5.6 caGrid 2.0 Platform and Terminology Integration

This section includes the following:

caGrid 2.0 Platform and Terminology Integration

The semantic infrastructure has to support seamless integration with the caGrid 2.0 platform. The following are some high-level platform and terminology requirements that are either supported or addressed by the semantic infrastructure.

5.6.1 Service Generation

Service generation is the ability to generate services from user defined service metadata. The semantic infrastructure provides this metadata and the platform leverages this metadata for service generation. The constraints and policies specified in the semantic infrastructure are inherited by the platform and are enforced as runtime policies.

Additional platform specific and runtime information is provided by the developer at the time of service generation.

5.6.2 Service Discovery and Utilization

This group of requirements focuses on enabling developers of composite services and applications to discover, compose, and invoke services. This includes the discovery of published services based on service metadata and the generation of client APIs in multiple languages to provide cross-platform access to existing services.

Discovery includes service discovery, data discovery, and policy discovery. Service discovery allows primary users as well as secondary users to locate a service specification and instances based on attributes in the service metadata (for example, via a search for specific microarray analysis services). Data discovery enables secondary users to find the types of data available in the ecosystem as well as summary-level information about available data sets. Policy discovery allows application developers to find and retrieve policies on services.

The platform will use the semantic infrastructure service metadata to address all the service discovery requirements. The semantic infrastructure relies on metadata about services and artifacts.

Link to use case satisfied from caGrid 2.0 Roadmap: As institutions share de-identified glioblastoma data sets, they are available to others via data discovery. The treatment recommendation service used by the oncologist is able to discover these new data sets and their corresponding information models, and include that data for subsequent use in recommendation of treatment.

Link to use case satisfied from caGrid 2.0 Roadmap: all of the data management and access services in the use case are utilized by application developers to build the user interfaces that the clinicians use during the course of patient care.

5.6.3 Service Orchestration and Choreography

Service orchestration and choreography allows both application developers and non-developers to discover service "building blocks" that can be composed dynamically to provide business capabilities. Special cases include the orchestration of multiple services for a distributed query, or for a transactional workflow. Service orchestration and choreography will leverage static and behavioral semantics from the Semantic Infrastructure 2.0.

The semantic infrastructure provides the behavioral semantics required for dynamic composibility of services or generation of distributed queries. This includes runtime contract discovery and negotiation to determine composibility of services based on service capabilities and constraints.

Another use case is dynamic retrieval and enforcement of the policies that are in effect for a service interaction in the areas of logging, validations, data transformation, or routing. This information can be used either during the design of the orchestration or during the execution of the defined flow.

Link to use case satisfied from caGrid 2.0 Roadmap: Federated query over The Cancer Genome Atlas (TCGA) data and other data sets is performed using a service orchestration.

5.6.4 Policy and Rules Management

Policy and Rules Management allow non-developer secondary users to create policies and rules and apply them to services. The scope of policies includes, but is not limited to, definition and configuration of business processing policy and related rules, compliance policies, quality of service policies, and security policies. Some key functional requirements for managing policies include capabilities to author policies and store policies, and to approve and validate policies and execute policies at runtime.

The semantic infrastructure will provide a mechanism to specify policies, including business processing policies and related rules, compliance policies, and quality of service policies. Tools and services for creating security specific policies will be provided by the caGrid 2.0 platform and will be used by the semantic infrastructure. All other policies specified in the semantic infrastructure will be enforced by the platform at runtime.

Link to use case satisfied from caGrid 2.0 Roadmap: Each institution has different data sharing needs, access control needs, and business rules for processing that are defined and customized. For example, policy at the pathologist's institution may state that the patient is scheduled for a visit when the review is complete.

5.6.5 Event Processing and Notifications

Event Processing and Notifications enables monitoring of services in the ecosystem and provides for asynchronous updates by services, effectively allowing a loose coordination of services that both provide and respond to conditions (possibly defined in business rules).

The semantic infrastructure will provide a placeholder to specify events and triggering conditions for data and services. The platform monitors these events at runtime and acts on these events.

Link to use case satisfied from caGrid 2.0 Roadmap: As patient care proceeds, the system notifies the designated clinicians that data (for example, images) are ready for review. Similarly, when notifications are received, event processing logic allows the appropriate parties to assign clinicians for care. In order to facilitate better treatment (a learning healthcare system), as new de-identified glioblastoma data is made available, notifications are sent that could indicate a recommended change in the treatment plan.

5.6.6 Data Representation and Information Models

This set of requirements includes providing an application developer with the ability to define application-specific attributes (for example, defined using ISO 21090 healthcare datatypes) and an information model that defines the relationships between these attributes and other attributes in the broader ecosystem. In particular, the last requirement suggests linked datasets, where application developers can connect data in disparate repositories as if the repositories are part of a larger federated data ecosystem. Additional requirements include the ability to publish and discover information models. Support is needed for forms data and common clinical document standards, such as HL7 CDA. To support the use of binary data throughout the system, the binary data must be typed and semantically annotated.

All information models, their representation and binding to datatypes and terminologies will be managed by the semantic infrastructure. The ability to publish and discover information models will be supported by the semantic infrastructure, and the platform will leverage these capabilities.

Link to use case satisfied from caGrid 2.0 Roadmap: The pathology, radiology and other data have various data formats which must be described, and the information model for the patient record must link between these various datatypes. The complete information model includes semantic links between datasets to build a comprehensive electronic medical record. Annotations on data are defined and included in the information model.

5.6.7 Data Management

Data management includes linking of disparate data sets and updates of data across the ecosystem. Data updates may include updates to multiple data sources, necessitating the need for transactions.

Linkages between the different disparate data sets will be managed by the semantic infrastructure. Data updates that trigger transactions are captured by the platform and are propagated upstream to the semantic infrastructure. An example would be the platform monitoring events to identify changes to data.

Link to use case satisfied from caGrid 2.0 Roadmap: the patient has an electronic medical record that spans multiple institutions. The clinical workup data (for example, genomics and proteomics data) is linked to the clinical care record; similarly pathology and radiology findings must be attached to the patient's electronic medical record.

5.6.8 Data Exploration and Query

The wealth of data must be accessible, resulting in the need for exploration of available datasets. This includes the ability to view seamlessly across independent data sets, allowing a secondary user to integrate data from multiple sources. In addition, the query capability must support sophisticated queries such as temporal queries and spatial queries.

The semantic infrastructure will provide metadata for discovery of these datasets. Complex temporal and spatial queries will be informed by the metadata but will be formulated and executed by the platform.

In order to also discover dataset contents exposed on the grid, the ECCF registry must have linkages from dataset metadata to the metadata about the data they contain. This is distinct from the metadata about the dataset (the owner, creation time, table structure of fields and attributes) and instead describes the type of data contents of the dataset so that a user can retrieve portions of a dataset of some type.

Link to use case satisfied from caGrid 2.0 Roadmap: The oncologist must be able to quickly find glioblastoma data sets, indicating the fields that he is interested in comparing from his clinical data in order to find similar disease conditions and associated treatment plans. Temporal queries allow clinicians to identify changes in patient condition and treatment over time.

5.6.9 Provenance

Provenance encompasses the origin and traceability of data throughout an ecosystem. This is a clear requirement directly from the use case in order to ensure that all steps of patient care and research are clearly linked via the patient record.

The semantic infrastructure will provide data provenance support.

Link to use case satisfied from caGrid 2.0 Roadmap: The origin of data is tied to the data creator, allowing the oncologist performing the match against TCGA data and other datasets to include and exclude data sets based on their origin.

5.6.10 Data Semantics

In a diverse information environment, semantics must be used to clearly indicate the meaning of data. This requirement is expected to be addressed by the semantic infrastructure, although there will be a touchpoint between caGrid 2.0 and Semantic Infrastructure 2.0 to annotate data with semantics. Integration with the semantic Infrastructure will enable reasoning, semantic query, data mediation (for example, ad hoc data transformation) and other powerful capabilities.

Data semantics are captured in the semantic infrastructure and the platform will leverage the semantic infrastructure interfaces for reasoning and analysis.

Link to use case satisfied from caGrid 2.0 Roadmap: The oncologist accesses the TCGA database to search for de-identified glioblastoma tumor data that is similar to the patient data exported from the hospital medical record. During this search, the semantics of the data fields are leveraged to indicate matches between TCGA data fields and the hospital medical record data fields.

5.6.11 External Data Repositories

There are numerous data repositories on the web today. These data repositories contain essential information that must be accessible to services in the ecosystem. As a result, caGrid 2.0 must provide capabilities to integrate these external repositories into the grid with the assumption that the remote service cannot be changed.

The semantic infrastructure will support integration with other metadata repositories, allowing the platform to leverage the semantic infrastructure for federated metadata discovery and analysis. The federated data query capabilities will be implemented by the platform.

Link to use case satisfied from caGrid 2.0 Roadmap: The oncologist searches both TCGA glioblastoma data as well as de-identified data that has been added by care providers around the country. The additional data sets are external data repositories.