



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

Workshop

“Semantic Enrichment of Scientific Literature”

Towards Advanced Library Services



Olivier 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

<http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi>

The screenshot displays the NCBI Entrez search engine interface. At the top left is the NCBI logo. The main header features the Entrez logo and the text "Entrez, The Life Sciences Search Engine". Below this is a navigation bar with links for HOME, SEARCH, SITE MAP, PubMed, All Databases, Human Genome, GenBank, Map Viewer, and BLAST. A search bar contains the text "NF2" and buttons for GO, CLEAR, and Help. The search results are presented in a grid of 20 items, each with a count, an icon, a title, a description, and a help link.

Count	Icon	Database Name	Description	Help
692		PubMed	biomedical literature citations and abstracts	?
166		PubMed Central	free, full text journal articles	?
1		Site Search	NCBI web and FTP sites	?
73		Books	online books	?
27		OMIM	online Mendelian Inheritance in Man	?
none		OMIA	Online Mendelian Inheritance in Animals	?
278		Nucleotide	sequence database (GenBank)	?
160		Protein	sequence database	?
1		Genome	whole genome sequences	?
1		Structure	three-dimensional macromolecular structures	?
none		Taxonomy	organisms in GenBank	?
790		SNP	single nucleotide polymorphism	?
35		Gene	gene-centered information	?
19		HomoloGene	eukaryotic homology groups	?
17		UniGene	gene-oriented clusters of transcript sequences	?
none		CDD	conserved protein domain database	?
8		3D Domains	domains from Entrez Structure	?
45		UniSTS	markers and mapping data	?
5		PopSet	population study data sets	?
1680		GEO Profiles	expression and molecular abundance profiles	?
1		GEO DataSets	experimental sets of GEO data	?
none		Cancer Chromosomes	cytogenetic databases	?

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

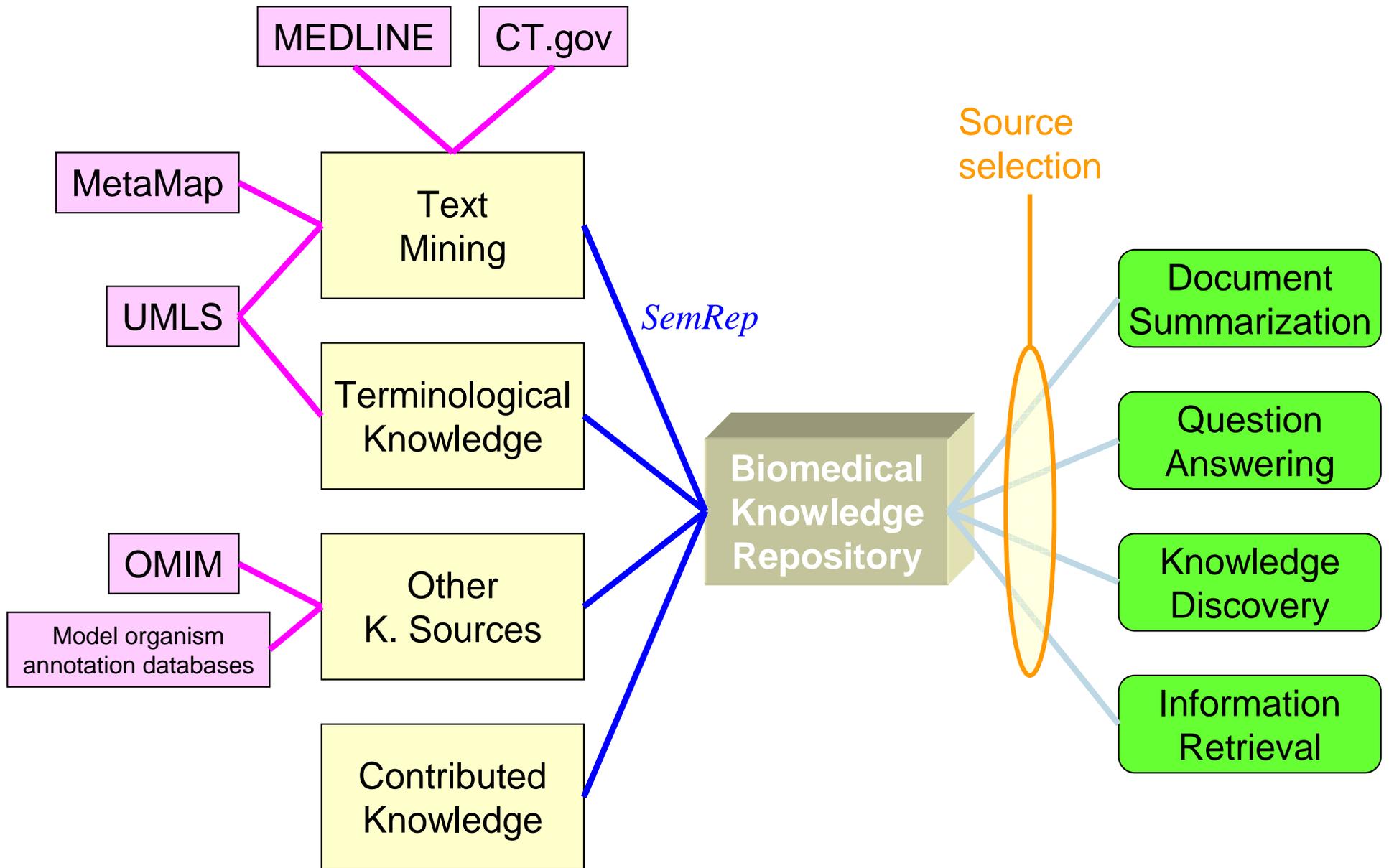
- ◆ Provenance information
 - Source (e.g., PMID)
 - Extraction mechanism
 - Timestamp
- ◆ Frequency information
 - Redundancy
- ◆ Collaborative annotation
 - “Was this information useful?”
 - Context of use/usefulness



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





Towards a
Biomedical Knowledge Repository

Cognitive Science Branch

◆ Semantic Knowledge Representation

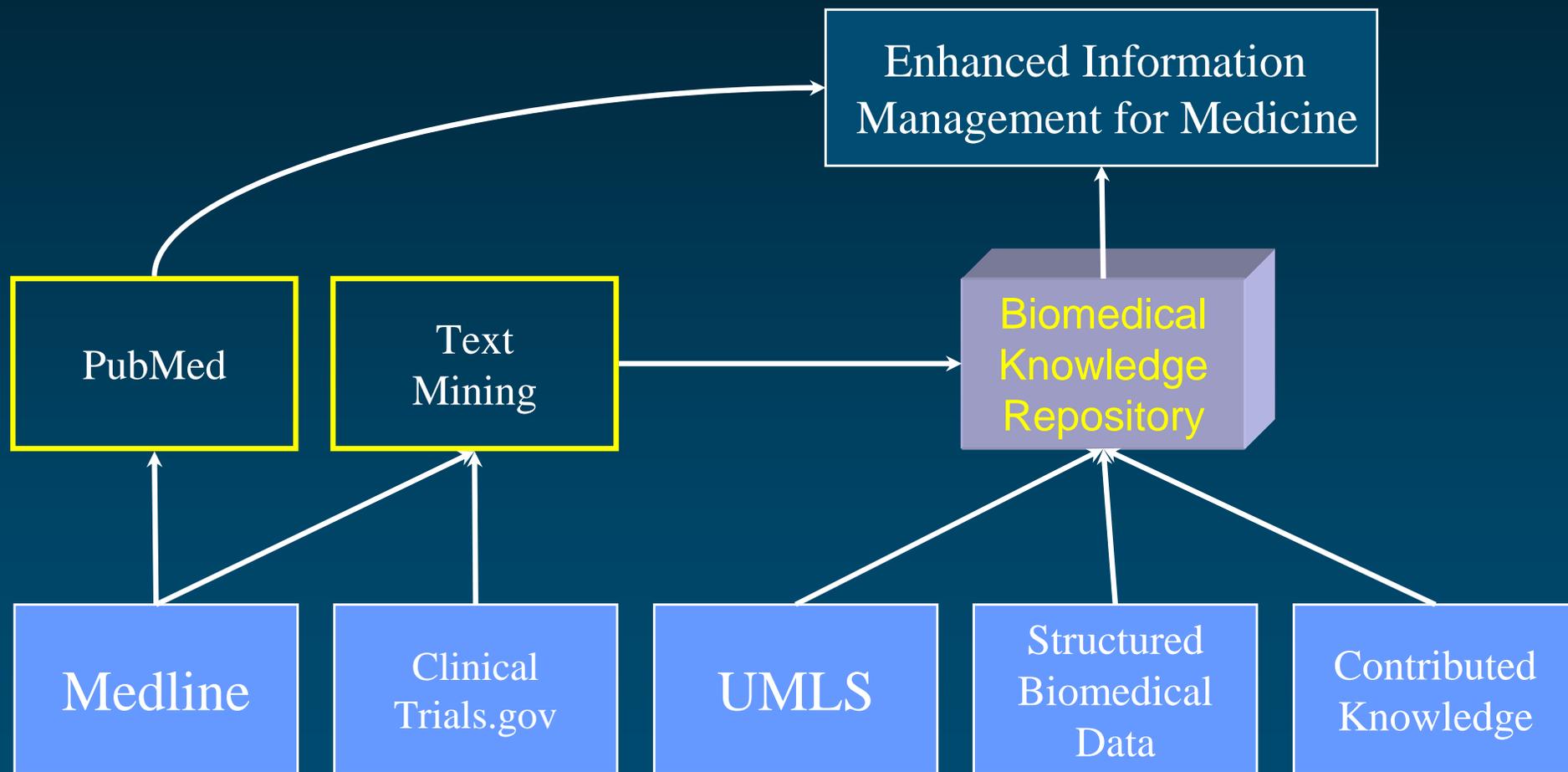
- Marcelo Fiszman
- Halil Kilicoglu
- François-Michel Lang
- *Thomas Rindflesch*

◆ Medical Ontology Research

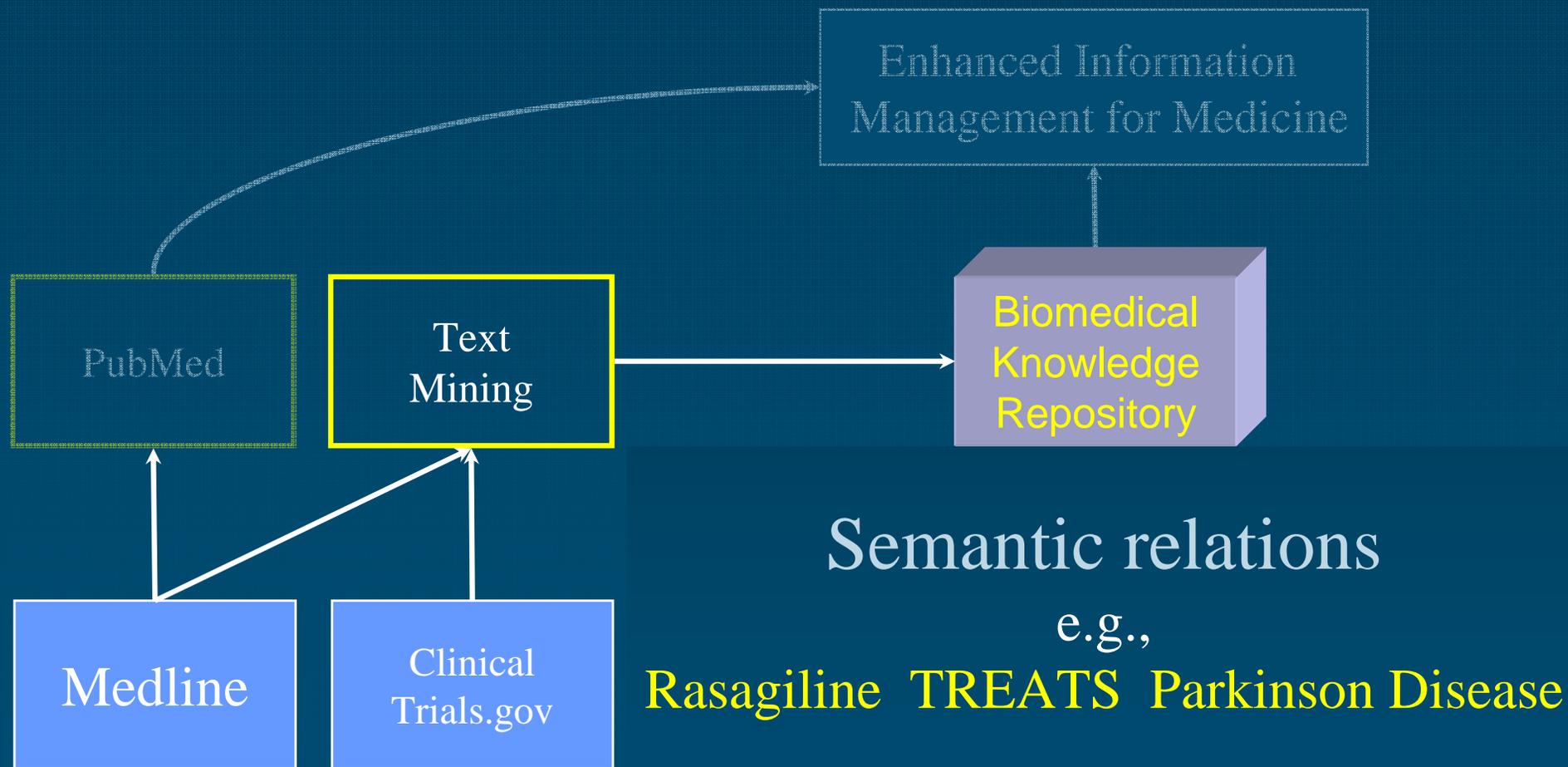
- *Olivier Bodenreider*
- Lee Peters
- Lowell Vizenor
- Kelly Zeng



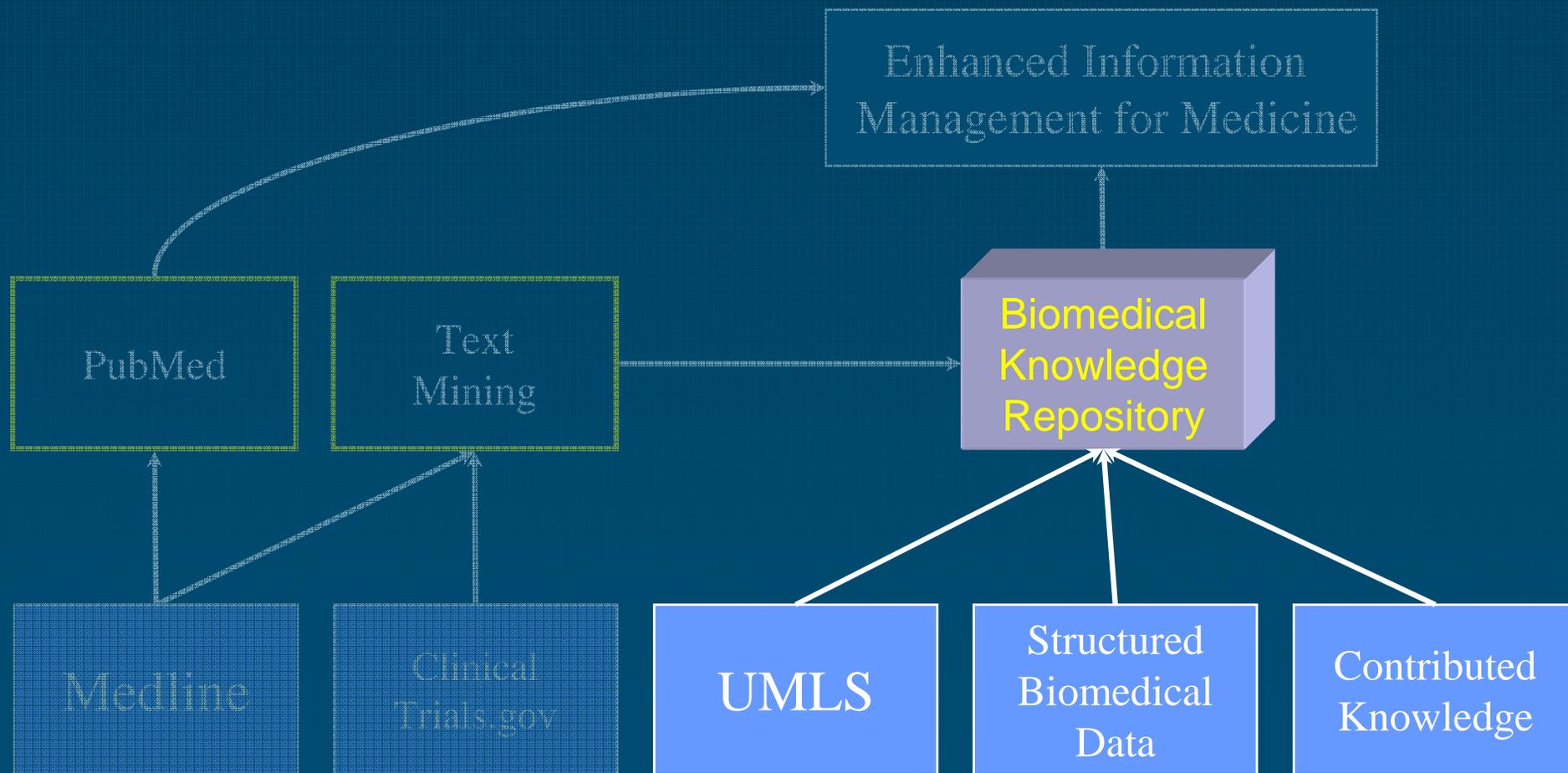
Creating the repository



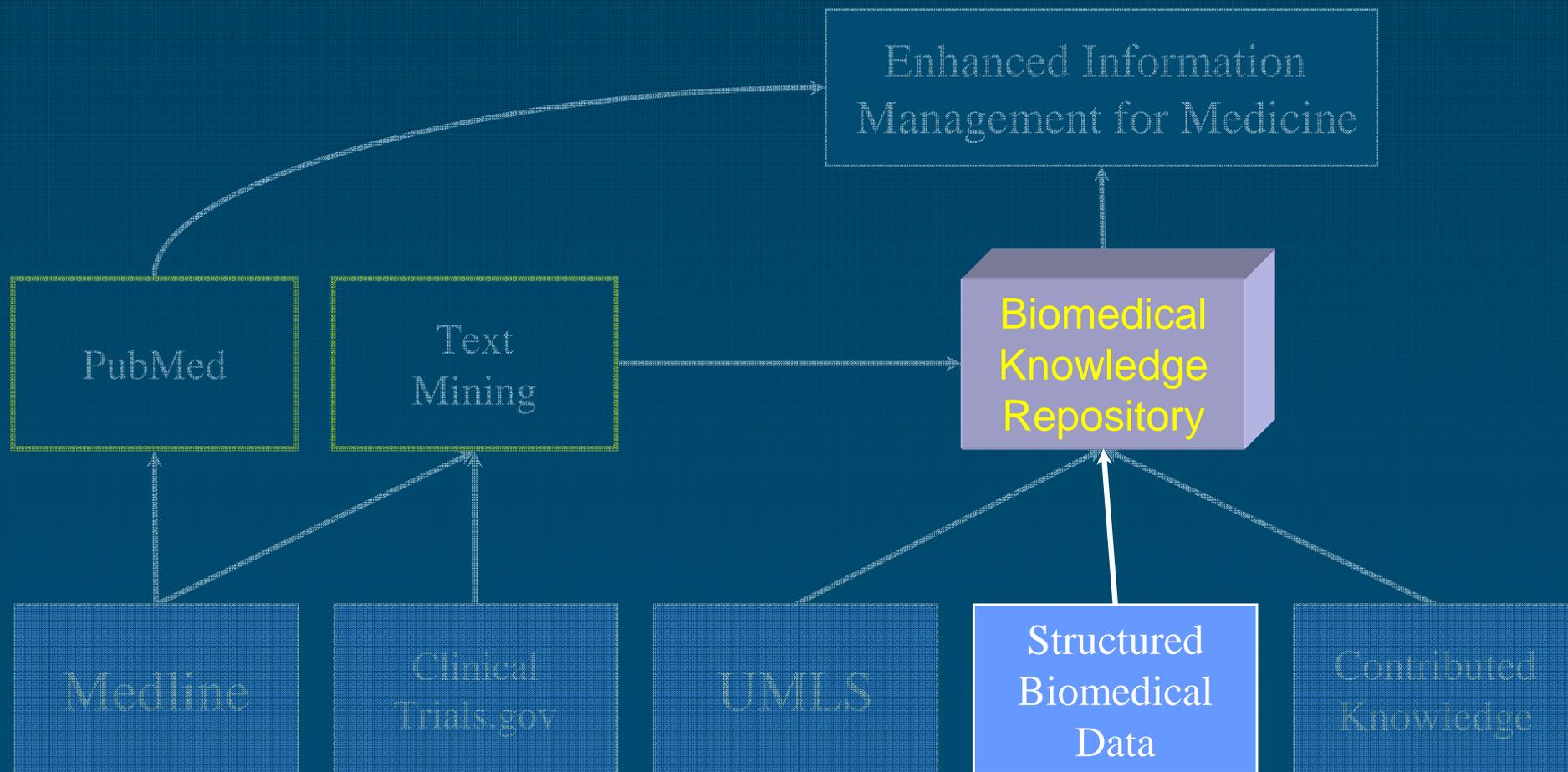
Creating the repository



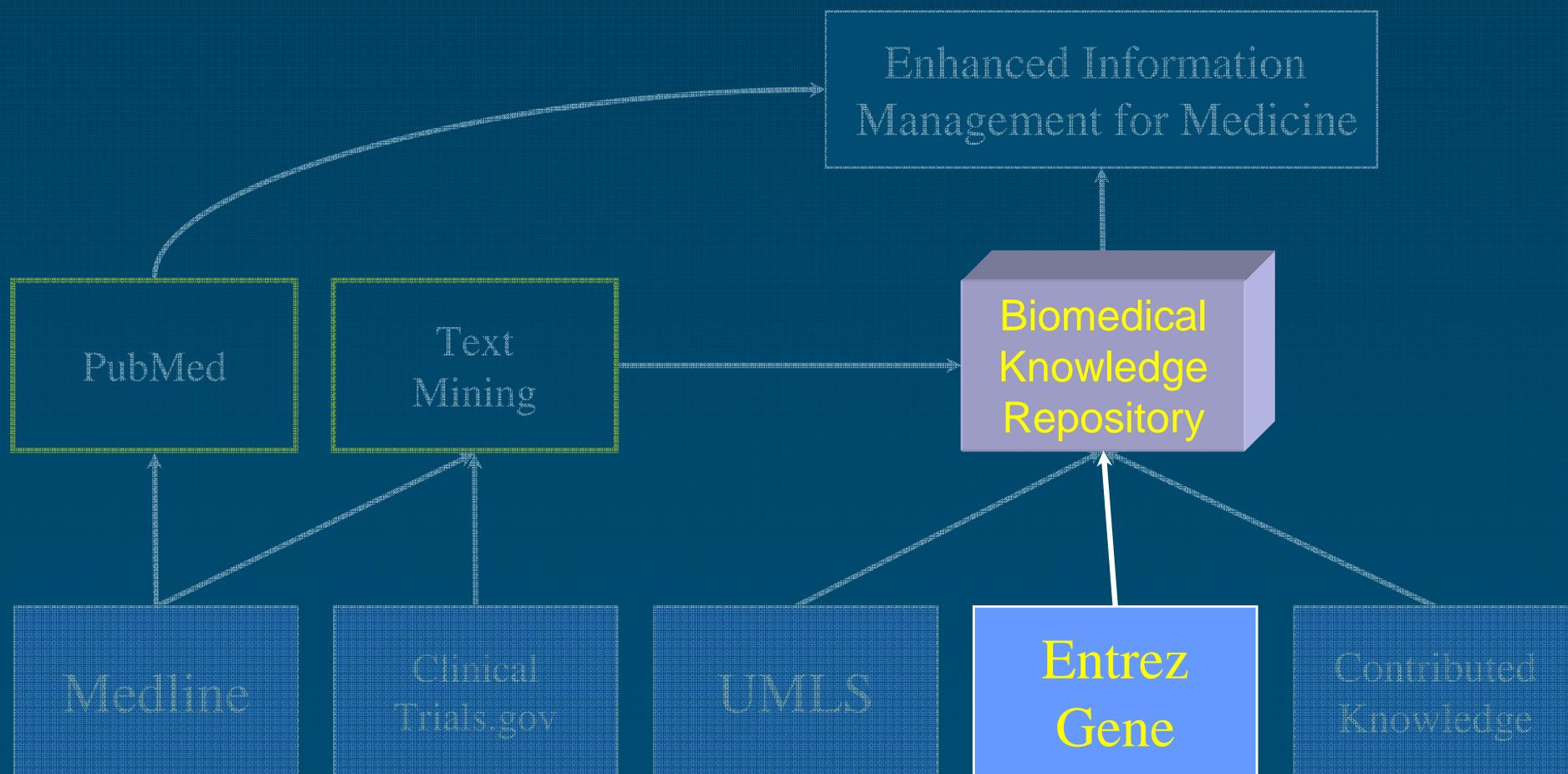
Creating the repository



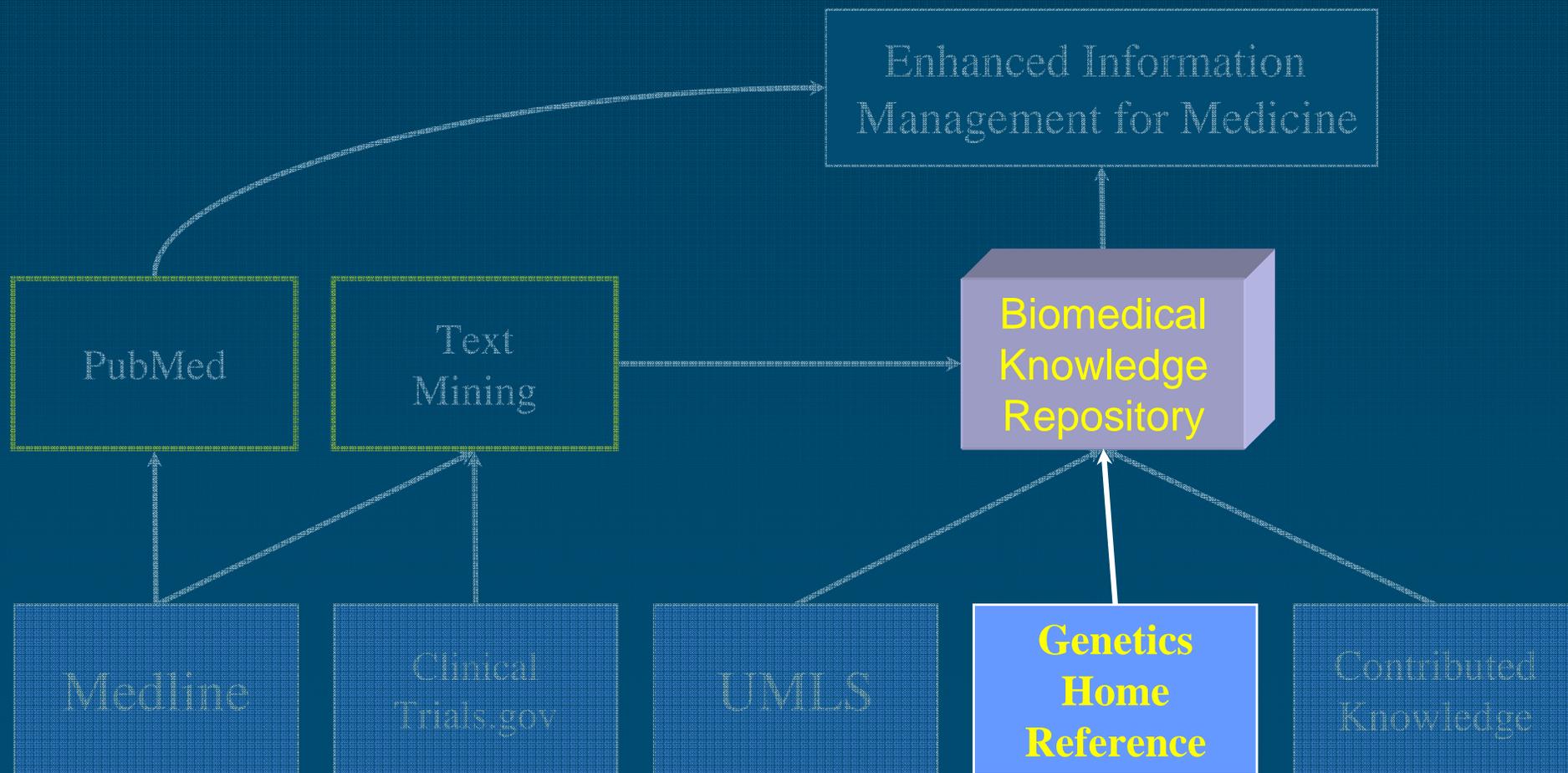
Creating the repository



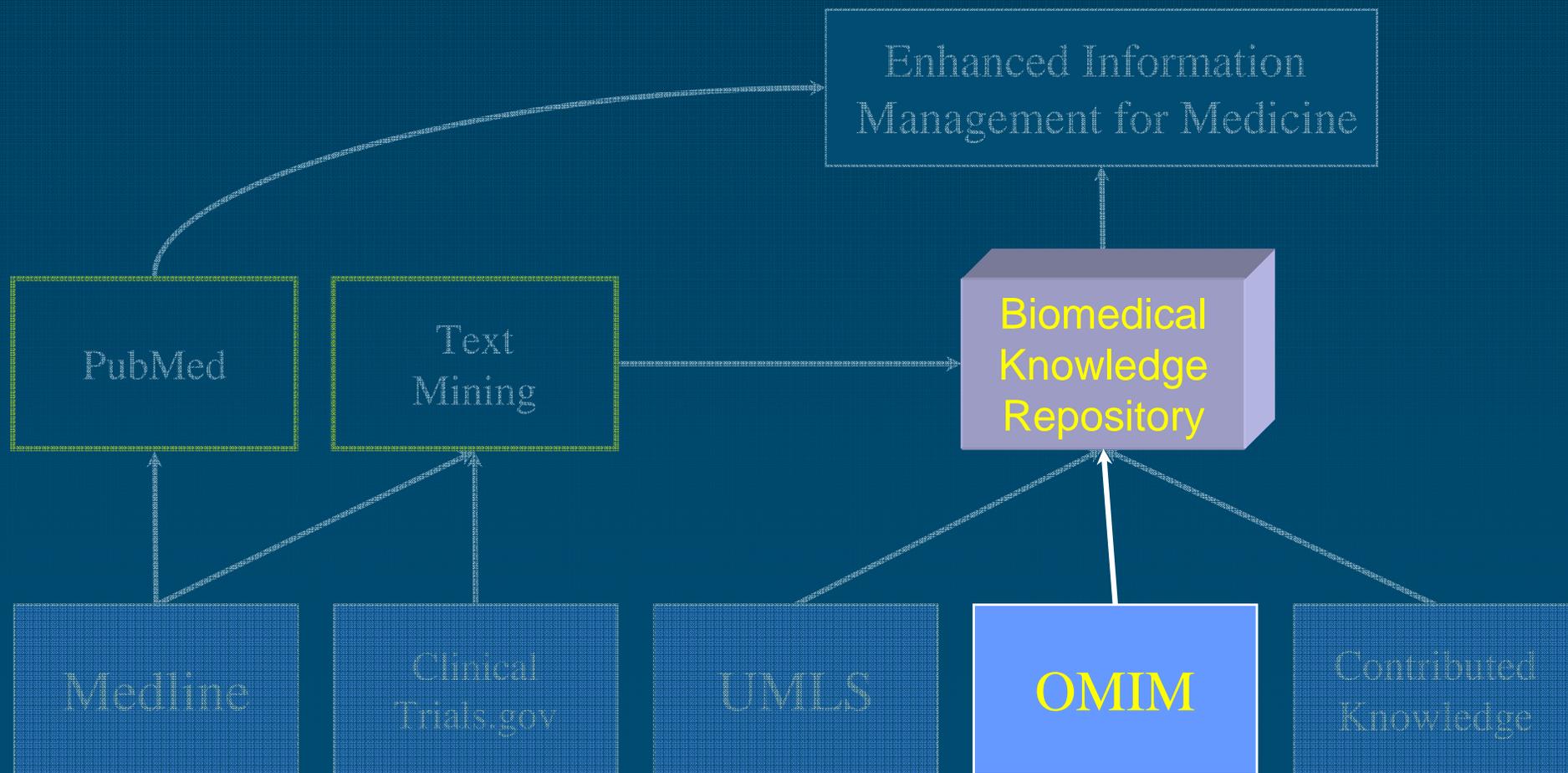
Creating the repository



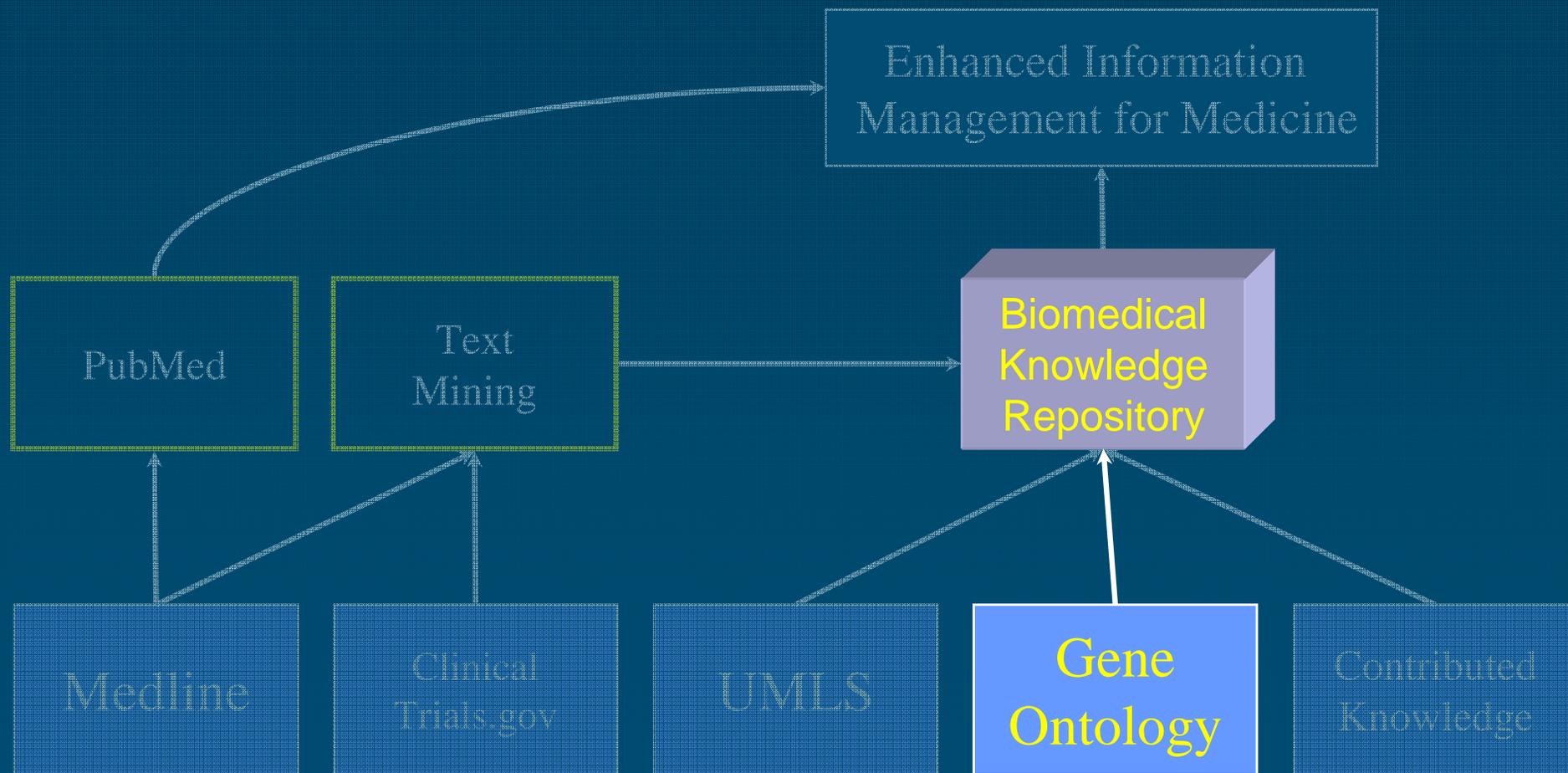
Creating the repository



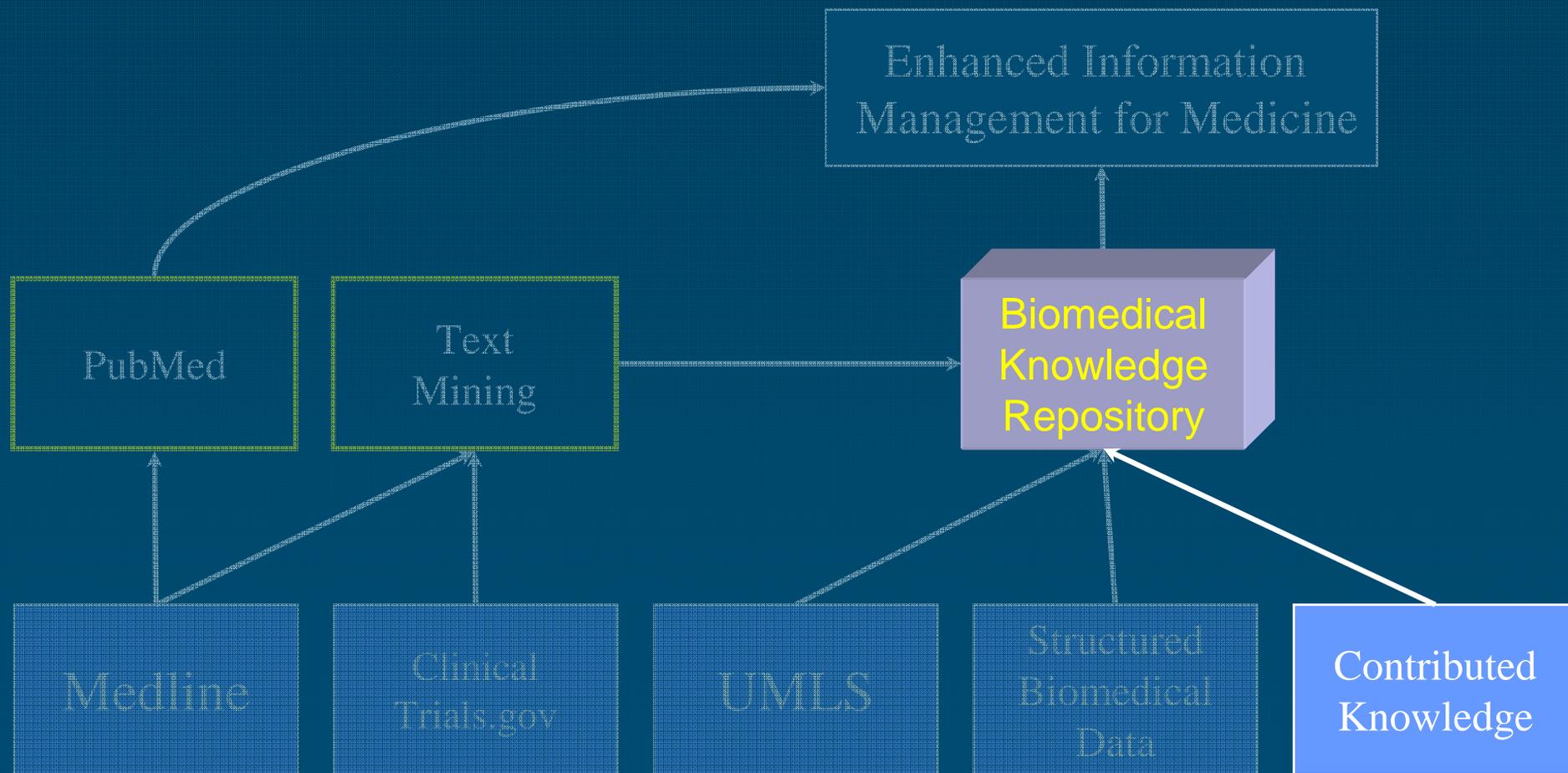
Creating the repository



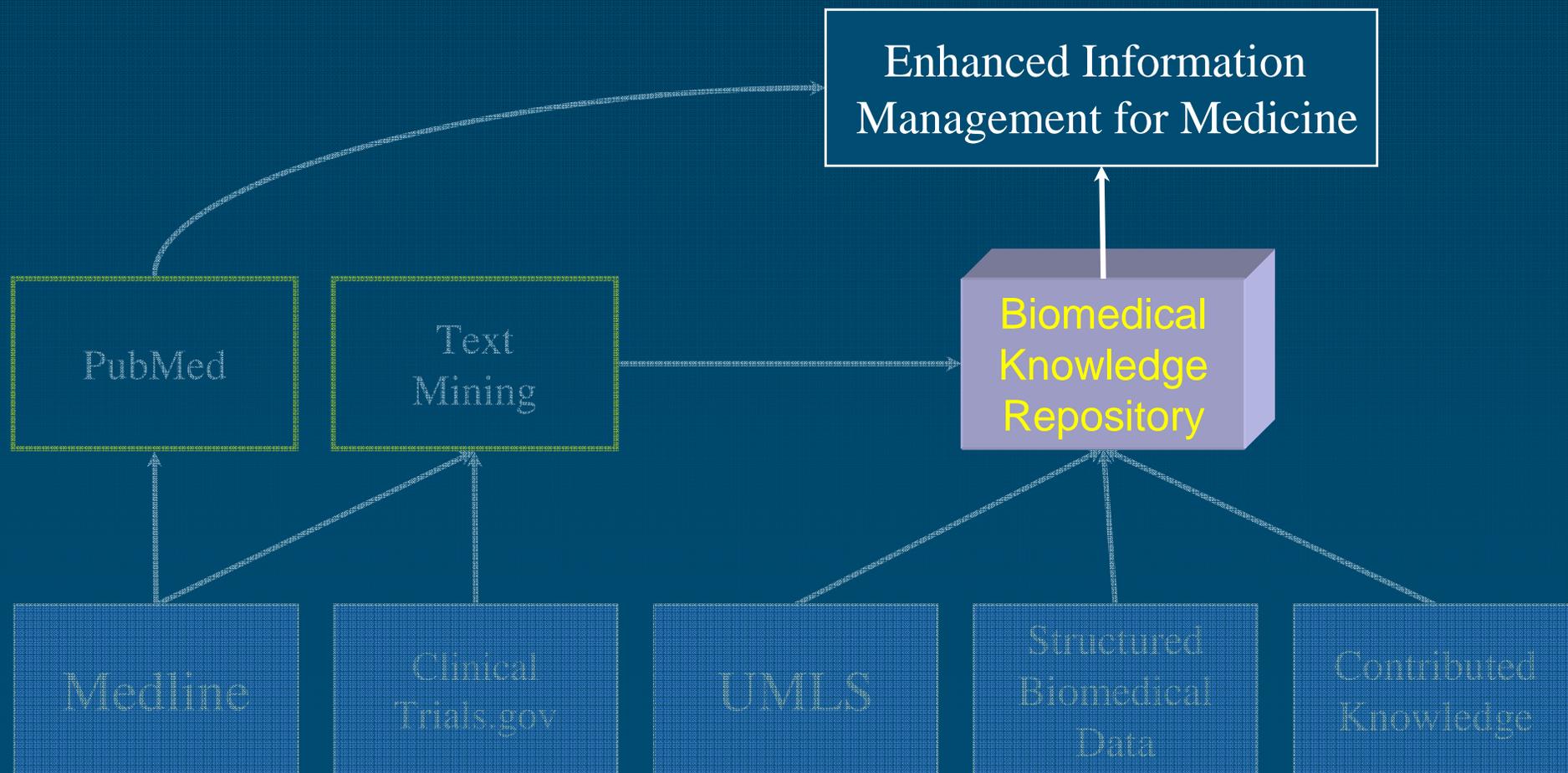
Creating the repository



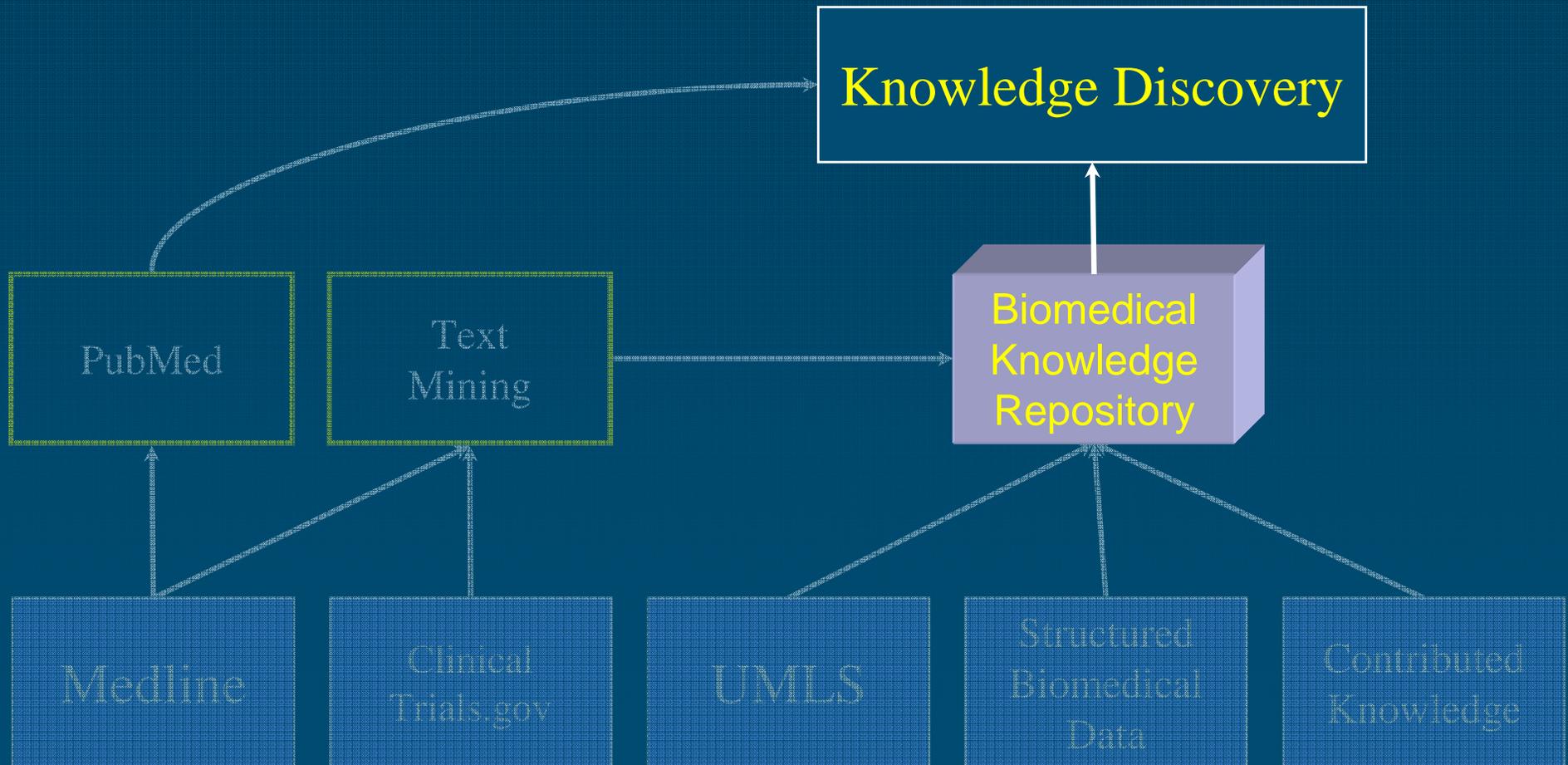
Creating the repository



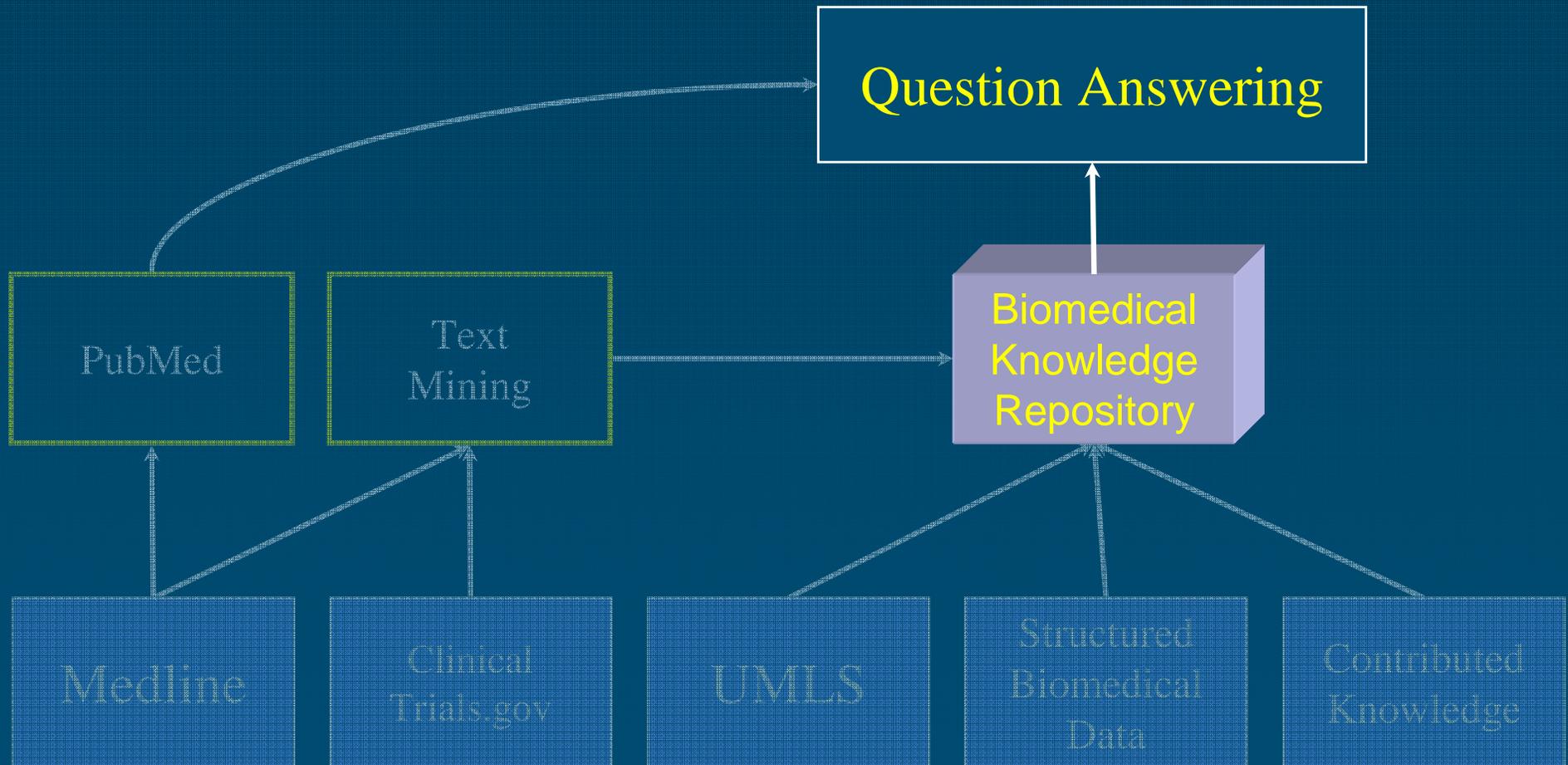
Advanced library services



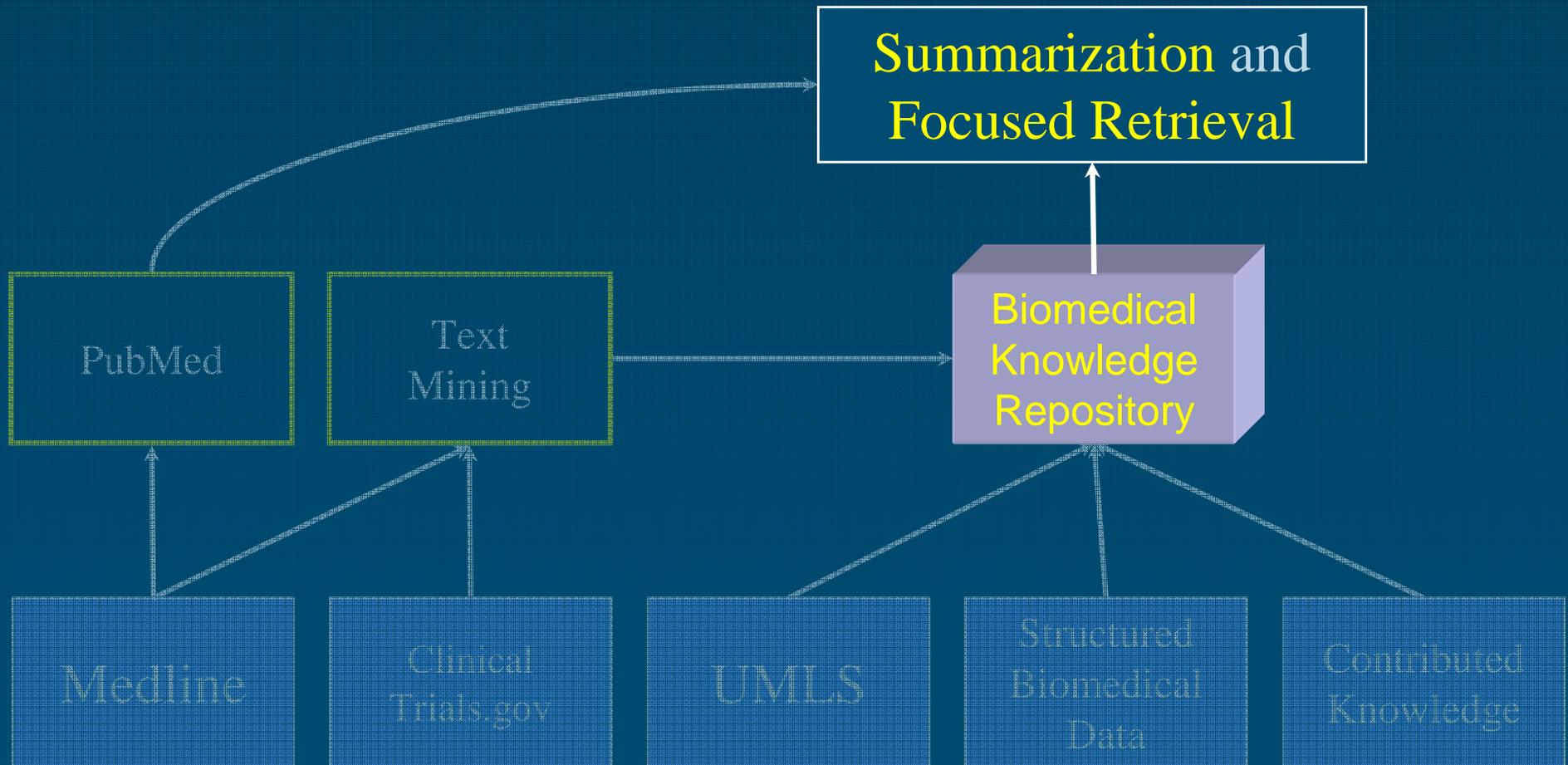
Advanced library services



Advanced library services



Advanced library services



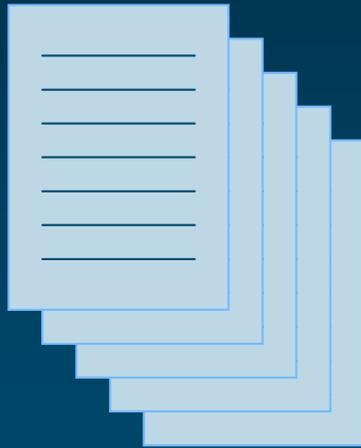
Summarizing Biomedical Text

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



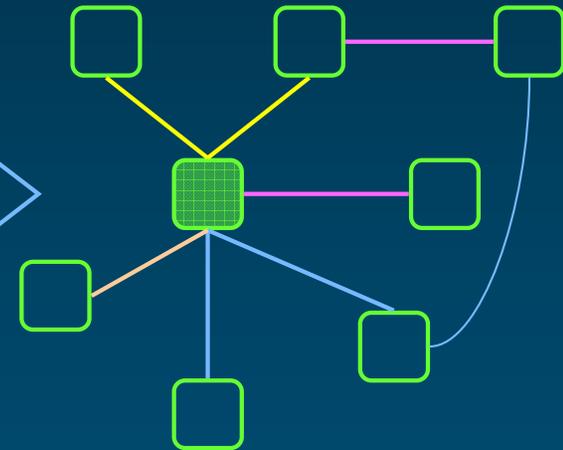
Text Mining Workflow



294 articles

Information retrieval

summarization

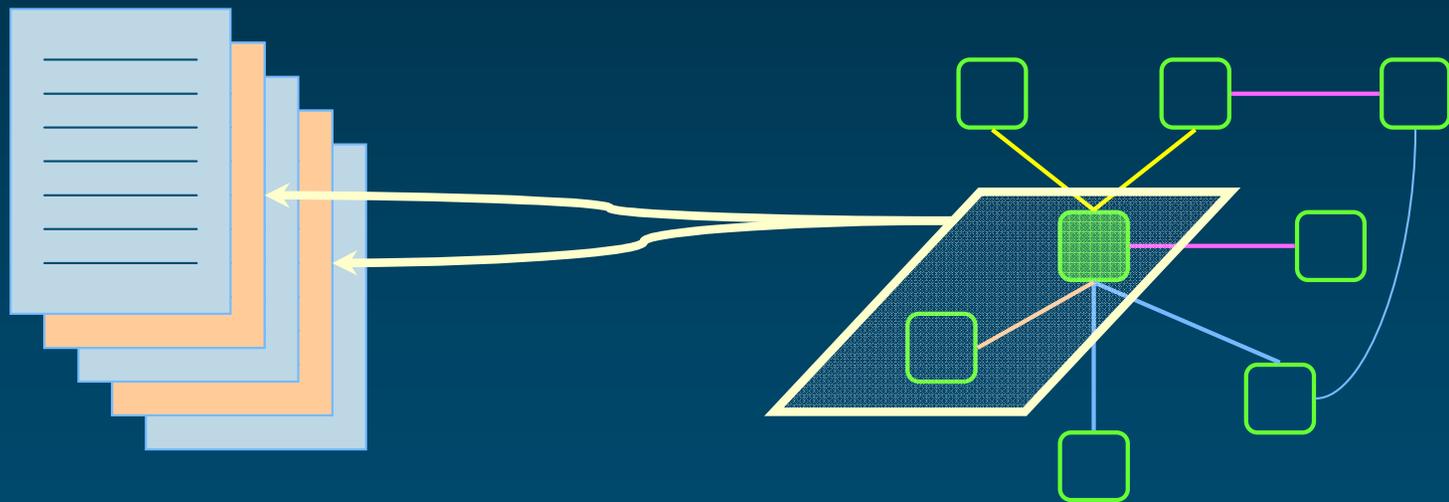


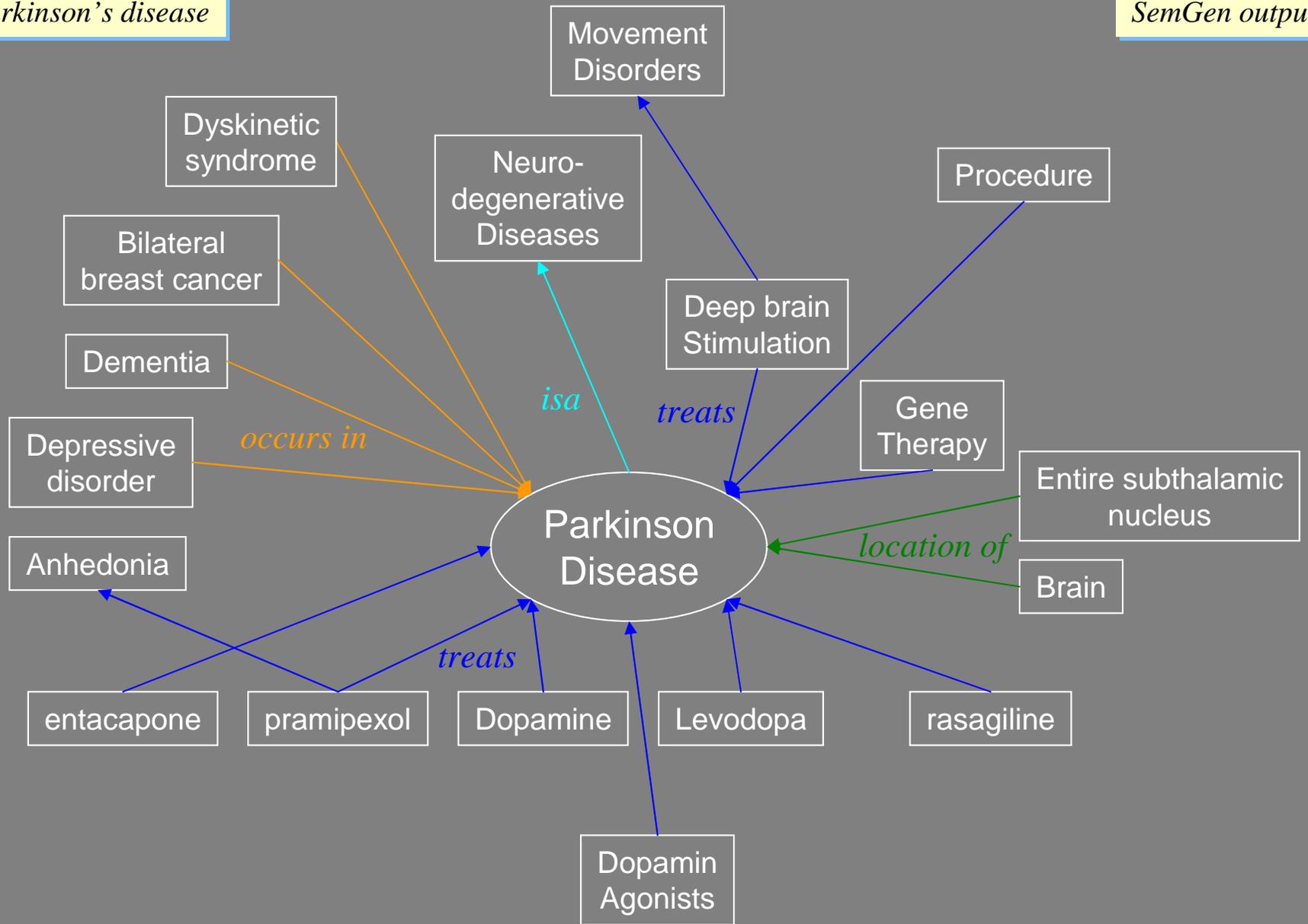
Network of relations

Text mining



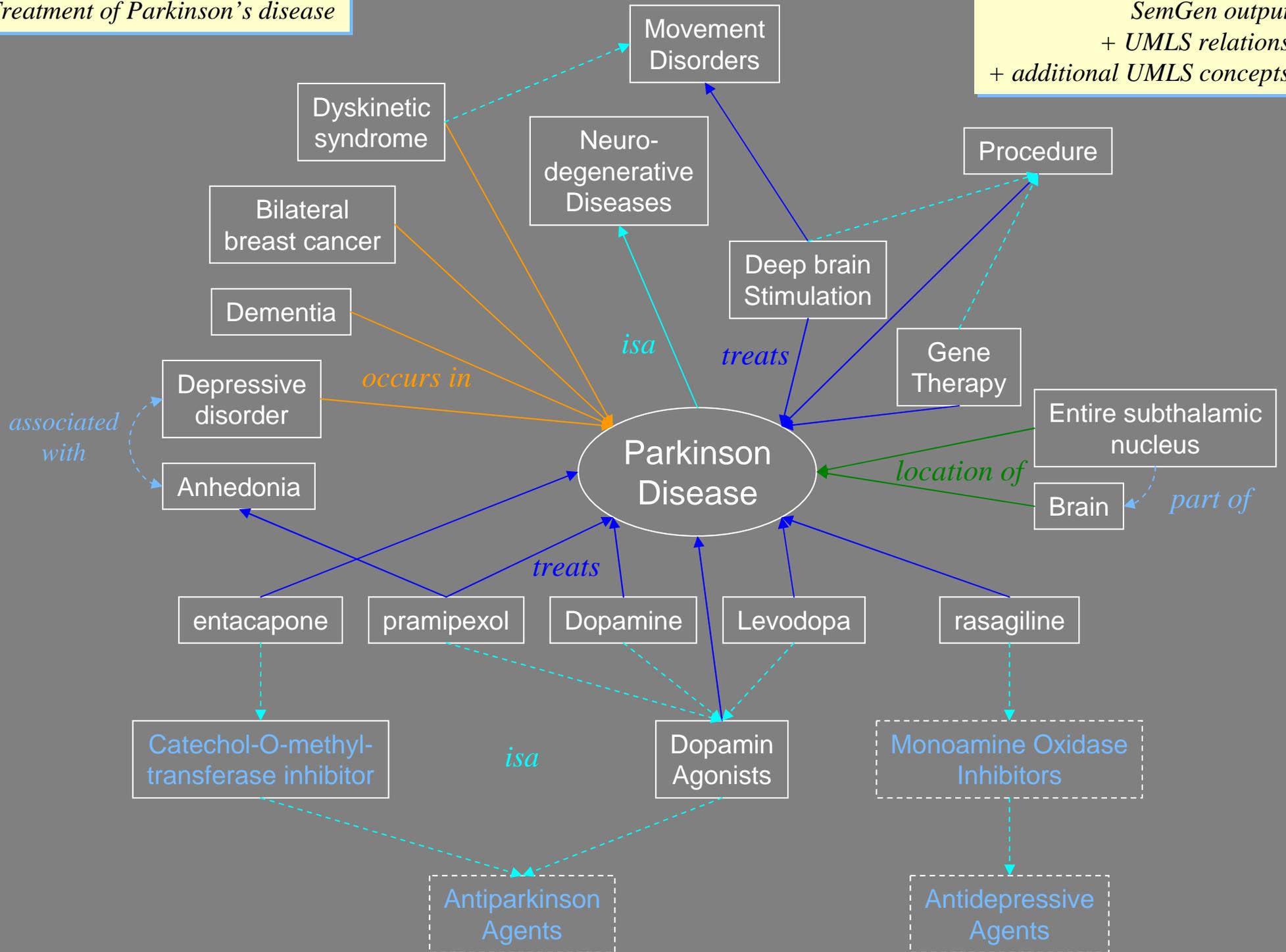
Text Mining Workflow





Treatment of Parkinson's disease

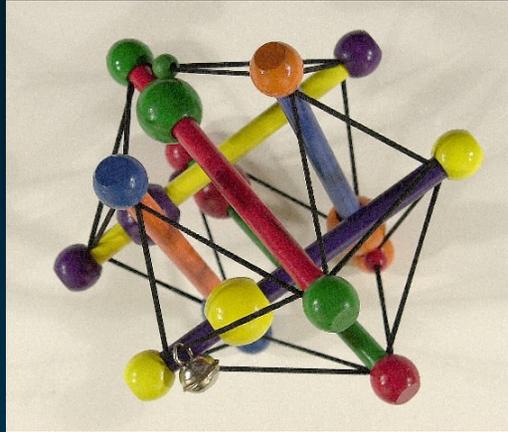
SemGen output
+ UMLS relations
+ additional UMLS concepts



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