caBIG™: Opportunities and Challenges to creating a federated global network of interoperable information systems

George A. Komatsoulis

***Abstract*–The cancer Biomedical Informatics Grid (caBIG™) was initiated by the US National Cancer Institute in 2004 to address the need for interoperable information systems to enable molecular medicine for oncology. With the successful completion of the pilot phase of caBIG™ in 2007, the NCI is in the process of expanding the program into the broader biomedical research and care delivery arena. To accomplish this goal it is necessary to address a series of challenges associated with common semantics, security, interoperability standards and politics. A partnership between caBIG™ and the UK National Cancer Research Institute (NCRI) is providing an initial model for addressing these challenges more globally.**

I. Introduction

Modern molecular technology has brought medicine to the edge of providing personalized medicine; that is, the ability to base care decisions based on the genetic characteristics of a patient. However, the promise of personalized medicine is critically dependent on successfully integrating classes of data (clinical outcomes information, molecular characterization data, biospecimens, medical images) that have traditionally been persisted in a variety of separate (and unfortunately closed) information systems.

Functionally, the current biomedical research informatics enterprise resembles the political system commonly described as feudalism; that is to say that there is weak central governance with most power residing in local authorities. Such a model almost guarantees that systems and the information that they contain will not be capable of the integration required for the personalized medicine paradigm as they tend to have idiosyncratic data access mechanisms and limited semantics that prevent easy integration. There are two alternative models for providing such integrated data. The first could be described as forced collectivization; that is the creation of highly centralized data repositories that are designed to aggregate the information in a single location. The flaws with such systems are that they tend to support only the small subset of analyses that they are designed around and, more importantly, that they limit the classes of information that can be collected and aggregated. The other alternative is federalism, where most authority remains local, but a central authority governs the standards and rules around interactions between local authorities. A federated system of information systems can only enable such data integration if the information systems are interoperable.

The IEEE Standard Computer Dictionary [8] defines interoperability as "ability of two or more systems or components to exchange information and to use the information that has been exchanged''. This definition encapsulates the reality that there are two components to interoperability; the ability to access a system (syntactic interoperability) and understand the information once it has been received (semantic interoperability). Unless both components are addressed, systems cannot interoperate. Consider two people who speak different languages; through speech they can pass messages (and likely identify the start and end of the messages as well as parse individual tokens) but they will not be able to utilize any of the information that is contained within the message; they are syntactically but not semantically interoperable. Similarly, consider two people fluent in English, one blind and one deaf. The deaf person could write and the blind person could speak but although the message is understandable by both if it could be received, they cannot successfully exchange the information; they are semantically but not syntactically interoperable.

Recognizing these issues, the National Cancer Institute (NCI) commissioned the cancer Biomedical Informatics Grid (caBIG™) program in 2004. The caBIG™ program was charged with creating the technical and sociological infrastructure that would enable interoperability between health information systems created in a federated environment. The caBIG™ program has created such an infrastructure for the NCI supported cancer research community and is poised to expand into the broader biomedical research community.

II. Technical Infrastructure

The technical challenges in creating an infrastructure that can meet the requirements of both syntactic and semantic interoperability are non-trivial and at a minimum include a mechanism to enable access to information systems and agreed upon semantics that are accessible to systems at runtime. Although caBIG™ has created a novel infrastructure to enable interoperability, it is built upon a wide variety of readily available and generally accepted standards in the information technology space.

Fundamentally there are (to paraphrase T.E. Lawrence) Four Pillars of Interoperability; interface integration, information models, controlled terminology and common data elements. Interface integration is the syntactic component; the ability to access resources and data of a

Manuscript received July 16, 2008. This work was supported by the direct operating expenditures of the National Cancer Institute, National Institutes of Health, US Department of Health and Human Services.

G. A. Komatsoulis is with the National Cancer Institute Center for Biomedical Informatics and Information Technology (CBIIT), 2115 E. Jefferson St., Suite 6000, Rockville MD 20852 (phone:301-451-2881; fax: 301-480-6641; e-mail: komatsog@mail.nih.gov).

system. Controlled terminology and common data elements are the semantic components, describing the data that is recorded and the context in which the data is recorded. Information models act as a bridge between the semantic and syntactic components. The caBIG™ technical architecture encapsulates all four of these pillars of interoperability.

To resolve the problem of interface integration, the caBIG™ program adopted the Object Oriented Programming paradigm that is a best practice within the software development community. Within caBIG™ two major classes of services exist: data services that provide a query interface to stored data and analytical services that manipulate data and return results. At a more detailed level, data services provide Query-By-Example (QBE) Application Programming Interfaces (APIs) that use instances of API objects to carry query requests and return results. Figure 1 shows a UML model of a portion of the caBIO v3.0 API specification. Each class in the API can be used as an input to the query service supported by caBIO. Results are returned as instances of objects in the API. Such query APIs are designed to allow complex queries, built up from multiple objects and the return of any class of data objects for which there is a traversable path between the query class and the proposed return value class.

Careful observers will note that the UML diagram shown in figure 1 is not a business API, but rather an API that describes the data. This is the result of a conscious decision to expose data to the world in a structure that mimics our understanding of the domain of science that is being represented; a practice known as domain information modeling. The reason for this choice was to maximize the likelihood of common classes and attributes in systems that were designed by different individuals with different business processes; but who were operating with a relatively common vision of how their domain worked in the real world. Thus, things in the real world become classes in the model and the associations in the model represent associations between those things in the real world where such associations have been implemented in the system. This information model thus drives the design of the API (a specialization of the Model Driven Architecture paradigm) and provides a bridge to the semantic components of a caBIG™ compatible system.

The use of controlled (and common) biomedical terminology for recording information is the easiest of the technical requirements to understand. Clearly, it will be difficult or impossible to aggregate information from multiple sources if different groups use different terms to represent the same value or use the same words to describe different values. Many standard terminologies (for example, SNOMED [3], LOINC [4], NCI-thesaurus [7], Gene Ontology [1], etc.) exist to describe information in various domains of knowledge; what is required is community acceptance of particular standards in various contexts. To that end, caBIG™ compatible applications are required to use controlled terminology that has been approved by the caBIG Vocabulary and Common Data Element (VCDE) workspace, which is composed of members of the caBIG™ community.

The final pillar of interoperability is the use of Common Data Elements or CDEs. CDEs provide context around information that is recorded by and transmitted from caBIG™ compatible applications. Consider a value of "anemia"; which could refer to a diagnosis (the patient has anemia and needs treatment) or an adverse event (the patient was being treated with taxol and suffered anemia as a consequence). A CDE provides a description of what is being recorded (say "name of adverse event") and what constitutes a valid response in that context. The former is often called a "Data Element Concept" or DEC and the latter a "Value Domain". There are two classes of value domains: enumerated, in which a list of acceptable responses is provided and non-enumerated, in which the response is flexible within certain boundaries (e.g. minimum and maximum values, length, or format). In a caBIG™ compatible system, each combination of a class and an attribute from the information model of a system (along with its valid values) are bound to a CDE. Fortuitously, an international standard for Common Data Elements, ISO 11179 was available and so caBIG™ adopted this standard with one modification. This modification is to add a requirement that CDEs be based on controlled biomedical terminology so that their meaning (as opposed to the meaning of the data that are recorded and described by the CDE) was clear and unambiguous.

The details of the technical implementation of the caBIG™ interoperability framework have been described elsewhere [2,5,6] and will not be further enumerated in this manuscript. It will suffice to say that the requirements for systems that wish to interoperate within this technical infrastructure build to a set of (generally technology neutral) compatibility guidelines. The guidelines are structured around the four pillars of interoperability with four different levels of compatibility starting from "Legacy" (not interoperable) through Bronze, Silver and finally Gold (maximally interoperable within the caBIG™ infrastructure). Generally speaking a "Silver" compatible system is considered ready to connect to the caBIG grid infrastructure, caGrid; to date (July 2008) information models for 78 applications have been loaded into our metadata repository and 30 systems have been connected to the caGrid infrastructure.

Before discussing the opportunities and challenges facing a larger global Grid infrastructure, it is necessary to describe caGrid, the service infrastructure of caBIG™. The overall structure of caGrid is shown in figure 2. There are four major categories of components: data and analytical services that provide access to data or analysis resources; metadata services such as the EVS, caDSR, Index and GME services that provide runtime access to (respectively vocabularies, CDEs, data and analytical service metadata and schemas), security services that provide infrastructure to support federated authentication and authorization, and finally workflow services such as the Federated Query Processor (FQP) and Workflow Service. It is beyond the scope of this

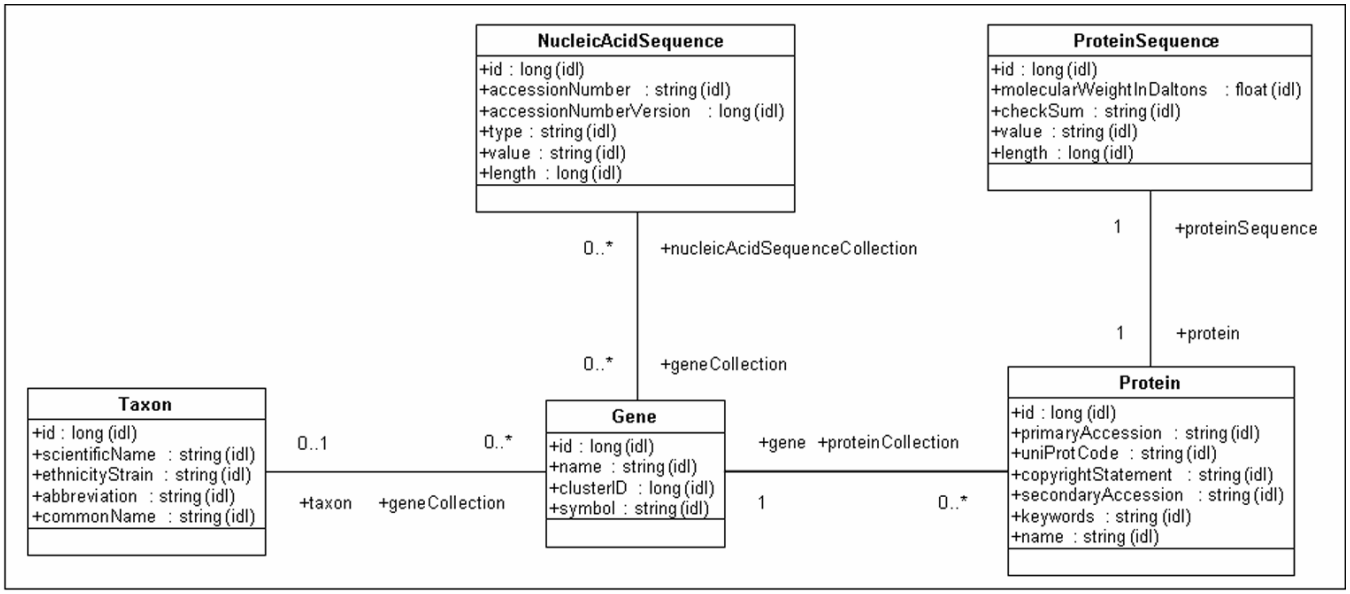

Fig. 1. A UML model of a portion of the caBIO version 3.0 Query API

article to fully describe all of these components [5,6], but several points are worth making. First, one of the primary additional capabilities provided by Grid technology (and caGrid in particular) is advertising/discovery; that is the ability to find services that have particular classes of information. This is provided in caGrid via the Index Service. Second, all of the components of caGrid are designed to operate in a federated environment; that is, there can be duplicates of each of these components that could be either independent or synchronized to allow the formation of subgrids or other independent Grids that organize outside of the direct control of the NCI's instance of caGrid (hereafter referred to as the NCI-caGrid).

I. Models for Expanding Interoperable Networks

The technical infrastructure described above is not in any way specific to cancer; in fact, the only cancer specific components of caBIG are the contents of the data repositories that are part of its ecosystem. Even this is entirely a function of the funding priorities of the NCI rather than any technical limitation of the caBIG™ infrastructure. This gives rise to three possible directions for the long-term evolution of caBIG™. First, it should be possible to expand caBIG™ beyond its current oncology community and expand into the more generic biomedical research community. Second, it should be possible to expand this architecture into the realm of care delivery, that is to say, linking the Electronic Health Record (EHR) with research. Third, it should be possible to expand beyond its current focus within the United States (or some combination thereof).

Each of these options would require addressing the needs and requirements of additional communities. There are several options for supporting these new needs, ranging from bringing these groups into the existing caBIG™ community to creating separate (but interoperable) communities that are partners with caBIG™. Regardless of the mechanism selected, there are four major categories of issues that would need to be resolved to enable such a global network: common semantics, security, interoperability standards and political/social issues. Overall, the most practical, scalable model for creating such a global network is an interoperable, global "Grid of Grids", that is a network of self-governing Grids created along the model of caBIG™ that share common models of interoperability. Such a structure allows for both enhanced community control and variations in regulatory frameworks across specialties and countries. The remainder of this manuscript will focus on these challenges, and the way that caBIG™ is currently working with partners (in particular the UK National Cancer Research Institute) to remove the barriers to such a network.

II. Issues Associated with Common Semantics

Two of the four pillars of interoperability are related to semantics, the first are vocabularies and the second common data elements. It is fairly clear why interoperability would be dependent upon collecting answers to the same questions and having a common *lingua franca* for coding responses to those questions. Unfortunately, there are a variety of reasons why this ideal is not always the case. First the tendency had been for differing areas of clinical practice (oncology, cardiology, etc.) to develop their own actual or *de facto* standards (both in terminology and common data elements) somewhat in isolation of other care areas. Lack of participation in the development of these standards can often lead to reluctance to adopt them, even if the standard will satisfactorily meet the needs of a new community. Second, regulatory requirements in different countries tend to develop in isolation, based on differing safety, privacy and intellectual property environments. Third, differences in

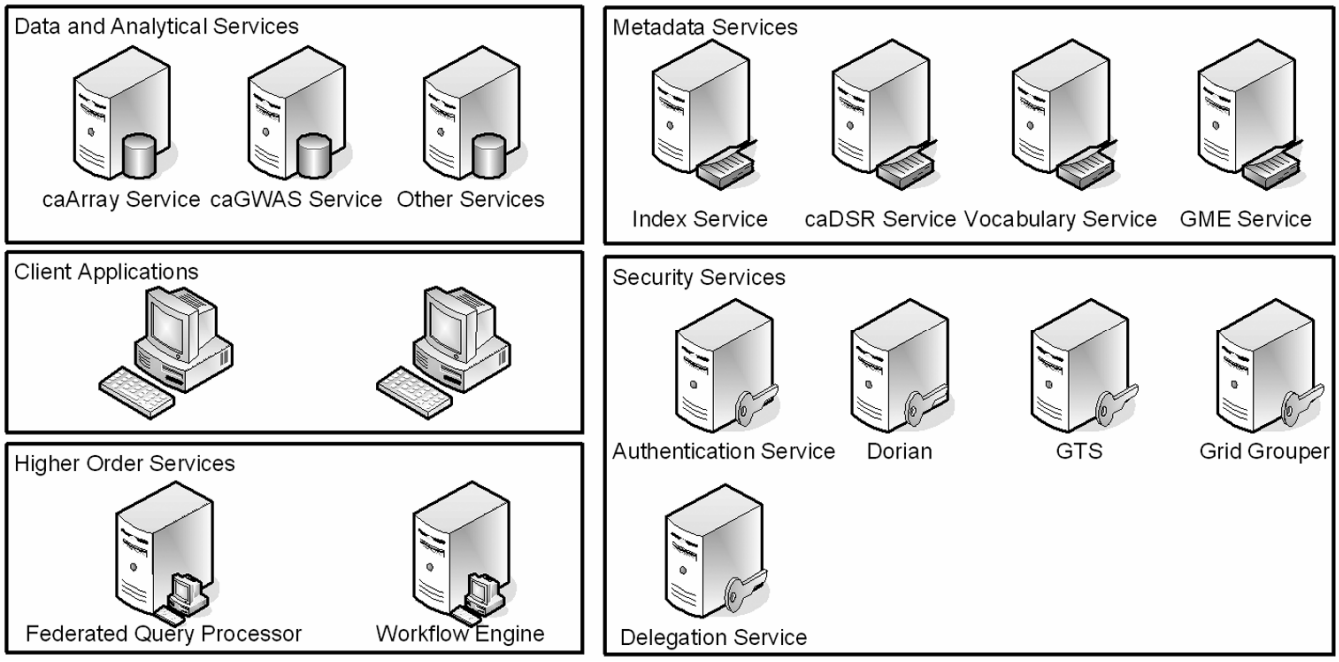

Fig. 2. Major components of caGrid

language between countries require (at a minimum) translations of common standardized terminologies; translations that are not always available, nor necessarily practical.

The general approach to problems of this nature is two-fold: harmonization and community involvement. Pragmatically, it will be impossible to create any standard terminology or any collection of standardized data elements that will meet the requirements of the global biomedical research and care delivery community. However, it is possible to identify a set of core data elements that are essential for aggregation of specific classes of data and agree upon a harmonized means of capturing this data. Such an effort clearly requires the involvement of the relevant communities of practice in the development of the harmonized standard so that an appropriate sense of ownership exists.

For vocabularies, a model currently exists for obtaining such a widely accepted terminology, federated terminology development as pioneered by the Gene Ontology (GO) [1]. GO has obtained a level of broad acceptance virtually unmatched among other biomedical terminologies by developing its content based on input from the people who will ultimately be using the terminology; fundamentally the terminology was authored by the community and organized by a small group of dedicated editors. Efforts organized in this way are typically more likely to have mechanisms for graceful evolution, since stakeholders can develop new content that is required due to changes in practice or understanding. Within caBIG™, the NCI's Enterprise Vocabulary Services (EVS) are creating a new federated terminology called Biomedical Grid Terminology (BiomedGT) that will utilize the GO model. The initial version of BiomedGT will be derived from the NCI Thesaurus (NCIt); however unlike NCIt, qualified subject matter experts from external organizations will be allowed to control the content of portions of the BiomedGT namespace. BiomedGT is deploying a variety of enabling technologies (such as a semantic media wiki, a specialized version of Protégé, and a new classifier) to allow for easy community input into content while still allowing a small group of editors to create a description logic based terminology.

Similarly, a number of examples of community based initiatives exist that are dedicated to the creation of standardized common data elements, or aggregations of common data elements such as clinical case report form (CRF) modules or information models. Organizations such as the Clinical Data Interchange Standards Council (CDISC) and its CDASH initiative, as well as the NCI's Clinical Trials Working Group have initiated such activities with regards to CRFs, while CDISC, Health Level 7 (HL7), the BRIDG project and caBIG™ have all begun work on common information models. All of these organizations operate based on community involvement and stakeholder consensus, and fortunately, have begun to work together to ensure common standards among these organizations. Recently, the UK NCRI and the US NCI (through the caBIG™ program) have begun investigating the possibility of jointly endorsing standards as a way to help drive a common consensus across international boundaries.

I. Issues Associated with Security

Issues associated with security (which broadly includes security, privacy and intellectual property protections) are some of the most vexing problems associated with the creation of broad (particularly transnational organizations).

Fundamentally, most of these issues resolve to locus of control questions (i.e. who is empowered to make decisions about authentication/authorization, data release, privacy, etc.) and variations in regulatory frameworks. The details of the GAARDS framework has been described elsewhere (see [5.6]). For purposes of this discussion, the important components of the framework are: Dorian, a system that accepts signed Security Access Markup Language (SAML) assertions from federated Authentications Services and provides Grid credentials in the form of X.509 certificates, the Grid Trust Service (GTS) that maintains the list of trusted certificate providers (including Dorian), and the Grid Grouper, a tool for creating and managing virtual organizations in a federated environment. The path to invocation of a secure service is indicated in Figure 3. The process begins when a client application authenticates to a local identity provider and receives a signed SAML assertion. This certificate is passed to Dorian, which validates the certificate against the list of trusted identity providers maintained in the GTS. If valid, Dorian issues a Grid credential in the form of an X.509 certificate, which is passed by the client to the secure service. The service then validates that the Dorian that issued the certificate is trusted (again by using a GTS). At that point, having validated identity, the service can make an authorization decision. GAARDS provides the Grid Grouper for role/group based access control, but this is not the only model. Authorization could be based on individual provisioning, or other mechanism, and the decision (as well as the mechanism for implementing that decision) is always made by the local service.

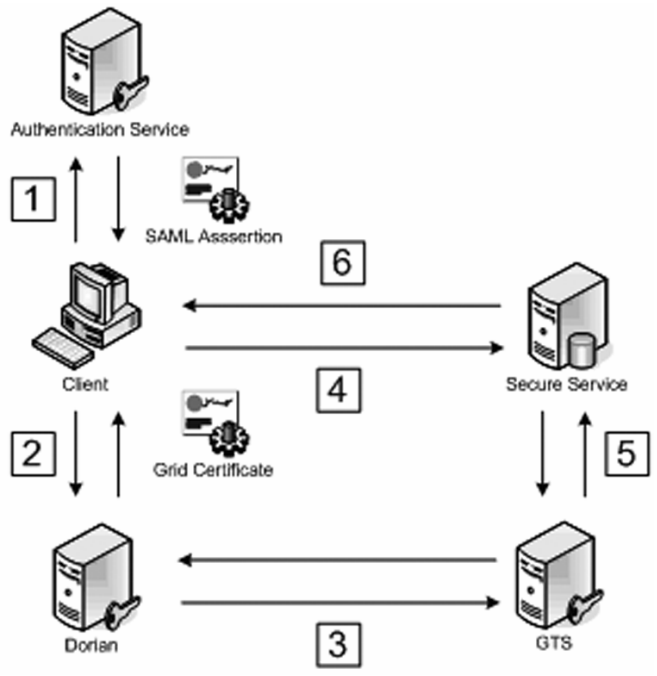

Fig. 3. Invocation of a secure service using caGrid

As indicated earlier, GAARDS (as well as all caGrid components) are designed to be deployed in a distributed fashion (much as the Domain Name Service is distributed). For example, most users of the NCI-caGrid are expected to utilize local identity providers (such as those provided by their university, Cancer Center or company) that supply appropriately configured SAML assertions. This is the only realistic arrangement to provision users across a potentially global environment. However, the ability to deploy multiple Dorian's provides the opportunity to create multiple trust fabrics within a common interoperable infrastructure. As an example, the NCI-caGrid may enter into a trust agreement with the inCommon Federation (this is actively being pursued in cooperation with the NIH) that would allow all users in inCommon to be authenticated in NCI-caGrid. Similarly, individual University systems, cooperative groups, etc.) may choose to create smaller trust fabrics at higher levels of assurance or for specific purposes (large multi-center trials for example). At a larger scale, the US NCI and UK NCRI are creating trust fabrics for their respective nations; specific bi-lateral negotiations can create a joint trust fabric even though the individual organizations maintain control of their own network of trusted identity providers.

In addition to the ability to create sub- or supersets of trusted organizations, the distributed nature of GAARDS allows for different policy frameworks in the various Grids. For example, the NCI-caGrid must obviously follow US law with regards to privacy, security and intellectual property. In the UK, different rules apply and the UK Grid will operate under the policies that implement their legal framework. Again specific bilateral negations will need to be applied to allow for authentication, authorization and data transfers possible.

II. Interoperability Standards

By definition, systems in a Grid of Grids will require some level of standards related to interfaces and common semantics in order to interoperate. Interoperability, however, is not binary, but rather there is a continuum of degrees of interoperability. Within caBIG™, this is captured in the four levels of compatibility (Legacy, Bronze, Silver, Gold). Obviously, however, this is not the only way to characterize interoperability; for instance, the UK NCRI has been examining a system that is more granular between the caBIG™ Bronze and Silver specification.

From the standpoint of the consumer of an information resource (and particularly from the standpoint of an organization looking to make a financial investment in such a resource) the level of interoperability is highly relevant. However, from the standpoint of a vendor of a piece of software having to demonstrate compatibility with standards to multiple organizations is inefficient and expensive, and could lead to fewer systems built to appropriate standards.

For this reason, it is highly desirable that interoperable Grids enter into cross-certification agreements allowing software that has met compatibility standards for one Grid to assert compatibility at an appropriate level for other interoperable Grids. The US NCI and UK NCRI are currently exploring such a cross-certification arrangement; the agreement may only apply to caBIG™ Silver systems and above (because of the above-mentioned granularity difference around systems that are comparable to caBIG™ Bronze).

III. Political/Social Issues

Particularly at an international level, political and community support for interoperability and data sharing are essential components of a successful Grid of Grids. In the United States, the US Department of State is required to negotiate most classes of international agreements. In the absence of support from the political components of a government, creation of such agreements is essentially impossible. Fortunately, the US has recognized the need for political leadership in this area, and an Office of the National Coordinator for Health Information Technology has been created as part of the Executive Office of the President.

IV. Conclusion – US/UK Partnership

The promise of personalized medicine is predicated on the aggregation and analysis of large amounts of data that has traditionally existed in "data stovepipes". The caBIG™ program sponsored by the US National Cancer Institute has created an infrastructure that allows for the creation of interoperable information systems for the purpose of enabling molecular medicine. However, the true promise of such technology is creating a global "Grid of Grids" that can bring such data together beyond oncology and beyond any single country. In 2006, the US NCI and UK NCRI formally partnered to achieve this global vision. Since initiation, the two groups have collaborated in a variety of areas. These include technical collaborations on infrastructure and portal technologies, cultural collaborations such as negotiations on trust fabrics, cross certification of compatibility and common semantics and an initial scientific collaboration involving data sharing with DICOM image data. The results of this collaboration will help inform further expansion of our Grid of Grids.

References

[1] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology." *Nat. Genet.* vol. 25, pp. 25-29, May 2000.

[2] G. A. Komatsoulis, D. B. Warzel, F. W. Hartel, K. Shanbhag, R. Chilukuri, G. Fragoso, S. de Coronado, D. M. Reeves, J. B. Hadfield, C. Ludet, and P. A. Covitz, "caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability." *J. Biomed Informatics*. vol. 40, pp. 106-123, Apr. 2008.

[3] K. M. Kudla, and M.C. Rallins, "SNOMED: a controlled vocabulary for computer-based patient records." *J. Ahima* vol. 69, pp. 40-44; quiz 45-46, May 1998.

[4] C. J. McDonald, S. M. Huff, J. G. Suico, G. Hill, D. Leavelle, R. Aller, A. Forrey, K. Mercer, G. D. DeMoor, J. Hook, W. Williams, J. Case, and P. Maloney, "LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update." *Clin Chem* vol. 49, pp. 624-633, April 2003.

[5] S. Oster, S. Langella, S. Hastings, D. Ervin, R. Madduri, J. Phillips, T. Kurc, F. Siebenlist, P. A. Covitz, K. Shanbhag, I. Foster, and J. Saltz, "caGrid 1.0: An Enterprise Architecture for Biomedical Research." *J. Am Med Inform Assoc.* vol. 15, pp. 138-149, Mar.-Apr. 2008.

[6] J. Saltz, S. Oster, S. Hastings, S. Langella, T. Kurc, W. Sanchez, M. Kher, A. Manisundaram, K. Shanbhag, and P. Covitz, "caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid." *Bioinformatics* vol. 22, pp. 1910-1916, Aug. 2006.

[7] N. Sioutos, S. de Coronado, M. Haber, F. Hartel, W. Shaiu, and L. Wright. "NCI Thesaurus: A Semantic Model Integrating Cancer-Related Clinical and Molecular Information." *J. Biomed Informatics* vol. 40, pp. 30-43, Feb. 2007.

[8] Staff, Institute of Electrical and Electronics Engineers, *IEEE Computer Dictionary-Compilation of IEEE Standard Computer Glossaries 1990*. pp. 610, 1990.