Towards Advanced Library Services

Wellcome Trust Genome Campus, Cambridge, Hinxton, UK
February 21, 2006

Workshop
“Semantic Enrichment of Scientific Literature”

Oliver Bodenreider
Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA
Disclaimer

- The project presented in this talk is being proposed as a new research initiative at the Lister Hill National Center for Biomedical Communications
- It has not been approved or reviewed by NLM yet
- The ideas presented here may not reflect NLM’s views

- In collaboration with Tom Rindflesch, NLM
Delivering Health Information

- Provide biomedical text to health care professionals and consumers
- Maintain NLM’s cutting edge
  - Support public health and healthy behavior
  - Assist clinical practice
  - Enable biomedical research and discovery
- Exploit current Library resources and advanced technology
Why additional services?

- Biomedical literature is growing at an increasingly faster pace
  - High-throughput approach to literature processing

- Integration between literature and other resources is insufficient
  - Adequate for navigating purposes
  - Insufficient for knowledge processing

- Information retrieval is the starting point, not the end of the journey for the researcher
Integration for navigation purposes


<table>
<thead>
<tr>
<th>PubMed</th>
<th>biomedical literature citations and abstracts</th>
<th>692</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubMed Central</td>
<td>free, full text journal articles</td>
<td>166</td>
</tr>
<tr>
<td>Site Search</td>
<td>NCBI web and FTP sites</td>
<td>1</td>
</tr>
</tbody>
</table>

| Books | online books | 73 |
| OMIM | online Mendelian Inheritance in Man | 27 |
| OMIA | Online Mendelian Inheritance in Animals | |

| Nucleotide | sequence database (GenBank) | 278 |
| Protein | sequence database | 160 |
| Genome | whole genome sequences | 1 |
| Structure | three-dimensional macromolecular structures | 1 |
| Taxonomy | organisms in GenBank | |
| SNP | single nucleotide polymorphism | 790 |
| Gene | gene-centered information | 35 |
| HomoloGene | eukaryotic homology groups | 19 |

| UniGene | gene-oriented clusters of transcript sequences | 17 |
| CDD | conserved protein domain database | |
| 3D Domains | domains from Entrez Structure | 3 |
| UniSTS | markers and mapping data | 45 |
| PopSet | population study data sets | |
| GEO Profiles | expression and molecular abundance profiles | 1680 |
| GEO DataSets | experimental sets of GEO data | |
| Cancer Chromosomes | cytogenetic databases | |

| none | none | none |
| none | none | none |
What additional services?

- Multi-document summarization
  - Extract and visualize the facts extracted from 250 recent abstracts on the treatment of Parkinson’s disease

- Question answering
  - Clinical and biological questions

- Knowledge discovery
  - Connect facts from heterogeneous resources

- Refined information retrieval
  - Indexing on relations in addition to concepts or association main heading/subheading
Fact-based vs. concept-based

- (concept, relationship, concept) triples are the common denominator to the various advanced services
  - Facts
  - Relations
  - Semantic predications
  - RDF triples
Biomedical knowledge repository

- **Knowledge integration**
  - Unique repository
  - Common format
  - Seamless environment
  - Phenotype and genotype information together

- **Enabling resource for the various services**
  - Summarization
  - Question answering
  - Knowledge discovery
  - Refined information retrieval
Sources of knowledge

- Biomedical literature
  - Facts extracted from MEDLINE abstracts and full-text publicly available articles using text mining techniques
  - Other corpora

- Structured databases / knowledge bases
  - NCBI resources
  - Model organism databases
  - Terminological knowledge
  - ...

- Contributed knowledge
  - The repository is open to collaborators outside NLM
Annotated knowledge

- **Provenance information**
  - Source (e.g., PMID)
  - Extraction mechanism
  - Timestamp

- **Frequency information**
  - Redundancy

- **Collaborative annotation**
  - “Was this information useful?”
  - Context of use/usefulness
Semantic Web perspective

◆ Common format for knowledge
  ● Resource Description Format (RDF)

◆ Common identification scheme
  ● Unified Resource Identifier (URI)

◆ Standard tools
  ● RDF browsers
  ● RDF “reasoners”

◆ High level of interest for biomedicine in the SW community
  ● Health Care and Life Sciences Interest Group
Biomedical Knowledge Repository

Information Retrieval

Question Answering

Knowledge Discovery

Document Summarization

Information Retrieval

Source selection

Text Mining

Terminological Knowledge

Contributed Knowledge

Other K. Sources

SemRep

Model organism annotation databases

MEDLINE

CT.gov

MetaMap

UMLS

OMIM
Towards a Biomedical Knowledge Repository
Cognitive Science Branch

- **Semantic Knowledge Representation**
  - Marcelo Fiszman
  - Halil Kilicoglu
  - François-Michel Lang
  - *Thomas Rindflesch*

- **Medical Ontology Research**
  - *Olivier Bodenreider*
  - Lee Peters
  - Lowell Vizenor
  - Kelly Zeng
Creating the repository

Enhanced Information Management for Medicine

Biomedical Knowledge Repository

PubMed

Text Mining

Medline

Clinical Trials.gov

UMLS

Structured Biomedical Data

Contributed Knowledge
Creating the repository

Semantic relations

e.g.,
Rasagilne TREATS Parkinson Disease

PubMed
Text Mining

Biomedical Knowledge Repository

Medline
Clinical Trials.gov

Enhanced Information Management for Medicine
Creating the repository

Enhanced Information Management for Medicine

Biomedical Knowledge Repository

- UMLS
- Structured Biomedical Data
- Contributed Knowledge

PubMed
Text Mining

Medline
Clinical Trials.gov
Creating the repository

Medline

ClinicalTrials.gov

UMLS

Structured Biomedical Data

Enhanced Information Management for Medicine

Biomedical Knowledge Repository

Text Mining

PubMed
Creating the repository

Enhanced Information Management for Medicine

Biomedical Knowledge Repository

PubMed
Text Mining

Medline
Clinical Trials.gov
UMLS

Entrez Gene

Contributed Knowledge

Lister Hill National Center for Biomedical Communications
Creating the repository

- PubMed
- Text Mining
- Medline
- Clinical Trials.gov
- UMLS
- Enhanced Information Management for Medicine
- Biomedical Knowledge Repository
- Genetics Home Reference
- Contributor Knowledge
Creating the repository

Enhanced Information Management for Medicine

Biomedical Knowledge Repository

OMIM

PubMed
Text Mining

Medline
Clinical Trials.gov
UMLS
Contributed Knowledge
Creating the repository

Enhanced Information Management for Medicine

Biomedical Knowledge Repository

Gene Ontology

PubMed
Text Mining

Medline
ClinicalTrials.gov
UMLS
Contributed Knowledge
Creating the repository

- PubMed
- Text Mining
- Medline
- Clinical Trials.gov
- UMLS
- Structured Biomedical Data

Enhanced Information Management for Medicine

Biomedical Knowledge Repository

Contributed Knowledge
Advanced library services
Advanced library services

- PubMed
- Text Mining
- Medline
- Clinical Trials.gov
- UMLS
- Structured Biomedical Data
- Contributed Knowledge

Knowledge Discovery

Biomedical Knowledge Repository

Lister Hill National Center for Biomedical Communications
Advanced library services

PubMed
Text Mining

Medline
ClinicalTrials.gov
UMLS
Structural Biomedical Data
Contributed Knowledge

Biomedical Knowledge Repository

Question Answering
Advanced library services

Summarization and Focused Retrieval

PubMed
Text Mining
Medline
ClinicalTrials.gov
UMLS
Structured Biomedical Data
Contributed Knowledge

Biomedical Knowledge Repository
Summarizing Biomedical Text
Summarizing Biomedical Text

- Search
  - Medline
  - ClinicalTrials.gov

- Summarize documents
  - Most salient semantic relations

- Visualize the summary

- Link the semantic relations to
  - Original text
  - Related structured knowledge
Text Mining Workflow

Information retrieval

retrieval

294 articles

summarization

Network of relations

Text mining

Parkinson disease/therapy

All Databases | PubMed | Nucleotide | Protein | Genome | Structure

Search | PubMed | for Parkinson disease/therapy | Go | Clear
Text Mining Workflow

Search: PubMed for Parkinson disease/therapy
Treatment of Parkinson’s disease

- Parkinson Disease
- Movement Disorders
- Neurodegenerative Diseases
- Deep brain Stimulation
- Gene Therapy
- Entire subthalamic nucleus
- Brain
- Antiparkinson Agents
- Antidepressive Agents
- Dopamine
- Dopamin Agonists
- Levodopa
- rasagiline
- entacapone
- pramipexol
- Catechol-O-methyltransferase inhibitor

- Dyskinetic syndrome
- Bilateral breast cancer
- Dementia
- Depressive disorder
- Anhedonia

- occurs in
- associated with
- location of
- part of
- isa
- treats
Conclusions

- Need to go beyond information retrieval
- Need to integrate multiple, heterogeneous knowledge sources to support knowledge processing, not only navigation
- Synergistic with the Semantic Web
  - Emerging standard framework
  - W3C Health Care and Life Sciences Interest Group
Medical Ontology Research

Contact: olivier@nlm.nih.gov
Web: mor.nlm.nih.gov

Olivier Bodenreider
Lister Hill National Center for Biomedical Communications
Bethesda, Maryland - USA