Can Semantic Web Technologies enable Translational Medicine? (Or Can Translational Medicine help enrich the Semantic Web?)

Vipul Kashyap[1], Tonya Hongsermeier[1], Samuel Aronson[2]
[1]Clinical Informatics R&D, Partners HealthCare System, Wellesley, MA 02481
[2]Harvard Partners Center for Genetics and Genomics, Cambridge, MA 02139

The success of new innovations and technologies are very often disruptive in nature. At the same time, they enable novel next generation infrastructures and solutions. These solutions often give rise to creation of new markets and/or introduce great efficiencies. For example, the standardization and deployment of IP networks resulted in introducing novel applications that were not possible in older telecom networks. The Web itself has revolutionized the way people look for information and corporations do businesses. Web-based solutions have dramatically driven down operational costs both within and across enterprises. In this position paper, we explore the area of Translational Medicine which aims to improve the communication between basic and clinical science so that more therapeutic insights may be derived from new scientific ideas - and vice versa. Translation research goes from bench to bedside, where theories emerging from preclinical experimentation are tested on disease-affected human subjects, and from bedside to bench, where information obtained from preliminary human experimentation can be used to refine our understanding of the biological principles underpinning the heterogeneity of human disease and polymorphism(s). Informatics in general, and semantic web technologies in particular, may have a big role to play in making this a reality. We present an application use case and investigate the technological questions it generates.

Translational Medicine: Use Case

The first goal of translational medicine is to accelerate the adoption of therapies and tests gleaned from genomics and clinical research into everyday clinical practice. The weak link in this chain is obviously the clinical practitioner as; so far the worlds of genomic research and clinical practice have been separate (though they have started to collide at present). Let's consider an example.

Consider a patient that presents with shortness of breath and fatigue. A clinical exam reveals abnormal heart sounds. The ultrasound ordered based on the clinical exam reveals cardiomyopathy. One could then screen for the following mutations:

- beta-cardiac Myosin Heavy Chain (MYH7)
- cardiac Myosin-Binding Protein C (MYBPC3)
- cardiac Troponin T (TNNT2)
- cardiac Troponin I (TNNI3)
- alpha-Tropomyosin (TPM1)
- cardiac alpha-Actin (ACTC)
- cardiac Regulatory Myosin Light Chain (MYL2)
- cardiac Essential Myosin Light Chain (MYL3)

The doctor in charge can select treatment based on all data. He can stratify for treatment by clinical presentation, imaging and non-invasive physiological measures in the genomic era, for e.g., non-invasive serum proteomics. The following challenges arise in the context of translational medicine

- How a clinical practitioner is made aware of the existence of these tests at the point of care?
- Even if a clinical practitioner is aware of these tests, how is he/she to determine the right candidates for these tests?
- What are the set of clinical care guidelines that specify when and how a physician should order, interpret and act on the test results?
- What are the observations required from a clinician in the ***context of*** other clinical findings and ongoing therapies so that he can indeed order, interpret and act on the test results? Some answers are:
  - Structured history of present illness, physical assessment, clinical impression/working diagnosis data
  - Structured family history data (crucial to identify candidates for certain diagnostic tests)
- How can research and clinical insights be transferred from clinical care to the genomics and clinical research worlds and vice-versa? These can be achieved by:
  - Study of phenotypic data in patients with markers for disease
  - Population segmentation with respect to molecular diagnostic testing and pharmacotherapeutics
  - Generation of new hypotheses, for instance to search for new markers for therapeutic differentiation
  - Suggestions for new genomics/proteomics experiments to explain clinical observations
  - Identification of targets for new drug development, segmented populations for new and existing drugs on the market

Research and Technological Issues: How can Semantic Web Technology Help?

In this section we discuss issues related to the informatics and technological infrastructure related to addressing the challenges of translational medicine enumerated in the previous section. In particular, we investigate the architectural components required to support a cycle of ***learning from*** and ***translation of innovation to*** the clinical care environment. Some components in the strawman architecture (**Figure 1**) could be:

- Workflow portals for clinical researchers, lab personnel, clinical trials, and clinical care providers
- An integration of genotypic and phenotypic patient data and reference information data to support transactions and analysis.
- Knowledge Discovery, Asset Management, and Publication/Decision Support Services and personnel
- Creation and/or derivation, validation, and publication of knowledge-bases for decision support systems used in healthcare delivery.

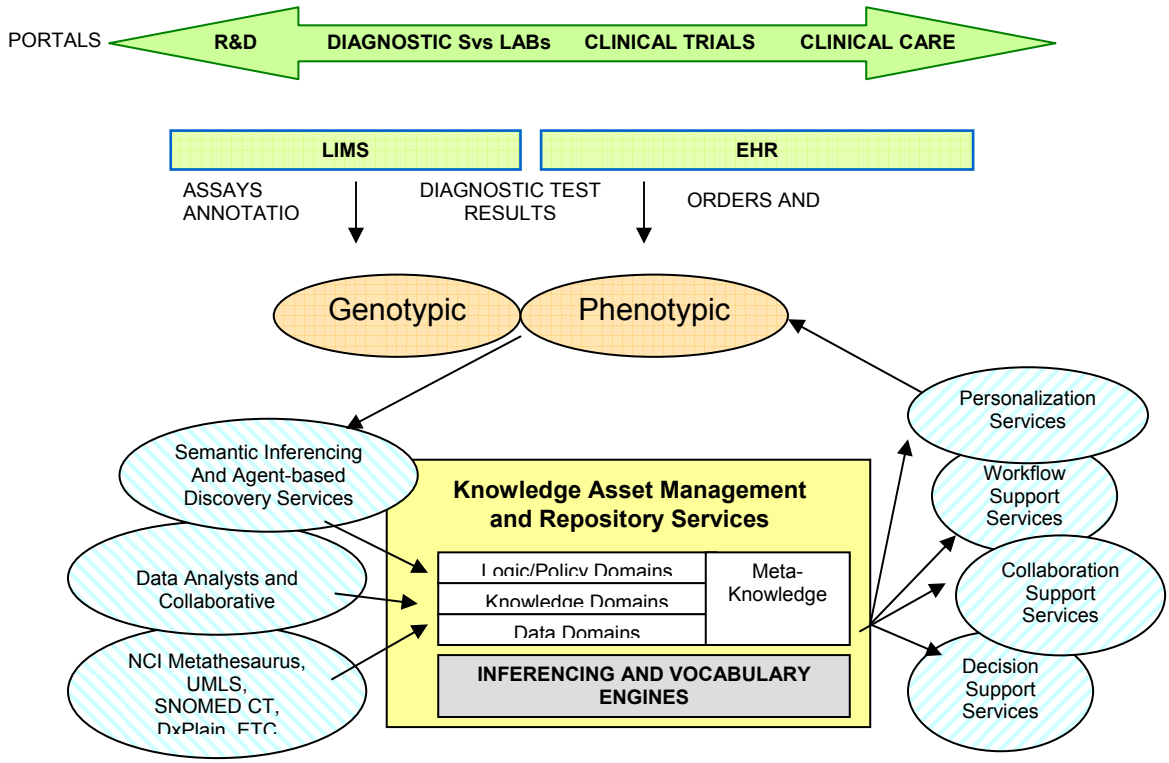

Figure 1: Strawman architecture for Translational Medicine

One of the fundamental issues in the architecture discussed above is that of integration of information across multiple information domains, one of genomics/proteomics research and the other of clinical practice. One of the sub goals of this is the creation of an integrated genomic data repository that could be deployed and used in the context of healthcare practice. Consider the data/information flow in a hypothetical translational medicine workflow as illustrated in **Figure 2**.

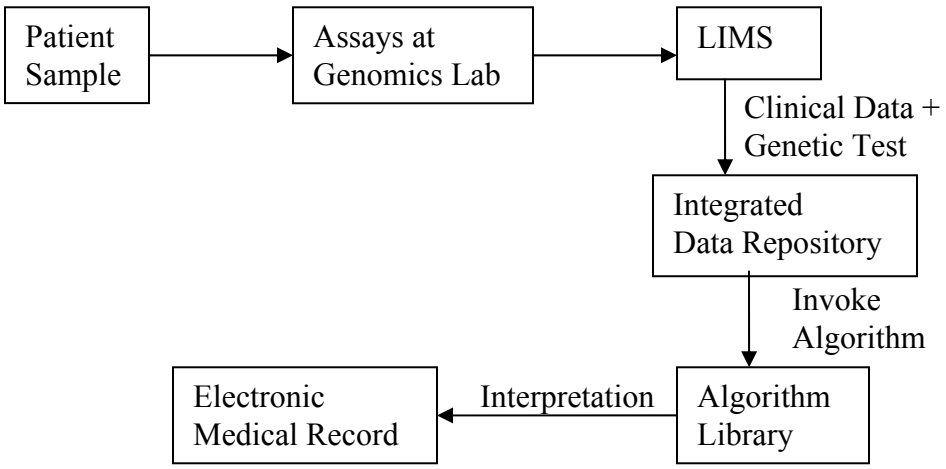

Figure 2: Hypothetical Information Flow in Translational Medicine

A prelude to integrating information across the clinical and genomic domains is the issue of creating a uniform common data model that is expressive enough to capture the wide variety of data types and artifacts created, invoked and consumed in the hypothetical information flow described above. Some of these data types are:

- Genomic/Proteomic

    - 2D representations of DNA and RNA sequences
    - Representation of Micro-array experiment data
    - 3D Protein structures
    - Protein Expression Pathways
    - Metabolic Pathways
    - Single Nucleotide Polymorphisms (SNPs)
- Clinical
    - Electronic Health Record
        - Structured Data Elements
        - Unstructured free form text fields
    - Biomedical concepts and codes from controlled vocabularies such as SNOMED, LOINC, etc.
    - Clinical care pathways and guidelines
    - Decision support and diagnostic rules
    - Alerts and Triggers
    - Specialized algorithms for interpretation of clinical test data

The wide spectrum of complex data types observed across the clinical-genomic spectrum make the design of a uniform common data model a challenging task. Some interesting characteristics observed across this spectrum give us a pointer to the requirements that must be supported by a semantic web for the life sciences.

- The wide variety of complex data types are not covered by any one W3C standard. For instance, graph based data such as pathways and protein structures can be represented using RDF, decision support rules can be represented using RuleML and SWIRL, whereas clinical care pathways and guidelines maybe best represented using the OWL-S standard. So the first question that arises is: Is it feasible to create a single W3C standard encompassing a common data model and format for representation of data for for translational medicine?
- Knowledge needs to represented at multiple levels: molecule → pathway → cell → organ → patient → population
- Clinical data is probabilistic or fuzzy in nature. For instance diagnoses are always specified with a degree of certainty. The current W3C semantic web standards need to be enhanced to represent uncertainty as a part of the underlying metamodel.
- Temporal information is an intrinsic component of clinical data, as it is important to track the progression of a disease in a patient over time. The ability of the semantic web standards to express temporal information and support temporal reasoning will be very useful.
- Spatial information is important in the context of proteomics data where the same set of atoms could have multiple orientations in 3D space. The ability of semantic web standards to support spatial reasoning and information will be very useful.
- Algorithms also need to be described using specialized information objects so that they can be invoked correctly in the right contexts for example for interpretation of test results.

Given the above characteristics, the following issues need to be considered:

- Are the RDF/OWL Data Models general enough to provide a common underlying representation for this wide variety of data types?
- Are the RDF/OWL data models expressive enough to capture and enable querying and reasoning with spatio-temporal information?
- Can the RDF/OWL data models support the probabilistic and fuzzy nature of clinical and genomics data?
- Can OWL-S be used to describe algorithms that can be invoked as services in the context of interpreting clinical test results?
- Given that the representational constructs required for the wide variety of data types are spread across the various W3C standards, does it make sense to combine these into a single standard to come up with a data model and format for translational medicine?
- How would this new standard compare with the Reference Information Model (RIM) proposed by the HL7 Standards body? How does the RIM meta-model compare with the RDF/OWL meta-model?

While a uniform common data model by itself doesn't guarantee effective information integration, it will definitely facilitate and make easier the task of achieving the goal. Of course there are significant challenges on creating a semantic web standard for translational medicine. Some other issues that need to be addressed that may be as important if not more as the issue of a uniform common data model are:

- New data stores that support storage, indexing and retrieval of these novel data types?
- New inference engines that perform OWL-based and probabilistic inferences in the context of clinical decision support
- Rapid Algorithmic evolution, collection of classification algorithms based on different areas of practice
- Efficient data mining operators for identification of new hypotheses, population segmentation, genetic markers, drug targets, etc.

Industrial acceptance and other issues

An important issue from a deployment perspective is the need to collaborate with other communities and incorporate current existing standards. There are large communities of researchers working in the areas of healthcare informatics and life sciences. The HL7 community has done significant work in the clinical informatics area and has developed standards such as the RIM. The W3C should reach out to this and other related standards bodies in an attempt to create common underlying standards.

The other issue is that of industrial acceptance. Rightly or wrongly, there is a wide spread impression that RDF/OWL standards have still not attained maturity. This impression is especially wide spread in the healthcare and maybe the life sciences industry. One may argue that this is most likely the case for every new technology, however to increase adoption and acceptance of semantic web technologies, a clear value proposition is required. The emerging area of translational medicine may indeed provide the value proposition and use case.