Adding Value to Scholarly Communications through Text Mining

Sophia Ananiadou

UK National Centre for Text Mining

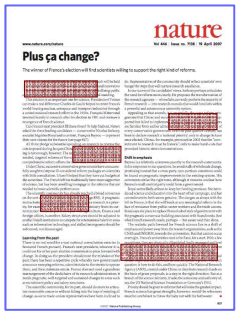
- first national text mining centre in the world
  www.nactem.ac.uk
- **Remit**: **Provision of text mining services to support UK research**
- **Funded by** JISC
- University of Manchester, collaboration with Tokyo 東京大学 THE UNIVERSITY OF TOKYO

From Text to Knowledge

nature

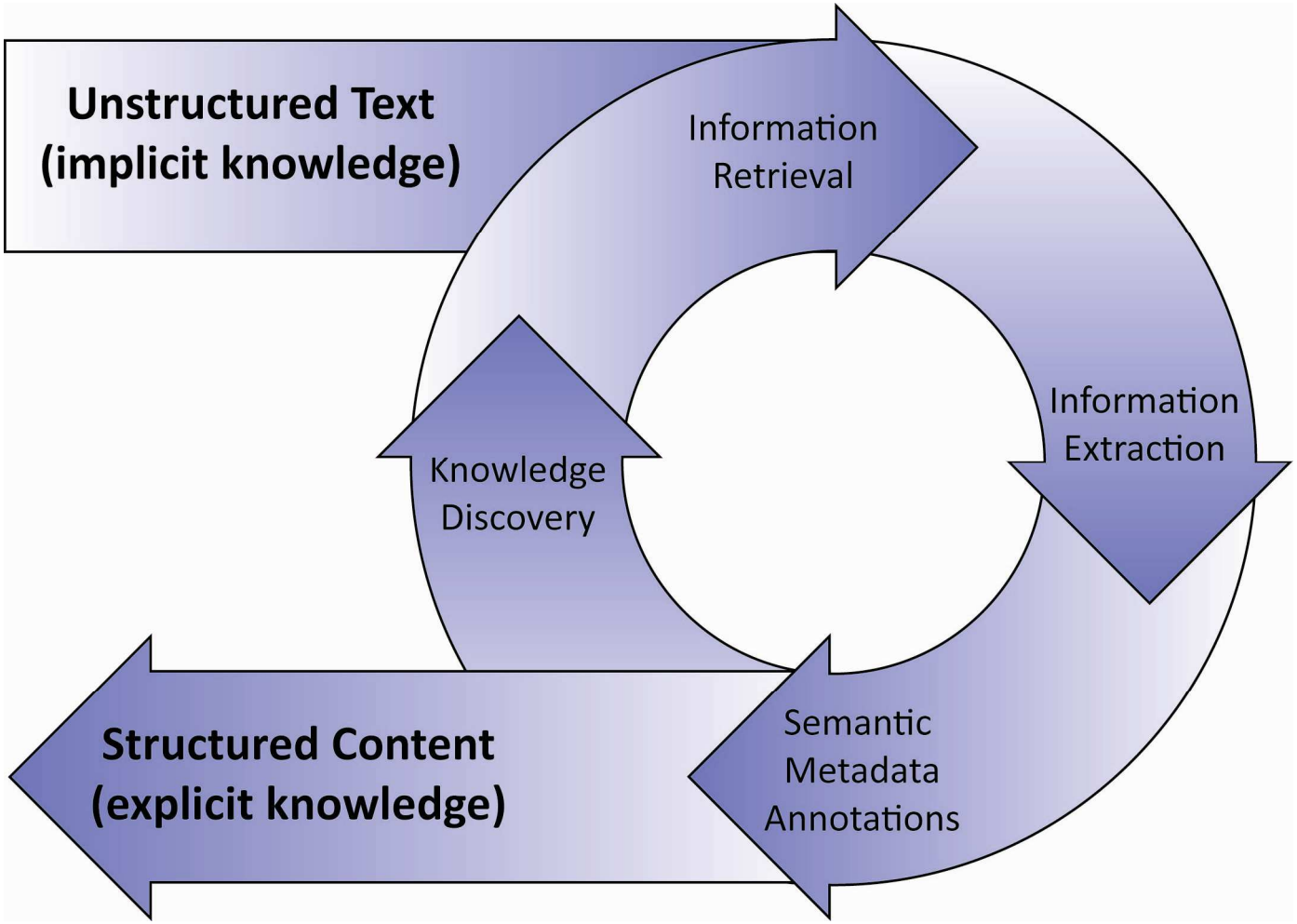
Plus ça change?

The winner of France's election will find scientists willing to support the right kind of reforms.

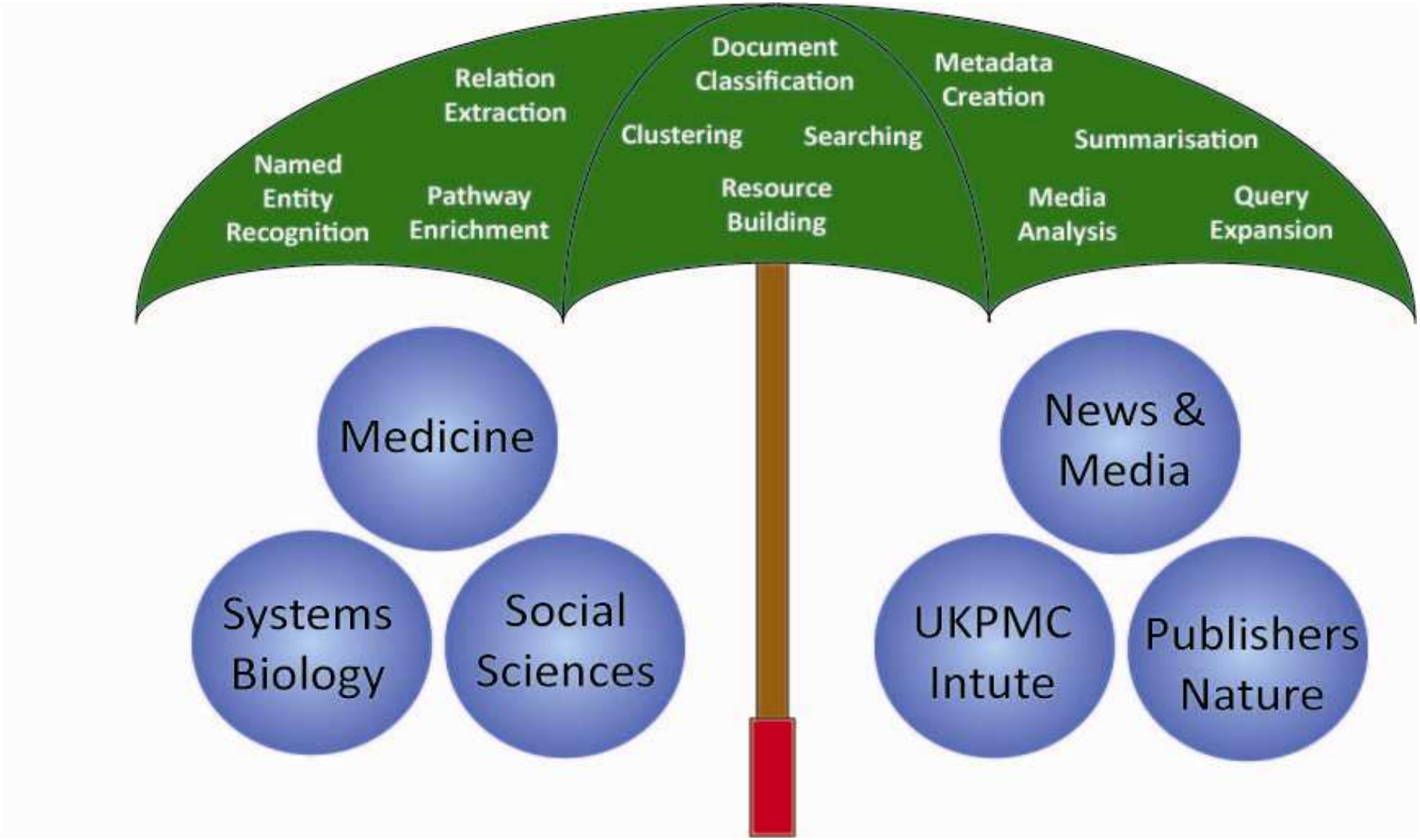

nature

Plus ça change?

The winner of France's election will find scientists willing to support the right kind of reforms.



Applications, users and techniques



Scholarly Communication Requirements

- What is needed in the repositories
  - Annotation and curation assistance
    - Creation of metadata, consistent manner
  - Name authorities
    - Merging and mapping existing resources
    - Prediction lists based on named entity recognition
    - Disambiguation
  - **Semantic metadata creation and enhancement**

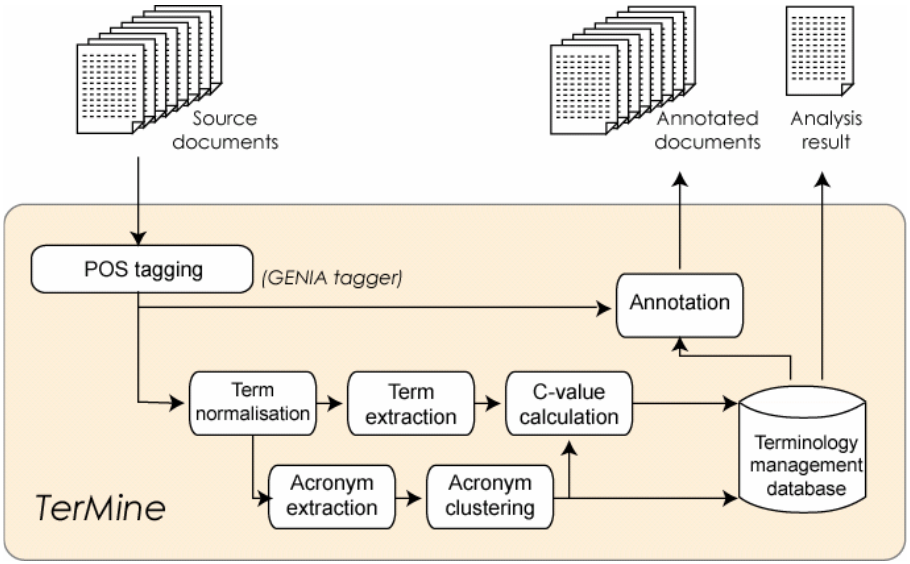
Provision of semantic metadata to support search

- Extraction of terms and named entities (names of people, organisations, diseases, genes, etc)
- Discovery of concepts *allows* semantic annotation and enrichment of documents
  - *Improves* information access by going *beyond* index terms, enabling semantic querying
  - *Improves* clustering, classification of documents
- Going a step further: extracting relationships, events from text
  - *Enables* even more advanced semantic applications

Semantic metadata for whom?

- end users
  - adds value to library content
  - allows enhanced searching functionalities
  - allows interaction with content, **living document**
- automated content aggregators
  - access to data-driven, quality metadata derived from text
- librarians
  - enhanced capability for semantic indexing, cross-referencing between Library collections and classification

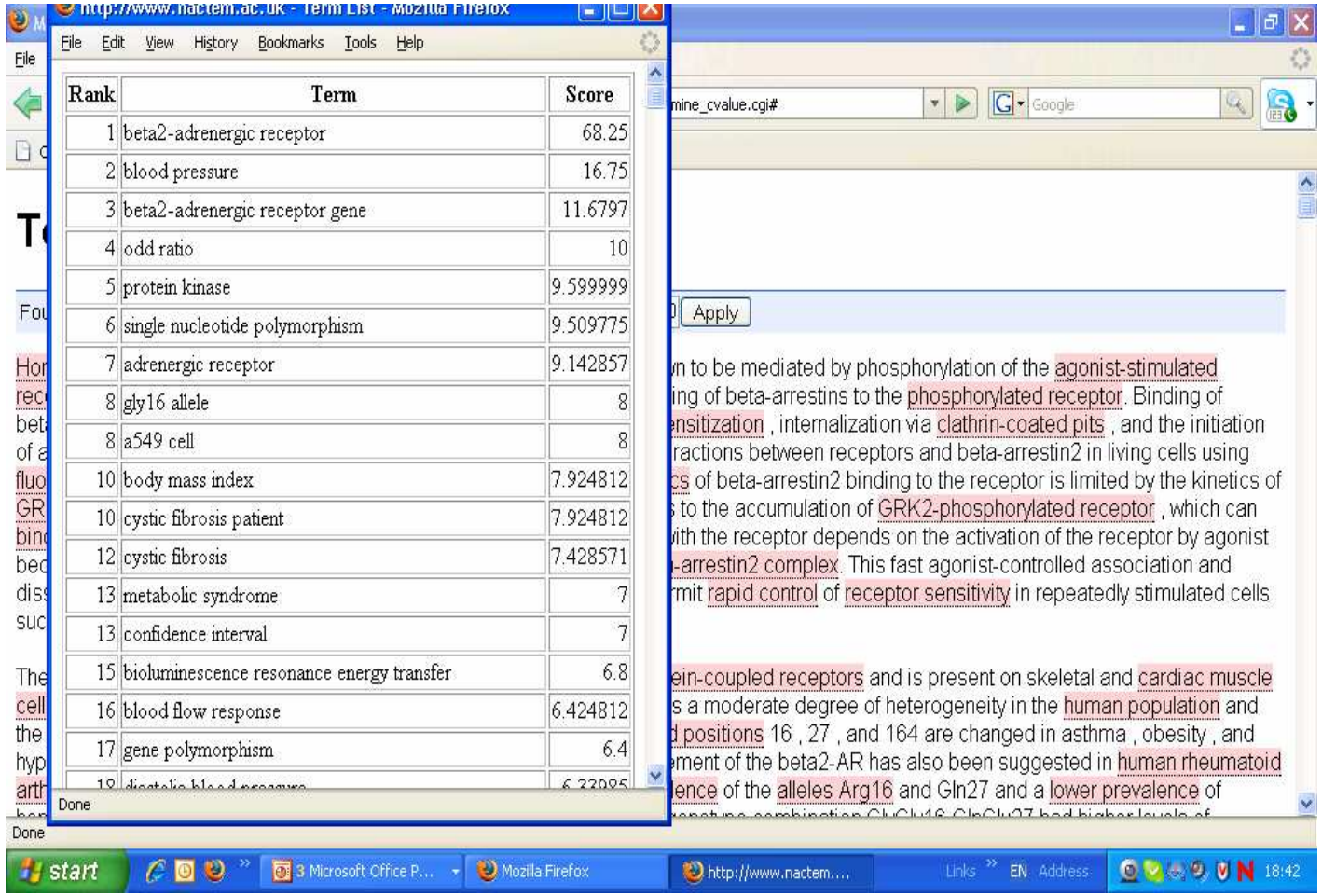
Terminology Services



TerMine

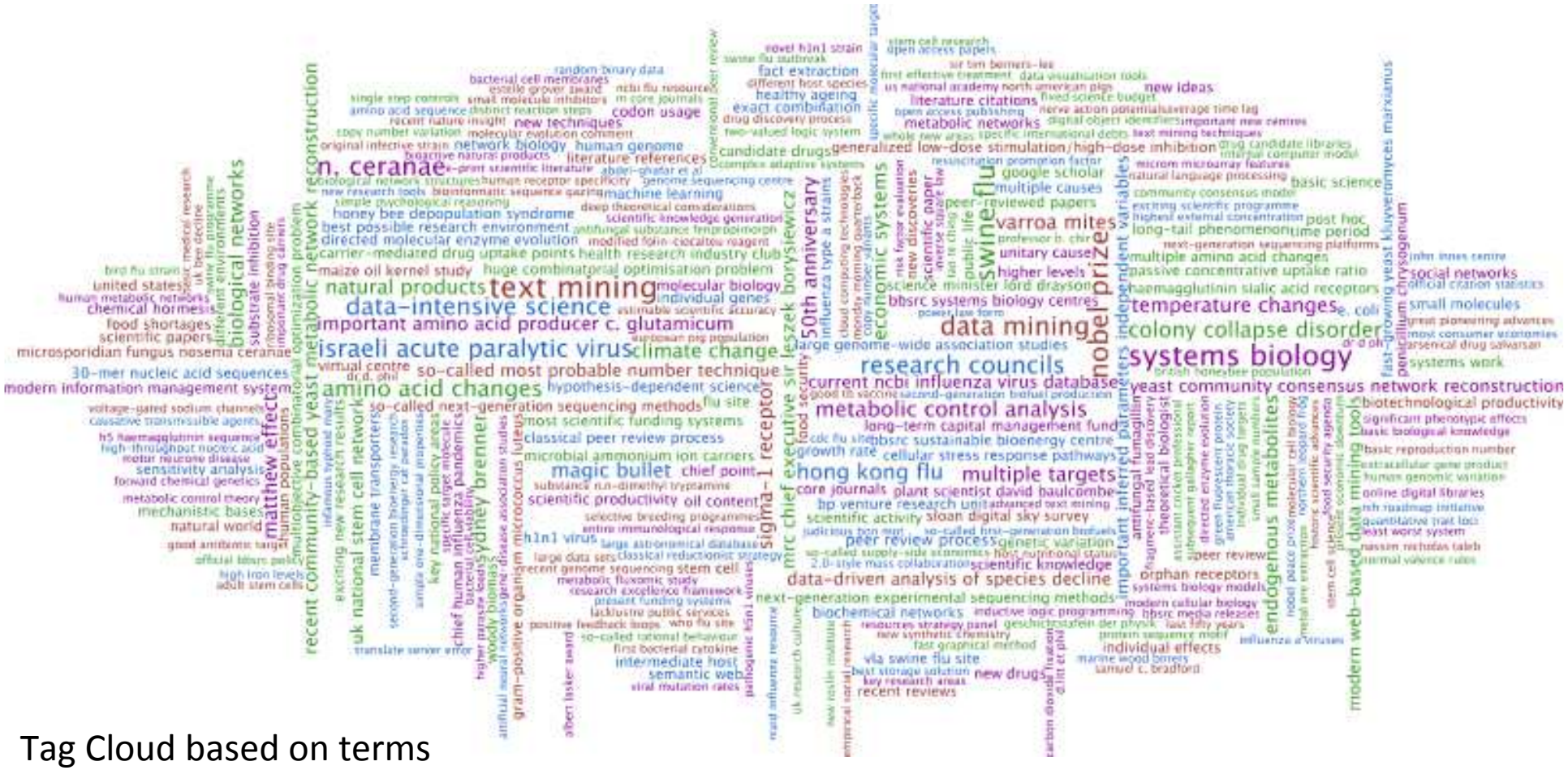
- Identifies the most significant terms
  - Used as metadata
- Suggests similar areas of interest
- Refines index terms for document classification
- Used for ontology building (Protégé TerMine plug-in)

Semantic metadata: terms



Rank	Term	Score
1	beta2-adrenergic receptor	68.25
2	blood pressure	16.75
3	beta2-adrenergic receptor gene	11.6797
4	odd ratio	10
5	protein kinase	9.599999
6	single nucleotide polymorphism	9.509775
7	adrenergic receptor	9.142857
8	gly16 allele	8
8	a549 cell	8
10	body mass index	7.924812
10	cystic fibrosis patient	7.924812
12	cystic fibrosis	7.428571
13	metabolic syndrome	7
13	confidence interval	7
15	bioluminescence resonance energy transfer	6.8
16	blood flow response	6.424812
17	gene polymorphism	6.4

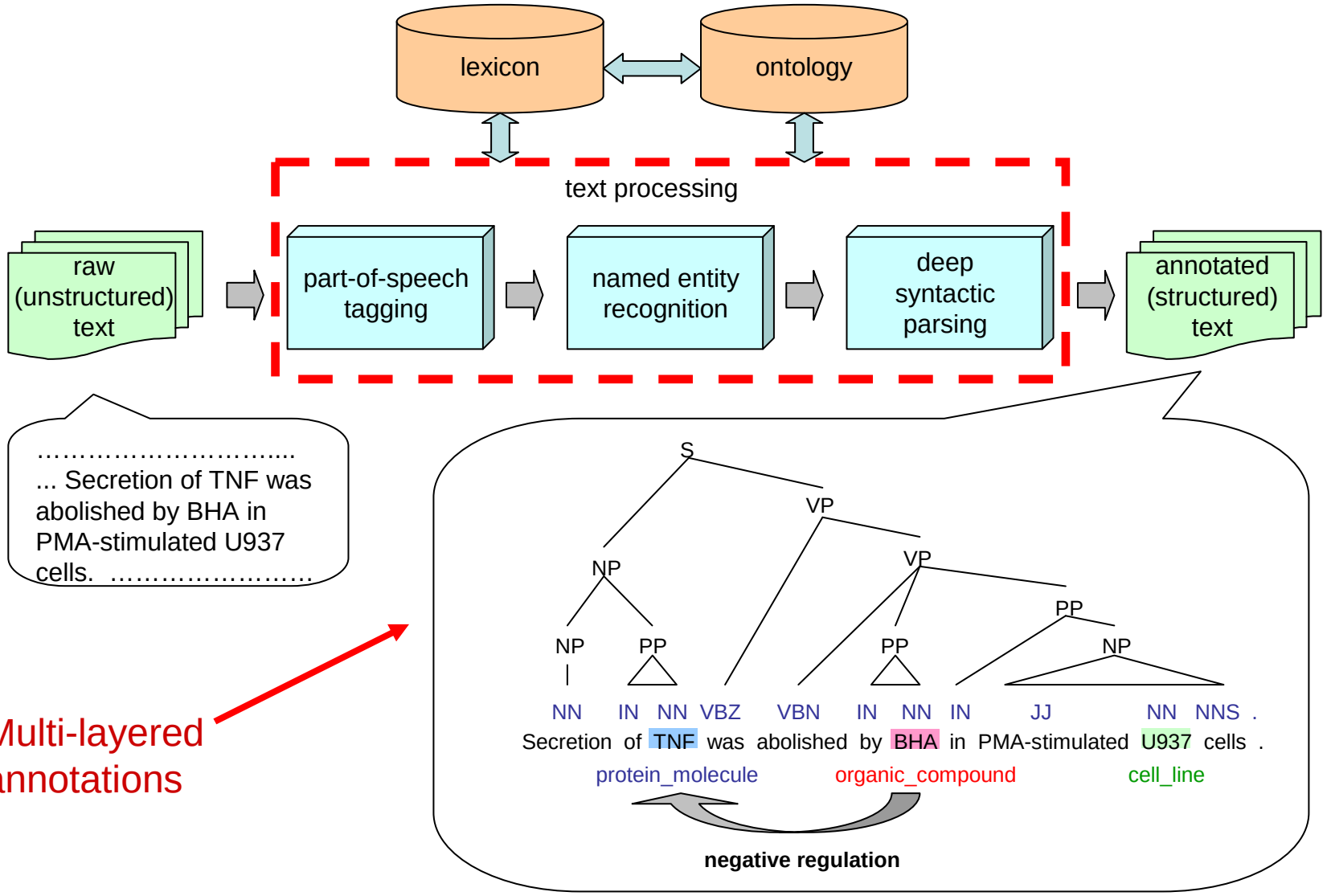
Term Based Applications



Tag Cloud based on terms automatically extracted from the blog of BBSRC Chief Executive Professor Kell.

Visualised in WORDL

Semantic metadata: facts



Text Annotations—generic to specific

- Task-oriented Annotation
  - Application annotated text
  - User system development
  - Defined by s
    - Specific curation tasks in specific environments
    - Mapping of Protein names to database IDs in specific text types
    - Specific event types such as Protein-Protein Interaction
    - Disease-Gene Association of specific diseases
    - Pathways

Linking text with knowledge

- Task-neutral Annotation [U-Tokyo, NaCTeM]
  - Development of generic tools
  - Defined by theories
  - s
  - ns
    - Phrase Structure
    - Dependency Structure
    - Deep Syntax (PAS)
  - Biology
    - Named Entities of various semantic types
    - Events
  - Linguistics + Biology
    - Co-reference

Institutional Repository Search

NaCTeM intute

Home | Advanced Query | Project Info | Academics | Developers | Organisations | Forums | Repository Search

Full Document Information

Similar Documents:
- Building Better Nurse Scheduling Algorithms
- On the job rotation problem
- Identifying and Assessing the Critical Risk Factors in an Underground Rail Project in Thailand: A Factor Analysis Approach
- Exploiting problem structure in a genetic algorithm approach to a nurse rostering problem
- An Estimation of Distribution Algorithm for Nurse Scheduling

Title:	Building Better Nurse Scheduling Algorithms
Authors:	Aickelin, Uwe, White, Paul
Subject:	NONE
Publication Date:	2004
Datestamp:	2007-10-30T14:37:54Z
Publisher:	Nottingham ePrints
Type:	Article, PeerReviewed
Identifier:	http://eprints.nottingham.ac.uk/662/1/04annals_nurse.pdf
Abstract:	The aim of this research is twofold: Firstly, to model and solve a complex nurse scheduling problem with an integer programming formulation and evolutionary algorithms. Secondly, to detail a novel statistical method of comparing and hence building better scheduling algorithms by identifying successful algorithm modifications. The comparison method captures the results of algorithms in a single figure that can then be compared using traditional statistical techniques. Thus, the proposed method of comparing algorithms is an objective procedure designed to assist in the process of improving an algorithm. This is achieved even when some results are non-numeric or missing due to infeasibility. The final algorithm outperforms all previous evolutionary algorithms, which relied on human expertise for modification.

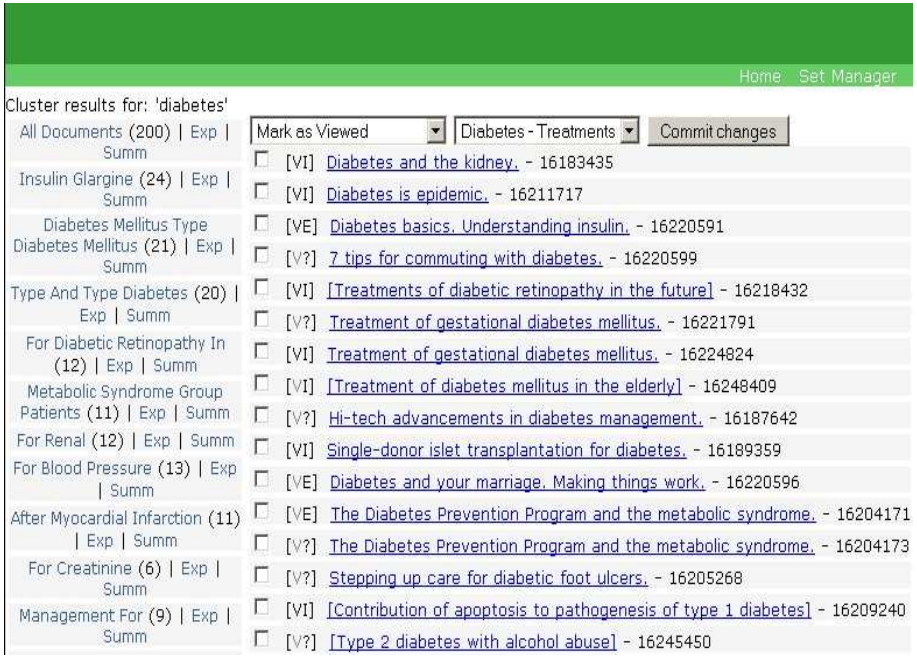
Terms from Document:
nurse scheduling problem/
scheduling problem/
nurse scheduling/
comparison method/
integer programming formulation/
traditional statistical technique/
evolutionary algorithm/
Dr Uwe Aickelin Dr Paul White School/
scheduling algorithm/
Building Better Nurse Scheduling Algorithms Annals/
integer programming/
Nottingham NG8 1BB UK Bristol/
complex nurse scheduling problem/
objective procedure/
single figure/
Computer Science University/

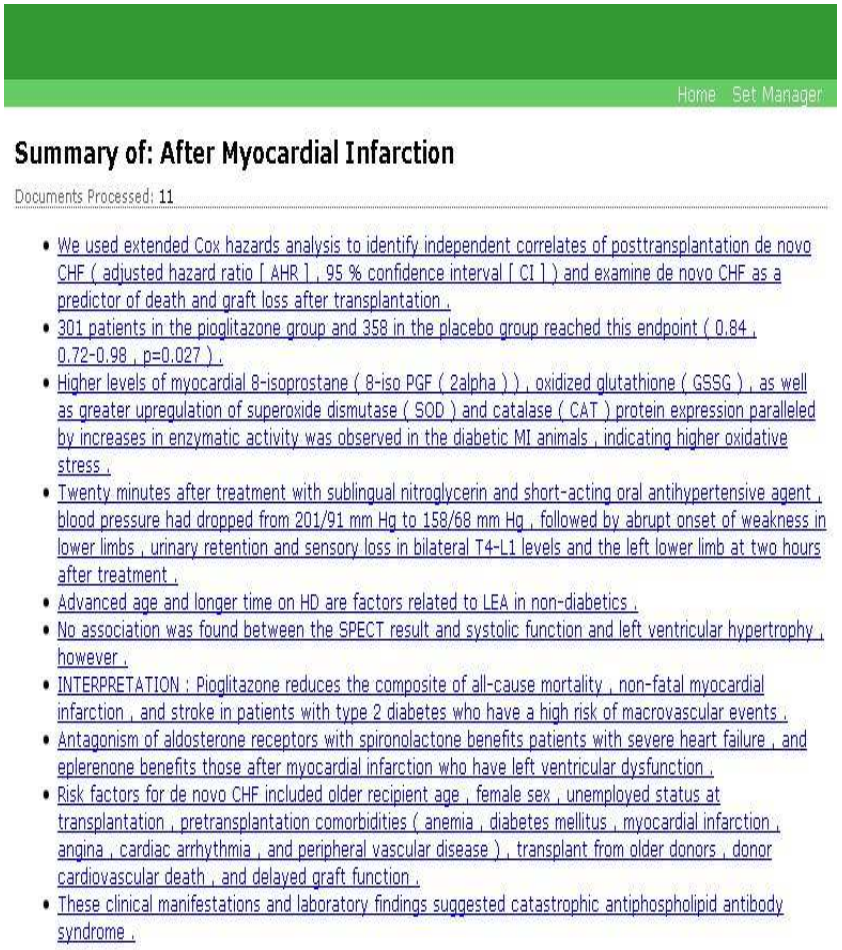
What NaCTeM provides to IRS

- Integration of full text harvesting and content extraction
- Concept/term based navigation and browsing
- Run time conceptual document similarity
- Run time document clustering

Assisting Systematic Reviews

- **Searching**
  - Query expansion, document clustering
- **Screening**
  - Document classification, sectioning
- **Synthesising**





UK PubMed Central

- Text mining derived annotations enriching an open access biomedical repository of full texts
- Focus on facts, events, named entities
  - extracted “behind the scenes”
- Enhancing user experience through semantic searching

NaCTeM services:

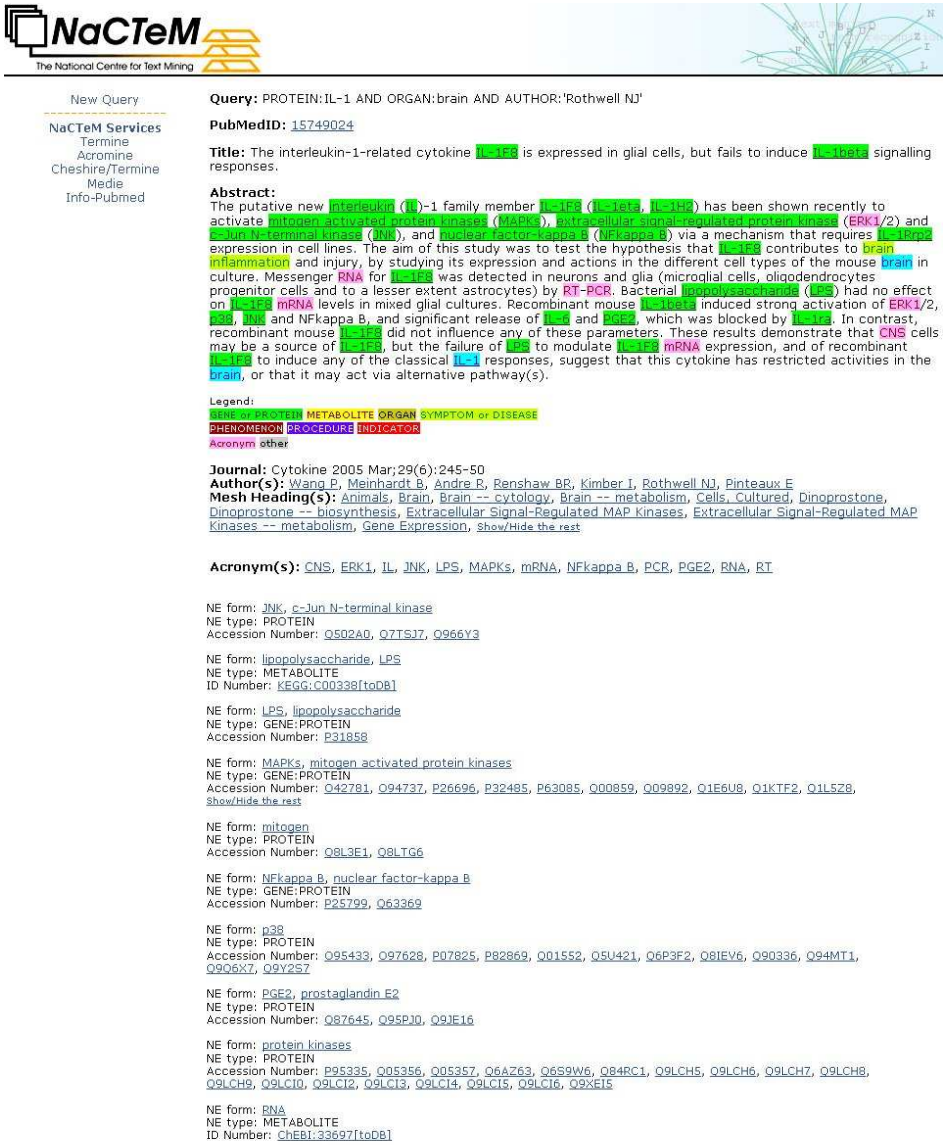
- KLEIO
- MEDIE
- FACTA

- British Library (lead), EBI, Mimas, University of Manchester

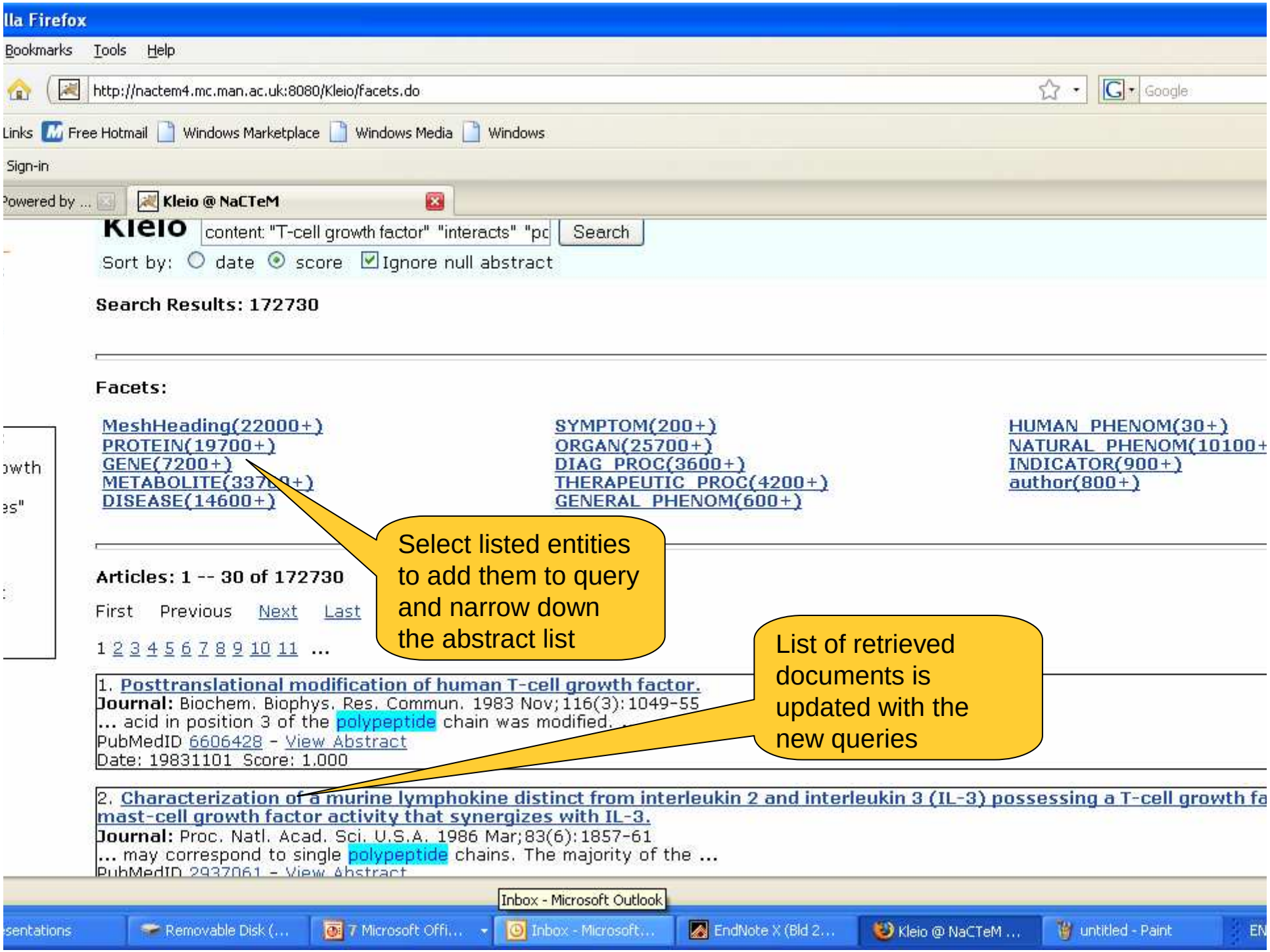
Arthritis Research Campaign, BBSRC, British Heart Foundation, Cancer Research UK, Chief Scientist Office, Department of Health - National Institute of Health Research, Medical Research Council, Wellcome Trust

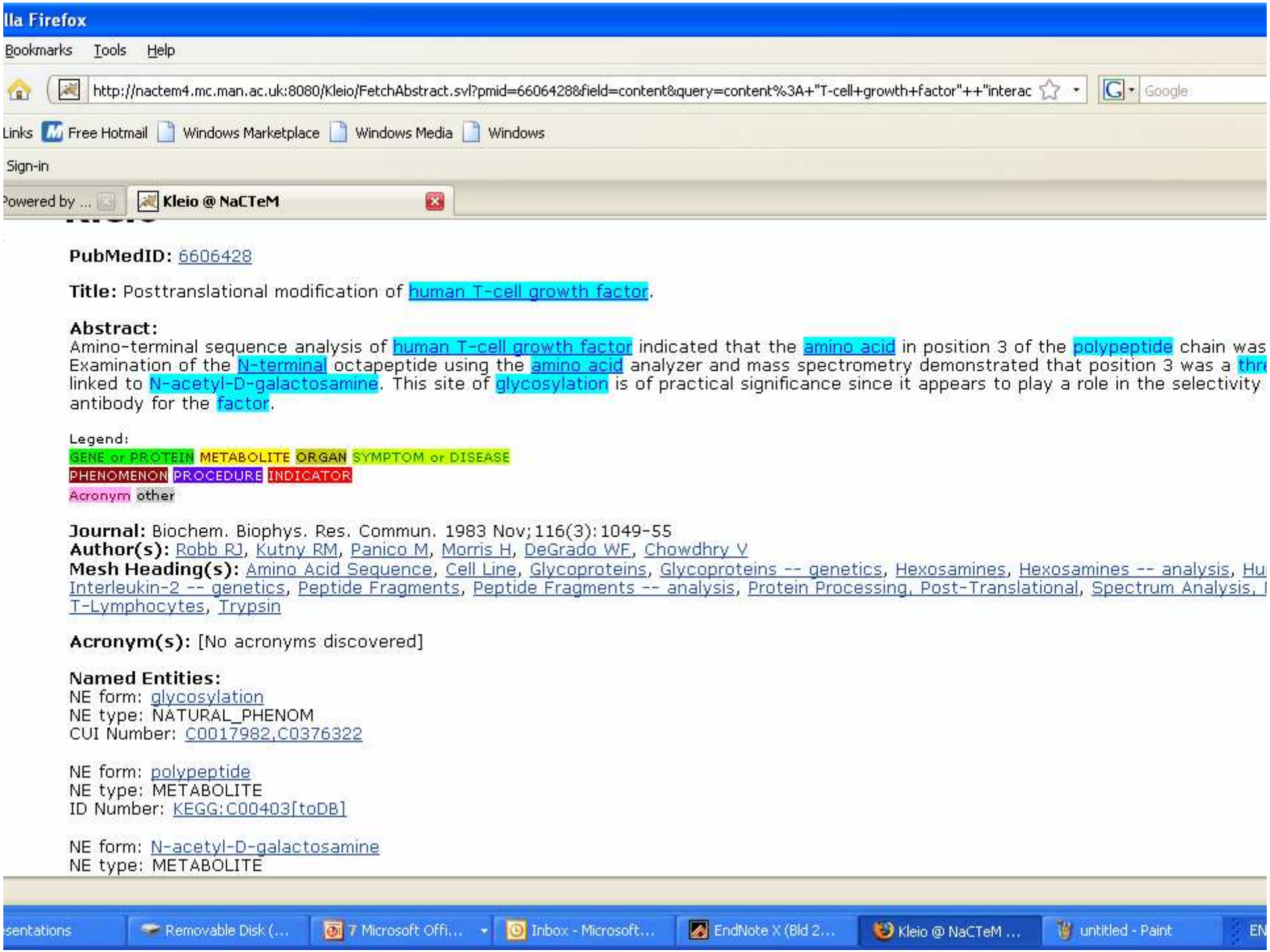
Semantic metadata: named entities

- Specialised biomedical named entities, e.g. protein, gene…
- Linked to external knowledge sources
- Allows typed searching
- Normalises terms to include acronyms, synonyms and variants
- Faceted browsing



KLEIO SERVICE


lla Firefox
Bookmarks Tools Help
http://nactem4.mc.man.ac.uk:8080/Kleio/facets.do
Google
Links Free Hotmail Windows Marketplace Windows Media Windows
Sign-in
Powered by ...
Kleio @ NaCTeM
Kleio
content: "T-cell growth factor" "interacts" "pc
Search
Sort by: date score Ignore null abstract
Search Results: 172730
Facets:
MeshHeading(22000+)
PROTEIN(19700+)
GENE(7200+)
METABOLITE(337
DISEASE(14600+)
SYMPTOM(200+)
ORGAN(25700+)
DIAG_PROC(3600+)
THERAPEUTIC_PROC(4200+)
GENERAL_PHENOM(600+)
HUMAN_PHENOM(30+)
NATURAL_PHENOM(10100+
INDICATOR(900+)
author(800+)
Select listed entities to add them to query and narrow down the abstract list
Articles: 1 -- 30 of 172730
First Previous Next Last
1 2 3 4 5 6 7 8 9 10 11 ...
List of retrieved documents is updated with the new queries
1. Posttranslational modification of human T-cell growth factor.
Journal: Biochem. Biophys. Res. Commun. 1983 Nov;116(3):1049-55
... acid in position 3 of the polypeptide chain was modified. ...
PubMedID 6606428 - View Abstract
Date: 19831101 Score: 1.000
2. Characterization of a murine lymphokine distinct from interleukin 2 and interleukin 3 (IL-3) possessing a T-cell growth fa
mast-cell growth factor activity that synergizes with IL-3.
Journal: Proc. Natl. Acad. Sci. U.S.A. 1986 Mar;83(6):1857-61
... may correspond to single polypeptide chains. The majority of the ...
PubMedID 2937061 - View Abstract
Inbox - Microsoft Outlook
Removable Disk (...
7 Microsoft Offi...
Inbox - Microsoft...
EndNote X (Bld 2...
Kleio @ NaCTeM ...
untitled - Paint

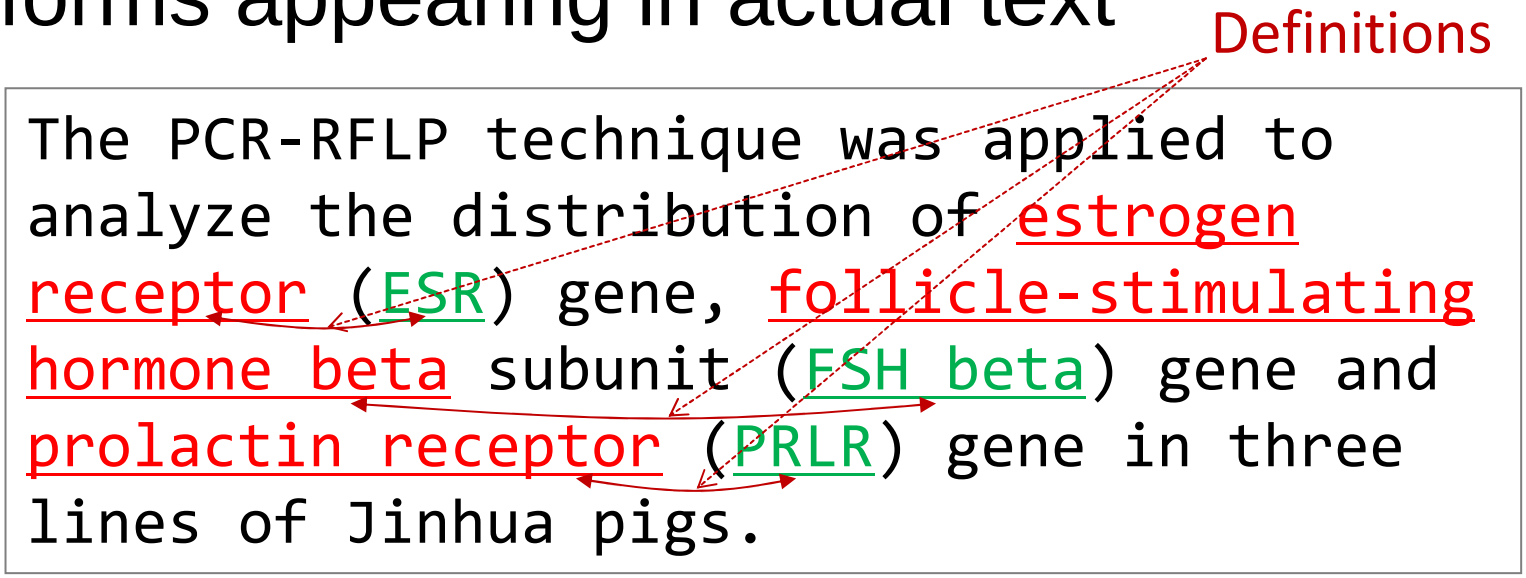

lla Firefox
Bookmarks Tools Help
http://nactem4.mc.man.ac.uk:8080/Kleio/FetchAbstract.svl?pmid=6606428&field=content&query=content%3A+"T-cell+growth+factor"++"interac
Google
Links Free Hotmail Windows Marketplace Windows Media Windows
Sign-in
Powered by ...
Kleio @ NaCTeM
PubMedID: 6606428
Title: Posttranslational modification of human T-cell growth factor.
Abstract:
Amino-terminal sequence analysis of human T-cell growth factor indicated that the amino acid in position 3 of the polypeptide chain was
Examination of the N-terminal octapeptide using the amino acid analyzer and mass spectrometry demonstrated that position 3 was a thr
linked to N-acetyl-D-galactosamine. This site of glycosylation is of practical significance since it appears to play a role in the selectivity
antibody for the factor.
Legend:
GENE or PROTEIN METABOLITE ORGAN SYMPTOM or DISEASE
PHENOMENON PROCEDURE INDICATOR
Acronym other
Journal: Biochem. Biophys. Res. Commun. 1983 Nov;116(3):1049-55
Author(s): Robb RJ, Kutny RM, Panico M, Morris H, DeGrado WF, Chowdhry V
Mesh Heading(s): Amino Acid Sequence, Cell Line, Glycoproteins, Glycoproteins -- genetics, Hexosamines, Hexosamines -- analysis, Hu
Interleukin-2 -- genetics, Peptide Fragments, Peptide Fragments -- analysis, Protein Processing, Post-Translational, Spectrum Analysis,
T-Lymphocytes, Trypsin
Acronym(s): [No acronyms discovered]
Named Entities:
NE form: glycosylation
NE type: NATURAL_PHENOM
CUI Number: C0017982,C0376322
NE form: polypeptide
NE type: METABOLITE
ID Number: KEGG:C00403[toDB]
NE form: N-acetyl-D-galactosamine
NE type: METABOLITE
sentations
Removable Disk (...
7 Microsoft Offi...
Inbox - Microsoft...
EndNote X (Bld 2...
Kleio @ NaCTeM ...
untitled - Paint

Named Entities as Index Terms

- Kleio uses standard indexing tools
  - but indexing happens at the level of semantic entities
    - identified
    - categorised
    - associated with a canonical concept name and identifier
- Index terms have many synonyms
  - Spelling variants, acronyms, etc

Querying with acronyms

- Extract abbreviations and their expanded forms appearing in actual text



PMID: 14986424

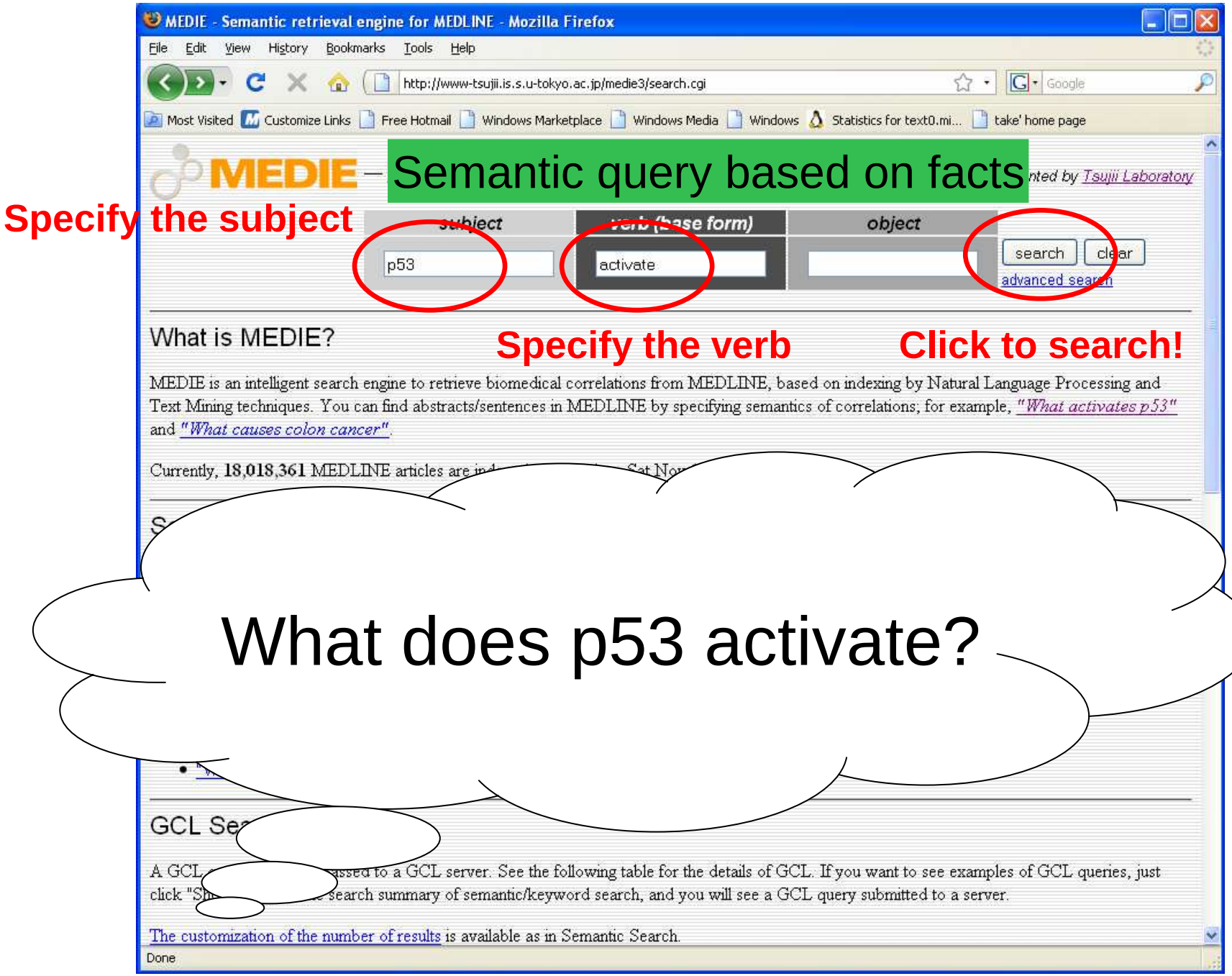
Abbreviations (short forms)

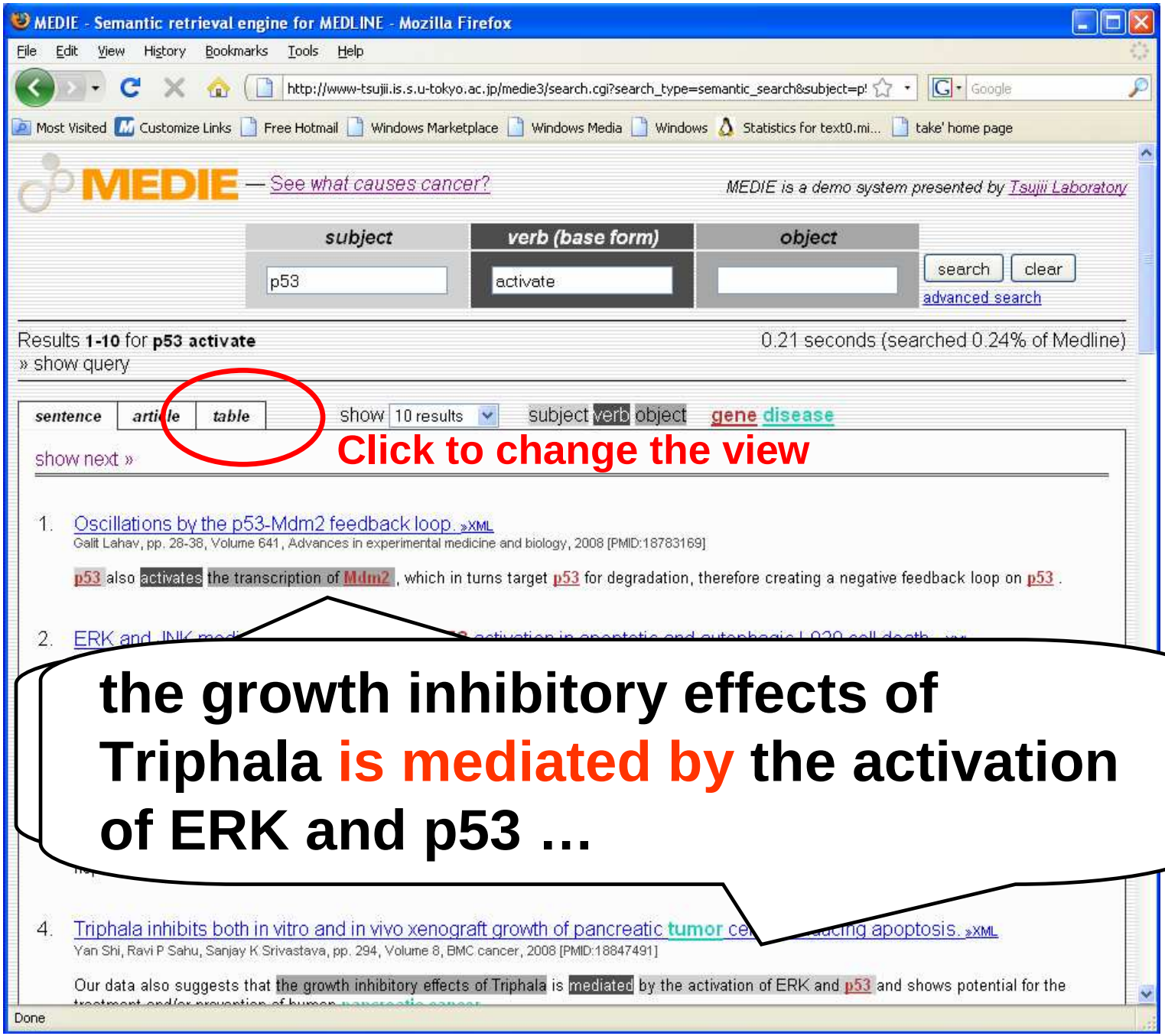
Expanded forms (long forms; full forms)

⇨ Term variation

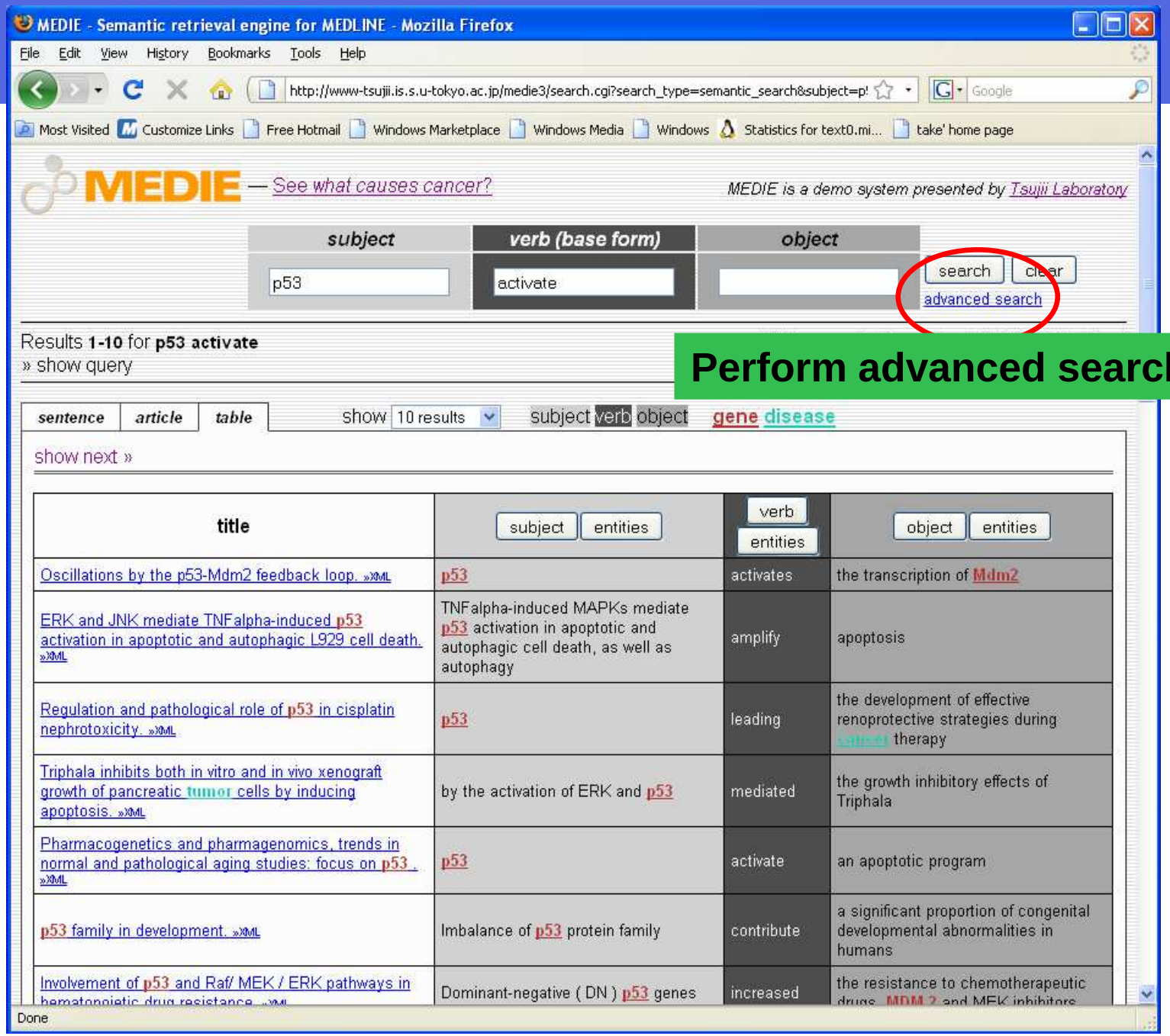
Semantic search based on facts

- MEDIE: an interactive advanced IR system retrieving **facts**
- ***Performs a semantic search***
- Core technology annotates texts
  - GENIA tagger → syntactic structures
  - Enju (deep parser) → facts
  - Dictionary-based named entity recognition

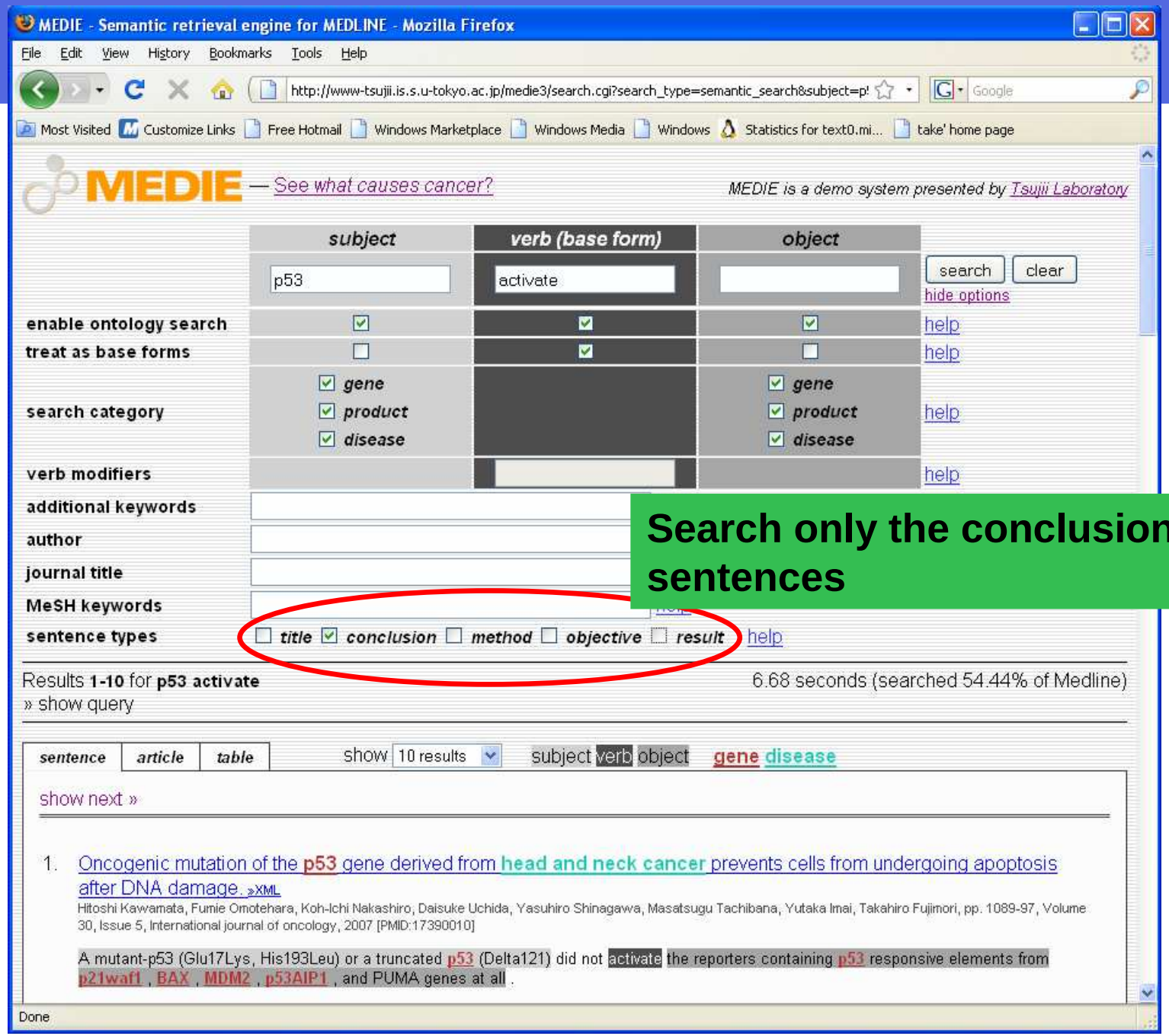

MEDIE - Semantic retrieval engine for MEDLINE - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www-tsujii.is.s.u-tokyo.ac.jp/medie3/search.cgi
Most Visited Customize Links Free Hotmail Windows Marketplace Windows Media Windows Statistics for text0.mi... take' home page
MEDIE
Semantic query based on facts
Specify the subject
subject
verb (base form)
object
p53
activate
search
clear
advanced search
What is MEDIE?
Specify the verb
Click to search!
MEDIE is an intelligent search engine to retrieve biomedical correlations from MEDLINE, based on indexing by Natural Language Processing and Text Mining techniques. You can find abstracts/sentences in MEDLINE by specifying semantics of correlations; for example, "What activates p53" and "What causes colon cancer".
What does p53 activate?
The customization of the number of results is available as in Semantic Search.
Done


MEDIE - Semantic retrieval engine for MEDLINE - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www-tsujii.is.s.u-tokyo.ac.jp/medie3/search.cgi?search_type=semantic_search&subject=p
Most Visited Customize Links Free Hotmail Windows Marketplace Windows Media Windows Statistics for text0.mi... take' home page
MEDIE — See what causes cancer?
MEDIE is a demo system presented by Tsujii Laboratory
subject
verb (base form)
object
p53
activate
search
clear
advanced search
Results 1-10 for p53 activate
0.21 seconds (searched 0.24% of Medline)
» show query
sentence
article
table
show 10 results
subject verb object
gene disease
show next »
Click to change the view
1. Oscillations by the p53-Mdm2 feedback loop. »XML
Galit Lahav, pp. 28-38, Volume 641, Advances in experimental medicine and biology, 2008 [PMID:18783169]
p53 also activates the transcription of Mdm2, which in turns target p53 for degradation, therefore creating a negative feedback loop on p53 .
2. ERK and JNK
4. Triphala inhibits both in vitro and in vivo xenograft growth of pancreatic tumor
apoptosis. »XML
Yan Shi, Ravi P Sahu, Sanjay K Srivastava, pp. 294, Volume 8, BMC cancer, 2008 [PMID:18847491]
Our data also suggests that the growth inhibitory effects of Triphala is mediated by the activation of ERK and p53 and shows potential for the
Done
the growth inhibitory effects of Triphala is mediated by the activation of ERK and p53 ...

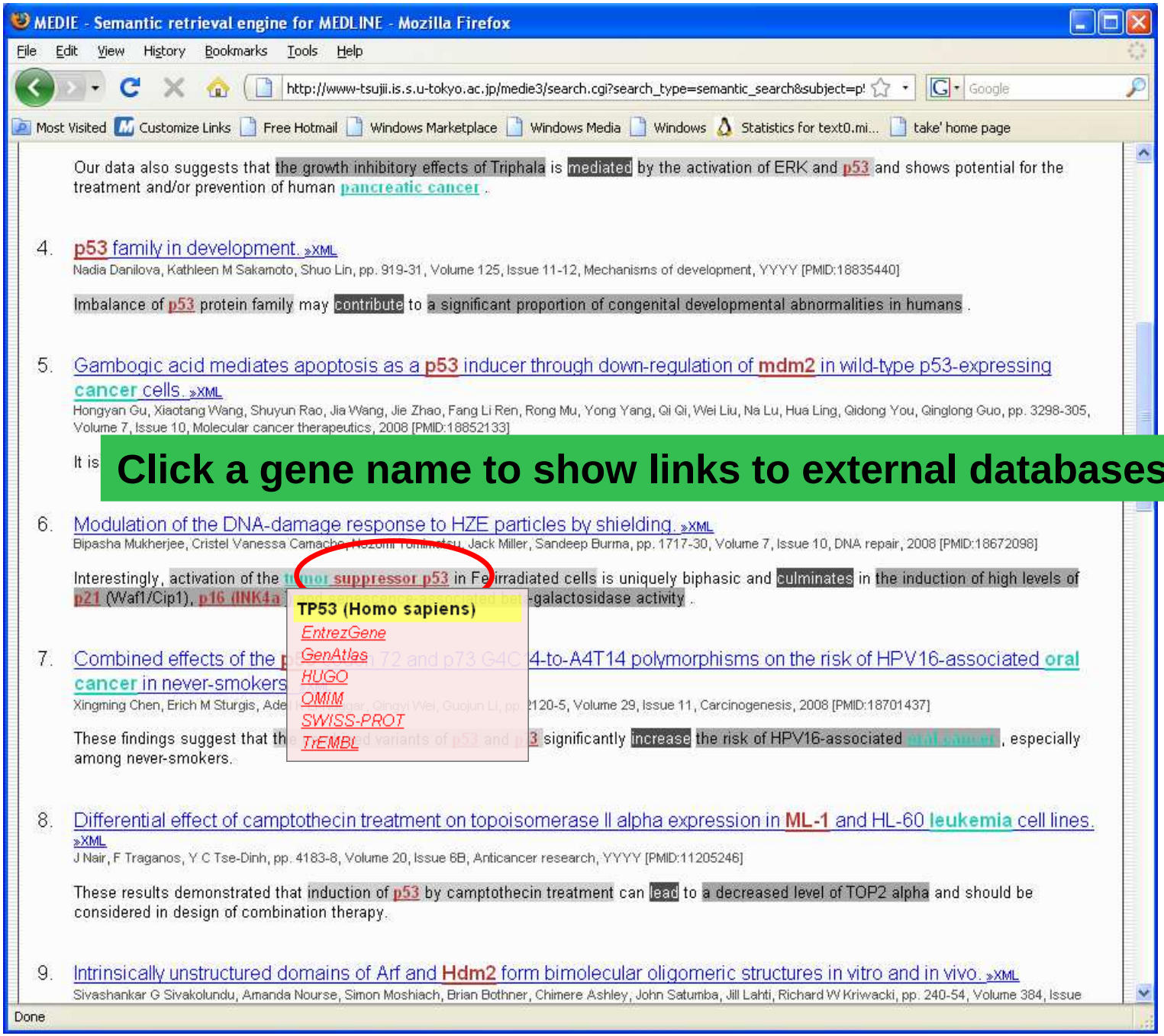
NaCTeM
The National Centre for Text Mining

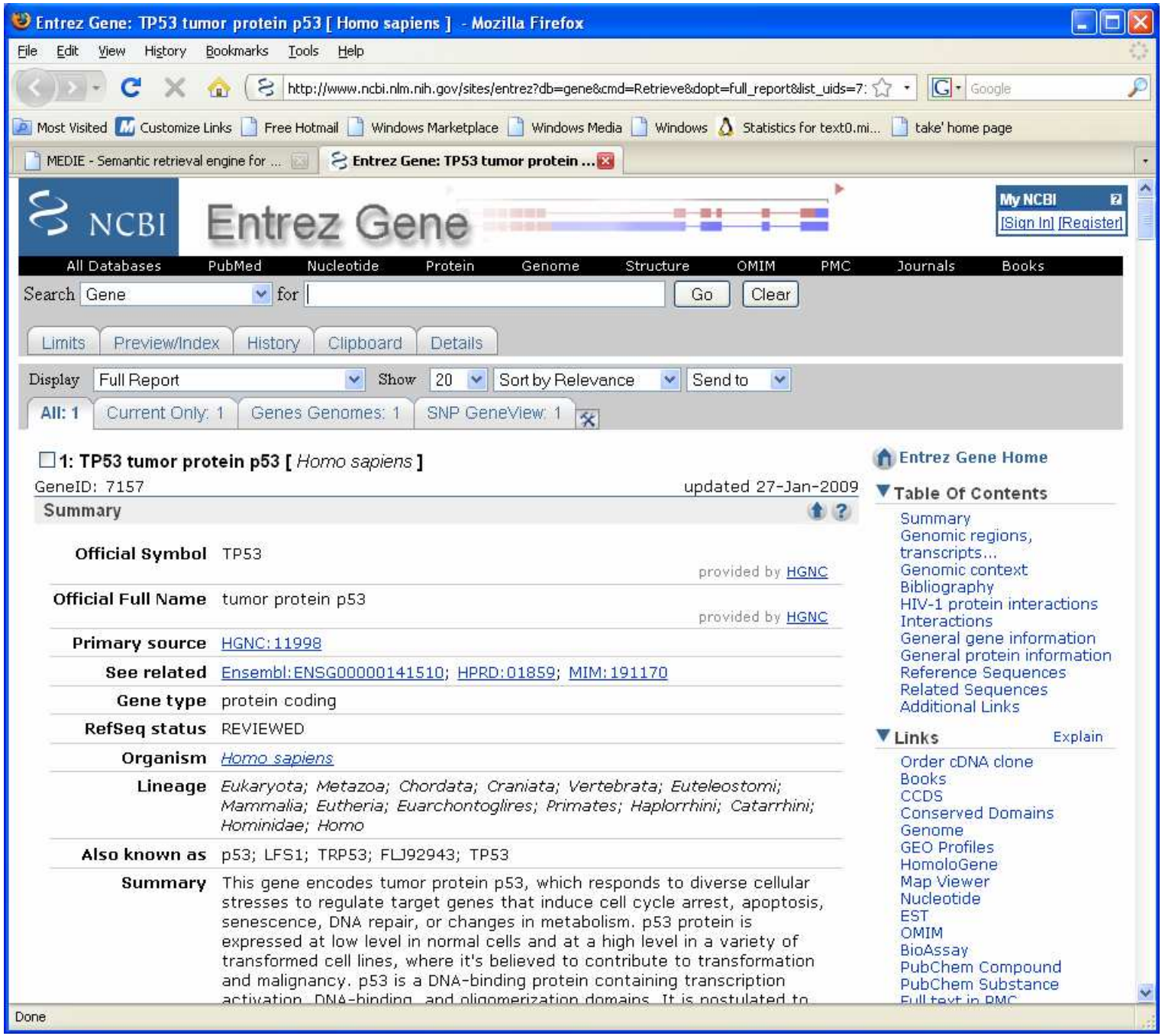


NaCTeM
The National Centre for Text Mining

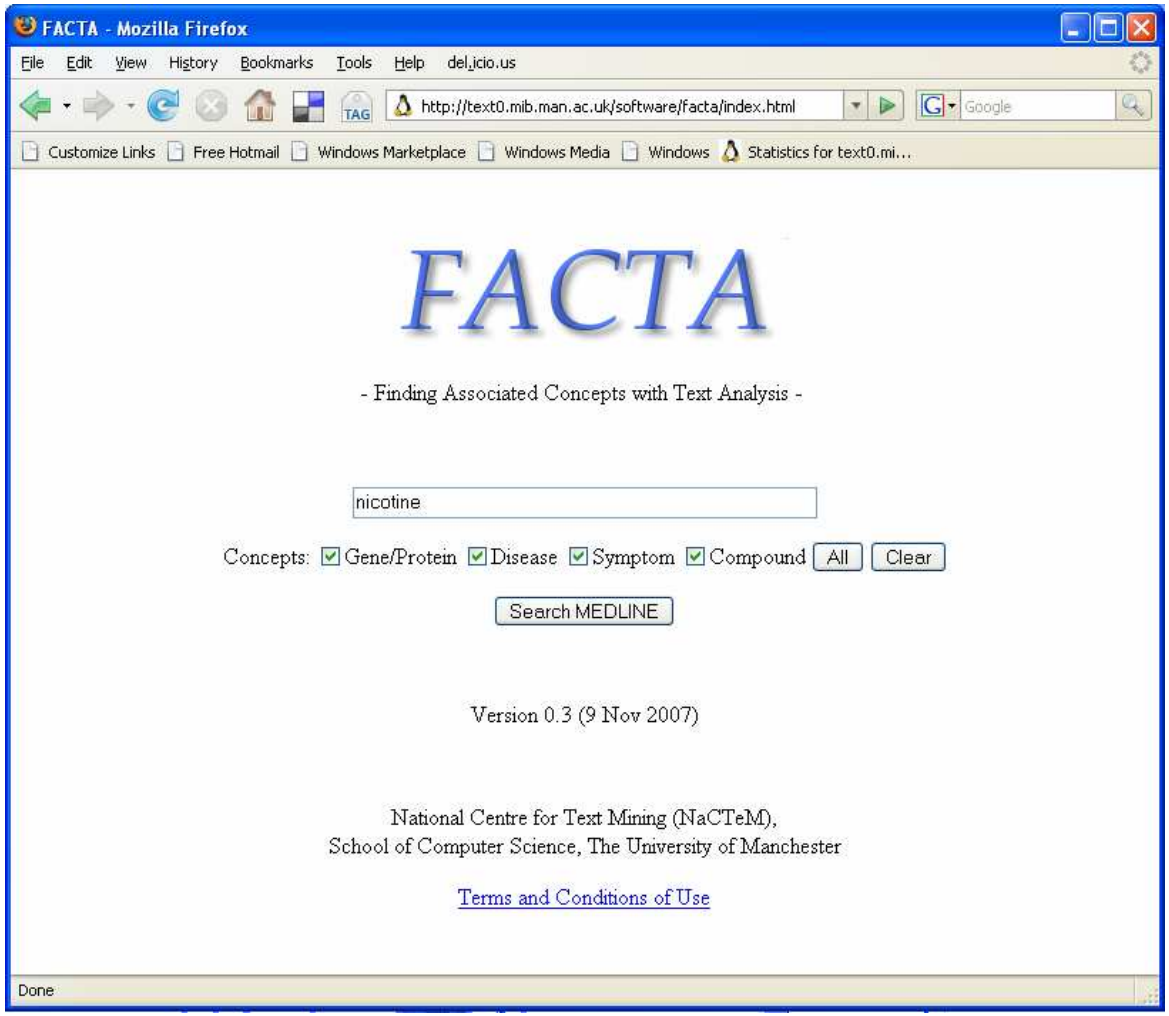


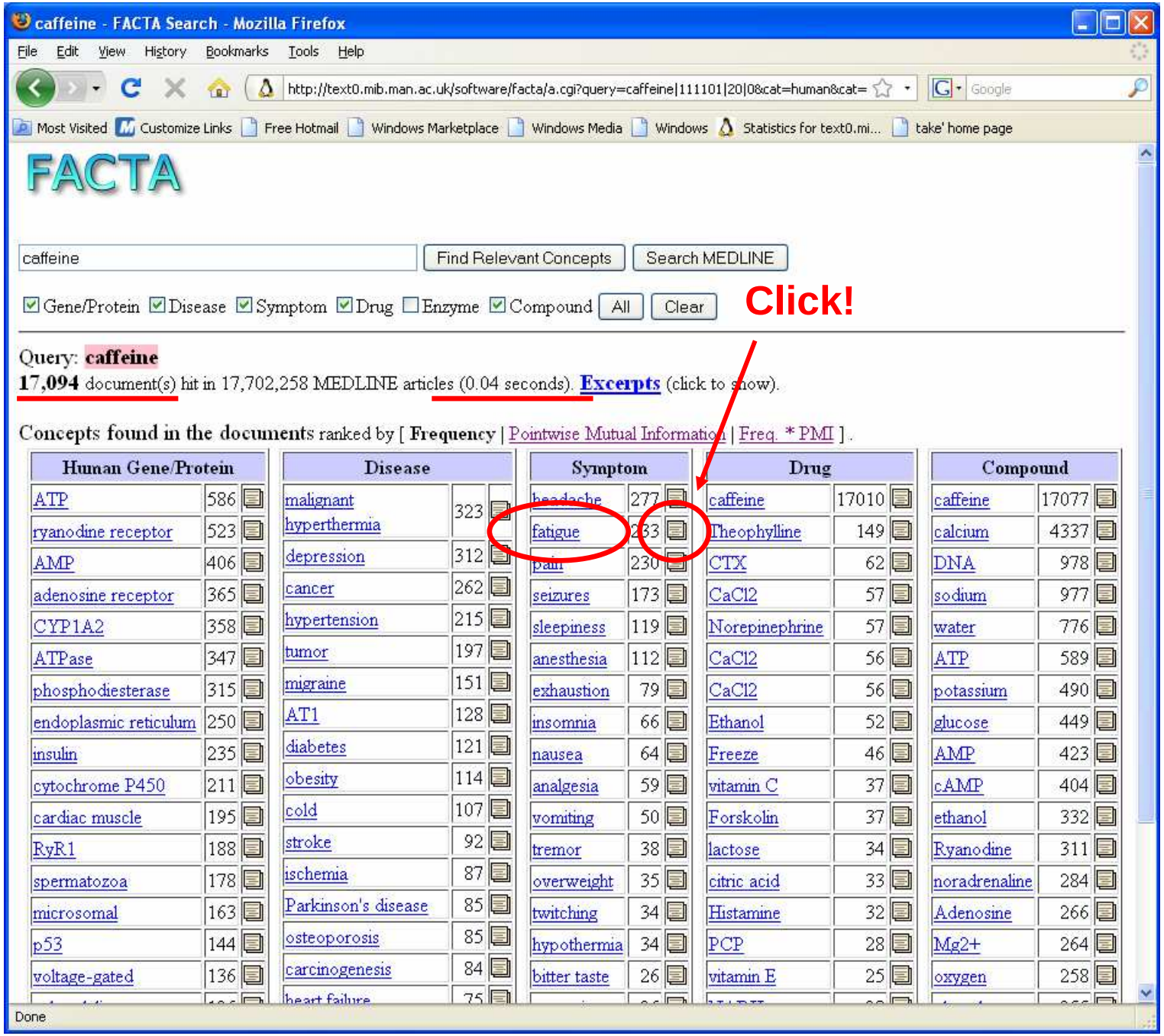
NaCTeM
The National Centre for Text Mining

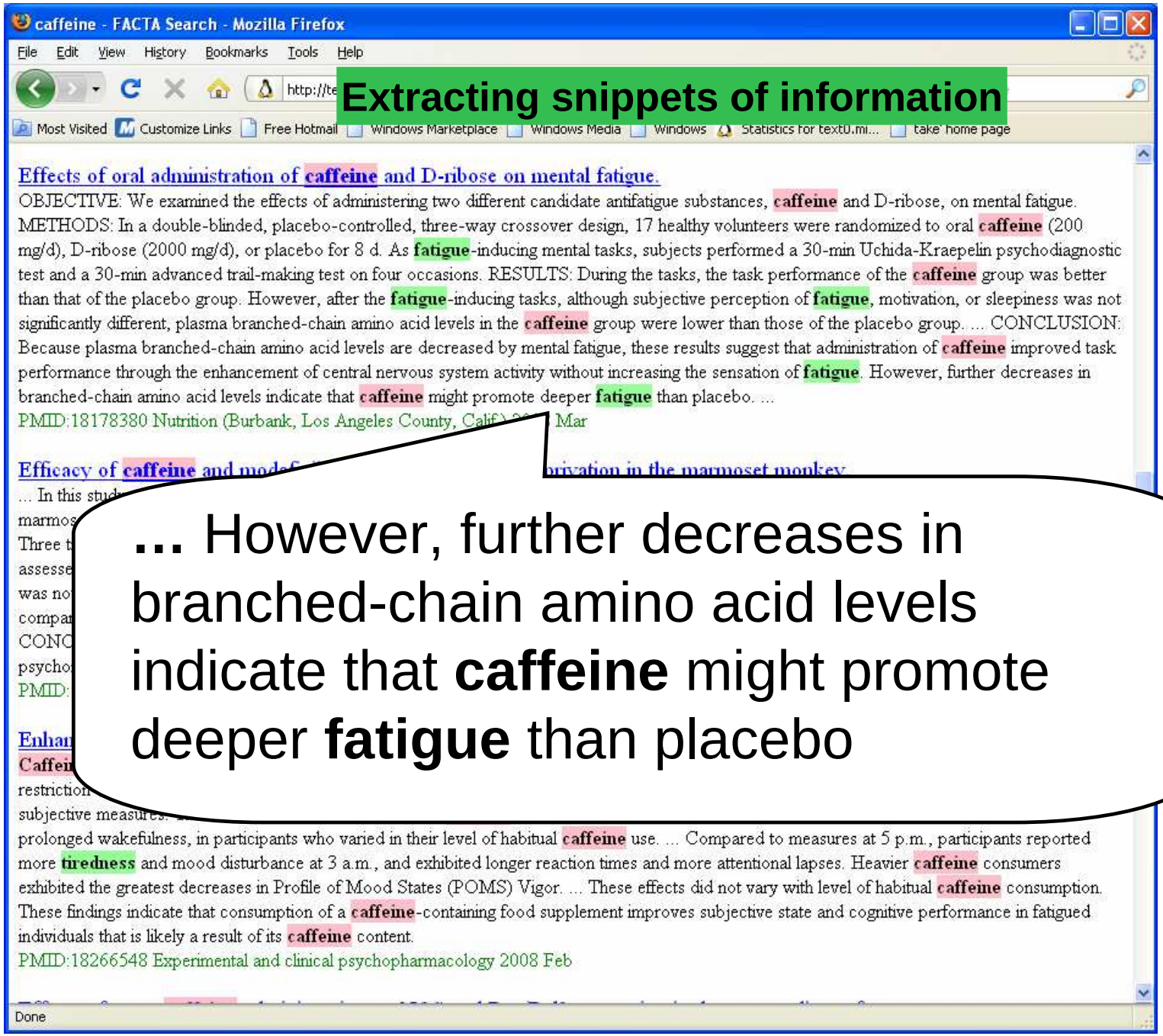

MEDIE - Semantic retrieval engine for MEDLINE - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www-tsujii.is.s.u-tokyo.ac.jp/medie3/search.cgi?search_type=semantic_search&subject=p!
Most Visited Customize Links Free Hotmail Windows Marketplace Windows Media Windows Statistics for text0.mi... take' home page
Our data also suggests that the growth inhibitory effects of Triphala is mediated by the activation of ERK and p53 and shows potential for the treatment and/or prevention of human pancreatic cancer .
4. p53 family in development. »XML
Nadia Danilova, Kathleen M Sakamoto, Shuo Lin, pp. 919-31, Volume 125, Issue 11-12, Mechanisms of development, YYYY [PMID:18835440]
Imbalance of p53 protein family may contribute to a significant proportion of congenital developmental abnormalities in humans .
5. Gambogic acid mediates apoptosis as a p53 inducer through down-regulation of mdm2 in wild-type p53-expressing cancer cells. »XML
Hongyan Gu, Xiaotang Wang, Shuyun Rao, Jia Wang, Jie Zhao, Fang Li Ren, Rong Mu, Yong Yang, Qi Qi, Wei Liu, Na Lu, Hua Ling, Qidong You, Qinglong Guo, pp. 3298-305, Volume 7, Issue 10, Molecular cancer therapeutics, 2008 [PMID:18852133]
It is
Click a gene name to show links to external databases
6. Modulation of the DNA-damage response to HZE particles by shielding. »XML
Bipasha Mukherjee, Cristel Vanessa Camacho, Jack Miller, Sandeep Burma, pp. 1717-30, Volume 7, Issue 10, DNA repair, 2008 [PMID:18672098]
Interestingly, activation of the tumor suppressor p53 in Fe irradiated cells is uniquely biphasic and culminates in the induction of high levels of p21 (Waf1/Cip1), p16 (INK4a
TP53 (Homo sapiens)
EntrezGene
GenAtlas
HUGO
OMIM
SWISS-PROT
TrEMBL
7. Combined effects of the
4-to-A4T14 polymorphisms on the risk of HPV16-associated oral cancer in never-smokers
Xingming Chen, Erich M Sturgis,
2120-5, Volume 29, Issue 11, Carcinogenesis, 2008 [PMID:18701437]
These findings suggest that th
significantly increase the risk of HPV16-associated
, especially among never-smokers.
8. Differential effect of camptothecin treatment on topoisomerase II alpha expression in ML-1 and HL-60 leukemia cell lines. »XML
J Nair, F Traganos, Y C Tse-Dinh, pp. 4183-8, Volume 20, Issue 6B, Anticancer research, YYYY [PMID:11205246]
These results demonstrated that induction of p53 by camptothecin treatment can lead to a decreased level of TOP2 alpha and should be considered in design of combination therapy.
9. Intrinsically unstructured domains of Arf and Hdm2 form bimolecular oligomeric structures in vitro and in vivo. »XML
Sivashankar G Sivakolundu, Amanda Nourse, Simon Moshiach, Brian Bothner, Chimere Ashley, John Satumba, Jill Lahti, Richard W Kriwacki, pp. 240-54, Volume 384, Issue
Done


Entrez Gene: TP53 tumor protein p53 [Homo sapiens] - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=Retrieve&dopt=full_report&list_uids=7:
Most Visited Customize Links Free Hotmail Windows Marketplace Windows Media Windows Statistics for text0.mi... take' home page
MEDIE - Semantic retrieval engine for ...
Entrez Gene: TP53 tumor protein ...
NCBI
Entrez Gene
My NCBI
[Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books
Search Gene for Go Clear
Limits Preview/Index History Clipboard Details
Display Full Report Show 20 Sort by Relevance Send to
All: 1 Current Only: 1 Genes Genomes: 1 SNP GeneView: 1
1: TP53 tumor protein p53 [Homo sapiens]
GeneID: 7157
updated 27-Jan-2009
Summary
Official Symbol TP53
provided by HGNC
Official Full Name tumor protein p53
provided by HGNC
Primary source HGNC:11998
See related Ensembl:ENSG00000141510; HPRD:01859; MIM:191170
Gene type protein coding
RefSeq status REVIEWED
Organism Homo sapiens
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as p53; LFS1; TRP53; FLJ92943; TP53
Summary This gene encodes tumor protein p53, which responds to diverse cellular stresses to regulate target genes that induce cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism. p53 protein is expressed at low level in normal cells and at a high level in a variety of transformed cell lines, where it's believed to contribute to transformation and malignancy. p53 is a DNA-binding protein containing transcription activation, DNA-binding, and oligomerization domains. It is postulated to
Entrez Gene Home
Table Of Contents
Summary
Genomic regions, transcripts...
Genomic context
Bibliography
HIV-1 protein interactions
Interactions
General gene information
General protein information
Reference Sequences
Related Sequences
Additional Links
Links Explain
Order cDNA clone
Books
CCDS
Conserved Domains
Genome
GEO Profiles
HomoloGene
Map Viewer
Nucleotide
EST
OMIM
BioAssay
PubChem Compound
PubChem Substance
Full text in PMC
Done

Mining associations from MEDLINE




caffeine - FACTA Search - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://text0.mib.man.ac.uk/software/facta/a.cgi?query=caffeine|111101|20|0&cat=human&cat=
Most Visited Customize Links Free Hotmail Windows Marketplace Windows Media Windows Statistics for text0.mi... take' home page
FACTA
caffeine
Find Relevant Concepts
Search MEDLINE
Gene/Protein Disease Symptom Drug Enzyme Compound All Clear
Click!
Query: caffeine
17,094 document(s) hit in 17,702,258 MEDLINE articles (0.04 seconds). Excerpts (click to show).
Concepts found in the documents ranked by [Frequency | Pointwise Mutual Information | Freq. * PMI] .
Human Gene/Protein
ATP 586
ryanodine receptor 523
AMP 406
adenosine receptor 365
CYP1A2 358
ATPase 347
phosphodiesterase 315
endoplasmic reticulum 250
insulin 235
cytochrome P450 211
cardiac muscle 195
RyR1 188
spermatozoa 178
microsomal 163
p53 144
voltage-gated 136
Disease
malignant hyperthermia 323
depression 312
cancer 262
hypertension 215
tumor 197
migraine 151
AT1 128
diabetes 121
obesity 114
cold 107
stroke 92
ischemia 87
Parkinson's disease 85
osteoporosis 85
carcinogenesis 84
heart failure 75
Symptom
headache 277
fatigue 233
pain 230
seizures 173
sleepiness 119
anesthesia 112
exhaustion 79
insomnia 66
nausea 64
analgesia 59
vomiting 50
tremor 38
overweight 35
twitching 34
hypothermia 34
bitter taste 26
Drug
caffeine 17010
Theophylline 149
CTX 62
CaCl2 57
Norepinephrine 57
CaCl2 56
CaCl2 56
Ethanol 52
Freeze 46
vitamin C 37
Forskolin 37
lactose 34
citric acid 33
Histamine 32
PCP 28
vitamin E 25
Compound
caffeine 17077
calcium 4337
DNA 978
sodium 977
water 776
ATP 589
potassium 490
glucose 449
AMP 423
cAMP 404
ethanol 332
Ryanodine 311
noradrenaline 284
Adenosine 266
Mg2+ 264
oxygen 258
Done

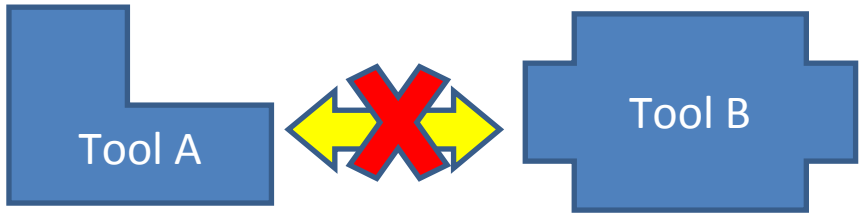

caffeine - FACTA Search - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Most Visited Customize Links Free Hotmail Windows Marketplace Windows Media Windows Statistics for textU.mi... take home page
Extracting snippets of information
Effects of oral administration of caffeine and D-ribose on mental fatigue.
OBJECTIVE: We examined the effects of administering two different candidate antifatigue substances, caffeine and D-ribose, on mental fatigue. METHODS: In a double-blinded, placebo-controlled, three-way crossover design, 17 healthy volunteers were randomized to oral caffeine (200 mg/d), D-ribose (2000 mg/d), or placebo for 8 d. As fatigue-inducing mental tasks, subjects performed a 30-min Uchida-Kraepelin psychodiagnostic test and a 30-min advanced trail-making test on four occasions. RESULTS: During the tasks, the task performance of the caffeine group was better than that of the placebo group. However, after the fatigue-inducing tasks, although subjective perception of fatigue, motivation, or sleepiness was not significantly different, plasma branched-chain amino acid levels in the caffeine group were lower than those of the placebo group. ... CONCLUSION: Because plasma branched-chain amino acid levels are decreased by mental fatigue, these results suggest that administration of caffeine improved task performance through the enhancement of central nervous system activity without increasing the sensation of fatigue. However, further decreases in branched-chain amino acid levels indicate that caffeine might promote deeper fatigue than placebo. ...
PMID:18178380 Nutrition (Burbank, Los Angeles County, Calif.)
Mar
Efficacy of caffeine and
... However, further decreases in branched-chain amino acid levels indicate that caffeine might promote deeper fatigue than placebo
prolonged wakefulness, in participants who varied in their level of habitual caffeine use. ... Compared to measures at 5 p.m., participants reported more tiredness and mood disturbance at 3 a.m., and exhibited longer reaction times and more attentional lapses. Heavier caffeine consumers exhibited the greatest decreases in Profile of Mood States (POMS) Vigor. ... These effects did not vary with level of habitual caffeine consumption. These findings indicate that consumption of a caffeine-containing food supplement improves subjective state and cognitive performance in fatigued individuals that is likely a result of its caffeine content.
PMID:18266548 Experimental and clinical psychopharmacology 2008 Feb
Done

Chemistry using Text Annotations (CheTA)

- Combining OSCAR, U-Compare, NaCTeM's TM tools, KLEIO service and Open Calais (Thompson Reuters)
- JISC funded "Metadata for Resource Generation"
- Partners
  - Peter Murray-Rust (Cambridge)
  - RSC
  - Thompson Reuters

Text Mining Tool Interoperability

- Connecting text mining tools is difficult
- Many types of annotation (words, phrases, relations…)
- Many layers of annotation (morphology, syntax, semantics…)
- U-Compare
- http://www.youtube.com/watch?v=Lo0il7UYL-M



Data format?
Data type?

Open common framework is ideal

UIMA, Unstructured Information Management Architecture (IBM)

Sharing resources

- U-compare: joint venture University of Tokyo, NaCTeM/UoM and University of Colorado
- Over 40 text mining tools are available to be used by the community…
- For specific data: Select best set of tools, observe similarities, domain adaptation
- GUIs, visualization tools

Text mining for the future of scholarly communications

- Researchers need applications enabling to:
  - Discover pertinent publications quickly
  - Annotate their collections with semantic entities and facts pertinent to their research: the living document
  - Relate entities, facts, relations extracted from literature
  - Share annotations, collections, facts and hypotheses generated from text mining