

Text Mining Activities at the National Centre

Sophia Ananiadou (Manchester)

Jun-ichi Tsujii (Manchester & Tokyo)

Paul Watry (Liverpool)

www.nactem.ac.uk

Outline

- Why text mining?
- Activities
 - Terminology Management
 - Information Extraction
 - Information Retrieval
- Core tools
 - TerMine, Medie, Info-PubMed, Cheshire
 - Taggers, Enju, term recogniser, annotation tools

How can text mining help?

- text mining can aid domain experts by automatically:
 - distilling information
 - extracting facts
 - discovering implicit links
 - generating hypotheses

Entities and concepts

- Extraction of named entities (names of organisations, people, locations), technical terms for any domain
 - Discovery of concepts allows **semantic annotation** of documents
 - Improves information access by going beyond index terms, enabling semantic querying
- Construction of **concept networks** from text
 - Allows clustering, classification of documents
 - Visualisation of knowledge maps

Relationships

- Extraction of relationships (**events and facts**) for knowledge discovery
 - Information extraction, more sophisticated annotation of texts
 - Beyond named entities: facts, events
 - Fact James Smith, Chief Research Scientist of XYZ Co.*
 - Event XYZ Co. announced the appointment of James Smith as Chief Research Scientist on 4th August 2005*
 - Enables more advanced semantic querying

Text Mining Tasks and Resources

- **Information retrieval**
 - Gather, select, filter, documents that may prove useful
- **Information extraction**
 - Partial, shallow, deep language analysis
 - Find relevant entities, facts about entities
- **Data Mining**
 - Discover new knowledge by associations
- **Resources:** ontologies, lexicons, terminologies, thesauri, grammars, annotated corpora
- **Tools:** tokenisers, taggers, chunkers, parsers, NE recognisers, semantic analysers

Term mining steps

Term recognition

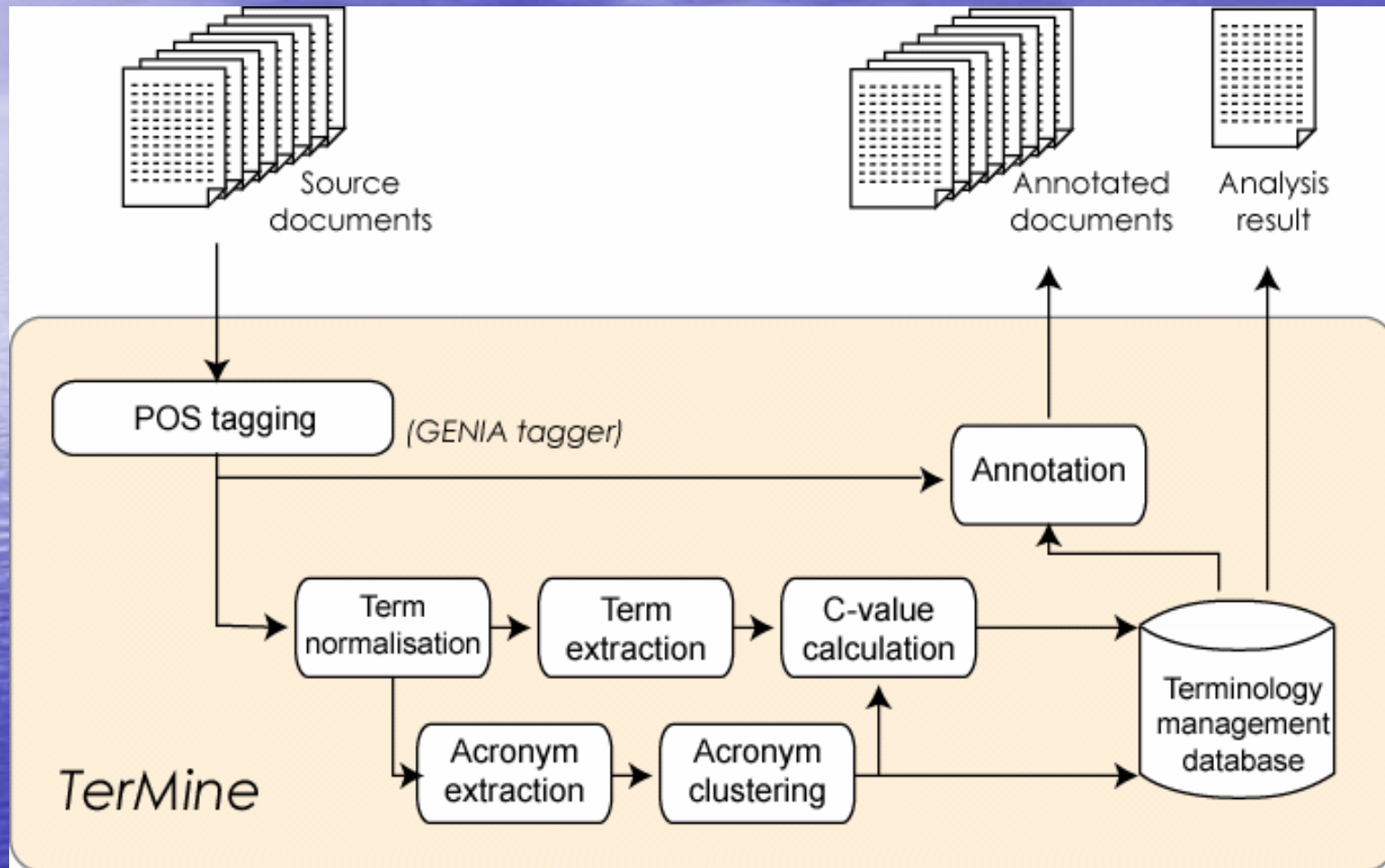


Term classification



Term mapping

TerMine: a term management system



DEMO

<http://www-tsuji.is.s.u-tokyo.ac.jp/termine/>



Source documents

Beta-arrestin binding to the beta2-adrenergic receptor requires both receptor phosphorylation and receptor activation.

Krasel C, Bunemann M, Lorenz K, Lohse MJ.

Institute for Pharmacology and Toxicology, Versbacher Strasse 9, D-97078 Wurzburg, Germany.

Homologous desensitization of beta2-adrenergic receptors has been shown to be mediated by phosphorylation of the agonist-stimulated receptor by G-protein-coupled receptor kinase 2 (GRK2) followed by binding of beta-arrestins to the phosphorylated receptor. Binding of beta-arrestin to the receptor is a prerequisite for subsequent receptor desensitization, internalization via clathrin-coated pits, and the initiation of alternative signaling pathways. In this study we have investigated the interactions between receptors and beta-arrestin2 in living cells using fluorescence resonance energy transfer. We show that (a) the initial kinetics of beta-arrestin2 binding to the receptor is limited by the kinetics of GRK2-mediated receptor phosphorylation; (b) repeated stimulation leads to the accumulation of GRK2-phosphorylated receptor, which can bind beta-arrestin2 very rapidly; and (c) the interaction of beta-arrestin2 with the receptor depends on the activation of the receptor by agonist because agonist withdrawal leads to swift dissociation of the receptor-beta-arrestin2 complex. This fast agonist-controlled association and dissociation of beta-arrestins from prephosphorylated receptors should permit rapid control of receptor sensitivity in repeatedly stimulated cells such as neurons.

beta2-adrenergic receptor gene single-nucleotide polymorphisms are associated with rheumatoid arthritis in northern Sweden.

Xu B, Arlehang L, Rantapaa-Dahlquist SB, Lefvert AK.

Department of Immunology, American Red Cross Biomedical Research and Development, MD 20855, USA. xubiy@usa.redcross.org

The beta2-adrenergic receptor (beta2-AR) belongs to the group of G-protein-coupled receptors and is present on skeletal and cardiac muscle cells and on lymphocytes. The gene encoding beta2-AR (ADRB2) displays a moderate degree of heterogeneity in the human population and the distributions of single-nucleotide polymorphisms (SNPs) at amino acid positions 16, 27, and 164 are changed in asthma, obesity, and hypertension and in the autoimmune disease myasthenia gravis. An involvement of the beta2-AR has also been suggested in human rheumatoid arthritis (RA) and its animal model. We describe here an increased prevalence of the alleles Arg16 and Gln27 and a lower prevalence of homozygosity for Gly16 and Glu27 in patients with RA. Patients having the genotype combination GlyGly16-GlnGlu27 had higher levels of rheumatoid factor (RF) and a more active disease than other patients. Patients having the genotype Arg16-Gln27+ had higher levels of RF when compared to those having Arg16+Gln27+, and patients who were carriers of Gln27 had a more active disease than non-carriers of Gln27. Our results show an association of beta2-AR SNPs with RA in a population from the northern part of Sweden. Our study also confirms the strong linkage disequilibrium of genotypes at amino acid

Result 1 - 50 of about 1131 terms

Rank	Term	Score
1	beta2-adrenergic receptor	65.7778
2	blood pressure	16.8
3	beta2-adrenergic receptor gene	14.8496
4	single nucleotide polymorphisms	9.50977
5	adrenergic receptor	9.14286
6	Gly16 allele	8
7	A549 cells	8
8	body mass index	7.92481
9	cystic fibrosis patients	7.92481
10	protein kinase	7.625
11	cystic fibrosis	7.33333
12	confidence interval	7
13	metabolic syndrome	7
14	allelic frequency	7
15	bioluminescence resonance energy transfer	6.8

Hybrid term recognition

- C-value is a hybrid, domain independent technique
- Combines linguistic filters and statistics
 - total frequency of occurrence of string in corpus
 - frequency of string as part of longer candidate terms (nested terms)
 - number of these longer candidate terms
 - length of string (in number of words)
- Output: automatically ranked terms

Extracting term associations

- Mining associations between terms for annotation
 - Hypothesis: similar terms tend to appear in similar contexts (patterns)
- combined various sources of similarity:
 - lexical associations
 - syntactic associations
 - contextual associations
 - ontological (using external resources)

Annotation & Information Extraction

Domain Knowledge

TM tools

For authors, tools for semantic enrichment

For researchers, tools for navigating relevant text and DBs

For curators, tools for the first stage of filtering text

For publishers, tools for new models of e-journals

For educators, tools for new models of e-learning

Part-Of-Speech tagging

The peri-kappa B site mediates human immunodeficiency

DT NN NN NN VBZ JJ NN

virus type 2 enhancer activation in monocytes ...

NN NN CD NN NN IN NNS

- Assign a part-of-speech tag to each token in a sentence.

Named-Entity Recognition

We have shown that interleukin-1 (IL-1) and IL-2 control
protein protein protein
IL-2 receptor alpha (IL-2R alpha) gene transcription in
DNA
CD4-CD8-murine T lymphocyte precursors.
cell_line

- Recognize named-entities in a sentence.
 - Gene/protein names
 - Protein, DNA, RNA, cell_line, cell_type

Chunking (shallow parsing)

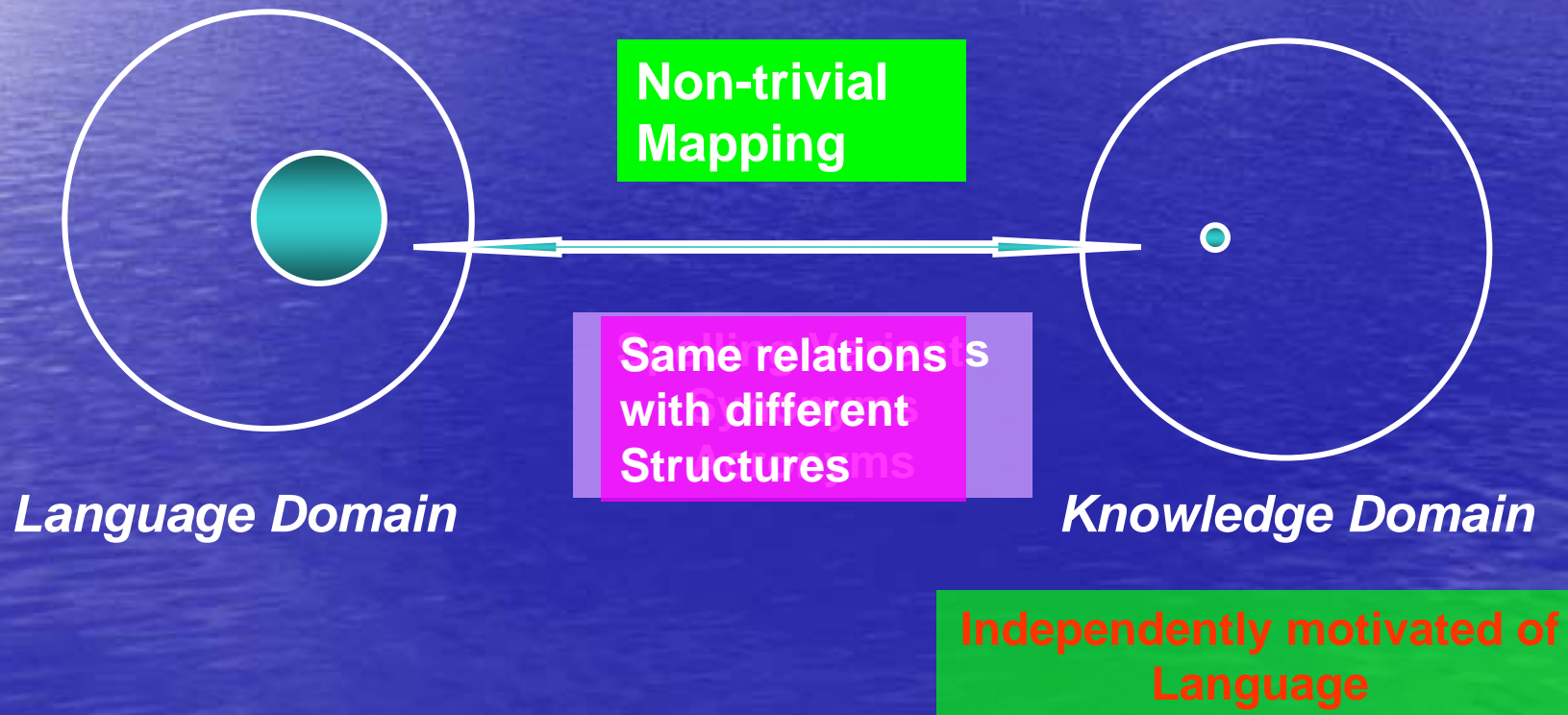
He reckons the current account deficit will narrow to
NP VP NP VP PP
only #1.8 billion in September.
NP PP NP

- A chunker (shallow parser) segments a sentence into non-recursive phrases.

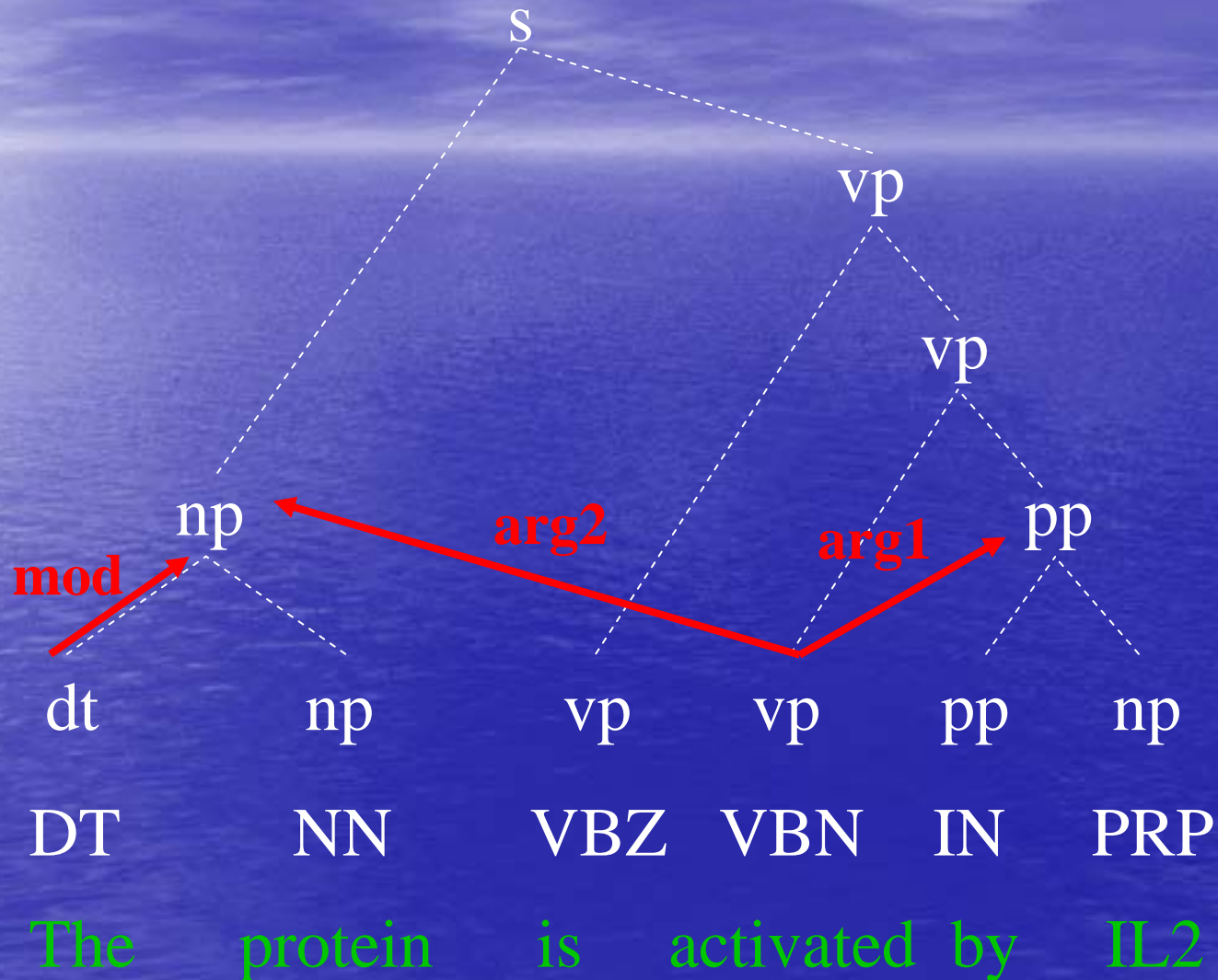
[A] protein activates [B] (Pathway extraction)

Transcription initiation by the sigma(54)-RNA polymerase holoenzyme requires an **enhancer-binding protein** that is thought to contact sigma(54) to **activate** transcription. **Full-strength** Straitee protein, we postulate, and only phosphorylated PHO2 protein could activate translation, but fails to PHO5 gene.

[sentence] > ([arg1_activate] > [protein])



Deep syntactic annotations



Our Policy for IE

- Distinguish task-independent part from task-specific part.

IE System

PAS = Predicate-Argument Structure

Task-independent

a full parser:
normalizes sentences
into PASs

Task-specific

extraction rules
on PASs

Learned automatically from corpus

Advantages of Full Parsing

- Normalization of syntactic variations into PASs

Entity1 activates Entity2

Entity2 is activated by Entity1

Entity1 cooperate to activate Entity2

Entity1 play key roles by activating Entity2

activate

ARG1 *Entity1*

ARG2 *Entity2*

We can construct more general extraction rules.

Fewer extraction rules



Smaller training corpus

MEDIE

- An interactive intelligent IR system retrieving *events*
- Performs a semantic search
- System components
 - GENIA tagger
 - Enju (HPSG parser)
 - Dictionary-based named entity recognition

DEMO <http://www-tsujii.is.s.u-tokyo.ac.jp/medie/>

Info-PubMed

- An interactive IE system and an efficient PubMed search tool, helping users to find information about biomedical entities such as genes, proteins, interactions between them.
- System components
 - MEDIE
 - Extraction of protein-protein interactions
 - Multi-window interface on a browser

Demo

<http://www-tsujii.is.s.u-tokyo.ac.jp/Info-pubmed>

Applications

- Hypothesis generation
- Sentiment analysis
- Aids to knowledge management
- Protection of the citizen (analysis of terrorism events)
- Reputation analysis
- Discovery of social impact of economic policies

Applications

- Aids for metadata creation (concepts instead of keywords)
- Semantic annotation not only based on concepts but also on facts, events extracted by IE
- Enables semantic querying
- Enables data mining

Applications

- Other text mining applications
 - Summarisation
 - Question answering
- Integration of IR with TM
 - Terms / concepts as index terms
 - Topic detection
 - Document clustering and classification

Application areas

- Geographic information through named entity recognition
- Temporal information extraction (historical documents)
- Systematic reviews
- Linking texts to factual databases

Cheshire Framework

- Combines advances in data grid, digital preservation, and digital library/text mining communities.
- Supports advanced information retrieval and knowledge generation through text mining.
- Supports presentation and reuse of documents and data.

Components

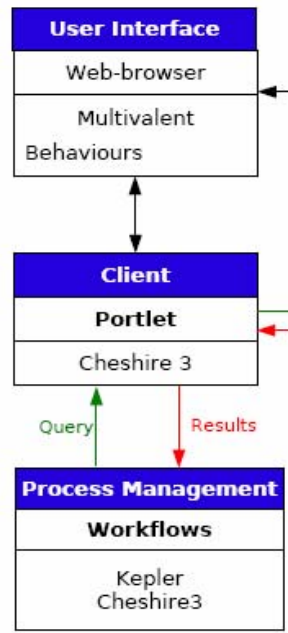
- Storage Resource Broker (SRB) to manage data collections
- Kepler workflow to ingest collections and apply eScience processes
- Cheshire system to retrieve information and Manchester/Tokyo text mining tools
- Multivalent preservation architecture to present and manipulate digital entities independently of infrastructure

Pages

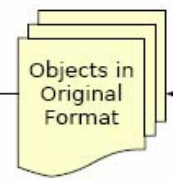
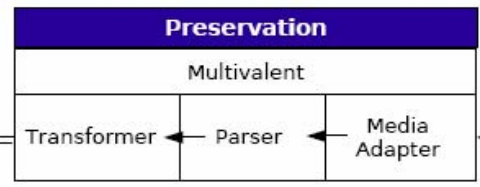
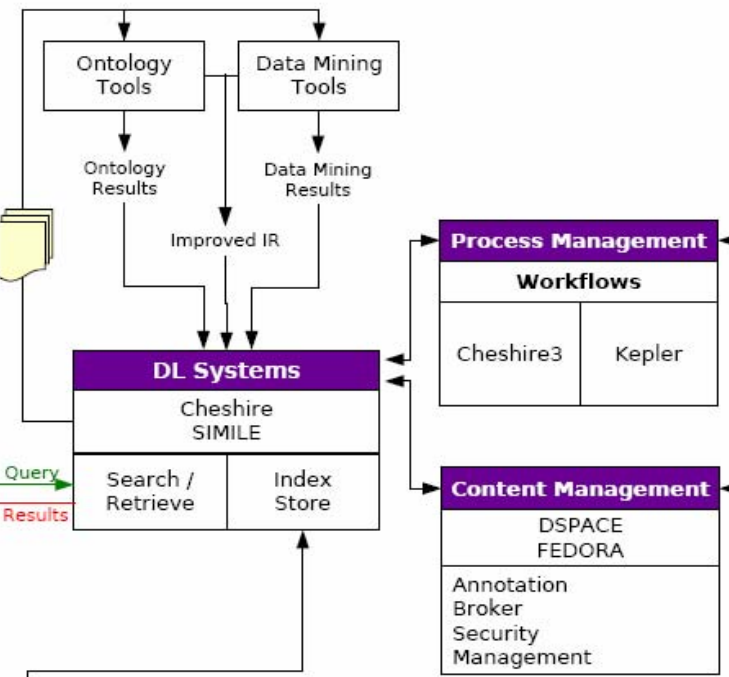
Attachments

Comments

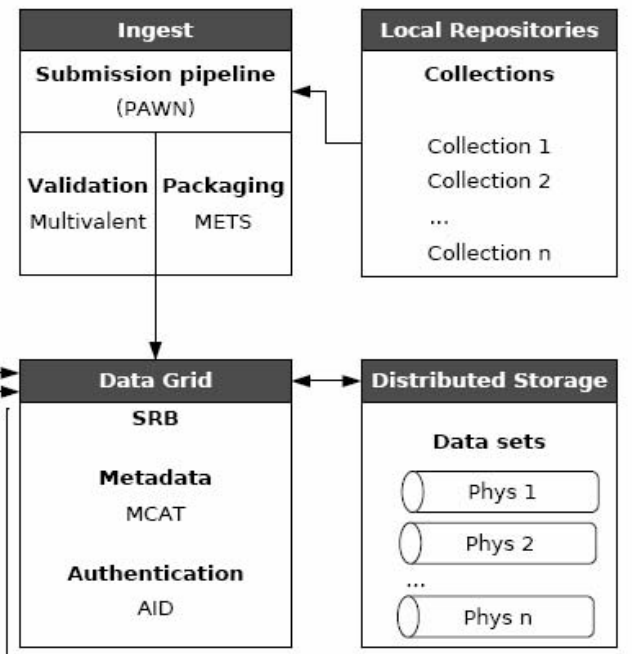
Application Specific Layer



Digital Library Layer



Data Grid Layer



Social Science eResearch agenda

- Link to the grid using Globus and OGSA
- Implement SRB for distributed access and management of resources
- Expand the concept of knowledge repository by investigating linking all types of research outputs
- Establish ontology research and application program

Grand challenge: generation of knowledge from data collections

- Involving the integration of digital libraries, data grids, and persistent archives in a common infrastructure
- Challenge: to identify and name physical relationships present within the data and then apply the discovered relationships to the processes that were used to collect or simulate the data.

Knowledge relationships

- Current infrastructure supports data management technologies
- Now we have to support equivalent abstractions for the manipulation of knowledge
- This is difficult because knowledge relationships are pervasive throughout the infrastructure

Our goal

- Build a knowledge generation infrastructure that uses abstractions to characterize the relationships in both digital entities and analysis applications so that the same infrastructure can be reapplied.