

# Emergent Semantics

## An organizing principle for Biomedical Informatics and Knowledge Management

Vipul Kashyap

Clinical Informatics R&D, Partners Healthcare System, Inc.

vkashyap1@partners.org

Biomedical and biological research has been transformed from a cottage industry, marked by scarce, expensive data generated manually, to a large-scale data-rich industry, marked by factory scale sequencing. This is in addition to vast amounts of biomedical research literature being generated at an increasing rate and available through various web based sources (e.g., PubMed [3]). Success in the life sciences will hinge critically on the availability of computational and data management tools to retrieve, fuse, interpret, analyze, classify, compare and manage the abundance of data. Biomedicine is fast becoming an information-based science with data/information playing a big role across the “research flow”, e.g., Genomics → Transcriptomics → Proteomics → Metabolomics → Final Products/Results. The final products may either be drugs and therapies or positive or negative research results in the field of biomedicine.

Scientific Data Integration has been identified as one of the most daunting challenges at the interface between computer science and biology [1], and has been seen as restraining rapid progress in biomedical research [2]. The standard paradigm in biology today is: Data → Hypotheses → Models → Experimentation → Data, and a solution framework for biomedical data integration should provide support for these artifacts. Biomedical data integration, poses a unique set of challenges [1,4]:

- Diversity of Information Objects, including Data Types and Queries:
  - Sequences, complex phenotypic and disease-relevant data, graphs, 3D structures, images
  - Similarity-based queries (e.g., sequence similarity), classification based queries (e.g., Papers about Gene X), what-if hypotheses generating queries (what if Gene X was suppressed, will Protein Z exist?)
- Diversity of information based computations and operations:
  - Experimental Plans and Protocols involving complex repetitive computations involving data retrieval, fusion and analytics.
  - Data Curation and Annotation Tasks.
  - Hypothesis validation across multiple scenarios, Federated search/query processing.
- Semantic Heterogeneity:
  - Multiple Controlled Vocabularies and Ontologies: Integration, Interoperation and Composition
  - Semi-automatic creation and verification of Mappings
    - Mappings and complex relationships across vocabularies and ontologies
    - Mappings, Annotations of data objects to concepts in controlled vocabularies and ontologies
- Dynamic and evolving nature of Biomedical Research
  - Evolution of schemas, ontologies and vocabularies and their impact on underlying mappings or annotations.
  - Evolution of data objects and their impact on associated mapping and annotations
  - Uncertainty and inconsistency of data
  - Support for pro-active data mining and hypothesis generation

Semantics-based approaches are being explored for information integration in the context of the Semantic Web [5]. However, “semantics” or “meaning” is not a fixed entity – it *emerges* from the interaction of people and applications in the context of performing biomedical research. An *emergent semantics-based* information infrastructure would be a pro-active platform where people and applications collaborate for creation of dynamic “semantics” reflecting the current state of knowledge in biomedical research. Some interesting properties of this infrastructure are:

- **Self-description:** The infrastructure shall enable self-description of biological data and content. This is currently the focus of various XML-based markup languages (e.g., BioPAX [6], OWL [17]) and ontologies/vocabularies (e.g., GeneOntology [7], UMLS® Semantic Network [8]) developed and to enable data sharing and interoperability.
- **Self-genesis:** The infrastructure shall proactively analyze data and content flowing through it, to create models, ontologies and concepts to capture semantics. This enables “bootstrapping” of the meanings prevalent in biological data and content.



**Metadata Extraction, Annotation and Mapping:** The ability to describe web resources using metadata descriptions constructed from domain ontologies is crucial in for curation and annotation of biomedical data and content. Some approaches that can help address this obstacle are:

- Association of a vector space generated from a document collection with ontologies [12] for semi-automatic annotation.
- Use of E-R models to describe and query text information and extraction of metadata from multimedia data [15, 16].
- Machine learning techniques to infer mappings between database schemas, web service descriptions and ontological concepts.

**Integration/Interoperation (III):** With a wide variety of communities and user groups on the web, there is a need for interoperation across heterogeneous overlapping ontologies (concepts from which might be used for creating information models and database schemas). **Self-integration/interoperation** is the key property that we seek to enable by addressing the III obstacle. Metadata annotations and mappings discussed above play a crucial role in this context. Approaches that use Description Logics to represent ontologies and their inference capabilities to support inter-ontology integration/interoperation [13,14] can be used to enable **self-integration/interoperation**. Specialized techniques for combining repetitive data retrieval and analysis operations/processes are required for enabling **self-provisioning**.

An emergent semantics-based platform enunciated above, enables a highly adaptable and flexible information infrastructure. The evolution of meaning and biomedical knowledge, supported by the underlying infrastructure, makes it easier to implement new applications for different purposes and requirements, and those that involve continuously evolving knowledge and information. A multi-disciplinary research agenda involving database and information systems research complemented with insights from Knowledge Representation, Machine Learning, Data Mining and NLP is crucial to realizing this vision. There is a need to go beyond traditional information systems and investigate approaches based on social networks and cultural anthropology to achieve a truly flexible *emergent semantics-based* Knowledge Management platform.

## References

1. Digital Biology: The Emerging Paradigm, November 6-7, 2003, NIH Natcher Conference Center, Bethesda, MD
2. NIH Roadmap: Accelerating Medical Discovery to improve Health, Bio-Informatics and Computational Biology, <http://nihroadmap.nih.gov/bioinformatics/index.asp>
3. PubMed, <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>
4. Workshop on Data Management for Molecular and Cell Biology, February 2-3, 2003, Lister Hill Center, NLM, NIH, Bethesda, MD, <http://pueblo.lbl.gov/~olken/wdmbio/>
5. T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities," Scientific American, May 2001.
6. BioPAX: Biological Pathways Exchange, <http://www.biopax.org>
7. Gene Ontology, <http://www.geneontology.org>
8. V. Kashyap and A. Borgida, "Representing the UMLS® Semantic Network using OWL (Or "What's in a Semantic Web Link?")", Proceedings of the 2<sup>nd</sup> International Semantic Web Conference (ISWC), October 2003, Sanibel Island, Florida.
9. V. Kashyap, C. Ramakrishnan, C. Thomas, D. Bassu, T. C. Rindflesch and A. Sheth, "TaxaMiner: An Experimentation Framework for Automated Taxonomy Bootstrapping", Technical Report, Computer Science Dept., University of Georgia, March 5<sup>th</sup> 2004, <http://lsdis.cs.uga.edu/~cthomas/resources/taxaminer.pdf>
10. C. Behrens and V. Kashyap, "The "Emergent" Semantic Web: A Consensus approach for Deriving Semantic Knowledge on the Web", Real World Semantic Web Applications, Frontiers in Artificial Intelligence and Applications, Vol 92
11. V. Kashyap, "Design and creation of Ontologies for Environmental Information Retrieval", Proceedings of the 12<sup>th</sup> International Conference on Knowledge Acquisition, Modeling and Management, October 1999, Banff, Canada.
12. V. Kashyap, C. Behrens and S. Dalal, "Professional Services Automation: A Knowledge Management Approach using LSI and Domain Specific Ontologies", Proceedings of the 14<sup>th</sup> International FLAIRS Conference (Florida AI Research Symposium), Special track on AI and Knowledge Management, May 2001, Key West, Florida, USA.

13. E. Mena, V. Kashyap, A. Illarramendi, and A. Sheth, Imprecise Answers in Distributed Environments: Estimation of Information Loss for Multi-Ontology Based Query Processing. International Journal of Cooperative Information Systems
14. E. Mena, A. Illarramendi, V. Kashyap and A. Sheth, "OBSERVER: An approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies", Distributed and Parallel Databases – An International Journal, Volume 8(2), April 2000
15. V. Kashyap, K. Shah and A. Sheth, "Metadata for building the MultiMedia Patch Quilt", MultiMedia Database Systems: Issues and Research Directions, Springer Verlag 1995, S. Jajodia and V. Subrahmanian (editors).
16. V. Kashyap and M. Rusinkiewicz, "Modeling and Querying Textual Data using E-R models and SQL", Proceedings of the Workshop on Management of Semi-Structured Data in conjunction with 1997 ACM International Conference on Management of Data (SIGMOD) Tucson, Arizona, May 1997
17. OWL Web Ontology Language Overview, <http://www.w3.org/TR/owl-features>