

Towards a National Health Knowledge Infrastructure (NHKI): The role of Semantics-based Knowledge Management

Vipul Kashyap and Tonya Hongsermeier
Clinical Informatics R&D, Partners HealthCare System
vkashyap1@partners.org

The healthcare system in the US is facing crucial challenges in delivering effective, efficient and high quality healthcare. These challenges are in part due to the volume of knowledge and information and their fragmentation across paper and heterogeneous clinical information systems. For instance, biomedical literature doubles every 19 years. AIDS literature in particular doubles every 22 months¹. A clinician needs approximately 2 million facts to practice. On the other hand, biomedical 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. Biology is fast becoming an information-based science with information and knowledge playing a critical role in the flow of research both within biology and into clinical research and practice.

Information overload and lack of access to knowledge have led to patient safety issues, with over 98,000 deaths due to medical error². A slow innovation-adoption curve has patients receiving only 54.9% of recommended care³. Our biomedical research is the envy of the world, but even our best hospitals fail to give some patients the latest treatments, years after they have been proven appropriate. The NIH says it takes from 10 to 17 years for new discoveries to be routinely used⁴. The problem is magnified with the advent of molecular medicine with the associated rich and complex types of knowledge and information. The goal of leveraging this knowledge into clinical care has given rise to the field of *translational medicine* which may be defined as⁶: (a) Validation of theories emerging from preclinical experimentation on disease-affected human subjects and; (b) Refinement of biological principles that underpin human disease heterogeneity and polymorphism(s) by using information obtained from preliminary human experimentation.

The National Health Information Infrastructure (NHII) initiative⁵ has proposed a comprehensive knowledge-based network of interoperable systems of clinical, public health, and personal health information. Improved decision-making will be enabled by making health information available when and where it is needed. This approach clearly identifies the role of knowledge management and informatics in addressing healthcare issues, but focuses only on the information infrastructure. We propose the National Health Knowledge Infrastructure (NHKI), a novel semantics-based knowledge management infrastructure for addressing healthcare delivery deficiencies and enabling translational medicine.

NHKI: Impacts

The NHKI when fully deployed will revolutionize the US healthcare delivery system and have a tremendous impact on the US economy:

- **Alignment:** Design, development and deployment of a common NHKI will enable functionalities and induce alignment across the *health ecosystem*, consisting of, among others, diverse market sectors such as healthcare delivery, drug discovery and life sciences. Rapid creation of novel diagnostics, personalized healthcare products and therapeutics focused on the end-consumer or patient, will enable introduction of synergistic market efficiencies reducing the cost of drugs, therapies and clinical care over time.

- **Rapid Innovation Adoption:** Making the right knowledge available at the right time to clinicians and consumers alike will enable rapid adoption of novel insights from life sciences research into clinical practice. This will reduce significantly, the time taken for adoption of innovative drugs and therapies.
- **Efficiency and cost reduction:** Use of preclinical experimentation on disease-affected human subjects will eliminate drug candidates earlier in the drug discovery process and reduce costs significantly. Phenotypic information from preliminary human experimentation can help focus the search for new biomarkers and drug candidates. Some of the cost reduction could be passed on to consumers, reducing significantly the cost of drugs in the market.

The role of Knowledge Management in Translational Medicine is illustrated in **Figure 1**.

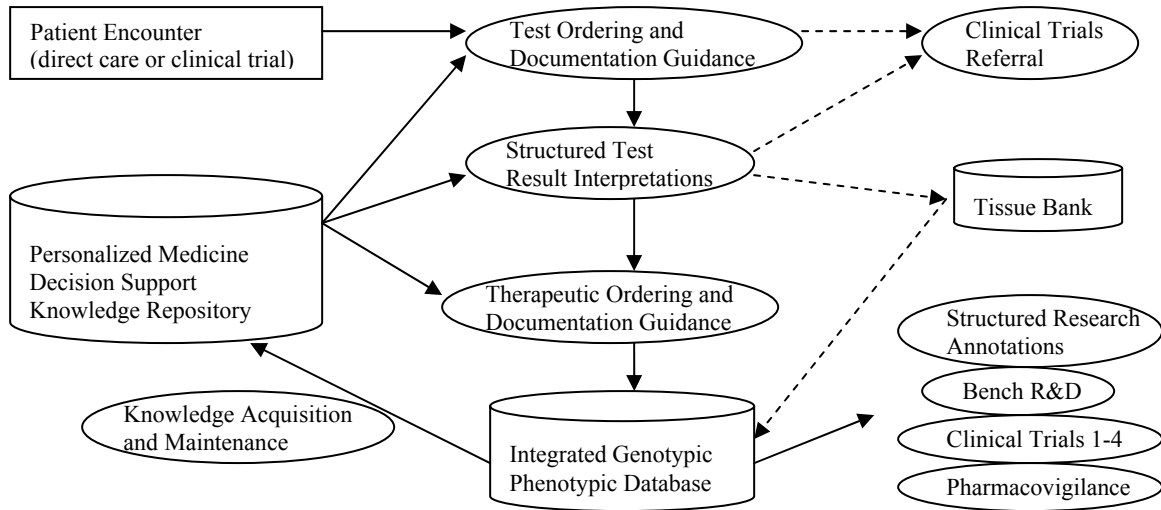


Figure 1: The Role of Knowledge Management in Translational Medicine

NHKI: A Semantics-based KM Platform for Translational Medicine

We propose a semantics-based KM platform²¹ as the underlying platform for the NHKI. It will provide computational and data management services to retrieve, fuse, interpret, analyze, classify, compare and manage the abundance of information and knowledge across the “health ecosystem”. The proposed KM Platform will support:

- Cross-domain, cross-organizational collaboration for creation and development of various types of knowledge objects and processes.
- Life cycle management of various data and knowledge objects, such as data models, to support the continuous refinement and evolution of knowledge.
- Business Process Management services that seek to align clinical care workflows/processes with genomics research and diagnostic processes
- Knowledge retrieval, aggregation and analysis to enable flow of insights (from the genomic to the clinical and vice versa), and for optimization for cost-effective decision making.

NHKI: Challenges

The design and development of the KM Platform presents us with a unique set of challenges^{7,8}:

Diversity of knowledge, information and computational processes: The KM Platform will deal with complex phenotypic and disease relevant data, genome sequences, biological and clinical care pathways/graphs, 3D protein structures and radiological images (**Figure 2**). It will support similarity queries for sequences, classification queries for literature search and what-if

queries such as “if Gene X is suppressed, will Protein Y be created?” Experimental plans, clinical care and guideline protocols, data curation and what-if analysis processes that string together repetitive computations involving data retrieval, fusion and analytics will also be supported.

Semantic Heterogeneity: The KM Platform will support “infrastructure knowledge components”

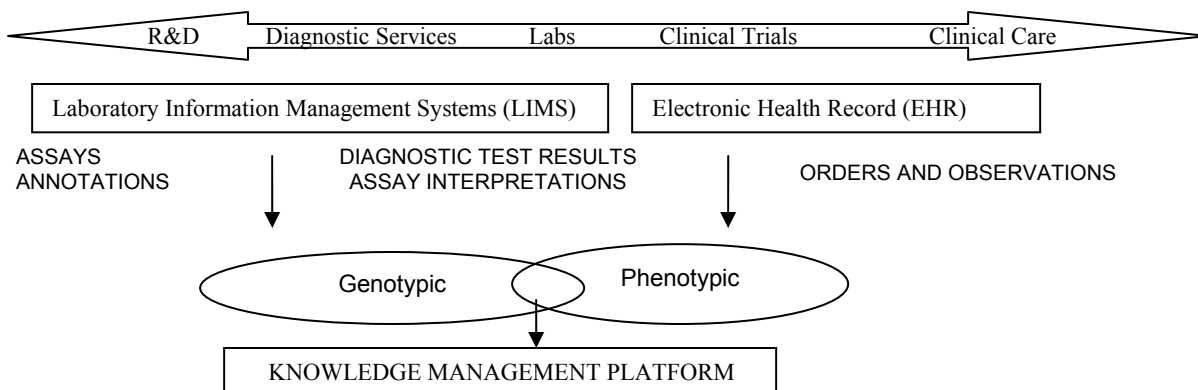


Figure 2: Translational Medicine: Knowledge and Information Diversity

such as controlled vocabularies and ontologies and interoperability across them for standardizing semantic specifications. Creation and verification of mappings, complex relationships across ontologies, and annotation of information with knowledge components will also be supported.

Dynamic and evolving nature of Knowledge: The KM Platform will support evolution of information, knowledge schemas, ontologies and vocabularies. It will also provide services and tools that propagate the impact of these changes on mappings, annotations and other knowledge objects (e.g., clinical guidelines, what-if hypotheses). The uncertainty and inconsistency associated with evolution and change will also be handled. Pro-active data mining and hypothesis generation will also be supported.

NHKI: An “Emergent Semantics” based Architecture for the KM Platform

The KM platform will need to capture and leverage the semantics of information, knowledge and processes for enabling integration and decision support across the health ecosystem. It will need to be configurable and extensible to address evolution of semantics and incorporation of new, emerging knowledge and information artifacts. “Semantics” or “meaning” is not a fixed entity – it *emerges* from the interactions of people with each other. We propose a pro-active platform and architecture (**Figure 3**), where people and applications collaborate in the creation of dynamic “emergent semantics”¹⁷ reflecting the current state of knowledge, and discuss its characteristics:

- **Self-description:** Enablement of knowledge and information artifacts to describe their own meanings, which is the focus of various XML-based markup languages (e.g., SBML⁹, OWL¹⁶) and biomedical ontologies and vocabularies (e.g., UMLS® Semantic Network¹⁰). This creates a substrate for using inference to link genotype with the phenotype and enable decision support. We have done initial work for representation and implementation of clinical guidelines¹⁸ in this context.
- **Self-genesis:** Enablement of proactive bootstrapping of models, ontologies and concepts based on analysis of information and knowledge (e.g., preclinical research literature, physician notes) flowing through the platform. This is in contrast to current manual and expensive approaches for creating and implementing knowledge. We have done initial work on bootstrapping medical vocabularies in the TaxaMiner project¹¹ and creating knowledge models and structured documentation using natural language processing techniques¹⁹.

- **Auto-emergence:** Monitoring of interactions and feedback between people and applications to identify, capture and describe new meanings/knowledge that “emerge” or “evolve” from these interactions. We are currently experimenting with collaboration tools to create clinical knowledge²⁰ and novel anthropological techniques¹² for consensus derivation. Auto-emergence will enable proactive linking of the genotype with the phenotype, and segmentation of patient populations for determination of new biomarkers and drug candidates.

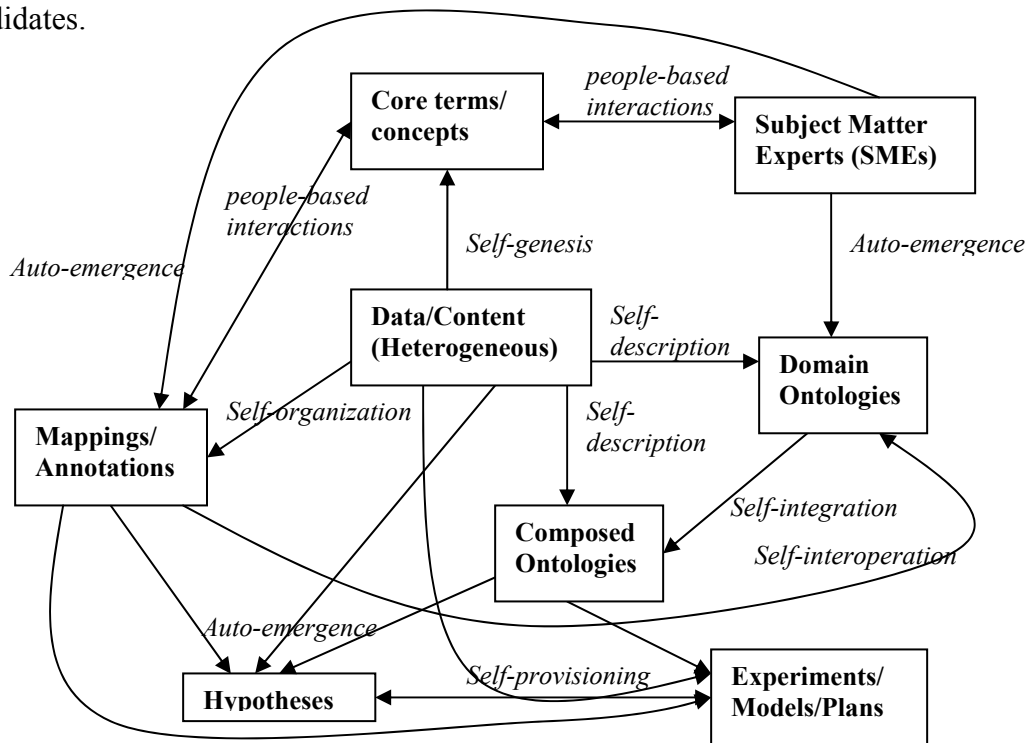


Figure 3: Emergent Semantics-based Platform and Architecture

- **Self-organization:** (Re-)organization in response to new requests for information or services; or due to emergence of new meanings (self-genesis, auto emergence).
 - **Self-interoperation/integration:** Interoperation/Integration across existing meanings may result in the creation and emergence of new meanings and services. We have proposed techniques to integrate/interoperate across ontologies¹⁴ which can be used to create new ontologies, minimizing human effort.
 - **Auto-annotation:** Automatic association of concepts, information and knowledge with each other, such as: classification of information into a taxonomy, metadata annotation or mapping of knowledge models and schemas. We have explored approaches for extraction of metadata from multimedia data¹⁵, and investigated issues related to object equivalences and context¹³ determine associations between knowledge objects.
 - **Self-provisioning:** Monitoring various information retrieval, analysis and computation operations to create experimental plans, models, protocols and guidelines. Plans and protocols may be combined in multiple ways to generate new experiments and protocols. that can be presented to researchers and practitioners across the health ecosystem. Specialized techniques for combining repetitive data retrieval and analysis operations/processes are required for enabling self-provisioning and we will be exploring semantic web services approaches such as OWL-S¹⁶.

Creating the NHKI: A Multi-disciplinary Endeavor

Making the NHKI a reality is a complex, challenging and multi-disciplinary endeavor requiring: (a) A deep emphasis on informatics and computer science research with cross-fertilization of requirements from various research and practice oriented disciplines across the health ecosystem; (b) A deliberate orientation towards design, deployment and system building efforts, with an incremental step-wise approach aligned to business needs; and (c) A solid understanding and cognizance of organizational, cultural and governance issues. The NHKI needs to be deployed in a cross-organizational, cross-industry manner across the health ecosystem for maximal impact, in a manner similar to the current deployment of the World Wide Web.

References

1. Covell DG, Uman GC, Manning PR. *Ann Intern Med.* 1985 Oct;103(4):596-9
2. Kohn LT, *To Err is Human, Building a Safer Health System*, Washington, DC, National Academy Press, 2000.
3. McGlynn EA et al. 2003 The Quality of Health Care Delivered to Adults in the U.S., *NEJM*, 348 (26): 2635-45
4. Tommy Thompson, Secretary, DHHS, National Health Information Infrastructure Conference, 2003. <http://aspe.hhs.gov/sp/nhii/Conference03/SpeechText.htm>
5. <http://aspe.hhs.gov/sp/nhii/FAQ.html>
6. <http://www.translational-medicine.com/info/about/>
7. Digital Biology: The Emerging Paradigm, November 6-7, 2003, NIH Natcher Conference Center, Bethesda, MD
8. NIH Roadmap: Accelerating Medical Discovery to improve Health, Bio-Informatics and Computational Biology, <http://nihroadmap.nih.gov/bioinformatics/index.asp>
9. Systems Biology Markup Language, <http://www.sbml.org>
10. V. Kashyap and A. Borgida. Representing the UMLS Semantic Network using OWL (Or "What's in a Semantic Web Link?") *Proceedings of the Second International Semantic Web Conference*, October 2003, Sanibel Island, Florida
11. V. Kashyap, C. Ramakrishnan, C. Thomas and A. Sheth. TaxaMiner; An Experimental Framework for Automated Taxonomy Bootstrapping *International Journal of Web and Grid Services, Special Issue on Semantic Web and Mining Reasoning*, September 2005 (to appear)
12. 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
13. V. Kashyap and A. Sheth, "Schematic and Semantic Similarities between Database Objects: A Context-based Approach," *Very Large Data Bases (VLDB) Journal*, 5(4), October 1996, pp. 276-304
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. OWL, <http://www.w3.org/TR/owl-features>, OWL-S, <http://www.w3.org/Submission/2004/07>
17. K. Aberer et. al., *Emergent Semantics Systems: Proceedings of the International Conference on Semantics of a Networked World: Semantics for Grid Databases*, Paris, June 2004.
18. V. Kashyap, A. Morales and T. Hongsermeier, Creation and maintenance of implementable clinical guideline specifications, ISWC 2005
19. A. Turcin, V. Kashyap, M. Palchuk, L. Morin, F. Chang and Q. Li, *Natural Language Processing: An Evaluation in the context of Structured Documentation and Knowledge Model Generation*, Poster, AMIA 2005.
20. T. Hongsermeier, V. Kashyap and J. Hanrahan. Collaborative Authoring of Decision Support Knowledge: A Demonstration, AMIA 2005.
21. T. Hongsermeier and V. Kashyap. A semantics-based Knowledge Management Platform for Translational Medicine, Session at Bio-IT World (to appear), May 2005.