Adoption and Adaptation of caGrid for CTSA
Shannon Hastings M.S., Scott Oster M.S., Stephen Langella M.S., Calixto Melean M.S., Tara Borlawsky M.S., Rakesh Dhavel M.S., Philip Payne PhD., The Ohio State University, Department of Biomedical Informatics, Software Research Institute, and Center for Clinical and Translational Science, Columbus, Ohio

Abstract
The field of informatics has been going through a rapid change over the past decade. New technologies such as grid computing[1-5] and knowledge anchored data, combined with major funding and growing community thrusts aimed at creating a richer multi-institutional research and clinical environment such as caBIG™[6-8] (Cancer Bioinformatics Grid), BIRN[9] (Bioinformatics Research Network), and CTSA(Clinical and Translational Science Awards) have lead to new ways to bring together information across institutional boundaries. This had lead to service oriented architectures based developments in creating semantically interoperable data and analytical services to increase speed, efficiency, and outcome of clinical and research efforts spanning the fields of medicine. The TRIAD (Translational Informatics and Data Management Grid) System, which will be used as the middleware system enabling the OSU CTSA to create a scalable, secure, and knowledge anchored data sharing environment, will adopt and adapt the caGrid infrastructure designed for the caBIG™ program.

Overview
The Software Research Institute (SRI) at the Ohio State University has been involved in various capacities with major efforts in creating and adopting services oriented science infrastructures in programs such as caBIG, CVRG, BIRN and the CTSA. For example, since the inception of caBIG™, SRI has provided the role of lead architect and lead development site for caGrid. caGrid is the middleware system designed for caBIG™ in order to create a loosely coupled yet highly interoperable grid service oriented architecture (SOA). caGrid is a generic software system comprised of grid middleware, tools, and services that can be leveraged to create a service oriented architecture that is secure, distributed, strongly typed, and semantically interoperable. The NIH/NCI caBIG™ program uses caGrid as the backbone to the cancer research grid infrastructure. Ohio State University has recently been awarded a Clinical and Translational Sciences Award (CTSA) from the National Center for Research Resources (NCRR). Within the OSU CTSA program we are faced with many of the same challenges encountered in the caBIG™ community:

- Physically and logically disparate community participants.
- Multi-institutional security interoperability and policy issues.
- Use of new technologies that require training, expertise, and process change.
- Complex federal and local data integrity and privacy constraints.
- Semantic and syntactic differences in data within and across research groups and institutions.

CTSA and CaGrid

The overlap of requirements for CTSA and caBIG™ has led to the decision to use caGrid as the core middleware for the OSU CTSA system called TRIAD as shown in Figure 1. caGrid will enable the CTSA programs to easily create a service oriented architecture that spans local institutions, collaborators, and researchers participating in the OSU CTSA along with the growth capabilities to create a nationwide CTSA grid based infrastructure in collaboration with other CTSA institutions. We believe this will be the future model leveraged across all CTSAs locally and as a whole across the program.
The caGrid infrastructure and tooling is a domain agnostic software system that can be leveraged for other research and practice domains beyond cancer. The caBIG™ program, which funded the development of caGrid, uses it as the core middleware system for the cancer research grid, however, caGrid is a generic set of services and tools that can be leveraged by other grid enablement efforts. caGrid is a large software system comprised of many cohesive yet decoupled components that will be adopted and adapted for reuse in the OSU CTSA TRIAD infrastructure. Using caGrid will enable TRIAD to immediately support many of its requirements by creating a grid-enabled service oriented architecture with semantic interoperability and strongly typed grid services. Using caGrid out of the box will enable TRIAD to leverage the following capabilities:

- Service oriented architecture utilizing Axis, Globus, and caGrid extensions.
- Data model management via the Global Model Exchange.
- Customizable security infrastructure leveraging GAARDS.
- Strongly typed and semantically annotated grid service creation using Introduce.
- Extendable and customizable grid service infrastructure.

Semantics

One of the fundamental aims of TRIAD is to create a semantically interoperable service oriented architecture that enables seamless creation and use of data and applications across all the critical areas of research and practice in medicine. In the caBIG™ program semantic interoperability is maintained by utilizing data models, represented as UML models, which are annotated with concepts and controlled vocabulary from the caBIG™ EVS system. The EVS manages the caBIG™ community’s global ontologies and controlled vocabularies. This use of this system and process comes with a rigorous and costly overhead consisting of upfront curation and human interaction in order to add new concepts and vocabularies to the single EVS system. This type of upfront expectation of knowledge creation and human processing to get ontologies and vocabularies into the EVS is prohibitive to the process of evolving local ontologies into standards over time. It seems to be most effective when all knowledge is either known upfront or will be created and curated upfront prior to its use. Earlier work[10] has shown that there are gaps in the knowledge supported by current standard ontologies and terminologies. In order to support creation and curation of new knowledge which may not yet be represented in current standard terminologies the TRIAD system aims to enable local ontology development coupled with global ontology publication and curation. TRIAD will support this type of knowledge development and management through software and practices that are not burdened with the requirements of large community consensus upfront. This will enable early use of potentially partially complete ontologies as well as the evolution of ontologies in the local and global areas into community accepted standards over time.

To support this process we leverage the LexBIG system developed at the Mayo Clinic. LexBIG is the main software system backing the most recent version of the caBIG™ EVS (Enterprise Vocabulary Services). This open source system, along with TRIAD customizations, extensions and tooling will enable the creation and curation of local ontologies created using tools such as LexBIG, Apelon DTS, Protégé[11], or others and enable those ontologies to be automatically published under controlled namespaces to a local ontology service (i.e. the OSU CTSA ontology and vocabulary service) and immediately used. Over time these disparate ontologies may be mapped together or curated into more standard ontologies as community adoption and consensus is reached. As the CTSA organizations begin to share data across themselves we see the evolution of these local ontologies into global ontologies or ontologies shared between subsets of disparate collaborators.

We also leverage BioPortal[12], an open source web portal developed by the National Center for Biological Ontologies, can leverage a LexBIG ontology system and provides a web-based view. This application provides the ability to search, browse, and view graphically the ontologies and terminologies available in the LexBIG. This will enable our user community to easily discover and learn about the semantics currently available as well as help them determine where extensions or modifications may need to be added for their particular domains.

Data Modeling and Management

In caBIG™, both the client and service programming interfaces are object oriented, and operate over well-defined and curated data types which are modeled in semantically annotated UML. Clients and services communicate through the grid using respectively grid clients and grid service infrastructure. The grid
communication protocol is XML, and thus the client and service APIs must transform the transferred objects to and from XML. This XML serialization of caGrid objects is restricted in that each object that travels on the grid must do so as XML which adheres to an XML schema[14] registered in the Global Model Exchange (GME)[15]; a repository of rules on the structure of XML documents used on the grid. This process creates an abstraction layer wherein clients and applications largely ignore the lower-level XML format, and work with information represented in an object-oriented fashion (as opposed to XML documents, or rows and columns in a database). This approach has been fairly successful in caBIG™ due to the tooling available, such as the caCORE SDK and Introduce toolkit, which can automatically handle the details of mapping to and from lower-level XML view. However, it provides additional infrastructural challenges in that there must be a formal and programmatically accessible mapping between the layers, such that tools and applications that need to work at differing abstraction layers can interoperate. For example, traditional workflow tools work purely with XML, but users in this environment expect to be able to describe analytical pipelines in the abstraction which they are normally presented: method invocation of object-oriented services. To address this in caBIG™, the Cancer Data Standards Repository (caDSR) provides a registry of semantically annotated UML models. Work is currently underway to add to that registration the mapping of those models to their respective XML formats, via binding to XML Schema components in the GME, a grid service for publishing and sharing XML schemas. Currently TRIAD leverages the NCI knowledge management applications such as LexEVS, caDSR, and GME but does not use the single point monolithic review process required to get data models, terminology, and ontology created and inserted into these systems. We envision extending or replacing the tools for data model and ontology management over time so that there is better support for enabling models and semantics to be created more in a more agile fashion by the domain experts themselves and evolve over time into standards as community expert consensus is reached.

Service Development

Services that make up the caGrid core service oriented infrastructure are architected and implemented with a tool called Introduce[13]. Introduce enables graphical creation of grid services and is extendable to create services that have specific features that specific grid deployments may desire. This tool enables TRIAD application developers to quickly design and build services without needing to train them to become grid computing experts. It will also empower the CTSA to create custom extensions that enable all services to have a common set of functionality. For example, a CTSA Introduce Service Extension, much like some custom caBIG™ extensions that, may enable the following capabilities:

- Common metadata that can be used to identify them.
- Ability to register their metadata to a CTSA Index Service so they can be discovered.
- Ability to easily leverage shared and standard data types and semantic annotations.
- Point-and-click service security configuration.

Using Introduce and leveraging its extension framework will enable TRIAD to provide tooling that will enable rapid creation and deployment of grid services. This will shorten the time it takes to gain results from using TRAID and will enable rapid adoption across the research and practice domains. Introduce will also lower the barrier to entry for building these services as it hides the complexity of the service oriented architecture as well as the complex security framework. It will enable development teams to build services and applications utilizing TRIAD without needing to become grid-computing experts or security experts and enable them to focus on domain they are working in. This has already been proven to work in the cancer research community with the wide adoption of Introduce within the caBIG™ program and we believe it will be successful in TRIAD as well.

Introduce will also enable services in TRIAD to be semantically annotated such that they will have terminology-based service metadata, leveraging the Metadata Extension available in caGrid, that is registered to an Index Service that can be discovered and searched. This metadata will contain semantic annotations that refer to common data elements, concepts, and vocabulary from the TRIAD knowledge management services described earlier. Annotations will be placed on items such as the data models used to represent that input and output parameters of each method the service provides as well as standard service metadata describing the service’s provider and its use.

Security

caGrid includes an extensive security framework known as GAARDS[17, 18]. The Grid Authentication and Authorization with Reliably Distributed Services (GAARDS) infrastructure
provides services and tools which enable the creation, administration, and enforcement, of internal and cross institution security policy in a service oriented architecture. Because caGrid utilizes the Globus Toolkit as the underlying service oriented toolkit, GAARDS is implemented on top of Globus and designed to leverage and enhance security capabilities in Globus. More specifically, it extends the Grid Security Infrastructure (GSI) to provide enterprise services and administrative tools for: 1) grid user management, 2) identity federation, 3) trust management, 4) group/VO management 5) Access Control Policy management and enforcement, and 6) Integration between existing security domains and the grid security domain. GAARDS is designed in a flexible way that enables services to be used individually or in combination together to meet the authentication and authorization needs for Grids.

TRIAD leverages many features of GAARDS in order to create a scalable and secure grid. Dorian, the GAARDS identity management service, is used as the main certificate authority and short-term certificate provider for our grid services and users. Certificates are used in the grid for the identification of the parties communicating on the grid. Dorian, in combination with an OSU local Identity Provider, enables Dorian to automatically issue certificates to users in TRIAD authorized by the OSU local LDAP system. The Identity Provider is responsible for checking the credentials of a user and certifying that they properly authenticated with the local identity framework and then sending an assertion to Dorian confirming a successful local authentication. For example, if TRIAD deploys an OSU Identity Provider which can authenticate users against the OSU campus wide LDAP, that Identity Provider can be used to let Dorian know if a user successfully authenticated to the OSU system and use that assertion to issue them short term TRIAD credentials to be used for grid wide communication. This enables quickly bringing in groups of users to TRIAD by simply trusting their Identity Provider and not requiring the addition, or central management of, every user to Dorian individually. Similarly, additional local Identity Providers can easily be deployed and trusted by Dorian. This federated identity management solution, provided by GAARDS, enables institutions to vet and control their own users locally yet provide them globally trusted grid credentials. As institutional participation in TRIAD grows, managing the trusted certificates and revocation lists becomes a challenging task of the grid administrator. TRAID uses the GAARDS Grid Trust Service (GTS) and SyncGTS framework to easily and programatically maintain the trust fabric of TRIAD as it changes over time. The combination of these GAARDS services enables TRIAD to span multiple institutional boundaries and still provide a safe and secure service oriented data sharing environment.

TRIAD

TRIAD is currently up in running at OSU, as shown in Figure 2, in its first prototype deployment. It is running in an enterprise virtualization environment in across several virtual machines each running grid service containers containing different core grid services. It contains its own GAARDS based grid security services (integrated with OSU Medical Center domain controller), LexEVS (pre-polulated with NCI Thesaurus), BioPortal, caDSR, and all accompanying caDSR web administration and user tools such as CDE Model Browser and the SIW. This installation has proven to be quite useful as we are now able to rapidly create translational informatics services that are strongly typed, semantically annotated, and secure. We have defined a process that enable our knowledge engineers to model and semantically annotate the data model. Then the model is processed in Introduce with the caCORE SDK grid service wizard to create the grid data service capable of storing and querying data represented by the model. The combination of these GAARDS services enables TRIAD to span multiple institutional boundaries and still provide a safe and secure service oriented data sharing environment.
semantics. The process by which we create and manage semantically annotated data models is still under research and development. Our current approach is to use the tools provided by the NCI but excluding the monolithic data model review process. We would like ot evolve this process into a more federated system so that model developed and managed at other institutions can easily be harmonized with other institutions or programs using potentially a different set of tools and applications.

Conclusion

Ultimately, the TRIAD architecture is intended to provide a robust, scalable, and extensible platform capable of supporting the exchange, management, integration, and dissemination of heterogeneous and multi-dimensional data endemic to the translational research setting. By adopting open standards and software engineering best practices throughout the design and implementation of TRIAD, our objective is to enable a seamlessinformatics pipeline that begins with the creation of semantically annotated domain models and culminates in the rapid deployment of strongly typed grid-services that are able to fully interoperate with shared services spanning the TRIAD grid-computing fabric. As the development of TRIAD continues, we envision the implementation of not only the previously described grid-computing infrastructure, but also the creation of an integrative “workbench” environment, leveraging the Eclipse framework, that is able to fully realize the informatics pipeline we have described. In doing so, our goal is to lower the barriers to adoption and use of TRIAD by providing domain experts and software engineers with a highly usable toolkit capable of supporting the rapid, standards-compliant deployment of grid services. We believe that the combination of TRIAD, better management and integration of disparate knowledge management tools, and such a “workbench” are necessary and highly desirable in terms of achieving the transformative goals of the CTSA program, and enabling interoperability of CTSA sites with the broader biomedical research community, thus supporting true, team-science approaches to previously intractable biological, clinical, and public health research problems.

References

1. Foster, I. and C. Kesselman, The grid : blueprint for a new computing infrastructure. 1999, San Francisco: Morgan Kaufmann Publishers. xxiv, 677 p.
2. Foster, I., Service-oriented science. Science, 2005. 308(5723): p. 814-817.
3. Foster, I., The grid: Computing without bounds. Scientific American, 2003. 288(4): p. 78-85.
4. Foster, I. and C. Kesselman, Globus: A metacomputing infrastructure toolkit. International Journal of Supercomputer Applications and High Performance Computing, 1997. 11(2): p. 115-128.
5. Foster, I., C. Kesselman, and S. Tuecke, The anatomy of the grid: Enabling scalable virtual organizations. International Journal of High Performance Computing Applications, 2001. 15(3): p. 200-222.
6. Saltz, J., et al., A roadmap for caGrid, an enterprise Grid architecture for biomedical research. Stud Health Technol Inform, 2008. 138: p. 224-37.
7. Oster, S., et al., caGrid 1.0: an enterprise Grid infrastructure for biomedical research. J Am Med Inform Assoc, 2008. 15(2): p. 138-49.
8. Saltz, J.H., et al., Integrating heterogeneous rule-engine technologies with caGrid. AMIA Annu Symp Proc, 2007: p. 1099.
9. Grethe, J.S., et al., Biomedical informatics research network: building a national collaboratory to hasten the derivation of new understanding and treatment of disease. Stud Health Technol Inform, 2005. 112: p. 100-9.
10. Richesson, R.L. and J. Krischer, Data standards in clinical research: gaps, overlaps, challenges and future directions. J Am Med Inform Assoc, 2007. 14(6): p. 687-96.
11. Protege. Available from: http://protege.stanford.edu/.
12. NCBO BioPortal. Available from: http://biportal.bioontology.org/.
13. Hastings, S., et al., Introduce: An Open Source Toolkit for Rapid Development of Strongly Typed Grid Services. J. Grid Comput., 2007. 5(4): p. 407-427.
14. XML Schema. Available from: http://www.w3.org/XML/Schema.
15. Hastings, S., et al., XML database support for distributed execution of data-intensive scientific workflows. Sigmod Record, 2005. 34(3): p. 50-55.
16. Semantic Annotations for WSDL and XML Schema. 2007; Available from: http://www.w3.org/TR/sawSDL/.
17. Langella, S., et al., Sharing data and analytical resources securely in a biomedical research Grid environment. J Am Med Inform Assoc, 2008. 15(3): p. 363-73.
18. Langella, S., et al., The Cancer Biomedical Informatics Grid (caBIG) Security Infrastructure. AMIA Annu Symp Proc, 2007: p. 433-7.