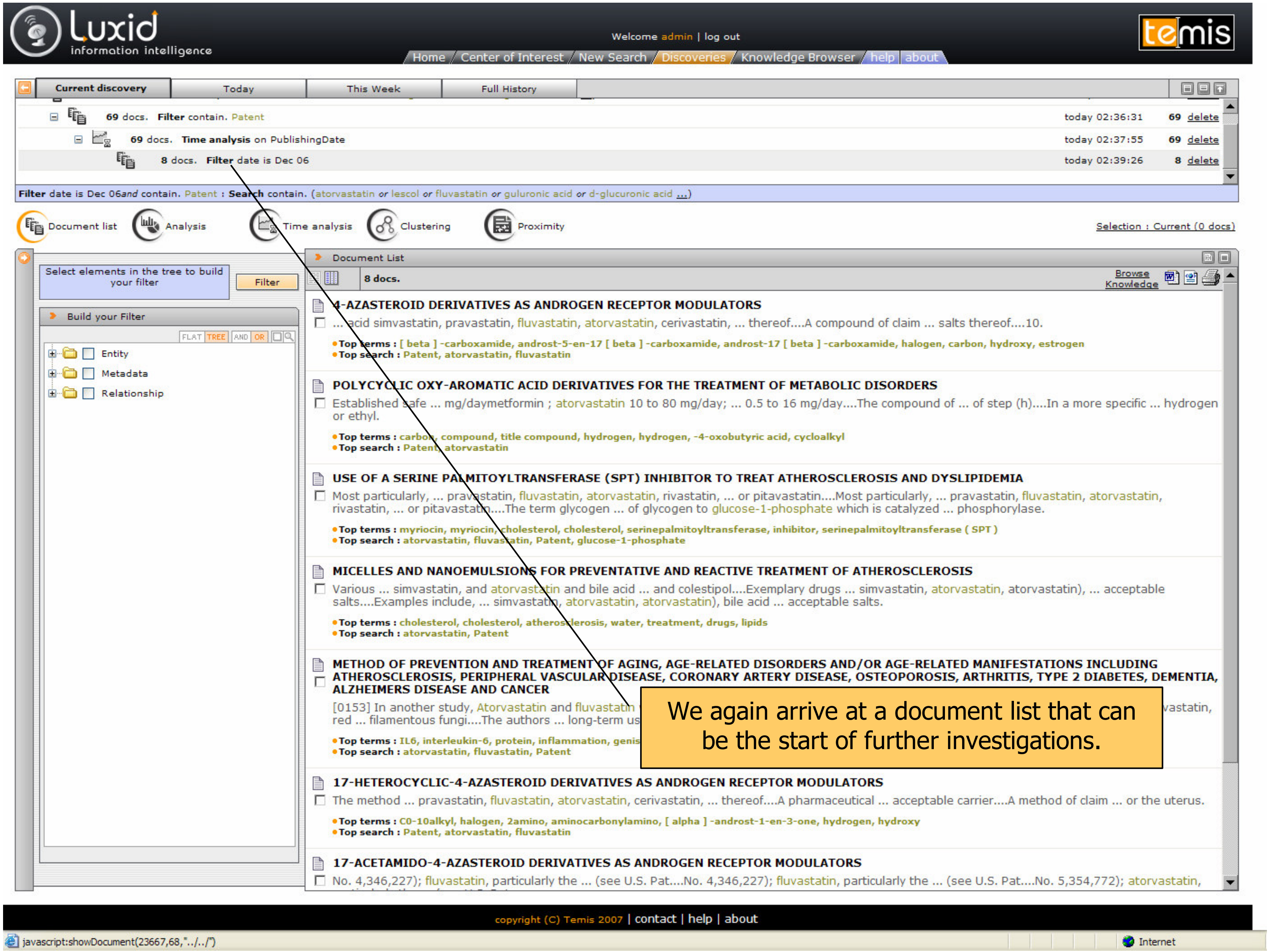
Sections: All

Select concepts to use

69 docs. Column 2D Monthly OK

javascript:timeDocument(67,0,'2006/12')

Pinpoint Dec '06 published patents



69 docs. **Filter** contain. Patent — today 02:36:31 — 69 delete

69 docs. **Time analysis** on PublishingDate — today 02:37:55 — 69 delete

8 docs. **Filter** date is Dec 06 — today 02:39:26 — 8 delete

Filter date is Dec 06*and* contain. Patent **: Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...)

8 docs.

4-AZASTEROID DERIVATIVES AS ANDROGEN RECEPTOR MODULATORS

□ ... acid simvastatin, pravastatin, fluvastatin, atorvastatin, cerivastatin, ... thereof....A compound of claim ... salts thereof....10.

- **Top terms :** [beta] -carboxamide, androst-5-en-17 [beta] -carboxamide, androst-17 [beta] -carboxamide, halogen, carbon, hydroxy, estrogen
- **Top search :** Patent, atorvastatin, fluvastatin

POLYCYCLIC OXY-AROMATIC ACID DERIVATIVES FOR THE TREATMENT OF METABOLIC DISORDERS

□ Established safe ... mg/daymetformin ; atorvastatin 10 to 80 mg/day; ... 0.5 to 16 mg/day....The compound of ... of step (h)....In a more specific ... hydrogen or ethyl.

- **Top terms :** carbon, compound, title compound, hydrogen, hydrogen, -4-oxobutyric acid, cycloalkyl
- **Top search :** Patent, atorvastatin

USE OF A SERINE PALMITOYLTRANSFERASE (SPT) INHIBITOR TO TREAT ATHEROSCLEROSIS AND DYSLIPIDEMIA

□ Most particularly, ... pravastatin, fluvastatin, atorvastatin, rivastatin, ... or pitavastatin....Most particularly, ... pravastatin, fluvastatin, atorvastatin, rivastatin, ... or pitavastatin....The term glycogen ... of glycogen to glucose-1-phosphate which is catalyzed ... phosphorylase.

- **Top terms :** myriocin, myriocin, cholesterol, cholesterol, serinepalmitoyltransferase, inhibitor, serinepalmitoyltransferase (SPT)
- **Top search :** atorvastatin, fluvastatin, Patent, glucose-1-phosphate

MICELLES AND NANOEMULSIONS FOR PREVENTATIVE AND REACTIVE TREATMENT OF ATHEROSCLEROSIS

□ Various ... simvastatin, and atorvastatin and bile acid ... and colestipol....Exemplary drugs ... simvastatin, atorvastatin, atorvastatin), ... acceptable salts....Examples include, ... simvastatin, atorvastatin, atorvastatin), bile acid ... acceptable salts.

- **Top terms :** cholesterol, cholesterol, atherosclerosis, water, treatment, drugs, lipids
- **Top search :** atorvastatin, Patent

METHOD OF PREVENTION AND TREATMENT OF AGING, AGE-RELATED DISORDERS AND/OR AGE-RELATED MANIFESTATIONS INCLUDING ATHEROSCLEROSIS, PERIPHERAL VASCULAR DISEASE, CORONARY ARTERY DISEASE, OSTEOPOROSIS, ARTHRITIS, TYPE 2 DIABETES, DEMENTIA, ALZHEIMERS DISEASE AND CANCER

□ [0153] In another study, Atorvastatin and fluvastatin vastatin, red ... filamentous fungi....The authors ... long-term us

- **Top terms :** IL6, interleukin-6, protein, inflammation, genis
- **Top search :** atorvastatin, fluvastatin, Patent

17-HETEROCYCLIC-4-AZASTEROID DERIVATIVES AS ANDROGEN RECEPTOR MODULATORS

□ The method ... pravastatin, fluvastatin, atorvastatin, cerivastatin, ... thereof....A pharmaceutical ... acceptable carrier....A method of claim ... or the uterus.

- **Top terms :** C0-10alkyl, halogen, 2amino, aminocarbonylamino, [alpha] -androst-1-en-3-one, hydrogen, hydroxy
- **Top search :** Patent, atorvastatin, fluvastatin

17-ACETAMIDO-4-AZASTEROID DERIVATIVES AS ANDROGEN RECEPTOR MODULATORS

□ No. 4,346,227); fluvastatin, particularly the ... (see U.S. Pat....No. 4,346,227); fluvastatin, particularly the ... (see U.S. Pat....No. 5,354,772); atorvastatin,

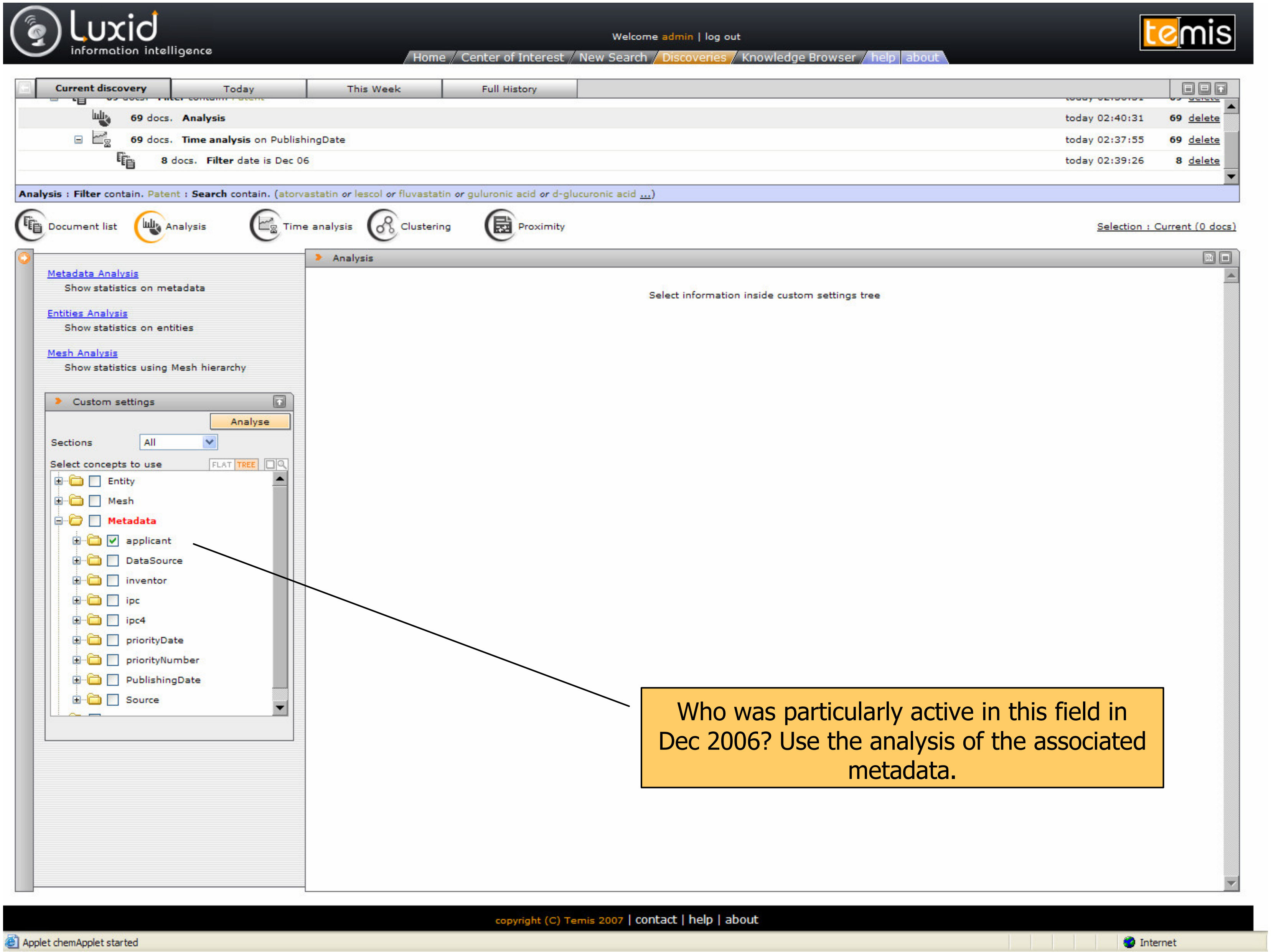
copyright (C) Temis 2007 | contact | help | about

javascript:showDocument(23667,68,"../../")

Internet

We again arrive at a document list that can be the start of further investigations.

Go back to Analysis



Luxid
information intelligence
temis
Welcome admin | log out
Home / Center of Interest / New Search / Discoveries / Knowledge Browser / help / about
Current discovery | Today | This Week | Full History

69 docs. **Analysis**	today 02:40:31	69 delete
69 docs. **Time analysis** on PublishingDate	today 02:37:55	69 delete
8 docs. **Filter** date is Dec 06	today 02:39:26	8 delete

Analysis : **Filter** contain. Patent : **Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...)

Document list
Analysis
Time analysis
Clustering
Proximity
Selection : Current (0 docs)

Metadata Analysis
Show statistics on metadata

Entities Analysis
Show statistics on entities

Mesh Analysis
Show statistics using Mesh hierarchy

Sections: All

Select concepts to use

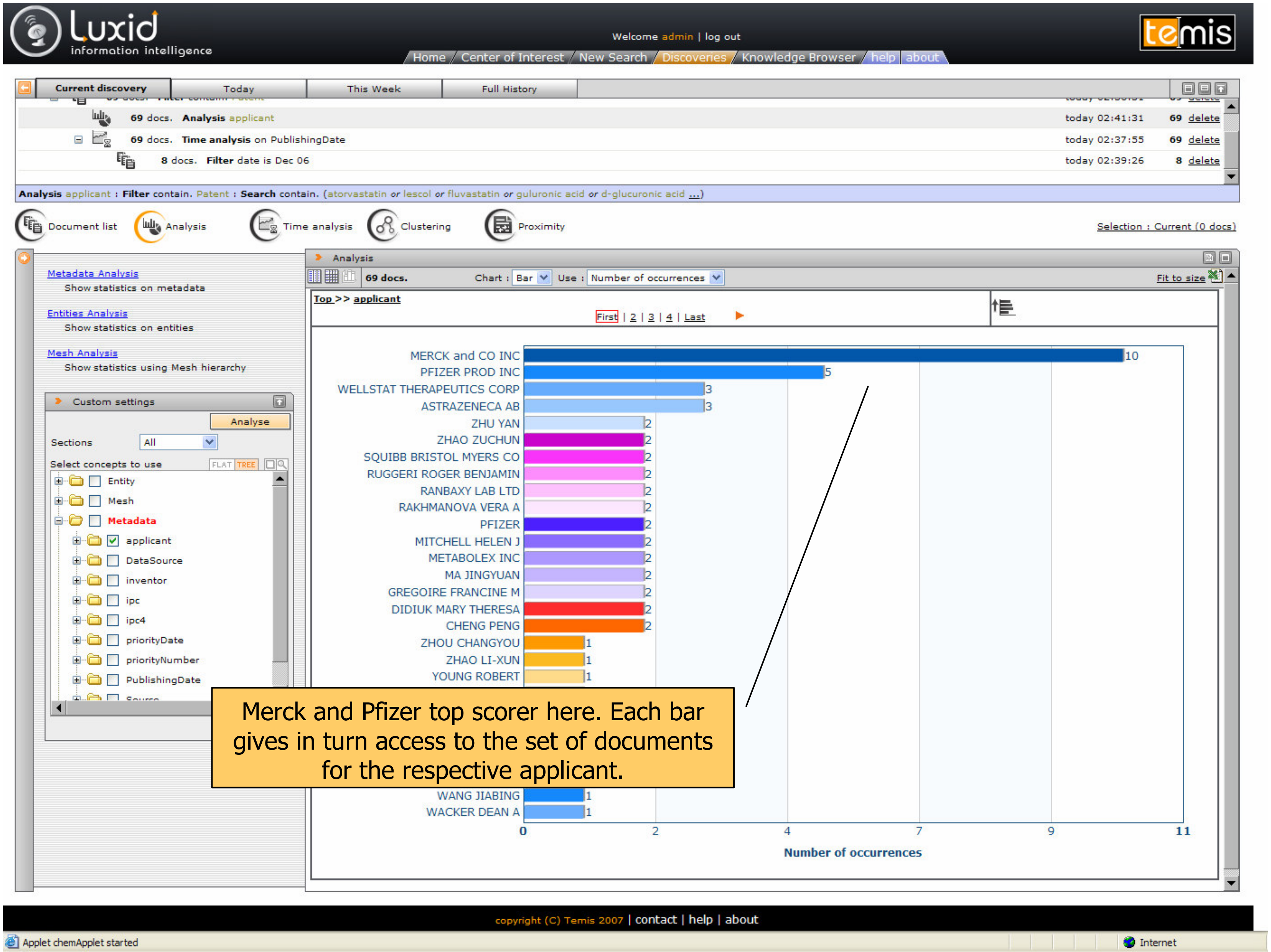
- □ Entity
- □ Mesh
- □ **Metadata**
  - ☑ applicant
  - □ DataSource
  - □ inventor
  - □ ipc
  - □ ipc4
  - □ priorityDate
  - □ priorityNumber
  - □ PublishingDate
  - □ Source

Select information inside custom settings tree

copyright (C) Temis 2007 | contact | help | about
Applet chemApplet started
Internet

Who was particularly active in this field in Dec 2006? Use the analysis of the associated metadata.

View top applicants in the field



69 docs. **Analysis** applicant — today 02:41:31 — 69 delete

69 docs. **Time analysis** on PublishingDate — today 02:37:55 — 69 delete

8 docs. **Filter** date is Dec 06 — today 02:39:26 — 8 delete

Analysis applicant : **Filter** contain. Patent : **Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...)

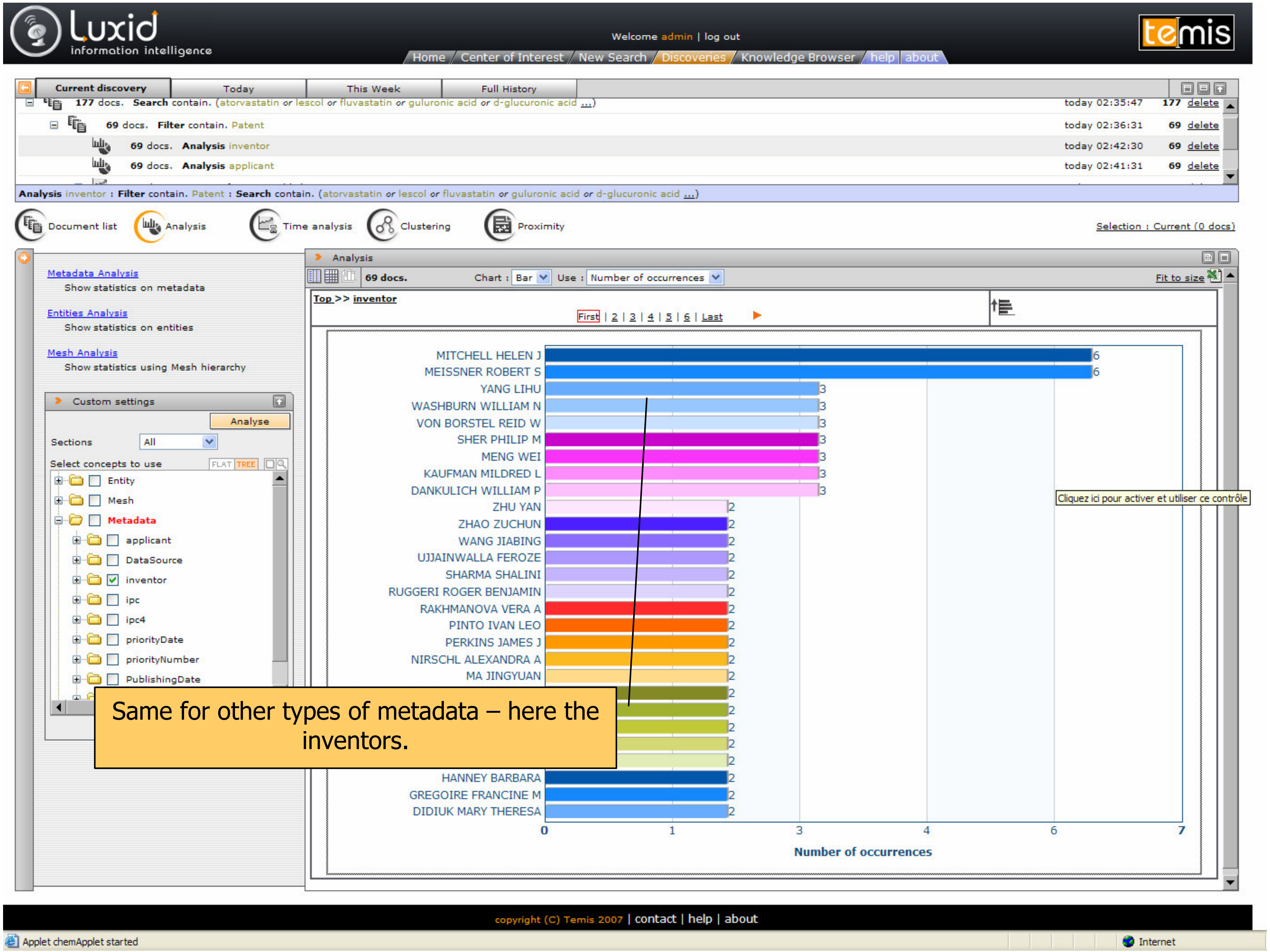
Sections: All

Select concepts to use: ☑ applicant

69 docs. Chart : Bar Use : Number of occurrences

Merck and Pfizer top scorer here. Each bar gives in turn access to the set of documents for the respective applicant.

View top inventors in the field



Welcome admin | log out

Home / Center of Interest / New Search / Discoveries / Knowledge Browser / help / about

Current discovery | Today | This Week | Full History

177 docs. Search contain. (atorvastatin or lescol or fluvastatin or guluronic acid or d-glucuronic acid ...) today 02:35:47 177 delete

69 docs. Filter contain. Patent today 02:36:31 69 delete

69 docs. Analysis inventor today 02:42:30 69 delete

69 docs. Analysis applicant today 02:41:31 69 delete

Analysis inventor : Filter contain. Patent : Search contain. (atorvastatin or lescol or fluvastatin or guluronic acid or d-glucuronic acid ...)

Document list | Analysis | Time analysis | Clustering | Proximity | Selection : Current (0 docs)

Metadata Analysis
Show statistics on metadata

Entities Analysis
Show statistics on entities

Mesh Analysis
Show statistics using Mesh hierarchy

Custom settings

Analyse

Sections: All

Select concepts to use FLAT TREE

- □ Entity
- □ Mesh
- □ Metadata
  - □ applicant
  - □ DataSource
  - ☑ inventor
  - □ ipc
  - □ ipc4
  - □ priorityDate
  - □ priorityNumber
  - □ PublishingDate

Analysis

69 docs. Chart : Bar Use : Number of occurrences Fit to size

Top >> inventor

First | 2 | 3 | 4 | 5 | 6 | Last

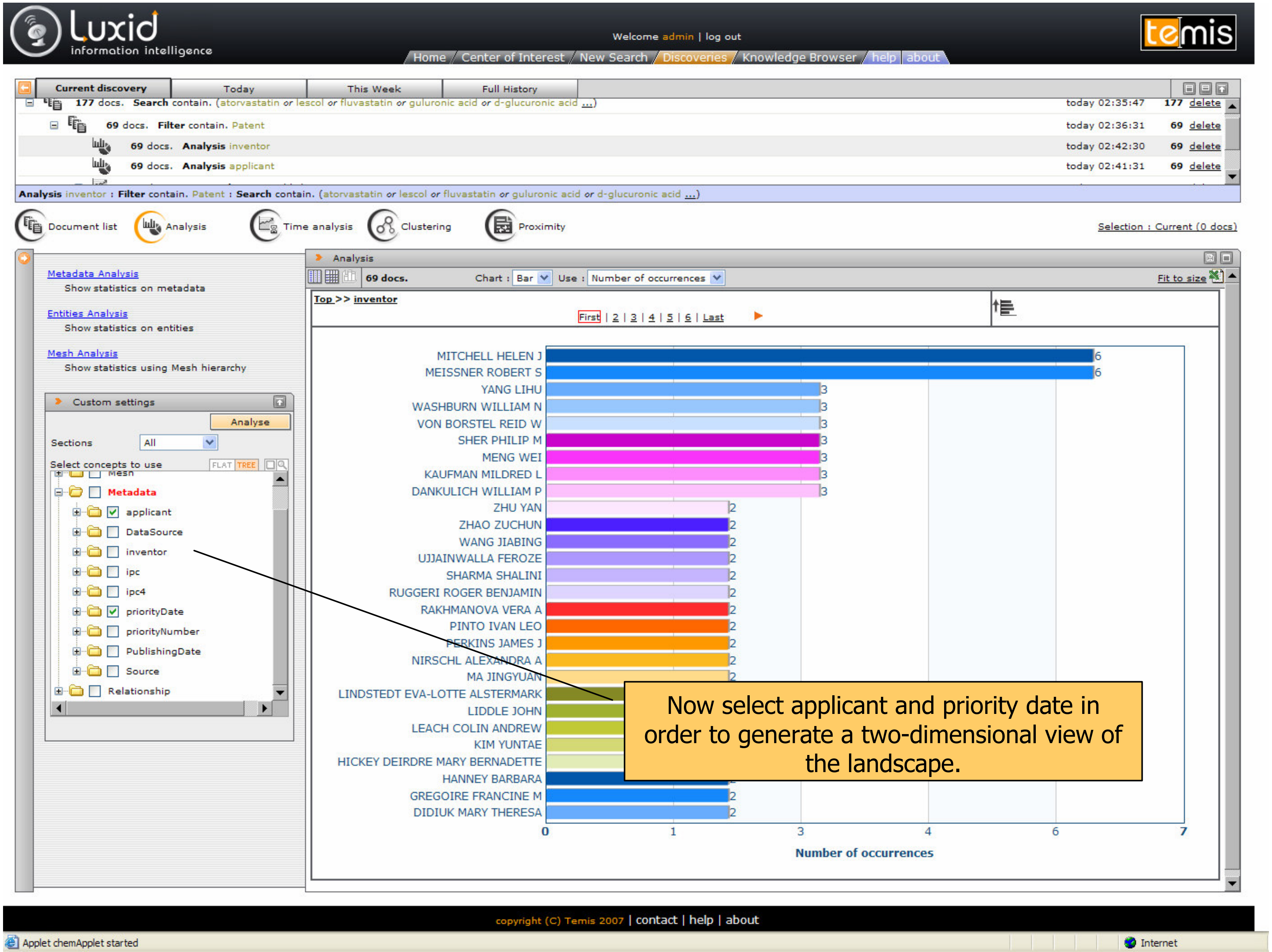
Same for other types of metadata – here the inventors.

copyright (C) Temis 2007 | contact | help | about

Applet chemApplet started

Internet

Prepare a cross-tab Priority vs Applicant



Welcome admin | log out

Home / Center of Interest / New Search / Discoveries / Knowledge Browser / help / about

Current discovery | Today | This Week | Full History

177 docs. **Search** contain. (atorvastatin or lescol or fluvastatin or guluronic acid or d-glucuronic acid ...)	today 02:35:47	177	delete
69 docs. **Filter** contain. Patent	today 02:36:31	69	delete
69 docs. **Analysis** inventor	today 02:42:30	69	delete
69 docs. **Analysis** applicant	today 02:41:31	69	delete

Analysis inventor : **Filter** contain. Patent : **Search** contain. (atorvastatin or lescol or fluvastatin or guluronic acid or d-glucuronic acid ...)

Metadata Analysis
Show statistics on metadata

Entities Analysis
Show statistics on entities

Mesh Analysis
Show statistics using Mesh hierarchy

Custom settings

Analyse

Sections: All

Select concepts to use

- Mesh
- Metadata
  - ☑ applicant
  - □ DataSource
  - □ inventor
  - □ ipc
  - □ ipc4
  - ☑ priorityDate
  - □ priorityNumber
  - □ PublishingDate
  - □ Source
- □ Relationship

Analysis

69 docs. Chart : Bar Use : Number of occurrences

Fit to size

Top >> inventor

First | 2 | 3 | 4 | 5 | 6 | Last

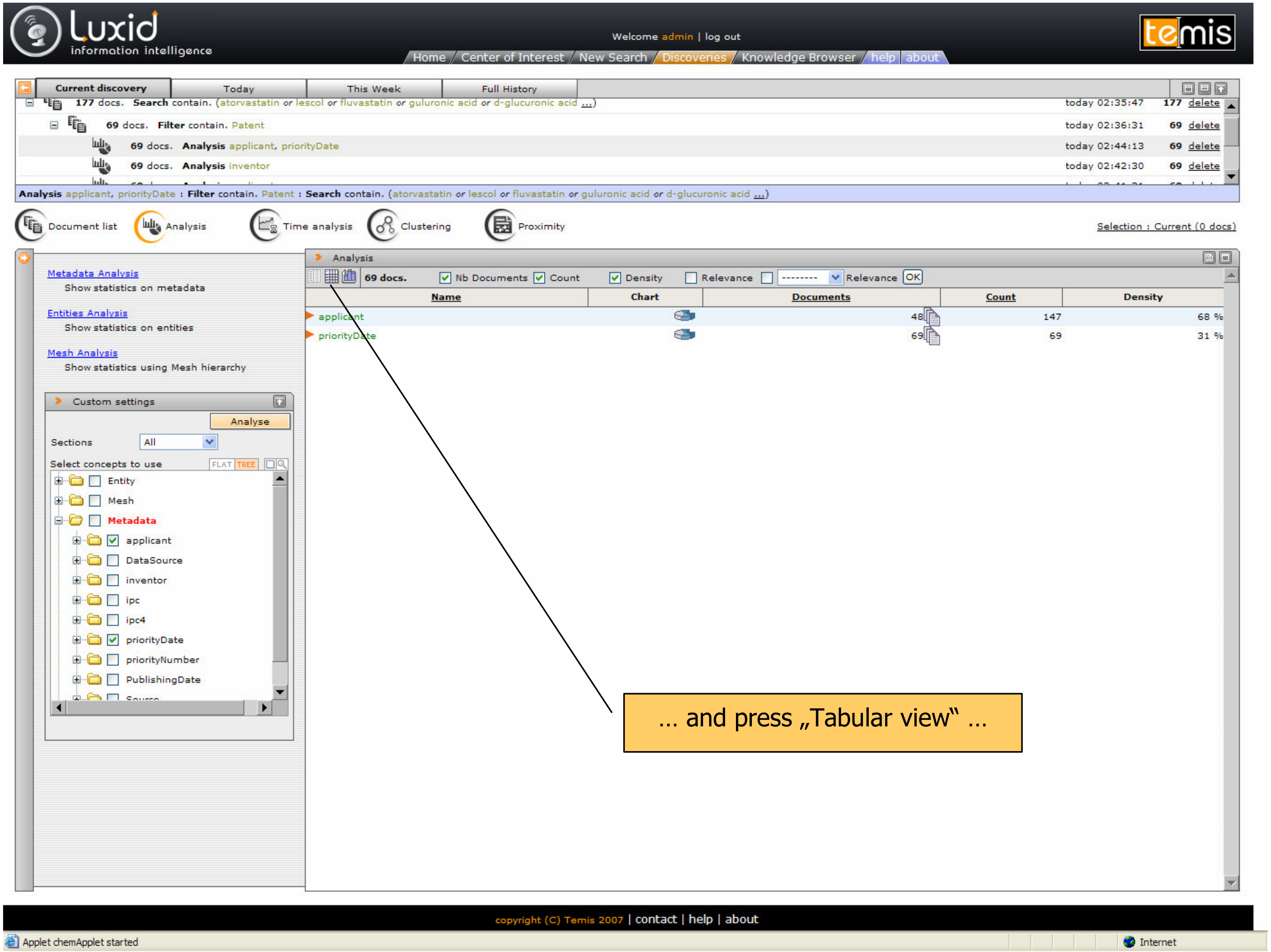
Now select applicant and priority date in order to generate a two-dimensional view of the landscape.

copyright (C) Temis 2007 | contact | help | about

Applet chemApplet started

Internet

A two dimensional analysis…



177 docs. **Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...)	today 02:35:47	177	delete
69 docs. **Filter** contain. Patent	today 02:36:31	69	delete
69 docs. **Analysis** applicant, priorityDate	today 02:44:13	69	delete
69 docs. **Analysis** inventor	today 02:42:30	69	delete

Analysis applicant, priorityDate : **Filter** contain. Patent : **Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...)

Metadata Analysis
Show statistics on metadata

Entities Analysis
Show statistics on entities

Mesh Analysis
Show statistics using Mesh hierarchy

Sections: All

Select concepts to use

- □ Entity
- □ Mesh
- □ Metadata
  - ☑ applicant
  - □ DataSource
  - □ inventor
  - □ ipc
  - □ ipc4
  - ☑ priorityDate
  - □ priorityNumber
  - □ PublishingDate
  - □ Source

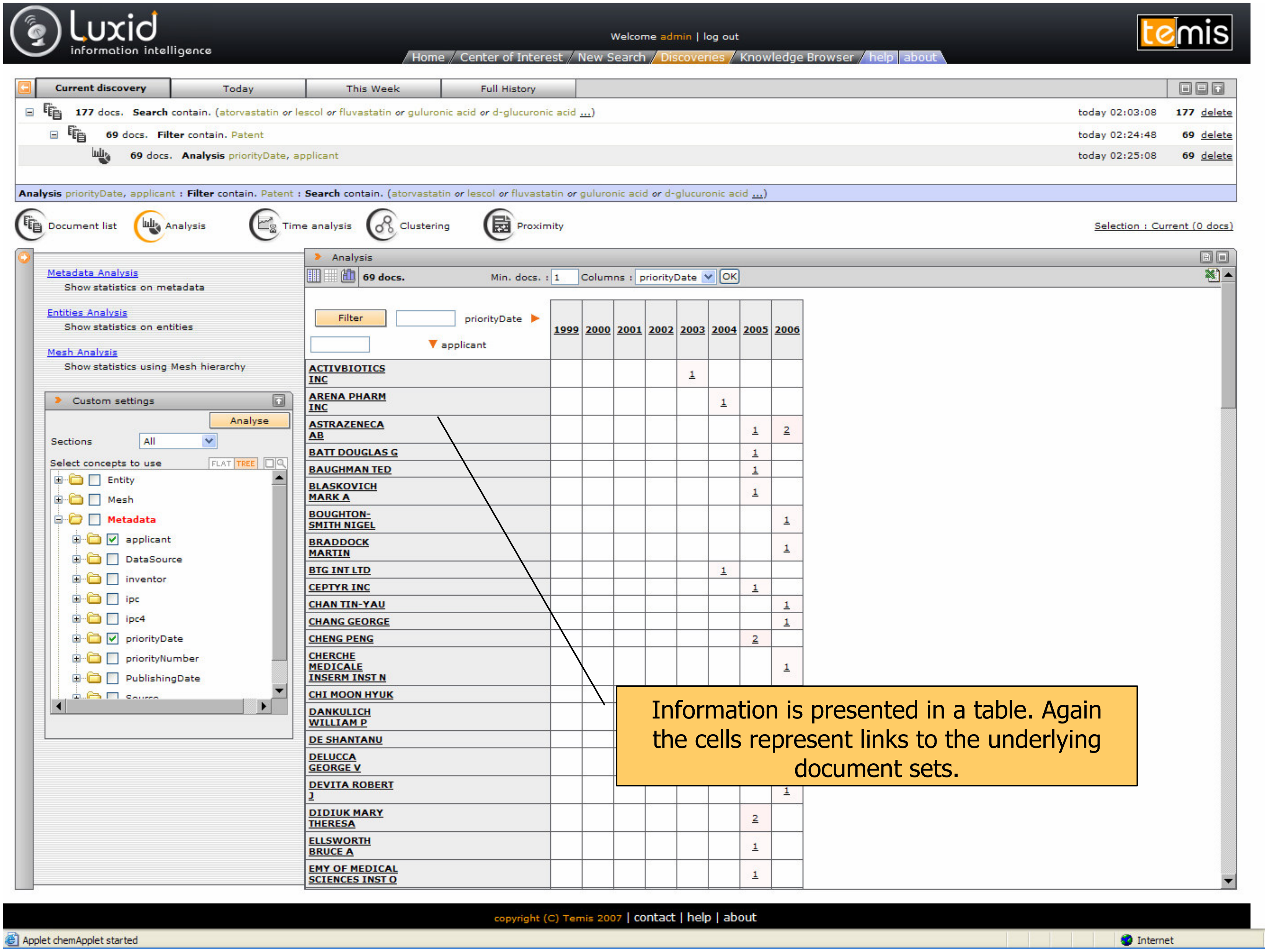
69 docs. ☑ Nb Documents ☑ Count ☑ Density □ Relevance □ -------- Relevance OK

Name	Chart	Documents	Count	Density
applicant		48	147	68 %
priorityDate		69	69	31 %

copyright (C) Temis 2007 | contact | help | about
Applet chemApplet started
Internet

… and press „Tabular view“ …

Turned into informative dashboard



177 docs. **Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...) today 02:03:08 177 delete

69 docs. **Filter** contain. Patent today 02:24:48 69 delete

69 docs. **Analysis** priorityDate, applicant today 02:25:08 69 delete

Analysis priorityDate, applicant : **Filter** contain. Patent : **Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...)

Metadata Analysis
Show statistics on metadata

Entities Analysis
Show statistics on entities

Mesh Analysis
Show statistics using Mesh hierarchy

Sections: All

Select concepts to use

- □ Entity
- □ Mesh
- □ Metadata
  - ☑ applicant
  - □ DataSource
  - □ inventor
  - □ ipc
  - □ ipc4
  - ☑ priorityDate
  - □ priorityNumber
  - □ PublishingDate
  - □ Source

Analysis

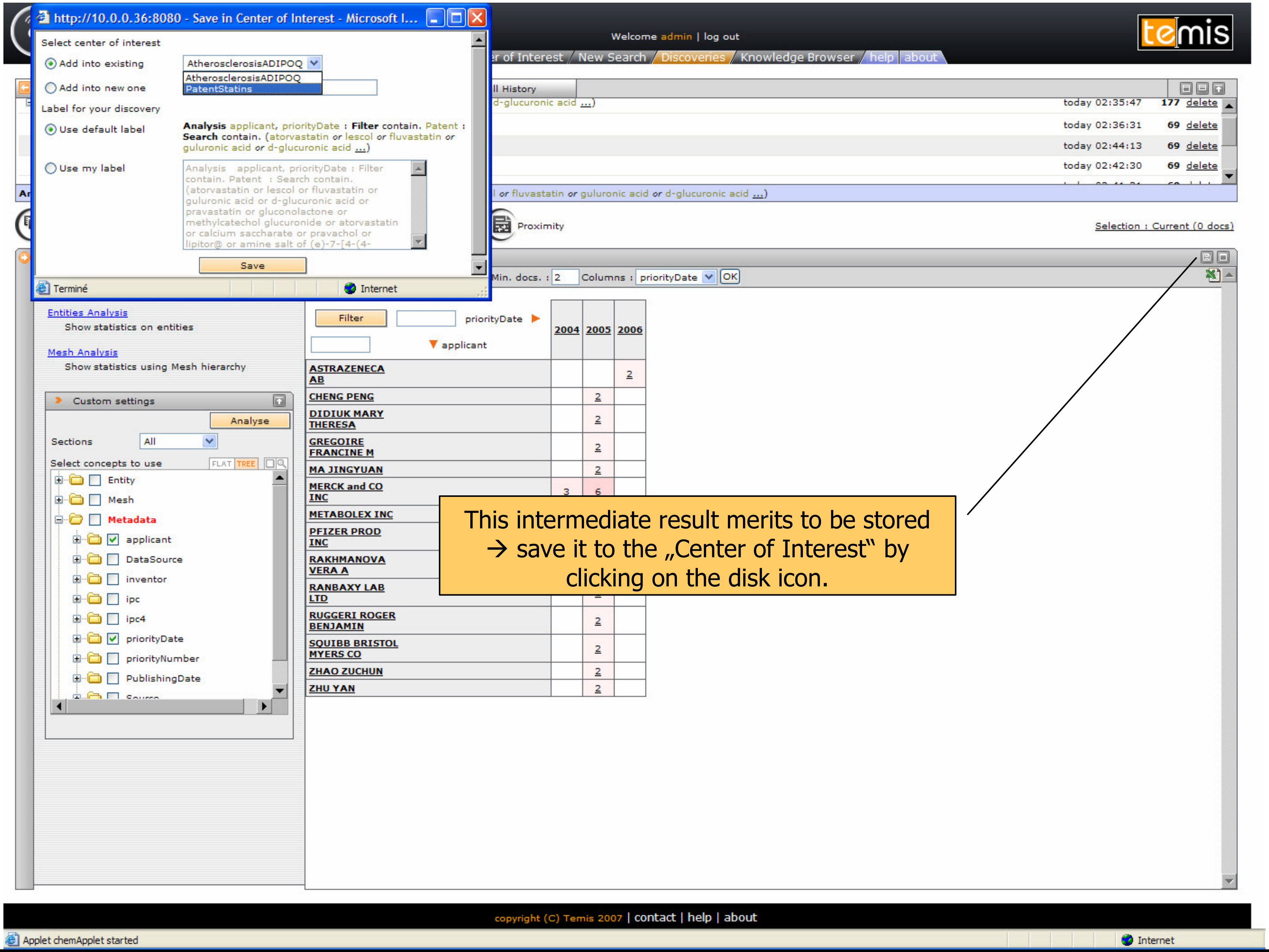
69 docs. Min. docs. : 1 Columns : priorityDate OK

applicant	1999	2000	2001	2002	2003	2004	2005	2006
ACTIVBIOTICS INC					1			
ARENA PHARM INC						1		
ASTRAZENECA AB							1	2
BATT DOUGLAS G							1	
BAUGHMAN TED							1	
BLASKOVICH MARK A							1	
BOUGHTON-SMITH NIGEL								1
BRADDOCK MARTIN								1
BTG INT LTD						1		
CEPTYR INC							1	
CHAN TIN-YAU								1
CHANG GEORGE								1
CHENG PENG							2	
CHERCHE MEDICALE INSERM INST N								1
CHI MOON HYUK								
DANKULICH WILLIAM P								
DE SHANTANU								
DELUCCA GEORGE V								
DEVITA ROBERT J								1
DIDIUK MARY THERESA							2	
ELLSWORTH BRUCE A							1	
EMY OF MEDICAL SCIENCES INST O							1	

Information is presented in a table. Again the cells represent links to the underlying document sets.

copyright (C) Temis 2007 | contact | help | about
Applet chemApplet started
Internet

Save to share with colleagues



http://10.0.0.36:8080 - Save in Center of Interest - Microsoft I...

Select center of interest

● Add into existing: AtherosclerosisADIPOQ (AtherosclerosisADIPOQ / PatentStatins)

○ Add into new one

Label for your discovery

● Use default label: **Analysis** applicant, priorityDate : **Filter** contain. Patent : **Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...)

○ Use my label: Analysis applicant, priorityDate : Filter contain. Patent : Search contain. (atorvastatin or lescol or fluvastatin or guluronic acid or d-glucuronic acid or pravastatin or gluconolactone or methylcatechol glucuronide or atorvastatin or calcium saccharate or pravachol or lipitor@ or amine salt of (e)-7-[4-(4-

Save
Terminé
Internet
Welcome admin | log out
...r of Interest / New Search / Discoveries / Knowledge Browser / help / about
...ll History

...d-glucuronic acid ...)	today 02:35:47	177	delete
	today 02:36:31	69	delete
	today 02:44:13	69	delete
	today 02:42:30	69	delete

...l *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...)

Min. docs. : 2 Columns : priorityDate OK

Entities Analysis
Show statistics on entities
Mesh Analysis
Show statistics using Mesh hierarchy
Custom settings
Analyse

Sections: All

Select concepts to use

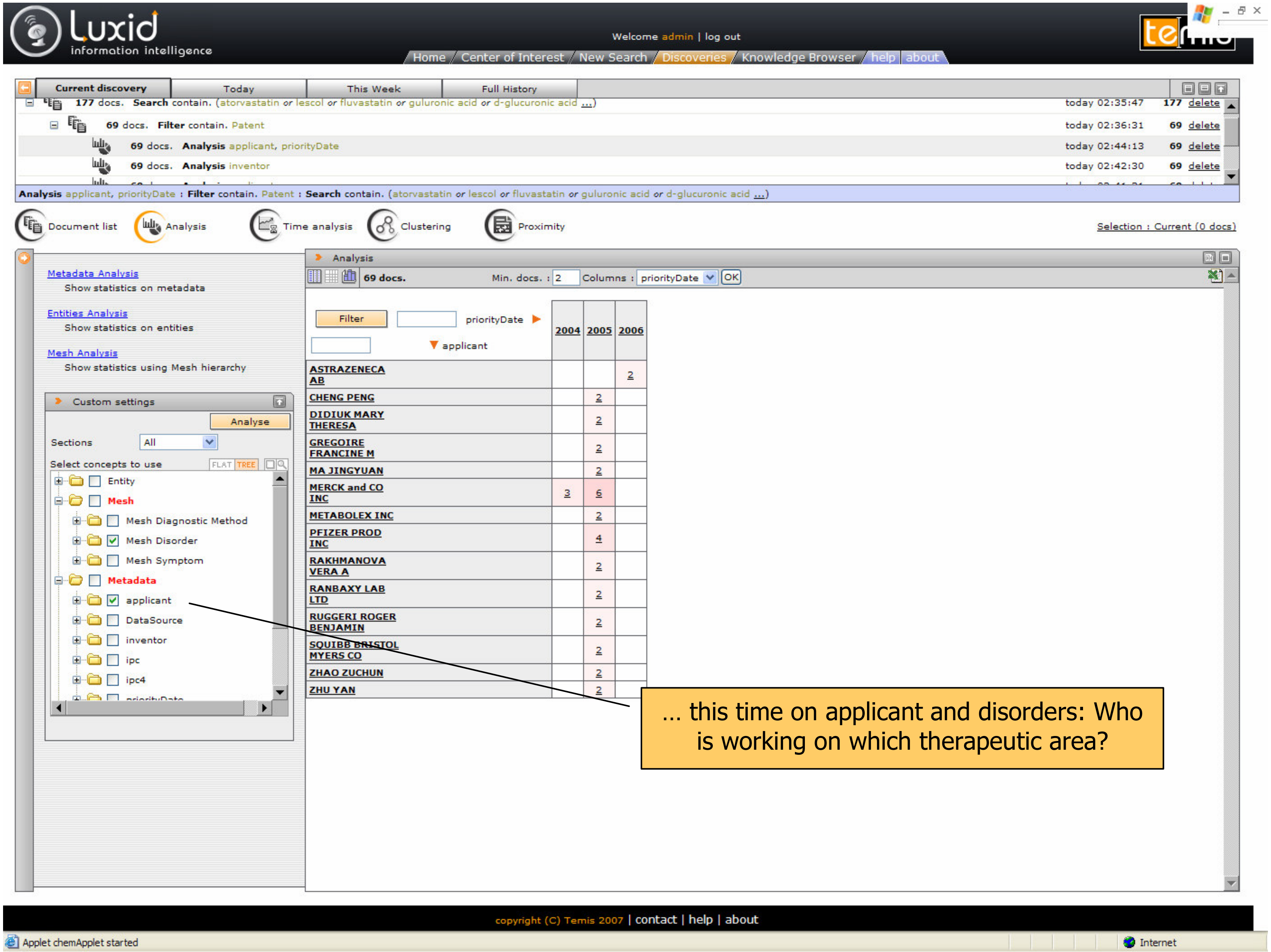
- □ Entity
- □ Mesh
- □ Metadata
  - ☑ applicant
  - □ DataSource
  - □ inventor
  - □ ipc
  - □ ipc4
  - ☑ priorityDate
  - □ priorityNumber
  - □ PublishingDate
  - □ Source

applicant	2004	2005	2006
ASTRAZENECA AB			2
CHENG PENG		2	
DIDIUK MARY THERESA		2	
GREGOIRE FRANCINE M		2	
MA JINGYUAN		2	
MERCK and CO INC	3	6	
METABOLEX INC			
PFIZER PROD INC			
RAKHMANOVA VERA A			
RANBAXY LAB LTD			
RUGGERI ROGER BENJAMIN		2	
SQUIBB BRISTOL MYERS CO		2	
ZHAO ZUCHUN		2	
ZHU YAN		2	

This intermediate result merits to be stored → save it to the „Center of Interest" by clicking on the disk icon.

| contact | help | about

Prepare a second cross analysis



177 docs. **Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...) today 02:35:47 177 delete

69 docs. **Filter** contain. Patent today 02:36:31 69 delete

69 docs. **Analysis** applicant, priorityDate today 02:44:13 69 delete

69 docs. **Analysis** inventor today 02:42:30 69 delete

Analysis applicant, priorityDate : **Filter** contain. Patent : **Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...)

Metadata Analysis
Show statistics on metadata

Entities Analysis
Show statistics on entities

Mesh Analysis
Show statistics using Mesh hierarchy

Sections: All

Select concepts to use

- □ Entity
- □ Mesh
  - □ Mesh Diagnostic Method
  - ☑ Mesh Disorder
  - □ Mesh Symptom
- □ Metadata
  - ☑ applicant
  - □ DataSource
  - □ inventor
  - □ ipc
  - □ ipc4

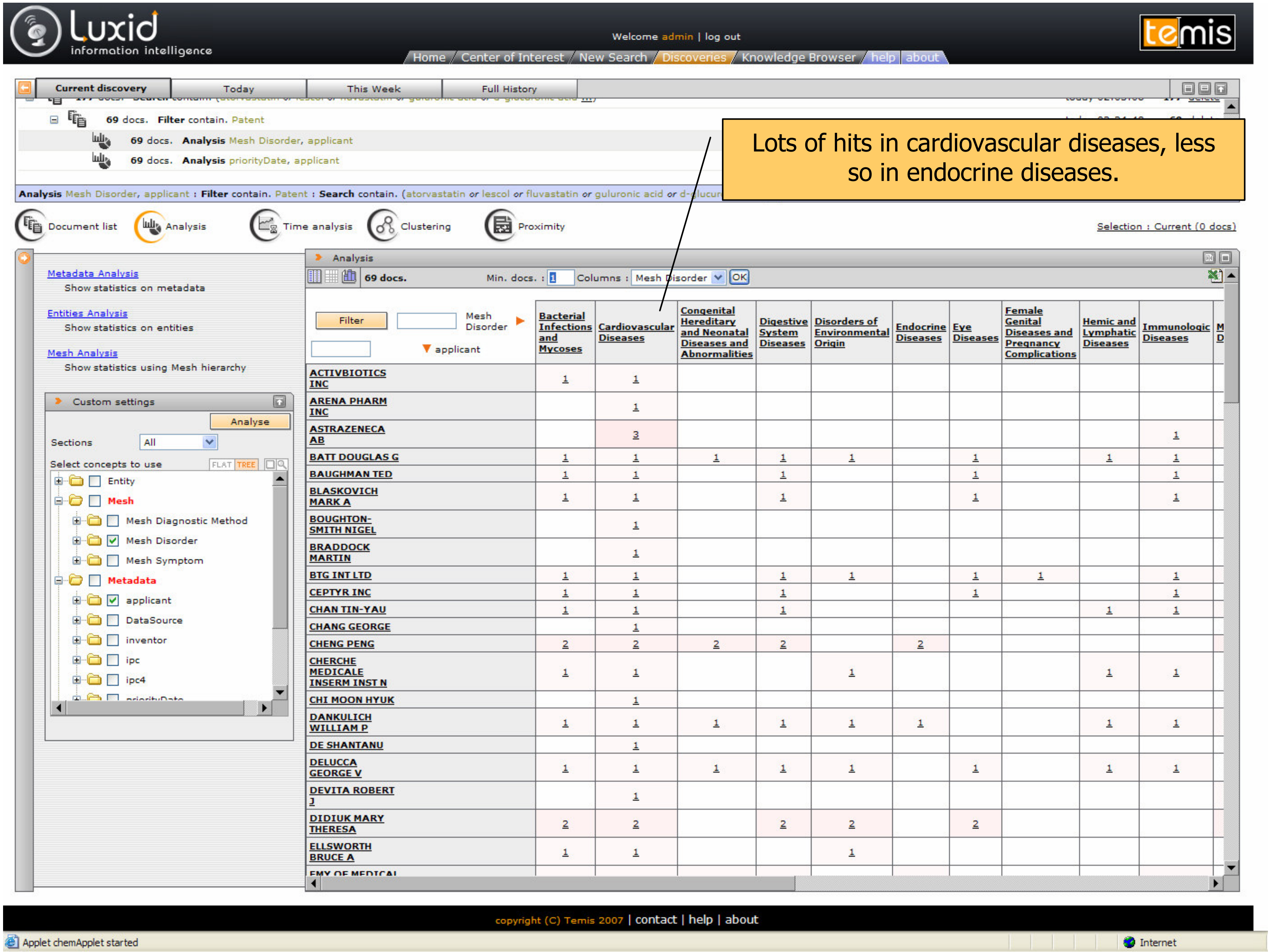
69 docs. Min. docs. : 2 Columns : priorityDate OK

priorityDate ▶ ▼ applicant	2004	2005	2006
ASTRAZENECA AB			2
CHENG PENG		2	
DIDIUK MARY THERESA		2	
GREGOIRE FRANCINE M		2	
MA JINGYUAN		2	
MERCK and CO INC	3	6	
METABOLEX INC		2	
PFIZER PROD INC		4	
RAKHMANOVA VERA A		2	
RANBAXY LAB LTD		2	
RUGGERI ROGER BENJAMIN		2	
SQUIBB BRISTOL MYERS CO		2	
ZHAO ZUCHUN		2	
ZHU YAN		2	

copyright (C) Temis 2007 | contact | help | about
Applet chemApplet started
Internet

... this time on applicant and disorders: Who is working on which therapeutic area?

Adjust layout parameters



69 docs. **Filter** contain. Patent

69 docs. **Analysis** Mesh Disorder, applicant

69 docs. **Analysis** priorityDate, applicant

Analysis Mesh Disorder, applicant : **Filter** contain. Patent : **Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucur

Metadata Analysis
Show statistics on metadata

Entities Analysis
Show statistics on entities

Mesh Analysis
Show statistics using Mesh hierarchy

Custom settings

Sections: All

Select concepts to use

- □ Entity
- □ Mesh
  - □ Mesh Diagnostic Method
  - ☑ Mesh Disorder
  - □ Mesh Symptom
- □ Metadata
  - ☑ applicant
  - □ DataSource
  - □ inventor
  - □ ipc
  - □ ipc4

Analysis

69 docs. Min. docs. : 1 Columns : Mesh Disorder OK

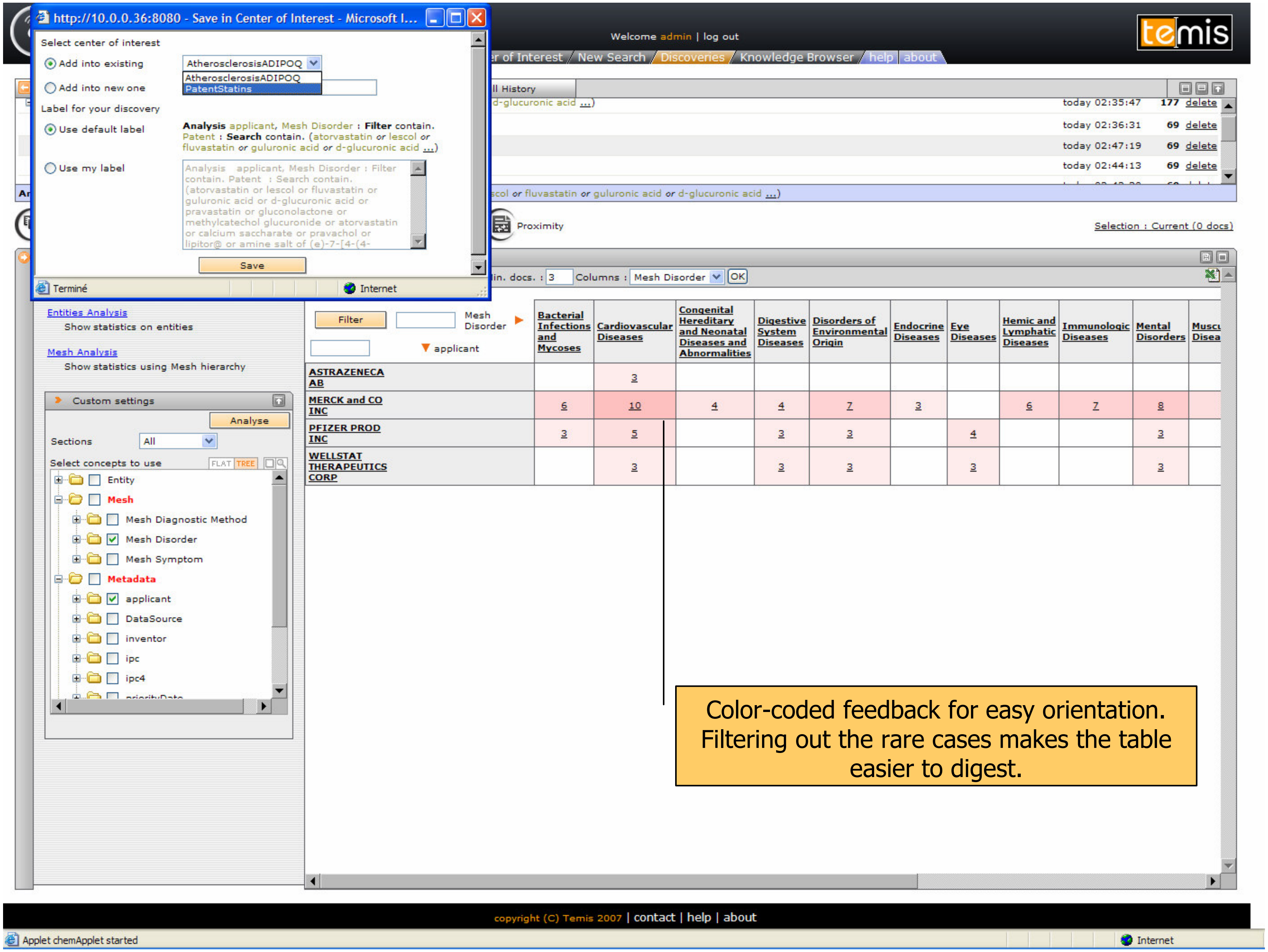
	Bacterial Infections and Mycoses	Cardiovascular Diseases	Congenital Hereditary and Neonatal Diseases and Abnormalities	Digestive System Diseases	Disorders of Environmental Origin	Endocrine Diseases	Eye Diseases	Female Genital Diseases and Pregnancy Complications	Hemic and Lymphatic Diseases	Immunologic Diseases
ACTIVBIOTICS INC	1	1								
ARENA PHARM INC		1								
ASTRAZENECA AB		3								1
BATT DOUGLAS G	1	1	1	1	1		1		1	1
BAUGHMAN TED	1	1		1			1			1
BLASKOVICH MARK A	1	1		1			1			1
BOUGHTON-SMITH NIGEL		1								
BRADDOCK MARTIN		1								
BTG INT LTD	1	1		1	1		1	1		1
CEPTYR INC	1	1		1			1			1
CHAN TIN-YAU	1	1		1					1	1
CHANG GEORGE		1								
CHENG PENG	2	2	2	2		2				
CHERCHE MEDICALE INSERM INST N	1	1			1				1	1
CHI MOON HYUK		1								
DANKULICH WILLIAM P	1	1	1	1	1	1			1	1
DE SHANTANU		1								
DELUCCA GEORGE V	1	1	1	1	1		1		1	1
DEVITA ROBERT J		1								
DIDIUK MARY THERESA	2	2		2	2		2			
ELLSWORTH BRUCE A	1	1			1					
EMY OF MEDICAL										

copyright (C) Temis 2007 | contact | help | about

Applet chemApplet started

Lots of hits in cardiovascular diseases, less so in endocrine diseases.

Save the dashboard again…



http://10.0.0.36:8080 - Save in Center of Interest - Microsoft I…

Select center of interest

● Add into existing: AtherosclerosisADIPOQ (options: AtherosclerosisADIPOQ, PatentStatins)

○ Add into new one

Label for your discovery

● Use default label: **Analysis** applicant, Mesh Disorder : **Filter** contain. Patent : **Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...)

○ Use my label: Analysis applicant, Mesh Disorder : Filter contain. Patent : Search contain. (atorvastatin or lescol or fluvastatin or guluronic acid or d-glucuronic acid or pravastatin or gluconolactone or methylcatechol glucuronide or atorvastatin or calcium saccharate or pravachol or lipitor@ or amine salt of (e)-7-[4-(4-

Save

Terminé — Internet

d-glucuronic acid ...)	today 02:35:47	177	delete
	today 02:36:31	69	delete
	today 02:47:19	69	delete
	today 02:44:13	69	delete

scol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...)

Entities Analysis
Show statistics on entities

Mesh Analysis
Show statistics using Mesh hierarchy

Custom settings

Analyse

Sections: All

Select concepts to use (FLAT | TREE)

- □ Entity
- □ Mesh
  - □ Mesh Diagnostic Method
  - ☑ Mesh Disorder
  - □ Mesh Symptom
- □ Metadata
  - ☑ applicant
  - □ DataSource
  - □ inventor
  - □ ipc
  - □ ipc4
  - □ priorityDate

Min. docs. : 3 Columns : Mesh Disorder OK

Filter — Mesh Disorder — applicant

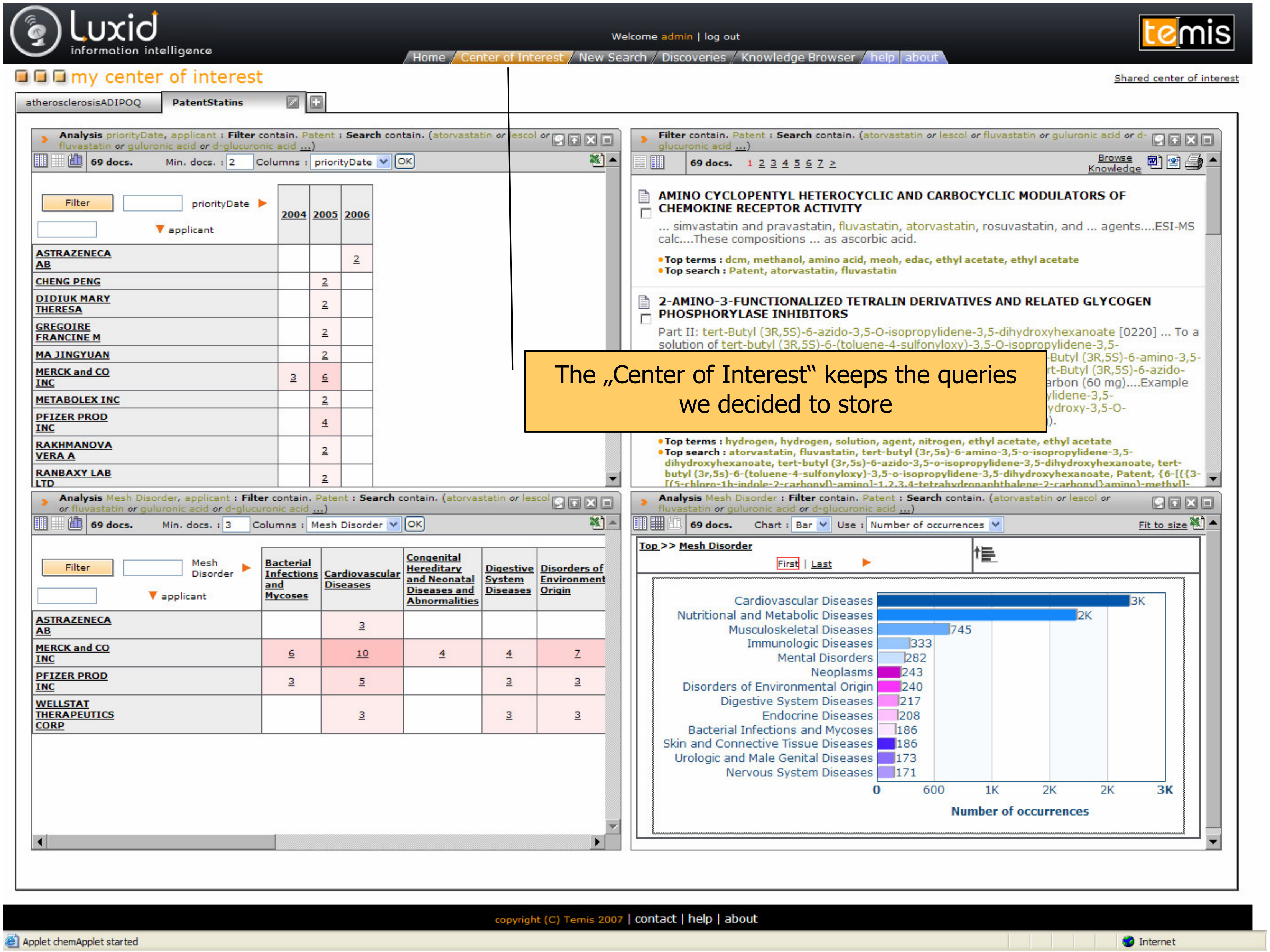
	Bacterial Infections and Mycoses	Cardiovascular Diseases	Congenital Hereditary and Neonatal Diseases and Abnormalities	Digestive System Diseases	Disorders of Environmental Origin	Endocrine Diseases	Eye Diseases	Hemic and Lymphatic Diseases	Immunologic Diseases	Mental Disorders	Muscu Disea
ASTRAZENECA AB		3									
MERCK and CO INC	6	10	4	4	7	3		6	7	8	
PFIZER PROD INC	3	5		3	3		4			3	
WELLSTAT THERAPEUTICS CORP		3		3	3		3			3	

copyright (C) Temis 2007 | contact | help | about

Applet chemApplet started — Internet

Color-coded feedback for easy orientation. Filtering out the rare cases makes the table easier to digest.

Instant access to stored information



Luxid
information intelligence
Welcome admin | log out
temis
Home / Center of Interest / New Search / Discoveries / Knowledge Browser / help / about

my center of interest

Shared center of interest

atherosclerosisADIPOQ | PatentStatins

Analysis priorityDate, applicant : **Filter** contain. Patent : **Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...)

69 docs. Min. docs. : 2 Columns : priorityDate OK

Filter priorityDate ▶ ▼ applicant

	2004	2005	2006
ASTRAZENECA AB			2
CHENG PENG		2	
DIDIUK MARY THERESA		2	
GREGOIRE FRANCINE M		2	
MA JINGYUAN		2	
MERCK and CO INC	3	6	
METABOLEX INC		2	
PFIZER PROD INC		4	
RAKHMANOVA VERA A		2	
RANBAXY LAB LTD		2	

Filter contain. Patent : **Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...)

69 docs. 1 2 3 4 5 6 7 > Browse Knowledge

AMINO CYCLOPENTYL HETEROCYCLIC AND CARBOCYCLIC MODULATORS OF CHEMOKINE RECEPTOR ACTIVITY

... simvastatin and pravastatin, fluvastatin, atorvastatin, rosuvastatin, and ... agents....ESI-MS calc....These compositions ... as ascorbic acid.

- **Top terms :** dcm, methanol, amino acid, meoh, edac, ethyl acetate, ethyl acetate
- **Top search :** Patent, atorvastatin, fluvastatin

2-AMINO-3-FUNCTIONALIZED TETRALIN DERIVATIVES AND RELATED GLYCOGEN PHOSPHORYLASE INHIBITORS

Part II: tert-Butyl (3R,5S)-6-azido-3,5-O-isopropylidene-3,5-dihydroxyhexanoate [0220] ... To a solution of tert-butyl (3R,5S)-6-(toluene-4-sulfonyloxy)-3,5-O-isopropylidene-3,5-
-Butyl (3R,5S)-6-amino-3,5-
rt-Butyl (3R,5S)-6-azido-
arbon (60 mg)....Example
ylidene-3,5-
ydroxy-3,5-O-
).

- **Top terms :** hydrogen, hydrogen, solution, agent, nitrogen, ethyl acetate, ethyl acetate
- **Top search :** atorvastatin, fluvastatin, tert-butyl (3r,5s)-6-amino-3,5-o-isopropylidene-3,5-dihydroxyhexanoate, tert-butyl (3r,5s)-6-azido-3,5-o-isopropylidene-3,5-dihydroxyhexanoate, tert-butyl (3r,5s)-6-(toluene-4-sulfonyloxy)-3,5-o-isopropylidene-3,5-dihydroxyhexanoate, Patent, {6-[({3-[(5-chloro-1h-indole-2-carbonyl)-amino]-1,2,3,4-tetrahydronaphthalene-2-carbonyl}amino)-methyl]-

Analysis Mesh Disorder, applicant : **Filter** contain. Patent : **Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...)

69 docs. Min. docs. : 3 Columns : Mesh Disorder OK

Filter Mesh Disorder ▶ ▼ applicant

	Bacterial Infections and Mycoses	Cardiovascular Diseases	Congenital Hereditary and Neonatal Diseases and Abnormalities	Digestive System Diseases	Disorders of Environment Origin
ASTRAZENECA AB		3			
MERCK and CO INC	6	10	4	4	7
PFIZER PROD INC	3	5		3	3
WELLSTAT THERAPEUTICS CORP		3		3	3

Analysis Mesh Disorder : **Filter** contain. Patent : **Search** contain. (atorvastatin *or* lescol *or* fluvastatin *or* guluronic acid *or* d-glucuronic acid ...)

69 docs. Chart : Bar Use : Number of occurrences Fit to size

Top >> Mesh Disorder

First | Last ▶

Cardiovascular Diseases 3K
Nutritional and Metabolic Diseases 2K
Musculoskeletal Diseases 745
Immunologic Diseases 333
Mental Disorders 282
Neoplasms 243
Disorders of Environmental Origin 240
Digestive System Diseases 217
Endocrine Diseases 208
Bacterial Infections and Mycoses 186
Skin and Connective Tissue Diseases 186
Urologic and Male Genital Diseases 173
Nervous System Diseases 171
0
600
1K
2K
2K
3K
Number of occurrences

copyright (C) Temis 2007 | contact | help | about

Applet chemApplet started

Internet

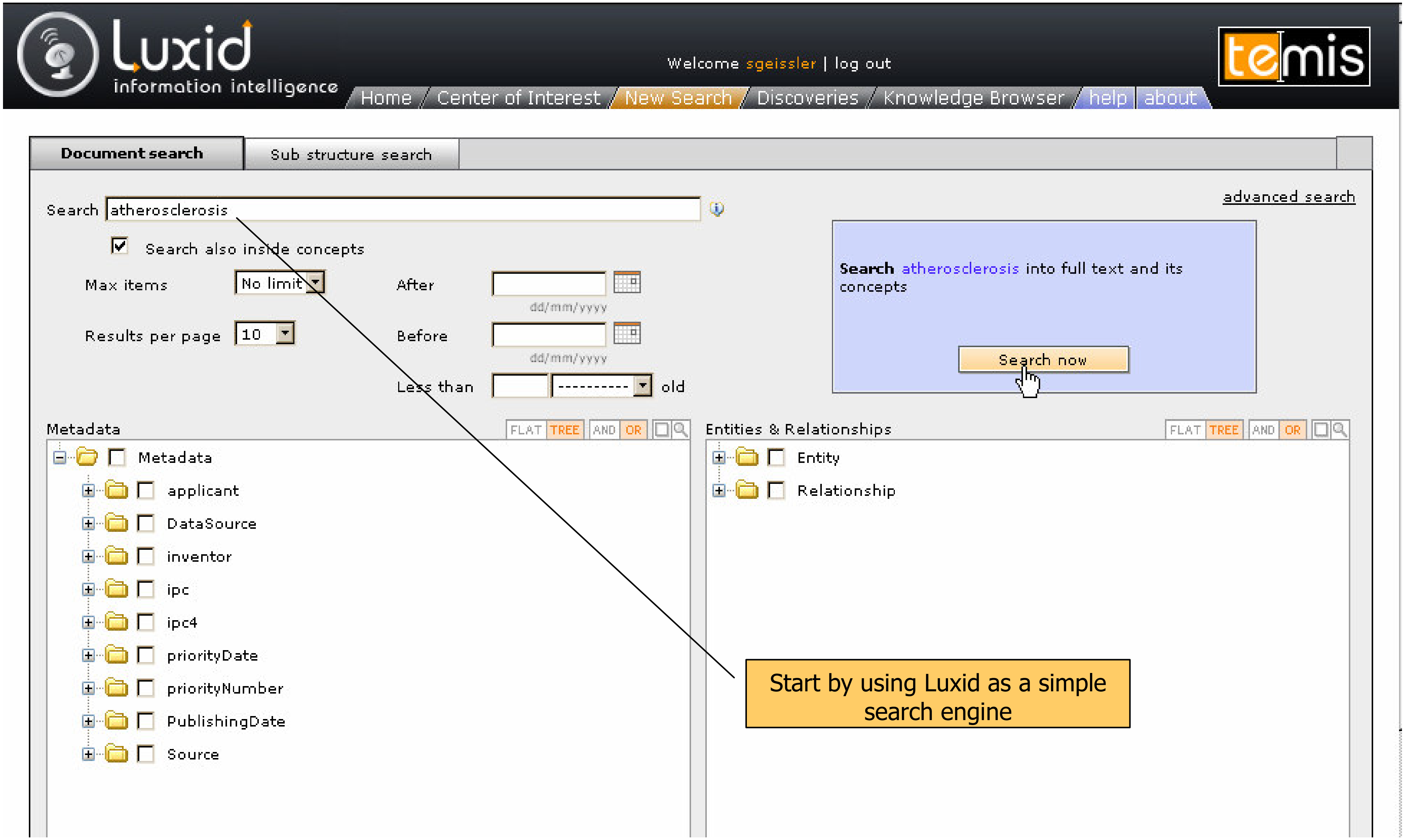
The „Center of Interest" keeps the queries we decided to store

Scenario #2

Explore potential targets

around a given disease

Search „atherosclerosis“



Search atherosclerosis

☑ Search also inside concepts

Max items: No limit

Results per page: 10

After: dd/mm/yyyy

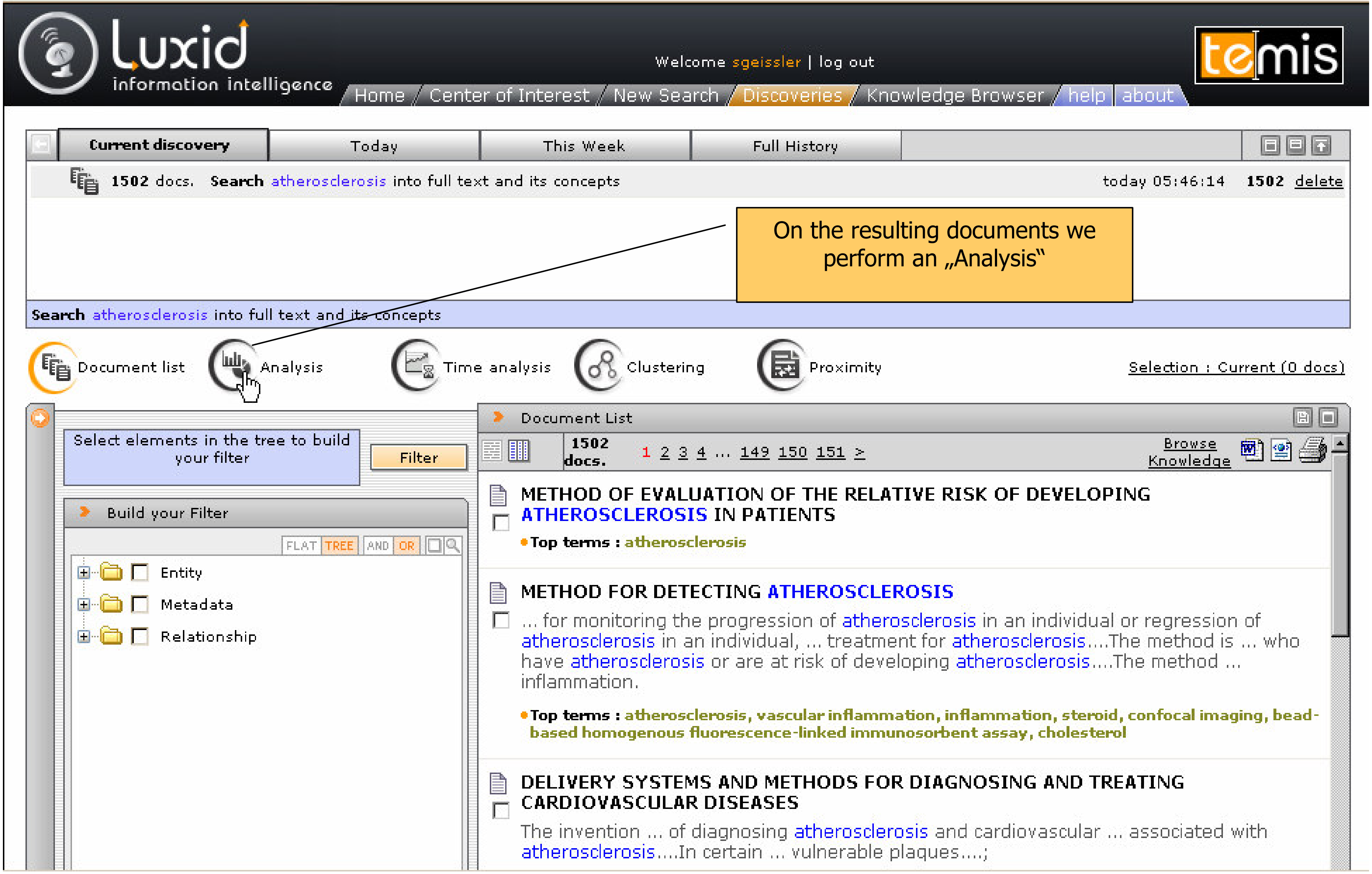
Before: dd/mm/yyyy

Less than ---------- old

Search atherosclerosis into full text and its concepts

Start by using Luxid as a simple search engine

Perform analysis on the result list



Welcome sgeissler | log out

Home / Center of Interest / New Search / Discoveries / Knowledge Browser / help / about

Current discovery | Today | This Week | Full History

1502 docs. **Search** atherosclerosis into full text and its concepts — today 05:46:14 — **1502** delete

Search atherosclerosis into full text and its concepts

Document list | Analysis | Time analysis | Clustering | Proximity — Selection : Current (0 docs)

Select elements in the tree to build your filter — Filter

Build your Filter

FLAT TREE AND OR

- □ Entity
- □ Metadata
- □ Relationship

Document List

1502 docs. 1 2 3 4 ... 149 150 151 > — Browse Knowledge

□ METHOD OF EVALUATION OF THE RELATIVE RISK OF DEVELOPING ATHEROSCLEROSIS IN PATIENTS

- **Top terms :** atherosclerosis

□ METHOD FOR DETECTING ATHEROSCLEROSIS

... for monitoring the progression of atherosclerosis in an individual or regression of atherosclerosis in an individual, ... treatment for atherosclerosis....The method is ... who have atherosclerosis or are at risk of developing atherosclerosis....The method ... inflammation.

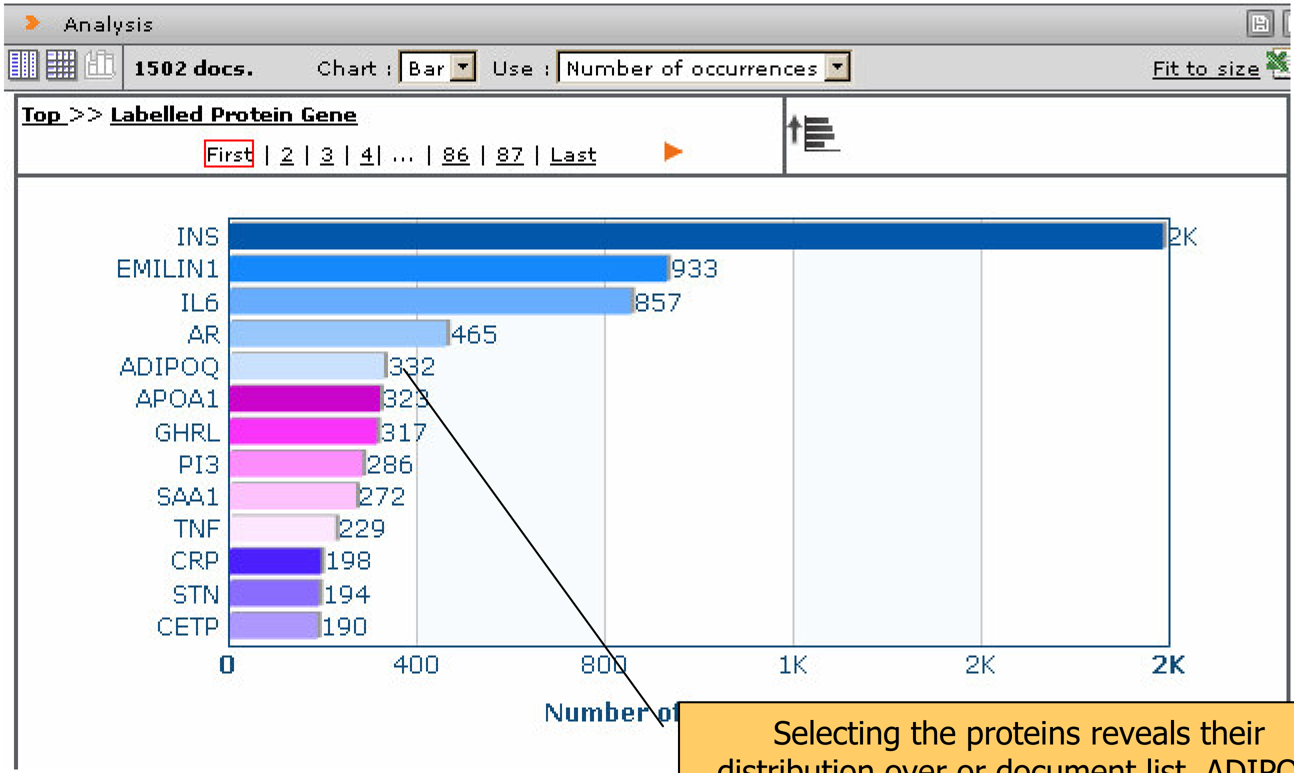
- **Top terms :** atherosclerosis, vascular inflammation, inflammation, steroid, confocal imaging, bead-based homogenous fluorescence-linked immunosorbent assay, cholesterol

□ DELIVERY SYSTEMS AND METHODS FOR DIAGNOSING AND TREATING CARDIOVASCULAR DISEASES

The invention ... of diagnosing atherosclerosis and cardiovascular ... associated with atherosclerosis....In certain ... vulnerable plaques....;

On the resulting documents we perform an „Analysis“

Inspect Protein names



Analysis

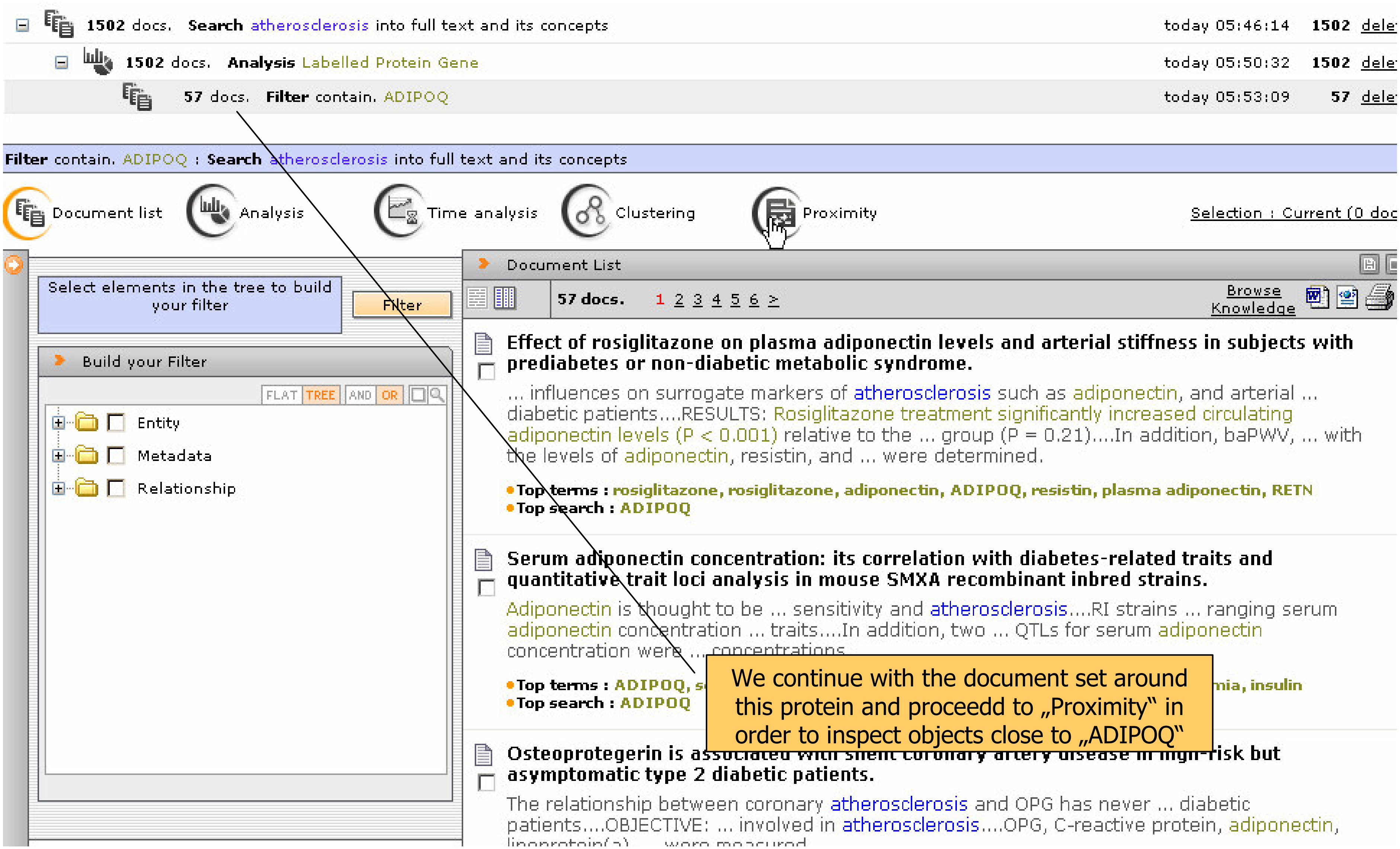
1502 docs. Chart : Bar Use : Number of occurrences Fit to size

Top >> Labelled Protein Gene

First | 2 | 3 | 4| ... | 86 | 87 | Last

Selecting the proteins reveals their distribution over or document list. ADIPOQ cathces our eye.

Let`s explore proteins that are mentioned together with ADIPOQ



1502 docs. **Search** atherosclerosis into full text and its concepts — today 05:46:14 — 1502 dele

1502 docs. **Analysis** Labelled Protein Gene — today 05:50:32 — 1502 dele

57 docs. **Filter** contain. ADIPOQ — today 05:53:09 — 57 dele

Filter contain. ADIPOQ : **Search** atherosclerosis into full text and its concepts

57 docs. 1 2 3 4 5 6 >

Effect of rosiglitazone on plasma adiponectin levels and arterial stiffness in subjects with prediabetes or non-diabetic metabolic syndrome.

... influences on surrogate markers of atherosclerosis such as adiponectin, and arterial ... diabetic patients....RESULTS: Rosiglitazone treatment significantly increased circulating adiponectin levels ($P < 0.001$) relative to the ... group ($P = 0.21$)....In addition, baPWV, ... with the levels of adiponectin, resistin, and ... were determined.

- **Top terms :** rosiglitazone, rosiglitazone, adiponectin, ADIPOQ, resistin, plasma adiponectin, RETN
- **Top search :** ADIPOQ

Serum adiponectin concentration: its correlation with diabetes-related traits and quantitative trait loci analysis in mouse SMXA recombinant inbred strains.

Adiponectin is thought to be ... sensitivity and atherosclerosis....RI strains ... ranging serum adiponectin concentration ... traits....In addition, two ... QTLs for serum adiponectin concentration were ... concentrations.

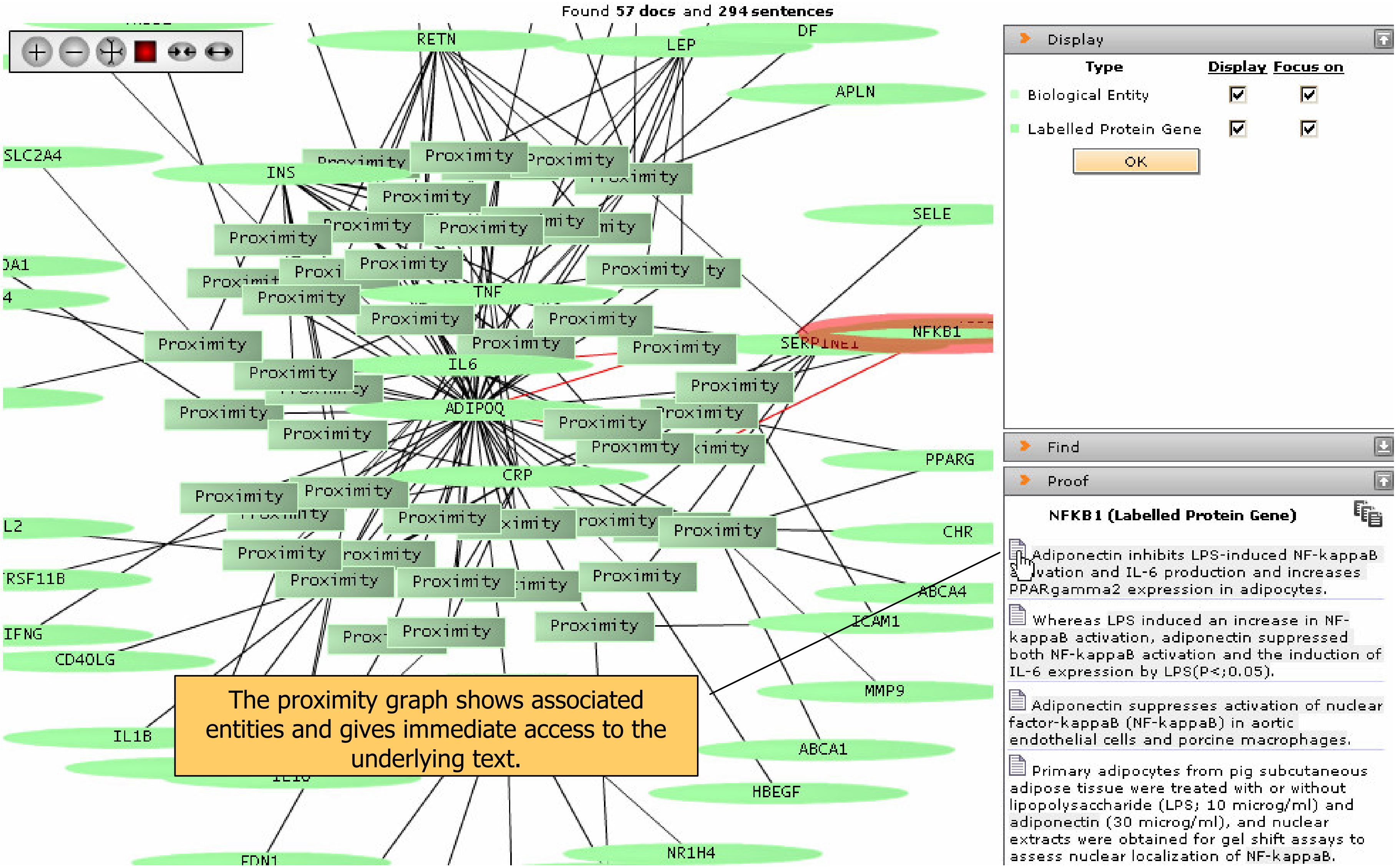
- **Top terms :** ADIPOQ, s... mia, insulin
- **Top search :** ADIPOQ

Osteoprotegerin is associated with silent coronary artery disease in high-risk but asymptomatic type 2 diabetic patients.

The relationship between coronary atherosclerosis and OPG has never ... diabetic patients....OBJECTIVE: ... involved in atherosclerosis....OPG, C-reactive protein, adiponectin, lipoprotein(a) ... were measured

We continue with the document set around this protein and proceedd to „Proximity" in order to inspect objects close to „ADIPOQ"

Proximity graph reveals associations



Found **57 docs** and **294 sentences**

Type	Display	Focus on
Biological Entity	☑	☑
Labelled Protein Gene	☑	☑

NFKB1 (Labelled Protein Gene)

Adiponectin inhibits LPS-induced NF-kappaB activation and IL-6 production and increases PPARgamma2 expression in adipocytes.

Whereas LPS induced an increase in NF-kappaB activation, adiponectin suppressed both NF-kappaB activation and the induction of IL-6 expression by LPS(P<;0.05).

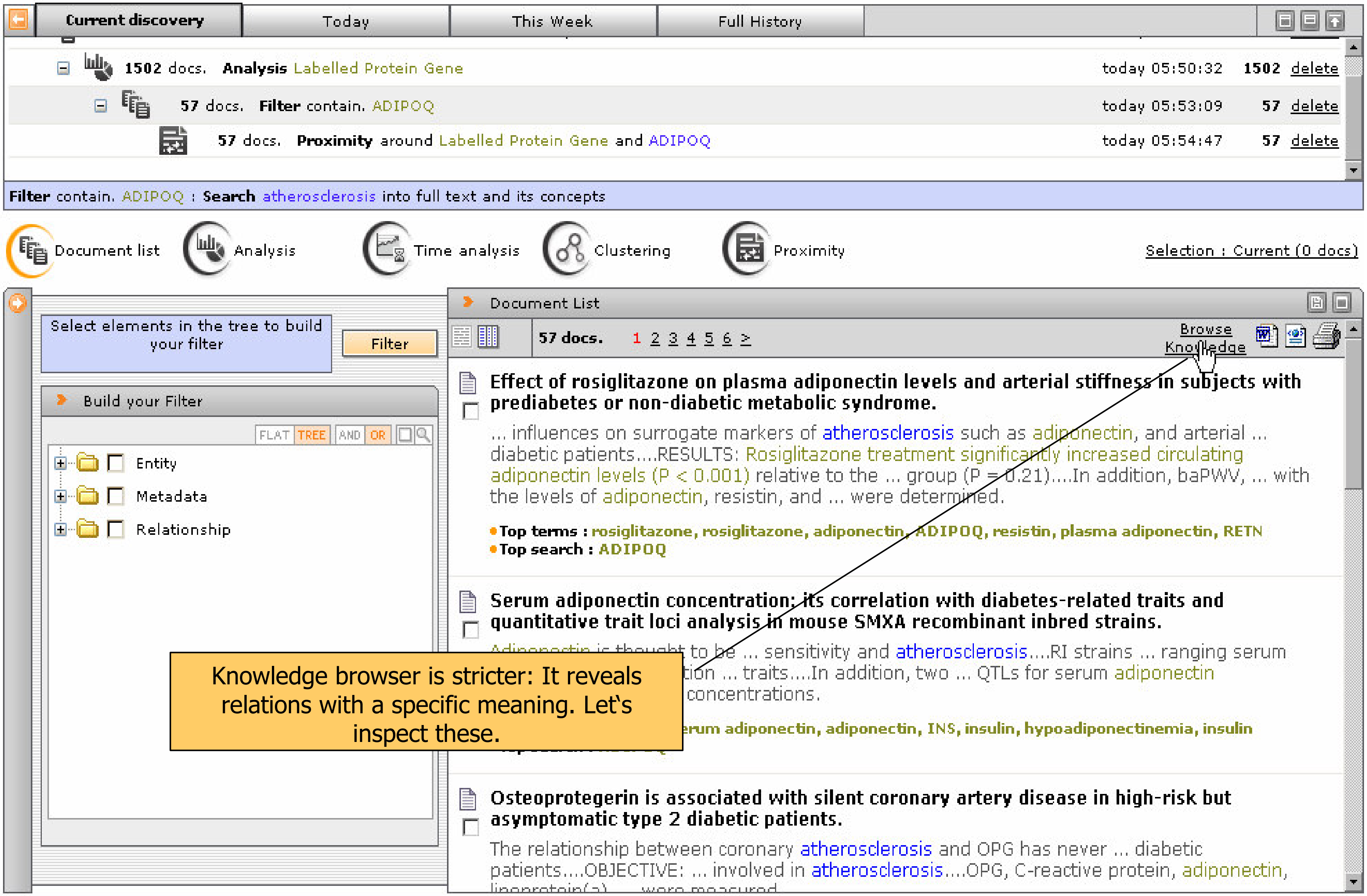
Adiponectin suppresses activation of nuclear factor-kappaB (NF-kappaB) in aortic endothelial cells and porcine macrophages.

Primary adipocytes from pig subcutaneous adipose tissue were treated with or without lipopolysaccharide (LPS; 10 microg/ml) and adiponectin (30 microg/ml), and nuclear extracts were obtained for gel shift assays to assess nuclear localization of NF-kappaB.

The proximity graph shows associated entities and gives immediate access to the underlying text.

One step further: Typed relations

temis



1502 docs. **Analysis** Labelled Protein Gene	today 05:50:32	1502	delete
57 docs. **Filter** contain. ADIPOQ	today 05:53:09	57	delete
57 docs. **Proximity** around Labelled Protein Gene and ADIPOQ	today 05:54:47	57	delete

Filter contain. ADIPOQ : **Search** atherosclerosis into full text and its concepts

Effect of rosiglitazone on plasma adiponectin levels and arterial stiffness in subjects with prediabetes or non-diabetic metabolic syndrome.

... influences on surrogate markers of atherosclerosis such as adiponectin, and arterial ... diabetic patients....RESULTS: Rosiglitazone treatment significantly increased circulating adiponectin levels (P < 0.001) relative to the ... group (P = 0.21)....In addition, baPWV, ... with the levels of adiponectin, resistin, and ... were determined.

- **Top terms : rosiglitazone, rosiglitazone, adiponectin, ADIPOQ, resistin, plasma adiponectin, RETN**
- **Top search : ADIPOQ**

Serum adiponectin concentration: its correlation with diabetes-related traits and quantitative trait loci analysis in mouse SMXA recombinant inbred strains.

Adiponectin is thought to be ... sensitivity and atherosclerosis....RI strains ... ranging serum ...tion ... traits....In addition, two ... QTLs for serum adiponectin concentrations.

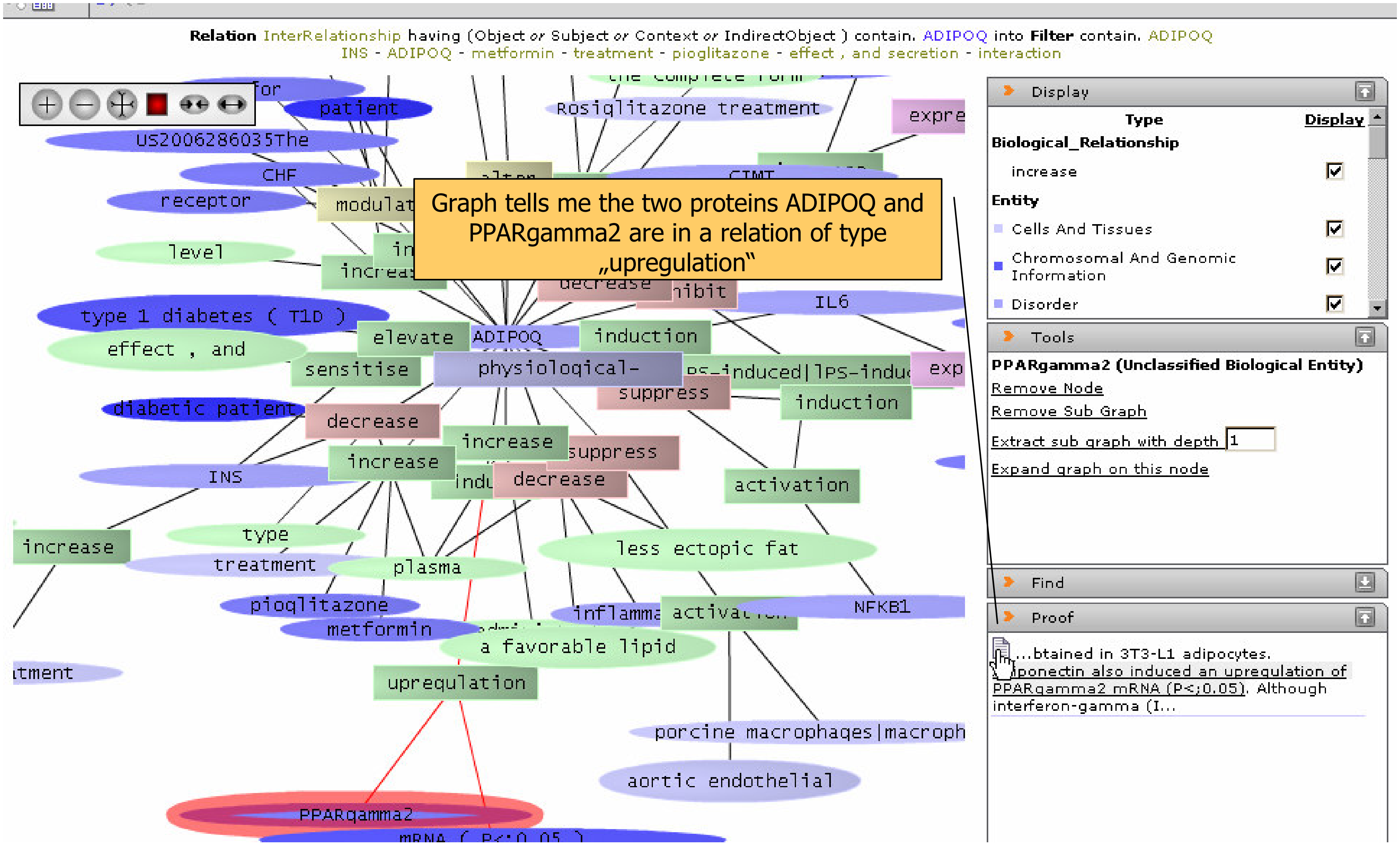
...erum adiponectin, adiponectin, INS, insulin, hypoadiponectinemia, insulin

Osteoprotegerin is associated with silent coronary artery disease in high-risk but asymptomatic type 2 diabetic patients.

The relationship between coronary atherosclerosis and OPG has never ... diabetic patients....OBJECTIVE: ... involved in atherosclerosis....OPG, C-reactive protein, adiponectin, lipoprotein(a) ... were measured

Knowledge browser is stricter: It reveals relations with a specific meaning. Let's inspect these.

The Knowledge Browser



Relation InterRelationship having (Object *or* Subject *or* Context *or* IndirectObject) contain. ADIPOQ into **Filter** contain. ADIPOQ
INS - ADIPOQ - metformin - treatment - pioglitazone - effect , and secretion - interaction

Graph tells me the two proteins ADIPOQ and PPARgamma2 are in a relation of type „upregulation"

Display

Type	Display
Biological_Relationship	
increase	☑
Entity	
Cells And Tissues	☑
Chromosomal And Genomic Information	☑
Disorder	☑

Tools

PPARgamma2 (Unclassified Biological Entity)

Remove Node

Remove Sub Graph

Extract sub graph with depth 1

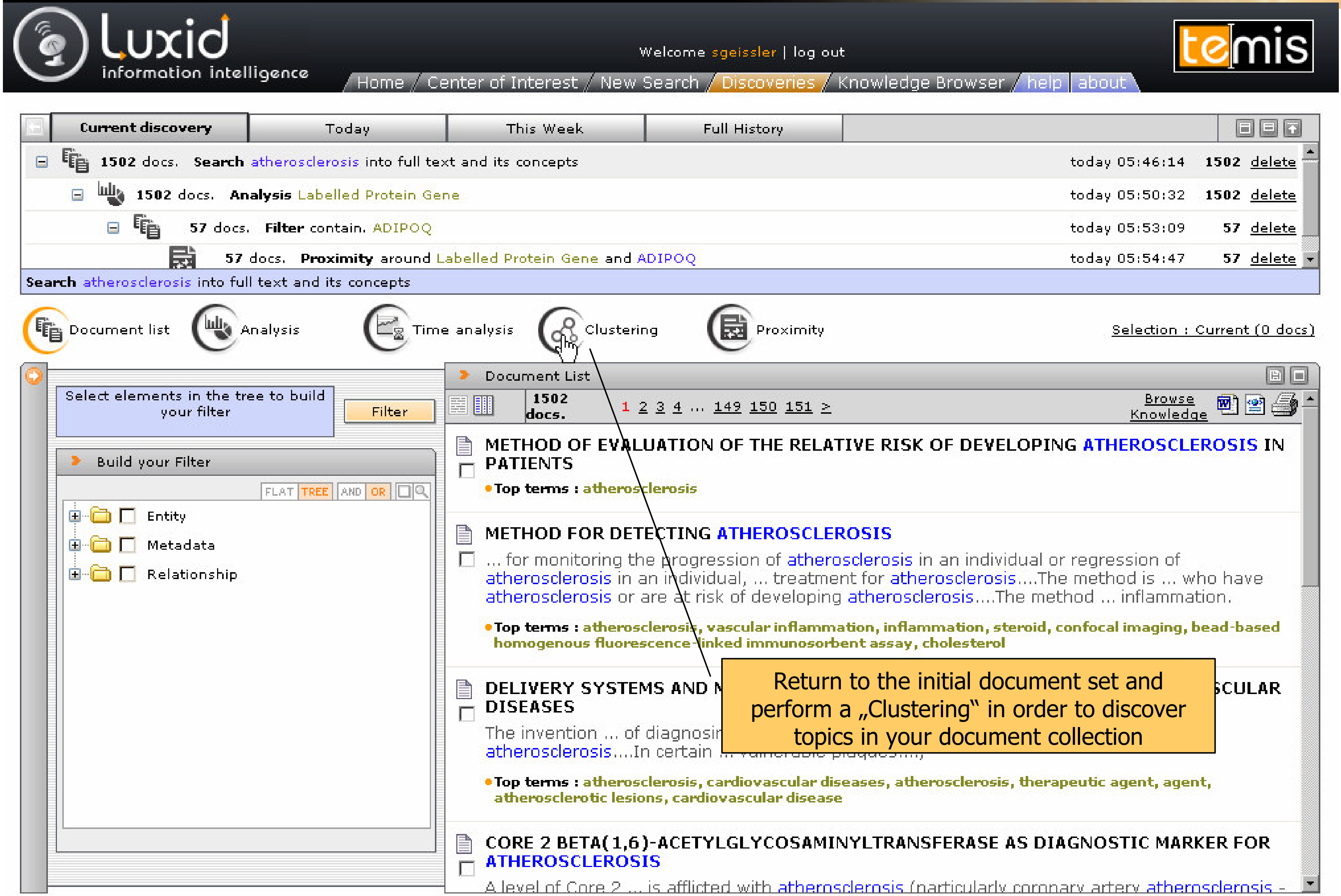
Expand graph on this node

Find

Proof

...btained in 3T3-L1 adipocytes. Adiponectin also induced an upregulation of PPARgamma2 mRNA (P<;0.05). Although interferon-gamma (I...

Clustering



Welcome sgeissler | log out

1502 docs. **Search** atherosclerosis into full text and its concepts	today 05:46:14	1502 delete
1502 docs. **Analysis** Labelled Protein Gene	today 05:50:32	1502 delete
57 docs. **Filter** contain. ADIPOQ	today 05:53:09	57 delete
57 docs. **Proximity** around Labelled Protein Gene and ADIPOQ	today 05:54:47	57 delete

Search atherosclerosis into full text and its concepts

Selection : Current (0 docs)

Select elements in the tree to build your filter

- ⊞ □ Entity
- ⊞ □ Metadata
- ⊞ □ Relationship

1502 docs.

□ METHOD OF EVALUATION OF THE RELATIVE RISK OF DEVELOPING ATHEROSCLEROSIS IN PATIENTS

- **Top terms :** atherosclerosis

□ METHOD FOR DETECTING ATHEROSCLEROSIS

... for monitoring the progression of atherosclerosis in an individual or regression of atherosclerosis in an individual, ... treatment for atherosclerosis....The method is ... who have atherosclerosis or are at risk of developing atherosclerosis....The method ... inflammation.

- **Top terms :** atherosclerosis, vascular inflammation, inflammation, steroid, confocal imaging, bead-based homogenous fluorescence-linked immunosorbent assay, cholesterol

□ DELIVERY SYSTEMS AND M... SCULAR DISEASES

The invention ... of diagnosi... atherosclerosis....In certain ... vulnerable plaques....)

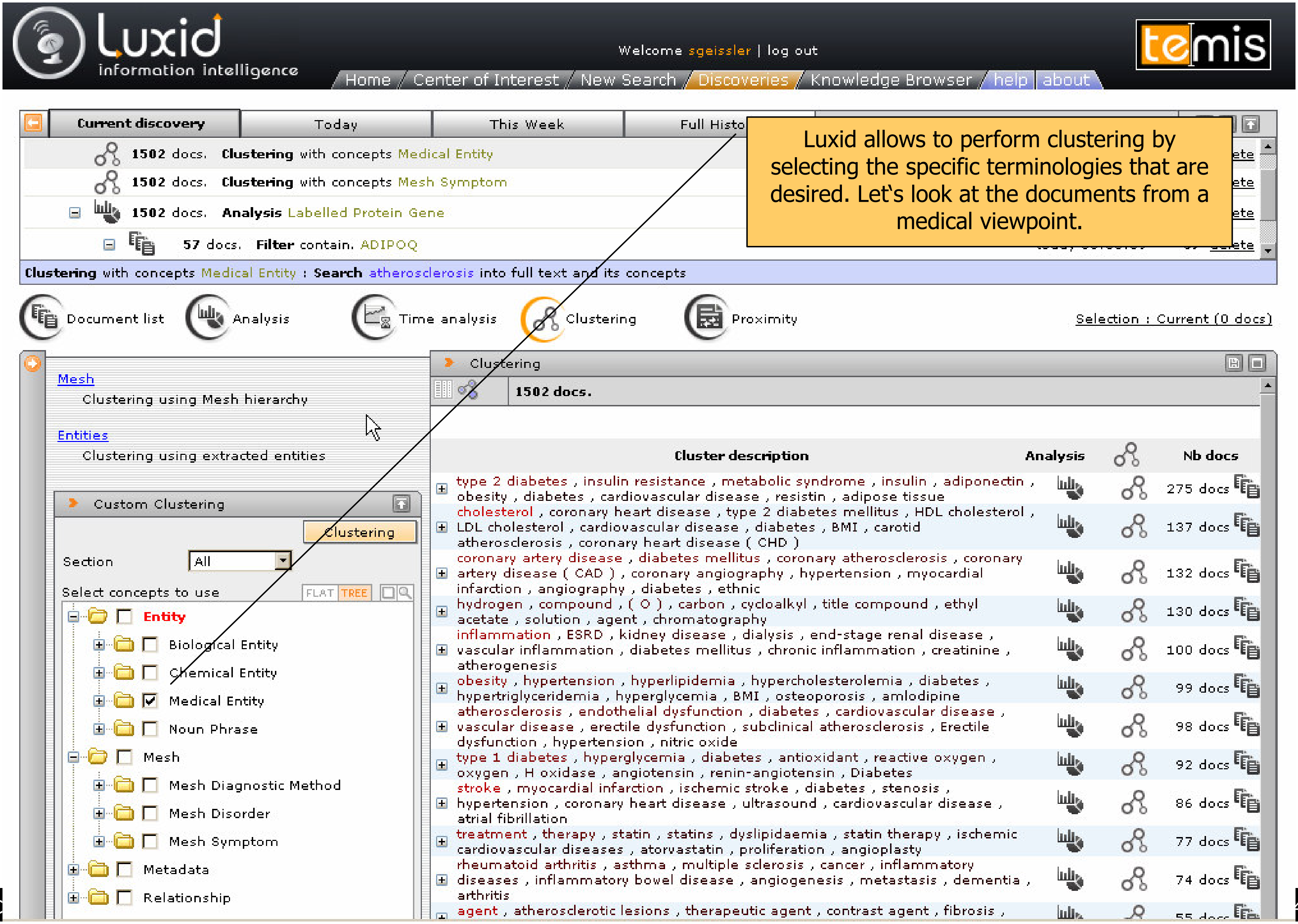
- **Top terms :** atherosclerosis, cardiovascular diseases, atherosclerosis, therapeutic agent, agent, atherosclerotic lesions, cardiovascular disease

□ CORE 2 BETA(1,6)-ACETYLGLYCOSAMINYLTRANSFERASE AS DIAGNOSTIC MARKER FOR ATHEROSCLEROSIS

A level of Core 2 ... is afflicted with atherosclerosis (particularly coronary artery atherosclerosis -

Return to the initial document set and perform a „Clustering" in order to discover topics in your document collection

Clustering



Luxid allows to perform clustering by selecting the specific terminologies that are desired. Let's look at the documents from a medical viewpoint.

1502 docs. **Clustering** with concepts Medical Entity

1502 docs. **Clustering** with concepts Mesh Symptom

1502 docs. **Analysis** Labelled Protein Gene

57 docs. **Filter** contain. ADIPOQ

Clustering with concepts Medical Entity : **Search** atherosclerosis into full text and its concepts

Mesh
Clustering using Mesh hierarchy

Entities
Clustering using extracted entities

Section: All

Select concepts to use

- □ Entity
  - □ Biological Entity
  - □ Chemical Entity
  - ☑ Medical Entity
  - □ Noun Phrase
- □ Mesh
  - □ Mesh Diagnostic Method
  - □ Mesh Disorder
  - □ Mesh Symptom
- □ Metadata
- □ Relationship

1502 docs.

Cluster description	Analysis		Nb docs
type 2 diabetes , insulin resistance , metabolic syndrome , insulin , adiponectin , obesity , diabetes , cardiovascular disease , resistin , adipose tissue			275 docs
cholesterol , coronary heart disease , type 2 diabetes mellitus , HDL cholesterol , LDL cholesterol , cardiovascular disease , diabetes , BMI , carotid atherosclerosis , coronary heart disease (CHD)			137 docs
coronary artery disease , diabetes mellitus , coronary atherosclerosis , coronary artery disease (CAD) , coronary angiography , hypertension , myocardial infarction , angiography , diabetes , ethnic			132 docs
hydrogen , compound , (O) , carbon , cycloalkyl , title compound , ethyl acetate , solution , agent , chromatography			130 docs
inflammation , ESRD , kidney disease , dialysis , end-stage renal disease , vascular inflammation , diabetes mellitus , chronic inflammation , creatinine , atherogenesis			100 docs
obesity , hypertension , hyperlipidemia , hypercholesterolemia , diabetes , hypertriglyceridemia , hyperglycemia , BMI , osteoporosis , amlodipine			99 docs
atherosclerosis , endothelial dysfunction , diabetes , cardiovascular disease , vascular disease , erectile dysfunction , subclinical atherosclerosis , Erectile dysfunction , hypertension , nitric oxide			98 docs
type 1 diabetes , hyperglycemia , diabetes , antioxidant , reactive oxygen , oxygen , H oxidase , angiotensin , renin-angiotensin , Diabetes			92 docs
stroke , myocardial infarction , ischemic stroke , diabetes , stenosis , hypertension , coronary heart disease , ultrasound , cardiovascular disease , atrial fibrillation			86 docs
treatment , therapy , statin , statins , dyslipidaemia , statin therapy , ischemic cardiovascular diseases , atorvastatin , proliferation , angioplasty			77 docs
rheumatoid arthritis , asthma , multiple sclerosis , cancer , inflammatory diseases , inflammatory bowel disease , angiogenesis , metastasis , dementia , arthritis			74 docs
agent , atherosclerotic lesions , therapeutic agent , contrast agent , fibrosis ,			55 docs

Agenda

1 **Welcome & Introduction**

Stefan Geißler - Guillaume Mazieres

2 **Luxid® for Life Sciences – Presentation**

Guillaume Mazieres - 5 ’

3 **Patent Analysis Scenarios with Luxid® for Life Sciences**

A. Patent landscape around a class of chemical substance
B. Explore potential targets around a therapeutic area

Stefan Geißler – 40 ’

4 **Q&A session**

All – 5 ’

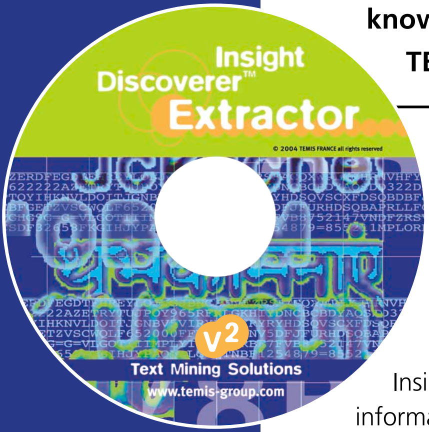
Molte Grazie!

Thank You For Your Time!

Detect strategic information with Insight Discoverer™ Extractor v2

Hunting for relevant information...

Companies today have to handle a growing volume of documents resulting from the increase in electronic communication. Because only those companies who know how to use this information will remain the leaders of tomorrow, TEMIS offers a new generation of text information processing tools.



Insight Discoverer™ Extractor, the information extraction solution

Insight Discoverer™ Extractor is an information extraction server dedicated to analysis of text documents. It detects pieces of information that are the most relevant to users: a merger announcement in an article for an analyst, a sales opportunity in an e-mail for a customer relationship manager, a specific skill in a resumé for a recruiter, for example.

Key applications of extraction

Used with specialized Skill Cartridges™, Insight Discoverer™ Extractor answers the needs for documentary analysis for departments as different as customer services, market watch units or human resources.

➜ A "CRM" Skill Cartridge™ to improve your customer understanding*.

It detects consumer satisfaction or discontent with a given product by analyzing customer e-mails, call center transcripts, discussion forums or even answers to open-ended questions in opinion surveys.

*the CRM Skill Cartridge™ is tailor-made for your business needs

➜ A "Competitive Intelligence" Skill Cartridge™ to facilitate strategic analysis.

It retrieves financial information (sales figures, profitability, growth), sales information (market shares, number of customers), stock market information (capitalization, trends). It also extracts all information concerning stock purchases, mergers, acquisitions, joint ventures, research areas and innovations by analyzing content from newswires, competitor Web sites, analysts' reports, scientific publications or patents.

➜ An "HR" Skill Cartridge™ to help you manage recruitment.

It identifies skills and know-how from applications (covering letters and resumés), and recruitment criteria from job offers.

Why choose Insight Discoverer™ Extractor?

Insight Discoverer™ Extractor offers an extended range of functions:

- Automatic identification of the document language
- Multilingual (12 languages available)
- Supports 50 formats (MS Word, MS Excel, MS PowerPoint, PDF, HTML, etc.)
- Compatible with Pre and Post Processing
- API provided for easy integration into your information system

Fine-tuning options...

Insight Discoverer™ Extractor is able to manage language complexity thanks to the flexibility of the Skill Cartridges™. To optimize extraction and obtain accurate information, Skill Cartridges™ can be fully customized with fine-tuning options such as:

- Extraction of negative or positive trends
- Differentiation between rumors and actions
- Enhancement of anaphora identification
- Resolution of acronyms
- Integration of ontologies

Text Intelligence™

TEMIS

For more information:

Contact

TEMIS France

>> Headquarters

Tour Gamma B
193-197 rue de Bercy
75582 Paris Cedex 12
France
Tel.: +33 (0)1 40 04 46 70
info@temis-group.com

>> Grenoble

Tel.: +33 (0)4 56 38 24 00
info@temis-group.com

TEMIS Germany

>> Heidelberg

Tel.: +49 - 6221 13753-12
info.de@temis-group.com

TEMIS USA

>> Philadelphia

Tel.: +1 571 235 83 95
info.us@temis-group.com

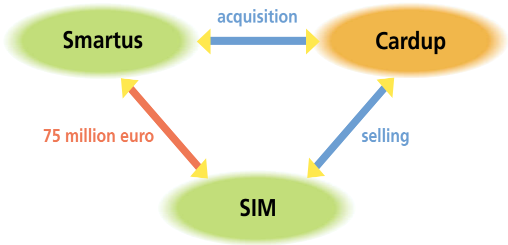
TEMIS Italy

>> Modena

Tel.: +39 - 059 237 634
info.it@temis-group.com

www.temis.com

Text Intelligence™



How Insight Discoverer™ Extractor works

Insight Discoverer™ Extractor performs a sequence of three linguistic analysis steps:

>> Corpus recognition: automatic language identification

>> Morpho-syntactic analysis:

- **Assigns** a grammatical category to each word in a document (noun, adjective, verb, etc.) as well as its morpho-syntactic characteristics (gender, number)
- **Lemmatization:** returns each word to its base form (singular for a plural, infinitive for a conjugated verb) so that it can be recognized independently of its inflected form

>> Knowledge extraction (runs extraction rules):

- **Recognition** of entities (name of companies, associations, organizations, products figures, dates, places, etc.)
- **Identification** of relationships between the entities (company-company, person-company, company-product, etc.)

The knowledge extraction is powered by Skill Cartridges™.

A **Skill Cartridge™** is a hierarchy of knowledge components describing the information to extract for a given business, specific field or topic.

A **knowledge component** is a lexicon and/or an extraction rule.

An **extraction rule** describes a sentence structure that characterizes a concept.

SMARTUS :

On Thursday, the Franco-German manufacturer announced it was to acquire the Cardup smart card manufacturer from Italian company SIM for 75 million euro in cash.

Who (guessed)	Smartus
Announcement	announced
Date	Thursday
Relation	acquisition
Whom (guessed)	Cardup
From whom (guessed)	SIM
Country	Italy

Discover the TEMIS product range

Insight Discoverer™ Extractor uses the XeLDA® engine's linguistic analysis technology and Skill Cartridges™. Its results can be used by the document portal solution Online Miner™, the information organization solution Insight Discoverer™ Clusterer and the information classification solution Insight Discoverer™ Categorizer.

NEW...

in Insight Discoverer™ Extractor V2

- Linux version available
- Extraction processing time cut by up to 50%
- New Skill Cartridge™ Compiler: 6 times faster than previous version and lightweight Skill Cartridges™ (size reduced by more than 90%)
- Easy conversion of existing Skill Cartridges™ to Insight Discoverer™ Extractor V2 compatible Skill Cartridges™ with SCtranslate
- More flexibility thanks to improved Skill Cartridge™ syntax
- Built-in linguistic engine XeLDA® upgraded to V2.5 (enhanced English and German processing)
- P3 architecture enables the integration of **P**re and **P**ost **P**rocessing in Skill Cartridges™. Pre Processing can be used to normalize input documents and Post Processing can be used to filter and rename concepts after an extraction process, for example.

This new version remains compatible with the previous one.

Specifications

>> Operating systems:

- Windows NT, 2000, XP workstation or server versions
- Linux

>> API:

- Java (RMI - Remote Method Invocation)

>> Source languages:

English, French, German, Italian, Dutch, Spanish, Portuguese, Czech, Greek, Hungarian, Polish, Russian.

>> Formats :

over 50 input formats (including MS Word, PDF and HTML).

Insight Discoverer™ Extractor V2

INTER PUBLI GRAPHIQUE 01 46 21 14 80

white paper

INSIGHT DISCOVERER™ EXTRACTOR

1. Overview

With the steady growth of business and scientific activities and the recent advances in Information Technology, huge amounts of electronically available but unstructured data have to be dealt with. New tools able to analyse and structure textual data need to be developed so that non-expert users can understand and evaluate the contents of their documents. In this context, a successful information extraction technology has a central role to play.
Information Extraction is the process of identifying relevant information where the criteria for relevance are predefined in the form of a template that is to be filled. The template pertains to actions between different actors related to events or situations and contains slots that denote who did what to whom, when and where, and possibly why. The template builder has to predict what information will be of interest to the customer according to his field of activity.

In an industrial context, the aim of Information Extraction is to build an extraction model adapted to the customer's application (its information assets: databases, documents and all the data issued by or received in a company). It requires the creation and validation of terminology resources specific to the described field as well as the definition of selection criteria using extraction rules.

2. Methodology

Our approach is based on discovering knowledge from a corpus in an iterative way. We use various text analysis processing: morpho-syntactic analysis, named entities recognition, pattern recognition using linguistics and/or semantic labels.
The main idea is to build patterns[1]. If, for instance, the aim of the application is to detect company mergers or acquisitions in press releases, it will seek expressions like '*company A* acquired *company B*' or '*company A* merged with *company C*'.
As *company* may be a company name and all the common words which refer to a company (firm, manufacturer, corporation, …). They are gathered under the semantic descriptor '*company*.' In the case of an application dedicated to the field of Competitive Intelligence, verbs like *buy*, *sell*, and *acquire*, which refer to a transfer of possessions, are coded together under the same descriptor.

In press releases, the economic actors are mentioned within their context. It is therefore necessary to associate specific terms under a common descriptor. This task is achieved through the use of extraction rules. An extraction rule is expressed by a regular expression which may refer to a lemma (canonical form), syntactic, or semantic label.

Example: "the first private label pasta maker" will be caught by the rule:

```
(company_Adj|Loc_Adj|#ORD)*/ (brand_product) / (Company)
```

Company_Adj refers to adjectives such as "leading", "industrial" or "public"
LocAdj refers to adjectives describing places such as "European" or "Italian"
#ORD refers to ordinals (first, second, …)
brand_product private label, branded + #NOUN
/ symbol separates two words
symbol denotes a syntactic tag (differentiates the syntactic labels from the semantic ones).

An expression like '*company A* acquired *company B*' will be extracted by the following rule:

Text Intelligence™

[1] A pattern is a regular expression whose purpose is to identify relevant phrases within a context. Combining an action to a pattern, a rule will add information to a word sequence, assigning a name concept which can be used by other rules.

```
company / possession_transfer / company
```

At this stage, it is possible to identify actors such as "who and which_company," which refer to the subject actor and the object actor respectively, taking the syntactic information into account (active vs. passive voice, modals, auxiliaries, and negation).

```
{company:who}   /   (HAVE|MODAL)*   /   possession_transfer   /{company
:which_company}
```

This rule will catch Company A would acquire Company B, Company A acquired Company B, Company A has bought Company B but Company B has not been acquired by Company A.

The following rule helps extract the roles filled by "who and which_company" (who buys whom).

company_acquisition	Spigadoro , Inc. Acquires Largest European Private Label Pasta Company
who	Spigadoro , Inc. [883,897]
which_company	Largest European Private Label Pasta Company

3. Processing

The extraction server proceeds in two stages:

1. the morpho-syntactic analysis – each entry is assigned a part of speech and a morphologic feature
2. the application of extraction rules.

3.1 Morpho-syntactic analysis

It gives the lemma and the morpho-syntactic label for each entry. It is coupled with a statistic model (Hidden Markov Model) , which helps in dealing with ambiguities.

3.2 Semantic analysis

Semantic tagging is performed with the following resources:

a) dictionaries, using existing resources (such as WordNet [FELDBAUM 1997, 1999] for English) and tailor-made dictionaries created by TEMIS' team of linguists

Brand dictionary	***Action dictionary***
<brandname_>	<communication_>
<Mineral_water>	<announcement_>
Arrowhead	announce
Badoit	announcement
Blue / Quellen	notify
Buxton	notification
Calistoga	proclamation
Contrex	promulgation
Vittel	statement
(Source)? / Perrier	talk
Poland / Spring	…
Quézac	
Salvetat	
….	

b) a set of contextual rules, which associate lemmas, syntax, and semantics.

```
<actor_in_agribusiness>

        ;; a seed and flour miller
(company_Adj|Loc_Adj|#ORD)* / (NOUN)* / (food_|brand_product) / (actor_|company)

       ;; maker of private label pasta
(company_Adj|Loc_Adj|#ORD)* / (actor_|company) / of / (food_|brand_product
```

Semantic labeling vs. morpho-syntactic labeling

A leading manufacturer of branded products in the Mediterranean food sector, today announced that it will acquire Pastificio Gazzola S.p.A., leading manufacturer of private label pasta.

		SEMANTIC LABELLING	MORPHO-SYNTACTIC LABELLING
A	<AT>	<a>	
leading	<NN>	<leading>	<company_Adj>
manufacturer	<NN>	<manufacturer>	<company>
of	<IN>	<of>	
branded	<VBN>	<brand>	
products	<NNS>	<product>	<brand_product>
in	<IN>	<in>	
the	<AT>	<the>	
Mediterranean	<JJ>	<Mediterranean>	<loc_Adj>
food	<NN>	<food>	<food_>
sector	<NN>	<sector>	
,	<CM>	<,>	
today	<NR>	<today>	
announced	<VBD>	<announce>	<announcement_>
that	<CS>	<that>	
it	<PPS>	<it>	
will	<MD>	<will>	
acquire	<VB>	<acquire>	<possession_transfer>
Pastificio	<NP>	<Pastificio><guessed>	
Gazzola	<NP>	<Gazzola><guessed>	
S.p.A.	<NP>	<S.p.A><guessed>	<company >
,	<CM>	<,>	
leading	<NN>	<leading>	<company_Adj>
manufacturer	<NN>	<manufacturer>	<company>
of	<IN>	<of>	
private	<JJ>	<private>	
label	<NN>	<label>	<brand_product>
pasta	<NN>	<pasta>	<food_>

At the end of the processing, a semantically tagged text, which can be read by other applications, is issued.

actor_	leading manufacturer of branded products
agribusiness_market	Mediterranean food sector
food_	Mediterranean food
communication_	announced
/information_	
/announcement_	
when time_/punctual_	today
what_announcement	company_acquisition
company_acquisition	will acquire Pastificio Gazzola S.p.A.
who	
which_company	Pastificio Gazzola S.p.A.
actor_in_agribusiness	leading manufacturer of private label pasta
food_	private label pasta

4. Transducers Technology

The extraction component, which performs semantic labeling from morpho-syntactically tagged texts, has been implemented using a dynamic composition of transducers. According to the specific processing, three levels of transducer components have been distinguished.

4.1 World level

At this stage, the transducer reads a word in the sentence and suggests an alternative for each character in the word. The word level processing is used to deal with accentuation, uppercase vs. lowercase, and hyphen phenomena. It acts as a pre-processor of information extraction and can recognize a word and assign it a semantic label, independent of how it is written in the source text.

Firstname *Valerie* refers to *Valerie*, *Valérie*, *VALERIE*, and *VALÉRIE*
Adjective *anglo-swiss* recognizes *anglo-swiss* and *anglo – swiss*

4.2 Lexical level

At the lexical level, the transducer deals with the semantic dictionaries. In input, the tagged sentence with suggestions. In output, it gives the semantic label associated to each lexical entry, independent of how it is expressed in the term databases (word vs. multiword):

Badoit
Blue / Quellen
(Source)? / Perrier ...

4.3 Rules level

At the rules level, the transducers takes the semantically labeled sentence and applies the rule dictionaries.

For more information:

Contact us
- By e-mail:
info@temis-group.com
- Or visit our Web site :
www.temis-group.com

TEMIS France
>> Headquarters
Tour Gamma B
193-197 rue de Bercy
75582 Paris Cedex 12
France
Tel. : +33 (0)1 40 04 46 70
Fax : +33 (0)1 40 04 46 80
info@temis-group.com

TEMIS France
>> Grenoble
Tel. : +33 (0)4 76 61 51 83
Fax : +33 (0)4 76 61 41 21
info@temis-group.com

TEMIS Germany
>> Heidelberg
Tel. : +49 - 6221 13753-12
Mob. : +49 - 174 6595 713
Fax : +49 - 6221 13753-14
info.de@temis-group.com

TEMIS Italy
>> Modena
Tel. : +39 - 059 237 634
Mob. : +39 - 349 4131 718
Fax : +39 - 059 220 093
info.it@temis-group.com

TEMIS UK
>> Bristol
Tel. : + 44 (0)117 949 8646
info.uk@temis-group.com

Text
Intelligence™

TEMIS

Drive your document flows with Insight Discoverer™ Categorizer

Managing information proactively…

Your company has to process an increasing volume of documents every day. And only information found in documents that have been classified can be exploited. Your company needs to organize its information system according to your organization and processes (sector-based, functional, theme-based).

Insight Discoverer™ Categorizer, the information classification solution

Insight Discoverer™ Categorizer is a document categorization server. It automatically classifies unstructured documents into pre-defined categories, combining statistical and linguistic analysis rules.

Key applications of categorization

Insight Discoverer™ Categorizer is used in three types of applications:

> **Categorization:** assignment of a document to one or more categories within a **taxonomy**

> **Indexing:** identification of topics in a document according to similarity with documents already indexed

> **Routing:** distribution of documents to certain people or departments according to their criteria of interest

TEMIS' document categorization server is a tool particularly suited to the needs of your Quality, R&D, Marketing, Methods and Documentation departments. It is used in three main fields in particular:

➜ **Competitive Intelligence**, to categorize financial, scientific or technical documents, press articles, patents or scientific publications

➜ **Knowledge Management**, to automatically load classification trees and knowledge databases or for routing the documents that are relevant to your teams

➜ **Customer Relationship Management**, for classification or routing of e-mail from customers and automating replies

Why choose Insight Discoverer™ Categorizer?

Insight Discoverer™ Categorizer has a number of advantages:

➜ **Productivity gains:** once their classification model has been validated, your users or departments have easier access to organized information, and can quickly access mission-critical information requiring immediate action.

➜ **Quality gains:** manual classification generates significant fluctuations in the quality of document assignment. Automatic classification, on the contrary, produces consistently high quality, whatever the form and syntax of the documents.

➜ **Increased motivation:** automation of categorization frees your teams from the tedious tasks of manual classification and enables them to dedicate their time to analysis tasks.

Text Intelligence™

For more information:

Contact

TEMIS France
>> Headquarters

Tour Gamma B
193-197 rue de Bercy
75582 Paris Cedex 12
France
Tel. : +33 (0)1 40 04 46 70
Fax : +33 (0)1 40 04 46 80
info@temis-group.com

TEMIS France
>> Grenoble

Tel. : +33 (0)4 76 61 51 83
Fax : +33 (0)4 76 61 41 21
info@temis-group.com

TEMIS Germany
>> Heidelberg

Tel. : +49 - 6221 13753-12
Fax : +49 - 6221 13753-14
info.de@temis-group.com

TEMIS Italy
>> Modena

Tel. : +39 - 059 237 634
Fax : +39 - 059 220 093
info.it@temis-group.com

TEMIS United Kingdom
>> Bristol

Tel. : + 44 (0)117 949 8646
info.uk@temis-group.com

www.temis-group.com

Text Intelligence™

TEMIS

How Insight Discoverer™ Categorizer works

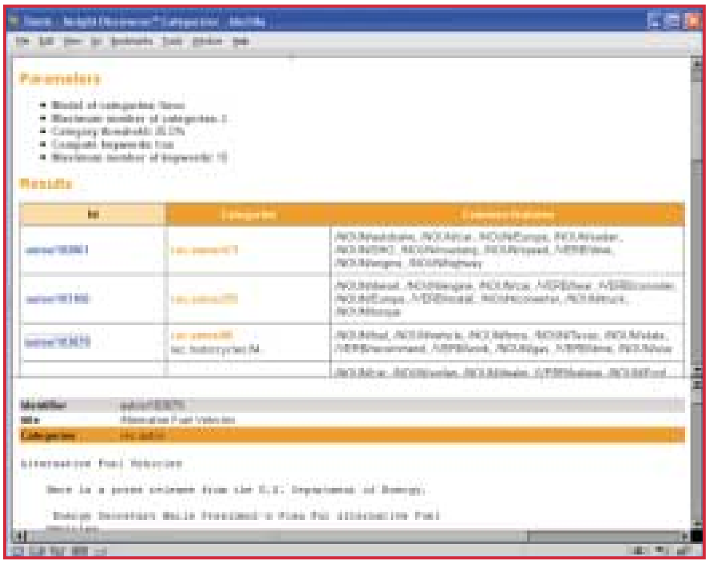
>> The learning phase: documents already categorized according to a company classification scheme undergo morpho-syntactic processing by Insight Discoverer™ Extractor, which associates a semantic descriptor to them (frequency of nouns, verbs, noun phrases, etc.). These documents are used as a basis for learning, and enable Insight Discoverer™ Categorizer to create the **categorization model** using an algorithm that combines the various semantic descriptors assigned to the same category. A minimum number of documents per category, between 25 and 50 depending on the document size, is required to guarantee the model quality.

>> The categorization phase: Insight Discoverer™ Categorizer can then assign all new documents to the various pre-defined categories. Each document to categorize is analyzed by Insight Discoverer™ Extractor. Its semantic descriptor is compared to that of previously categorized documents. One or more categories are proposed for each document, with a confidence indicator.

>> Evaluation of automatic categorization: the categorization template is assessed as follows: 90% of documents already categorized are used for learning. The remaining 10%, which were not used in the learning phase, are used as a test corpus. The quality (precision and recall) can thus be measured for each category. Another method consists of validating the categorization model by using test sets.

Discover the TEMIS product range

For optimum categorization quality, you must first ensure that the classification model is relevant. In some cases, automatic document assignment reveals inappropriate categories and unbalanced distribution. TEMIS' Insight Discoverer™ Clusterer classification suggestion tool provides an excellent solution to this problem.



Insight Discoverer™ Categorizer uses linguistic analysis performed by Insight Discoverer™ Extractor, which describes documents by generating their semantic profile. Insight Discoverer™ Categorizer then uses this for the learning and assignment phases.

Specifications

>> Operating systems:
- Windows NT, 2000, XP workstation or server versions.
- Linux.

>> API :
- Java (RMI - Remote Method Invocation).
- Documentation.

>> Source languages (supported by Insight Discoverer™ Extractor):
English, French, German, Italian, Dutch, Spanish, Portuguese, Czech, Greek, Hungarian, Polish, Russian.

>> Formats :
Over 50 input formats (including MS Word, PDF and HTML).
The document information is generated in XML.

Insight Discoverer™ Categorizer

INTER PUBLI GRAPHIQUE 04 76 21 14 01

white paper

INSIGHT DISCOVERER™ CATEGORIZER

1. Introduction

DocCat is a project that uses Insight Discoverer™ technologies (ID Extractor and ID Categorizer), which is a system for the automatic categorization, indexing, and archival of textual data for large commercial text archives. DocCat was one of the first systems for automatic indexing that was put in production in a real-world environment and demonstrates the usefulness of this approach through its use on many hundreds of new documents every day. Moreover DocCat is an ongoing project and the properties of the system are subject to continuous refinement and adaptation to specific requirements.

2. Description

The term *automatic indexing* here refers to the process of assigning terms to a document that correspond to its content, and not to the process of setting up an index for full text retrieval. While the latter typically contains virtually all the words of a document, with their lemmatised forms and without the stop words, the former means annotating a document with only a few terms from a thesaurus or some other kind of controlled set of index terms. DocCat assigns index terms (here, “topics”) to each document.

3. The making of DocCat

In the summer of 1998, a team at the IBM Institute for Logic and Linguistics in Heidelberg, Germany, was engaged in studying the automatic indexing of speech data, i.e. the annotation of audio data with phrases and keywords that allow a text-based search in a sound archive. It became apparent that a large proportion of the functionality of such a system might be of interest for the handling of textual data, too. After all, a textual representation (possibly infected with errors from the speech recognition component) was to be the input for the indexing component. Consequently, our work aroused interest not only in places where large audio archives had to be managed, but also in the text archive field, especially at the Gruner+Jahr (G+J) press database in Hamburg, where a number of projects involving automatic indexing had already been considered.

In the first pilot project we were asked to demonstrate the practicability of our approach in a limited scenario with German documents. When the results of this pilot study turned out to be encouraging, a continued cooperation was agreed upon where DocCat was to be extended towards a system that could be applied to everyday tasks in a commercial text archive.

The idea in both the pilot study and the second phase was that the G&J archive should form a training corpus of manually annotated documents that would enable specific machine learning algorithms to collect correlations between a document’s textual data and the presence or absence of a specific label. The following were the various treated labels:

- ✓ **Broad topic:** The manual annotations assign up to four labels to a document, indicating the broad topic of the document (Examples: Sports, Theatre, Economics, Science, ...). A total of 44 such topics had to be learned.
- ✓ **Specific topics:** A finer classification of a document’s topics is achieved by assigning one or more of a fixed vocabulary of some 2000 thesaurus terms to it (Examples: Doping, Trade Unions, Market Research, ...).
- ✓ **Keywords:** While a document need not literally contain a thesaurus term in order to receive this term as a label, some terms that are not part of the fixed vocabulary (i.e. the thesaurus) may also serve as an indication of the document’s contents. These

Text Intelligence™

keywords were extracted from and assigned to a document alongside the thesaurus terms.

- ✓ **Person Names:** Since the information about what individuals are mentioned together with which topics or events are of specific interest for documentation purposes. Person names have to be identified and, as far as possible, rated according to their importance in the text.
- ✓ **Names of Companies & Organizations:** The same holds for the names of other entities such as companies or organizations.
- ✓ **Geographical labels:** Finally, each document receives a label indicating the geographic region (continent, country, province, etc.) that can be used to locate the document's contents.

The question which DocCat had to answer was: Is it possible to set up an automatic procedure which assigns such labels to new, unseen documents in a way that the quality is comparable or even superior to that of the manual annotation, or which can at least be used as a starting point for manual post processing, thus making the manual annotation more effective?

4. The Insight Discoverer™ components within DocCat

A fundamental requirement in setting up the DocCat system was that the amount of effort that had to be invested in building corpus- or domain-specific knowledge bases such as dictionaries, grammars or thesauri should be kept as low as possible. This not only reduces the workload necessary for the development of the system, but, at the same time, ensures that new classifiers for new corpora, domains or documentation requirements could be built (i.e. "trained") with minimal effort. If no new types of labels need to be considered, building a new classifier will require only a training corpus of annotated documents as input.

4.1 Classification tasks: Insight Discoverer™ Categorizer

The tasks of assigning the labels for broad, specific, and geographical topics are treated as classical text classification tasks. In the learning phase, more than 100,000 terms, the "features" of the classification task, and their occurrence together with the respective classes, are collected. In the case of the specific topic learning phase, this process requires the handling of a matrix with several hundreds of millions of cells. The large number of features considered in DocCat is a consequence of the ability of the German language to build a virtually infinite number of compound nouns.

The information about the number of occurrences of a term within a class makes it possible to calculate a relevance measure for this term/class pair. Table 1 shows an excerpt from the ranked list of features for the thesaurus category "Bergsteigen" (mountain climbing).

Setting a sensible threshold in order to cut this list off avoids having to keep irrelevant information about the correlation of a specific class and a feature which is not significant for this class.

Relevance measure	Feature
9,888.81	Bergsteiger
6,943.51	Besteigung
6,633.86	Everest
4,754.4	Kletterer
3,222.71	Eiger
3,072.6	Sherpa
2,844.18	Mount
2,818.94	Eiger-Nordwand
1,875.14	Bergführer
1,037.98	Fels
...	...
0.7	Mark
0.63	Ziel
0.51	Deutschland
0.47	Leben

Table 1: Example of a list of terms and their relevance for the category "Bergsteigen" (mountain climbing)

The resulting classifier then makes the assumption of independence of the occurrence of the various features in a document and ranks the respective classes according to a confidence value. See section *Results* for a table of results on the broad topic classification task.

Since the classification process gives quantitative results, the behavior of the system can be influenced by tuning a variety of parameters concerning ranking, threshold level, total number of resulting codes, etc. As with most classification problems, there is a trade off between ID Categorizer's recall (the proportion of correct codes that the system has assigned) and its precision (the proportion of assigned codes that are correct). In such a situation, tuning the parameters allows a user to emphasize one over the other, by maximizing recall at the expense of precision, vice versa, or finding a good compromise between the two. This compromise, however, has been found to be suitable for the implementation of a workflow that processes documents automatically (i.e. without manual post-editing) in a professional documentation environment.

4.2 Keywords

Often, newly coined terms acquire high relevance for a certain topic long before the thesaurus is updated in order to reflect this development. The more dynamic a domain, the harder it is to maintain an appropriate controlled vocabulary. Thus many potentially relevant and useful indexing terms would be lost if a document only received labels from the predefined set of thesaurus terms.

DocCat reflects this by annotating a document with a selection of the highest ranking terms from the text as keywords. These keywords not only turn out to be, in many cases, good extensions to the thesaurus-based annotation, but they may also be considered as a rudimentary kind of summary (or a starting point for the generation of such a summary) that facilitates the browsing of larger document collections, e.g. search results. DocCat is able to lemmatize the identified keywords (i.e. reduce them to their non-inflected form) by drawing upon a large underlying dictionary and a morphological rule set for the processing of compound nouns (Insight Discoverer™ Extractor).

4.3 Information Extraction tasks

Naturally, identifying names in free text is not a classification task. On the contrary, it requires a specific kind of parsing. People's names (and to a lesser extent, names of organizations) bear a specific structure that can be recognized by pattern matching with sufficient accuracy.

Insight Discoverer™ Extractor uses a restricted form of a regular grammar to perform this recognition process. Some 15 rules specify the way in which titles, first names, family names and job titles are typically combined to form people's names in texts. Large collections of first names, job titles etc. are used as knowledge sources for the recognition component.

After having recognized proper names in a document, Insight Discoverer™ Extractor also tries to find canonical forms for the identified names (i.e. if someone is mentioned with a title once, with the first name a second time and with neither in a third position, Insight Discoverer™ Extractor recognizes that the person which the three forms refer to, has been mentioned three times in this document). Insight Discoverer™ Extractor even performs a rudimentary kind of anaphora resolution in order to count the number of occurrences of a name. That means that in (the corresponding German translation of) a sentence like (1)

(1) Madonna was not only one of the most controversial artists of the nineties,
she was also one of the most successful.

Madonna will be treated as if *she* had been mentioned twice. Counting the number of occurrences in the document is used as the prime criterion to assign a relevance measure to a name.

Apart from their importance as index terms for a document, people's names also have a considerable effect on the classification process. Failing to recognize them and treat them as one token in the text (rather than treating first and family names as separate tokens) would mean that important information is ignored and would therefore lead to poorer classification results. The following hypothetical example might illustrate this point.

(2) Helmut Schmidt (former German chancelor): politics
Martin Scorsese (US film director): film
Martin Schmidt (German ski athlete): sports

Assuming that we had no name recognition component and assuming that the classifier has learnt that the tokens *Helmut* and *Schmidt* have some significance for the topic *politics* and *Martin* and *Scorsese* have some relevance for *film*, then the tokens *Martin* and *Schmidt* will falsely signal some influence towards *politics* and *film* even if they occur as parts of another name, here that of an athlete. The name recognition component on the other hand will treat these names as tokens, exploiting the fact that in most domains, most names tend to be mentioned with a limited set of topics and thus often serve as good features for a classification component.

5. Results

DocCat currently runs in a specific setting where documents that are assigned to one or several broad topics *(TOP = {sports, theatre, economics, ...})* have to be annotated with finer labels from a fixed vocabulary of thesaurus terms *(SUBTOP = {..., aids, demonstration, university, theft, cancer, ... })*. Of these, DocCat has learnt to assign some 600 terms[1].
The results on the different broad topics reflect the internal homogeneity of the respective sub corpora: On a well-defined topic such as *sports* (recall 60%, precision 70%), DocCat performed much better than in other areas like *society/celebrities* (recall 51%, precision 40%) in assigning the correct specific topic terms.

[1] The quantitative results reported here have been obtained in an extended evaluation period at the G+J press database. The author would like to thank Günter Peters and Volker Gaese from G+J for making their findings available for this report.

Topic	Recall	Precision
sports	*60*	*70*
energy	*65*	*66*
tourism	*70*	*59*
transportation	*63*	*62*
fashion	*63*	*46*
…	…	…
society	*51*	*40*
family	*43*	*47*

Table 2: Precision, Recall for the assignment of SUBTOP labels for a number of TOP corpora

The first line for instance in table 2, reflects the fact that on the sub-corpus annotated with sports DocCat performed the task of selecting the correct assignments with 60% recall and 70% precision from a set of some 600 SUBTOP labels.
It has to be noted that these results reflect the number of cases where the DocCat result differs from the initial manual annotation found in the test set, no matter how this difference should be judged. Besides the cases, however, where we have to admit a clear DocCat error, there are also many cases where one has to conclude that both results (the manual and the automatic) are equally valid and even a considerable fraction, where the results of DocCat can or have to be considered superior. While we have yet to evaluate this question quantitatively, we are confident that such a study will prove that the quality of DocCat is even better than that reflected in the tables above.

6. System specifications

The DocCat system is implemented using C++ for the number crunching part of the calculations and Perl for the text processing parts and as a glue language. It currently runs on AIX and Windows (95 and NT). A minimum of 128MB of RAM is necessary for the training of moderately large corpora (> 50,000 documents). The resulting classifier then is able to process a document in less than a second and annotate it in the way described above.

DocCat runs as a separate process that is called over a simple interface (file or TCP/IP) from the respective text documentation environment and delivers its results back over the same interface. Thus DocCat has minimal requirements concerning the target environment and can be employed in any documentation system where external functionalities can be integrated.

7. Conclusion and Outlook

The DocCat automatic indexing system can be applied in a real world setting, yielding results that enhance the efficiency of manual indexing. For the setting up of a classifier for a new domain or a new classification system, nothing but the annotated training corpus is needed - no hand coding of grammar or correlation rules is required.

For more information:

Contact us
- By e-mail:
info@temis-group.com
- Or visit our Web site :
www.temis-group.com

TEMIS France
>> Headquarters
Tour Gamma B
193-197 rue de Bercy
75582 Paris Cedex 12
France
Tel. : +33 (0)1 40 04 46 70
Fax : +33 (0)1 40 04 46 80
info@temis-group.com

TEMIS France
>> Grenoble
Tel. : +33 (0)4 76 61 51 83
Fax : +33 (0)4 76 61 41 21
info@temis-group.com

TEMIS Germany
>> Heidelberg
Tel. : +49 - 6221 13753-12
Mob. : +49 - 174 6595 713
Fax : +49 - 6221 13753-14
info.de@temis-group.com

TEMIS Italy
>> Modena
Tel. : +39 - 059 237 634
Mob. : +39 - 349 4131 718
Fax : +39 - 059 220 093
info.it@temis-group.com

TEMIS UK
>> Bristol
Tel. : + 44 (0)117 949 8646
info.uk@temis-group.com

Text Intelligence™

Structure your information with Insight Discoverer™ Clusterer

The quest for information...

Your employees may spend up to 80% of their time looking for the information they need to do their jobs. Information science provides reliable content storage solutions (databases, document management) and retrieval solutions (search engines), but appropriating information is a different challenge.

Insight Discoverer™ Clusterer, the information organization solution

Insight Discoverer™ Clusterer is an automated classification server that dynamically groups documents according to their semantic similarity. It proposes the most relevant classification for a given document collection. Users can then browse through their documents organized according to theme and sub-theme. They have both an overview of the information and different avenues to explore. It is therefore easier to find and appropriate relevant information.

Key applications of clustering

Insight Discoverer™ Clusterer provides companies with advanced solutions for Competitive Intelligence and optimization of Customer Relationship Management. In the field of Competitive Intelligence, Insight Discoverer™ Clusterer offers a high-performance solution for organizing information related to competitors and the market. It dynamically organizes collections of documents captured by a search engine according to topics. Users can then browse through their document collections and find the most interesting competitive behaviors. In the field of Customer Relationship Management, Insight Discoverer™ Clusterer enables marketing departments to get to know the behavior of their consumers and to orient their strategies accordingly. The software segments your customer database according to e-mail messages, customer correspondence or opinion surveys, for example. This segmentation is innovative because it is based on customers' writing and what they express, rather than models based on descriptive statistics.

➜ **Researchers:**
You obtain state-of-the art research information by organizing scientific publications and patents in a given field.

➜ **Product managers:**
You identify customer segments and isolate characteristic behavior or representative comments for a given range or product, by analyzing e-mail messages or customer satisfaction surveys, for example.

➜ **Financial analysts:**
You isolate the major themes and trends in your sector and detect high value added information by analyzing your flows from the economic, financial and trade press.

Text Intelligence™

TEMIS

For more information:

Contact

TEMIS France

>> Headquarters

Tour Gamma B
193-197 rue de Bercy
75582 Paris Cedex 12
France
Tel. : +33 (0)1 40 04 46 70
Fax : +33 (0)1 40 04 46 80
info@temis-group.com

TEMIS France

>> Grenoble

Tel. : +33 (0)4 76 61 51 83
Fax : +33 (0)4 76 61 41 21
info@temis-group.com

TEMIS Germany

>> Heidelberg

Tel. : +49 - 6221 13753-12
Fax : +49 - 6221 13753-14
info.de@temis-group.com

TEMIS Italy

>> Modena

Tel. : +39 - 059 237 634
Fax : +39 - 059 220 093
info.it@temis-group.com

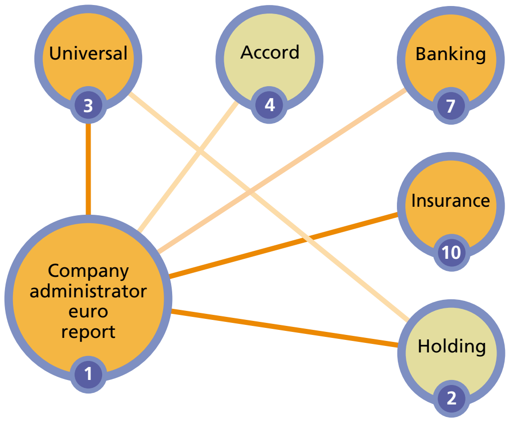
TEMIS United Kingdom

>> Bristol

Tel. : + 44 (0)117 949 8646
info.uk@temis-group.com

www.temis-group.com

Text Intelligence™



Why choose Insight Discoverer™ Clusterer?

Insight Discoverer™ Clusterer structures information so the content can be used. It offers excellent visibility for large sets of documents and complex issues. It thus enables considerable productivity gains in knowledge management and especially facilitates decision-making.

There are many advantages to using this solution:

> Rapid viewing of information

> Ease of processing of large volumes of documents for wide-ranging searches

> Easy to browse through information spheres

> Increased productivity through grouping of similar information

> Easier to appropriate information

How Insight Discoverer™ Clusterer works

Insight Discoverer™ Clusterer uses an innovative classification process based on a combination of linguistic and statistical analyses. It uses the morpho-syntactic analysis performed by Insight Discoverer™ Extractor, which can be customized with a Skill Cartridge™, to generate document descriptors. The algorithm that groups similar documents into classes is specially adapted for text analysis. Users can configure the depth of the classification model and the number of classes per level. A heading is assigned to each class which uses, in hierarchical order, the terms and expressions that are most characteristic of the class. Insight Discoverer™ Clusterer is a flexible tool that can satisfy different levels of requirements:

>> **Viewing:** a mapping module enables users to view classes in the form of spheres whose sizes vary according to the quantity of documents they represent. It also identifies the links between classes in the form of segments of varying widths.

>> **Analysis:** By projecting descriptive variables onto the mapping result, you can instantly identify the spheres for action according to color codes.

>> **Customization:** using a Skill Cartridge™ enables you to integrate business vocabulary or rules to homogenize the proposed classification.

Discover the TEMIS product range

Insight Discoverer™ Clusterer uses the morpho-syntactic analysis performed by Insight Discoverer™ Extractor. Insight Discoverer™ Extractor describes the documents upstream by generating their semantic profile. This profile is then used by Insight Discoverer™ Clusterer to propose classification schemes.

Specifications

>> **Operating systems:**
- Windows NT, 2000, XP workstation or server versions.
- Linux.

>> **API :**
- Java (RMI - Remote Method Invocation).
- Documentation.

>> **Source languages (supported by Insight Discoverer™ Extractor):**
English, French, German, Italian, Dutch, Spanish, Portuguese, Czech, Greek, Hungarian, Polish, Russian.

>> **Formats :**
over 50 input formats (including MS Word, PDF and HTML).

Insight Discoverer™ Clusterer

INTER PUBLI GRAPHIQUE 60 14 21 46 01