A Distributed Information Services Architecture to Support Biomarker Discovery in Early Detection of Cancer

Daniel Crichton[1], Sean Kelly[1], Chris Mattmann[1, 2], Qing Xiao[1], J. Steven Hughes[1], Jane Oh[1], Mark Thornquist[3], Donald Johnsey[4], Sudhir Srivastava[4], Laura Essermann[5], William Bigbee[6]

[1]*Jet Propulsion Laboratory*
California Institute of Technology
Pasadena, CA, USA 91109
{crichton,kelly,qxiao,hughes,joh}@jpl.nasa.gov

[2]*Computer Science Department*
University of Southern California
Los Angeles, CA 90089
mattmann@usc.edu

[3]*Fred Hutchinson Cancer Research Center*
Seattle, WA 98109
thornquist@fhcrc.org

[4]*National Cancer Institute*
National Institutes of Health
Bethesda, MD 20892
{johnseyd, srivasts}@mail.nih.gov

[5]*University of California, San Francisco*
San Francisco, CA 94143
laura.esserman@ucsfmedctr.org

[6]*University of Pittsburg*
Pittsburgh, PA 15213
bigbeewl@upmc.edu

Abstract

Informatics in biomedicine is becoming increasingly interconnected via distributed information services, interdisciplinary correlation, and cross-institutional collaboration. Partnering with NASA, the Early Detection Research Network (EDRN), a program managed by the National Cancer Institute, has been defining and building an informatics architecture to support the discovery of biomarkers in their earliest stages. The architecture established by EDRN serves as a blueprint for constructing a set of services focused on the capture, processing, management and distribution of information through the phases of biomarker discovery and validation.

1. Introduction and Scientific Drivers

The Early Detection Research Network (EDRN) of the U.S. National Cancer Institute (NCI) has established a research network of collaborating scientists from over 30 institutions focused on identifying and validating cancer biomarkers at their earliest stages. "The work of the EDRN is concentrated on … the discovery of markers, the validation of markers in distinguishing the presence or absence of cancer, and testing markers for the ability to detect preclinical and early-stage disease" [1]. If such markers can be identified and validated, researchers are hopeful that it will help in treatment of cancer. Coordinated discovery of biomarkers across cancer research centers provides an opportunity for the EDRN to increase the accuracy of the results of studies. However, the distributed nature of the EDRN coupled with the lack of data standards represents a modern challenge for building a bioinformatics infrastructure to capture and distribute the science and ancillary data acquired during biomarker studies within the enterprise. The EDRN has a scientific need for an infrastructure for linking highly diverse systems together into a virtual *data grid* [2] to support new analysis mechanisms ultimately identifying and validating new biomarkers. NASA's Jet Propulsion Laboratory (JPL), one of the key players of the EDRN Informatics Working Group, has been leading the ongoing effort in constructing the data grid architecture and infrastructure that supports the science-driven research needs of the EDRN [3]. The first application developed for EDRN focused on providing a common informatics framework for accessing heterogeneous bio-specimen repositories located at participating EDRN sites across the U.S. As the infrastructure has evolved, the core principles of building services that integrate general client applications with heterogeneous, distributed data resources have not changed. The EDRN has recognized the need to build a

"knowledge system" where bio-specimens, scientific data, study-specific data, and biomarker data can be captured, accessed, and shared at a national level via a transparent, grid-type architecture [4]. As a result, EDRN is focused on addressing five critical informatics goals: (1) defining an information model for describing the EDRN problem space; (2) enabling all components of the knowledge system to be distributed; (3) providing software interfaces for capture, discovery, and access of data resources across the knowledge system; (4) providing a secure transfer and distribution infrastructure to meet U.S. federal regulations for data sharing; and (5) providing an integrated portal environment across the distributed EDRN.

The rest of this paper is organized as follows: Section 2 describes the overall system architecture of EDRN, its motivations, and its high level components. Section 3 highlights EDRN information model and its relationship to the EDRN architecture. Section 4 discusses efforts within EDRN to warehouse scientific data. Section 5 describes how EDRN addresses distributed resource discovery. Section 6 overviews the EDRN Portal. Section 7 surveys related work and Section 8 concludes the paper.

2. Architecture

One of the critical characteristics of the architecture has been leveraging architectural patterns across very different science environments [2, 4]. Due to JPL's involvement, the software architectural models for supporting *planetary science* have been leveraged to develop the EDRN architecture. The architecture definition focused on definition of both the *information* and *functional* portions of the EDRN knowledge system.

As part of defining the architecture for any science-oriented data system, we identified these particular functions as having architectural patterns [5] for which a set of services could be implemented. The services include data capture, discovery, access, retrieval, processing, and distribution. Because of EDRN's distributed nature, these services should allow for distributed, independent deployment yet still work in concert with one another allowing virtual systems that span organizational boundaries to be constructed. To meet U.S. federal regulations for sharing health science data, each service requires the ability to limit access to trusted peers.

In addition to the above described functional architecture, there is also the *information* architecture that is critical to forming an integrated informatics platform. As part of designing the information architecture, the EDRN has been developing an information model for the representation of information associated with data objects managed within the knowledge system. This includes models for such objects as biomarkers, studies, participants and organs. To support definition of the information model, the Fred Hutchinson Cancer Research Center (FHCRC) led development of a set of common data elements (CDEs) for describing these data and their associated attributes, thus providing a common language for communicating information across the EDRN scientific discovery lifecycle. The EDRN CDEs have been agreed upon by the EDRN investigators as critical attributes of information that must be collected by all EDRN sites.

The EDRN worked with JPL to adopt the Object Oriented Data Technology (OODT) framework [6, 7] as the foundation for the EDRN informatics infrastructure. OODT provides a set of core services that implement the above functions which themselves are driven by a domain model (i.e. the biomarker information model). The loose-coupling between each service and its associated domain model allows for the services to support other domains besides cancer research (e.g., planetary science). Each of the OODT services can be deployed independently and then integrated using XML-based interfaces over a distributed, grid architecture. This service independence makes it possible to query multiple institutional repositories concurrently, compiling the results into a unified view, and making them available for analysis. The OODT service framework is based on the software architectural notion of components [8, 9]. Each component has well known interfaces that enable them to be plugged together in a distributed manner. The components themselves sit on top of off-the-shelf middleware (such as SOAP [10] services) so that they can be deployed into an enterprise topology.

The *Catalog and Archive Service* component provides the ability to catalog, process, and store information objects in a distributed environment. The *Profile Service* component provides a registry of information about managed information objects necessary to discover them. Multiple profile services can be distributed and integrated into a directed graph topology in order to crawl the registries and locate critical information objects. The *Product Service* provides a mechanism for access, retrieval, and transformation of science data products and information from remote repositories. The *Query Service* provides an interface to distributed components so that they work together.

EDRN is working towards building the knowledge system (recall Section 1) on top of these services. A series of use cases were developed that identify various scientific discovery scenarios that drive the construction of the knowledge system. These scenarios

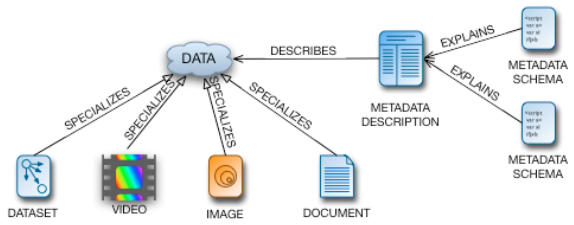

Figure 1. EDRN Information Model

identified the need to develop applications across the EDRN to support the critical functions for managing the data and the workflow associated with EDRN's various types of information (e.g., see Figure 1). Figure 2 demonstrates the complete knowledge environment that provides virtualized access to the distributed applications via a centralized scientific portal. It is important to recognize that the EDRN information model is used as an over-arching data architecture for inter-relating the data objects captured in the components shown in the knowledge system in Figure 2. The portal provides secure access in compliance with U. S. federal regulations.

3. Information Model

The EDRN domain information model drives the system by describing the data objects that are captured and managed within the EDRN knowledge system. At the core of this model is the identification of a standard set of metadata elements that can be used for annotating these objects. Without metadata, there is little chance of discovering data, interpreting it correctly, and reusing it in an automated fashion. As an example, consider Figure 1 that identifies how certain data object classes within the knowledge system are

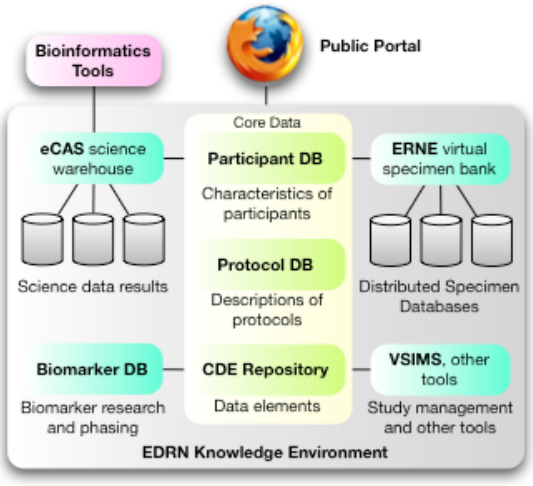

Figure 2. EDRN Knowledge Environment

identified by a metadata description derived from a standard schema:

Multiple metadata schemata provide machine usable explanations of a metadata description, which serves to describe the inception and composition of data. Data can come in a variety of flavors, including tabular datasets, videos, images, documents, and other formats. The set of EDRN CDEs is a standard data dictionary of data elements implemented with the ISO/IEC 11179 standard [11]. The data dictionary provides the elements by which these metadata schemas can be constructed.

The CDEs are vital towards making EDRN applications interoperable. CDEs serve as a standard vocabulary for describing each element including data types, units, representation, enumerated legal values, legal ranges of values, and so forth. With CDEs, software can perform automatic validation, form generation, and scientific correlation.

As an example, consider a typical CDE for "cell count"; a sampling of the attributes for this element is as follows:

Attribute	**Definition**
Contexts	EDRN, validation studies
Object class	Specimen
Property	Bone Marrow
Name	SPECIMEN_BONE_MARROW-CELL_COUNT
Version	1.0
Title	Cell Count
Description	The number of cells (in millions) in a stored cell specimen taken from bone marrow.
Data type	Integer
Minimum value	0
Maximum value	Unlimited
Units	count
Status	Approved

When a piece of data such as "196" is presented with metadata "cell count", software automatically knows how to (1) validate the data when coming from unreliable sources (hand data entry, for example); (2) store the data in short-term storage in an object; (3) persist the data into long-term storage in a database with correct typing; and (4) display the data with appropriate labels and help text.

Every CDE also has its own unique ID in the form of a Uniform Resource Identifier (URI) [12]. These take the forms of URLs that point to the CDE's definition pages. Because they are URIs, they can be used in XML-based standards that expect URIs, such as XHTML, XLink, and so forth.

4. Warehousing and Workflow

A fundamental goal of the EDRN knowledge system has been the capture, processing and warehousing of scientific data generated during a biomarker validation study. As stated earlier, researchers develop validation studies in an effort to identify biomarkers and disease markers, which may indicate the existence of or the potential for cancer [13]. Much of a biomarker study centers on the use of sophisticated screening technologies that obtain data using instrumentation from various populations of study. A large set of data samples means more study power in determining correlations in marker and disease activity allowing proposed markers to be validated. Often studies are conducted at multiple laboratories increasing the power associated with a correlation. As a result, capturing and sharing the information across the EDRN network is critical.

The *EDRN Catalog and Archive Service* (eCAS) is a distributed metadata-driven system for the capture, tracking, processing and retrieval of scientific data from biomarker validation studies. eCAS promises to be an invaluable tool that will make it possible for doctors, scientists, clinicians, and researchers to share their results, correlate their data, discover promising knowledge of new biomarkers, and more. As an example, a researcher tabulating results from a spectrograph has access to hundreds of data points. Finding relative minima and maxima creates new data points. Combining that data with calibration information for the instrument refines that data. Correlating those results with other spectrograph runs creates information for tracking changes over time, between specimens sampled, between instruments manufactured, and more. Each of these states requires a workflow-oriented system [14] for managing the information and the associated processing steps. eCAS implements four specific strategies: (1) eCAS tags every datum with an open-ended set of extensible metadata, (2) eCAS tracks every datum such that the system knows its location, version, and distribution status at every moment, (3) eCAS employs open web standards, including Transport Layer Security (TLS), Resource Description Framework (RDF) [15], XML, and HTTP, and (4) eCAS implements a underlying workflow engine for managing each task. In the envisioned deployment of eCAS, institutions run eCAS software servers that participate in a peering network grid, creating a virtual organization that transcends institutional boundaries.

The eCAS enables scaleable and transparent replication of data and metadata, improving availability and reliability. This architecture is mirrored within institutions, creating virtual departments; and within departments, creating virtual workgroups. Flexible access controls enable researchers to designate that a particular datum be available to specific users, workgroups, departments, organizations, and so forth.

In addition to mechanisms for capturing and virtualizing access to data, an integrated cancer platform requires the introduction of mechanisms for data processing and generation of scientific data results. Based on our experience of delivering these types of systems for NASA's planetary and earth science missions, we believe this will prove to be a critical capability for the NCI. Such a capability will require that researchers be able to reprocess lower level science products in order to verify research results. In addition, EDRN will be able to integrate production of science data products all the way to the instrumentation. This will ensure consistency in how data sets are produced. We are working to make such capabilities available within a grid environment such that remote and server-side processing can occur as part of the integrated platform for EDRN. In addition, high performance and cluster computing for computationally constrained algorithms can be executed remotely on devices that can handle such demands. This will be integrated as part of the workflow infrastructure of the eCAS software.

5. Discovery and Access

In addition to eCAS described above, there are other applications being integrated into the knowledge environment which provide access to critical information necessary for biomarker research including access to information about stored specimens along with annotated information about biomarkers and their phase of study. The *EDRN Resource Network Exchange*, or ERNE, is a virtual specimen bank unifying disparate and scattered specimen repositories. The *Biomarker Database* is a tracking system for biomarker research, including collection of such data as phase of development, studies and related trials, and other data. Other applications are already deployed and in development by both the FHCRC and JPL

The eCAS, Biomarker Database, and ERNE applications are all *realized* using combinations of the aforementioned OODT basic data grid components: the Catalog and Archive Service, Profile Service, Product Service, and Query Service discussed in Section 2. These realizations are made possible using a thin layer of "glue code" that provides for the communications pathways between the components and visible user interfaces. In addition, by leveraging the biomarker domain information model, these applications are interoperable, supporting discovery of new relationships between disparate data and information sources and allowing for queries and different

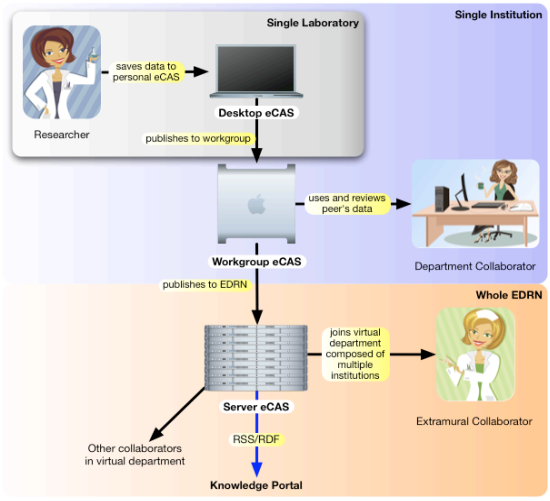

Figure 3. eCAS Deployment Scenario

application views of the information available within the knowledge system.

5.1. Component Architecture

The eCAS is built on the OODT Catalog and Archive Service. It also uses the Profile Service to search and return XML descriptions of the science data captured within eCAS. In addition, the Product Service provides eCAS with the ability to retrieve data from the system providing on-demand transformation functions that can be executed in real-time. The OODT Profile and Product server components enable distributed eCAS deployments to be plugged into the knowledge system to share the science data results in the warehouse.

The EDRN Resource Network Exchange (ERNE) uses the OODT Profile Service to optimize its query before being sent to every participating product server. The Profile Service contains metadata that describes what kinds of specimens are available at each ERNE site and various other parameters. For example, when a user searches for blood specimens, ERNE uses the Profile Service to determine which sites have blood. It then queries the Product Service running at each site. At each site, the Product Service's job is to translate the specimen query from the common format into its site-specific format, gather the results, and then translate them into the common format before returning back to ERNE. This will be discussed in detail in the next section describing the ERNE deployment.

The Biomarker Database, being a rather unique application, makes use only of the Profile Service. Its use is in providing a metadata lookup facility in order to attach an open-ended set of metadata to biomarkers being tracked.

Figure 4 depicts the relationship between the EDRN application components and the core OODT data grid services.

5.2. Application Deployment

Each of the EDRN applications is deployed to participating sites in different ways, reflecting the unique nature of their interactions with users. Where possible, we strive to make installation as simple as possible given constraints on platforms, existing data, and existing processes.

For eCAS, installation may take place on a desktop, workgroup, or server system. Any system running an eCAS installation becomes an eCAS server, capable of accepting data, cataloging it, running post-processing tasks, versioning data, and so forth. Within EDRN, we are implementing a centralized eCAS managed at the National Cancer Institute along with deplyoing eCAS directly at the EDRN cancer research centers. Figure 3 demonstrates the deployment scenario.

The ERNE application takes a different form from eCAS. Any site that would like to share its specimens first constructs a map between the EDRN Common Data Elements (CDEs) and their local data model for their specimen repository. The CDEs provide a uniform vocabulary for querying for and describing specimens. Using a web application, a site registers the details of its specimen database, including such information as database type and operating system platform, table schema, logical organization of specimens, participants, histology, and other records, and the relationship of such organization to the CDEs.

Researchers pose queries for specimens using a centralized web application. That web application then sends the query (in CDE format) to product servers that

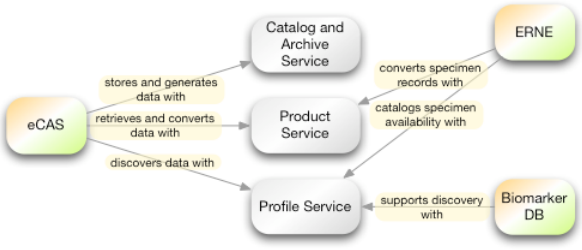

Figure 4. EDRN's use of OODT components

can answer it. Product servers convert the query, access their local databases, convert the results, and return them to the web application for display. It is important to point out that the system will continue to evolve as researchers identify more attributes for granularly describing specimens. Figure 5 depicts the deployment.

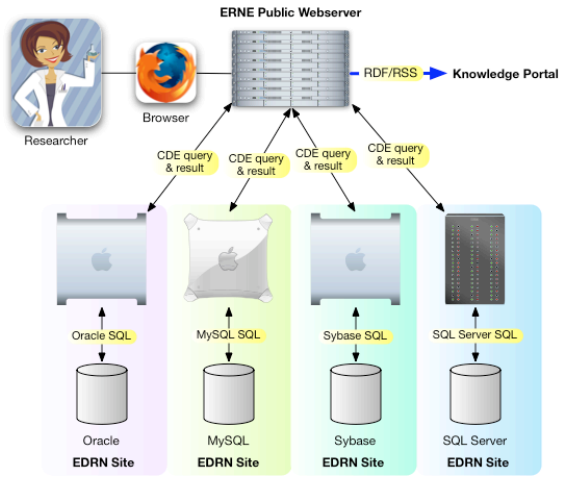

Figure 5. ERNE Deployment Scenario

While eCAS is deployed using the peer-to-peer architectural style [16] and ERNE is more of a client-server style, the Biomarker Database is a centralized application. The Biomarker Database enables researchers to track the progress of biomarker development through its various phases [17]. Search functions enable researchers to find interesting biomarkers, participate in clinical trials in support of biomarkers, and share data regarding biomarkers. Figure 6 shows the deployment.

Biomarkers are entities tracked within the database whose states change over time based on a specific

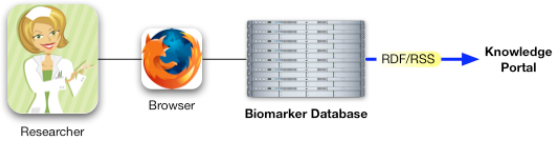

Figure 6. Biomarker DB Deployment Scenario

biomarker development workflow. Each phase of a biomarker is tracked throughout the discovery process from "Preclinical Exploratory Studies" to "Cancer Control Studies" [17]. In this regard, the Biomarker Database is much like an issue tracking system which tracks the current state of a biomarker. The biomarkers themselves are annotated using the EDRN CDEs which allow them to be related to data objects captured in other EDRN applications. Because each of the EDRN applications uses the same set of CDEs, they can interoperate and automatically correlate information. This forms the logical EDRN Knowledge System. For example, a biomarker tracked in the Biomarker Database can link to a cell count in a specimen record in ERNE; the result of an analysis can reference that cell count stored in eCAS; etc. Providing access to this information through a shared mechanism provides an integrated view of the information within the EDRN enterprise.

6. Semantic Knowledge Environment

Each of the EDRN applications described thus far—eCAS, ERNE, and the Biomarker Database—all draw from the set of CDEs defined for EDRN. This enables those applications to *interoperate* by sharing a common vocabulary for describing the entities relevant to each application and correlating them with cooperating applications. In addition, these and future applications *publish their collective knowledge* to a unified knowledge system portal.

EDRN runs a fourth application in the form of a *public portal* that enables doctors, researchers, clinicians, cancer sufferers, patient advocates, and the general public to examine, browse, search, and track EDRN activities. By publishing the content of the eCAS deployments, of ERNE, and of the Biomarker Database to the public portal, researchers gain a "one stop shopping" location for EDRN data and information with a unified search that presents opportunities for correlation and discovery of cross-application knowledge.

The various eCAS deployments, ERNE, and the Biomarker Database publish their content using the Resource Description Framework (RDF), a semantic web standard. RDF describes content (called "subjects") by first identifying them with URIs. For EDRN, the URIs to eCAS data are the URLs to the eCAS installation plus the path to the data provided by the eCAS server. For ERNE, the URIs are URLs to individual specimen records. For the Biomarker Database, URIs are URLs to single biomarkers tracked in the database.

RDF makes statements about subjects using properties (identified by URIs) and values for those properties. For EDRN, property URIs are URLs to the Common Data Element definitions Disparate resources in eCAS, ERNE, and the Biomarker Database can thus be correlated.

Publication to the EDRN Knowledge System portal

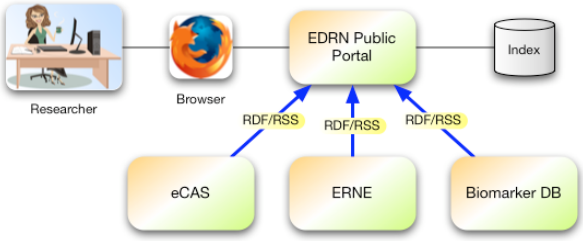

Figure 7. EDRN public portal architecture

is planned to happen with a periodic cycle using RSS, a pull-based technology. The portal will request RSS feeds from participating eCAS deployments, from ERNE, and from the Biomarker Database. Using the RDF provided over RSS, the portal will update its internal indexes of what knowledge is available. Researchers posing queries to the portal query those internal indexes. Figure 7 demonstrates the architecture.

Queries may be posited to just one type of application's data or to all three. Both forms-based searches and Google-like free-text searches are available. The Google-like free-text search is enhanced with AJAX [18] technology that provides instant feedback on search terms as the user types them, character-by-character. Such feedback is eminently useful in providing hints to the user on the usefulness on search terms without the slow drill-down of search refinement.

7. Related Work

There are several areas work related to EDRN's informatics. The work more or less falls into three fundamental principalities: grid computing, large-scale, data-intensive systems, and bioinformatics. In this section we overview a small cross-section of representative projects in each of the above three areas, comparing and contrasting each project to the EDRN project.

Grid computing is an exciting new paradigm focusing on architecture and the technology for large-scale, parallel computation, and data-intensive operations across loosely connected, large organizations distributed across the world. Grids are divided into two sub-categories. *Computational grids* [4] are highly complex software entities that utilize powerful computing resources and capacious networking capabilities in support of solving extremely challenging scientific tasks. *Data grids* [2], on the other hand, are (also) highly complex software entities, focused conversely on sharing, distributing, and managing large-amounts (terabytes and petabytes) of scientific data.

Several projects have attempted to define architectures for large-scale, distributed data-intensive systems, including grids. Gomaa et al. [19] present a novel architecture for describing large-scale data-intensive information systems, specifically applied to NASA's EOSDIS science domain. This work focuses only on a single style—federated client-server [16], in contrast to our own work, which has tried to support both federated client-server, and peer-to-peer.

The conclusions of two independently conducted research studies [20, 21] identify three key areas that the current grid implementations must address in order to promote data and software interoperability: formality in grid requirements specification, rigorous architectural description, and interoperability between grid solutions. The EDRN project represents a natural leap forward in each of these areas. Its focus on architectures for data-intensive, "grid-like" systems naturally addresses architectural description. Additionally, requirements specification is an integral part of architectural description. Finally, software architecture plays a key role in system interoperability, and the goal of EDRN is to take advantage of the architectural style provided by OODT through separation of the functional and information architectures to ensure interoperability between data-intensive systems constructed using the model-driven methodology.

Bioinformatics involves the use of information technology in furtherance of biology-related scientific tasks. Proteomics studies, specimen tracking, clinical trials all benefit from the support of bioinformatics technologies, e.g., study validation and visualization tools, and patient tracking systems. There are several related bioinformatics efforts to EDRN. Finkelstein et al. [22] describe three key challenges that software engineers must faced in support of bioinformatics and systems biology tasks: (1) defining and managing views of bioinformatics model*s*, (2) providing model checking capabilities and validation, and (3) maintaining consistency amongst distributed information models. In this paper, we have described how *each of these three* key challenges was addressed within the context of the EDRN project. Begent et al. [23] identify six key complementary challenges involved with the large-scale integration of biomedical information systems: (1) decentralized technology construction, (2) unobtrusiveness, (3) construction of a flexible system architecture, (4) data model and element variety, (5) data ownership and (6) training of users and engineers. The EDRN project has been a leader in addressing each of these areas through the use of the OODT software technology. In a recent paper [7], we described how OODT addresses nine key software engineering challenges, including unobtrusiveness, data ownership, and flexible system architectures.

The National Cancer Institute (NCI) has recently begun an initiative known as the Cancer Biomedical Informatics Grid, or caBIG [24]. The goal of the system is to create "a common, extensible informatics platform that integrates diverse data types and supports interoperable analytic tools" to "allow research groups to tap into the rich collection of emerging cancer research data while supporting their individual investigations". Consequently, our recent work within EDRN has involved the integration of one of caBIG's

core software technologies [25] into the EDRN Knowledge System.

8. Conclusion

The EDRN Knowledge System promises to dramatically improve the capability for scientific research by enabling real-time access to a variety of information that crosses research center boundaries. While there are clear scenarios for how such a system can improve the discovery process, the system needs to be agile enough to support new approaches to discovering cancer biomarkers. An important measure of success going forward will be working with the research community to ensure that integrated informatics capabilities help to transform the way in which discovery and scientific research is performed. We have found that decomposing the knowledge system into a set of communicating information services based on a domain information model allows us to better deliver specialized components that meet specific scientific needs while supporting the evolution towards an integrated scientific network. Clearly, virtualized data grids are in their infancy, but the needs of programs like the EDRN are demonstrating the benefits and the criticality for bringing scientific research endeavors together into a secure, integrated enterprise to support collaboration and discovery.

9. Acknowledgements

This work was performed at the Jet Propulsion Laboratory managed by the California Institute of Technology under a contract with the National Aeronautics and Space Administration.

10. References

[1] T. Reynolds, "Validating Biomarkers: Early Detection Research Network Launches First Phase III Study," *Journal of the National Cancer Institute*, vol. 95, pp. 422-423, 2003.

[2] A. Chervenak, et al., "The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Data Sets," *Journal of Network and Computer Applications*, vol. 23, pp. 187-200, 2000.

[3] D. Crichton, et al., "Creating a National Virtual Knowledge Environment for Proteomics and Information Management," in *Informatics and Proteomics*: Marcel Dekker Publishers, 2005.

[4] C. Kesselman, et al., "The Anatomy of the Grid: Enabling Scalable Virtual Organizations," *Intl' Journal of Supercomputing Applications*, pp. 1-25, 2001.

[5] E. Gamma, *Design patterns : elements of reusable object-oriented software*. Reading, Mass.: Addison-Wesley, 1995.

[6] C. Mattmann, et al., "Software Architecture for Large-scale, Distributed, Data-Intensive Systems," in *Proc. 4th IEEE/IFIP Working Conference on Software Architecture (WICSA-4)*, pp. 255-266, 2004.

[7] C. Mattmann, et al., "A Software Architecture-Based Framework for Highly Distributed and Data Intensive Scientific Applications," in *Proc. International Conference on Software Engineering (ICSE)*, pp. 721-730, 2006.

[8] M. Shaw, et al., *Software architecture : perspectives on an emerging discipline*. Upper Saddle River, N.J.: Prentice Hall, 1996.

[9] N. Medvidovic, et al., "A Classification and Comparison Framework for Software Architecture Description Languages," *IEEE Transactions on Software Engineering*, vol. 26, pp. 70-93, 2000.

[10] M. Gudgin, et al., "Simple Object Access Protocol Version 1.2," 2003.

[11] ISO/IEC, "Framework for the Specification and Standardization of Data Elements," ISO/IEC, Ed. Geneva, 1999.

[12] T. Berners-Lee, et al., "Uniform Resource Identifiers (URI): Generic Syntax," 1998.

[13] D. Crichton, et al., "An Interoperable Data Architecture for Data Exchange in a Biomedical Research Network," in *Proc. CBMS*, pp. 65-72, 2001.

[14] J. Yu, et al., "A Taxonomy of Scientific Workflow Systems for Grid Computing," *SIGMOD Record Special Issue on Scientific Workflows*, vol. 34, pp. 2005.

[15] D. Beckett, "RDF/XML Syntax Specification (Revised)," W3C Recommendation, http://www.w3.org/TR/rdf-syntax-grammar/ 2004.

[16] R. Fielding, "Architectural Styles and the Design of Network-based Software Architectures," Ph.D., University of California, Irvine, 2000.

[17] M. Pepe, et al., "Phases of Biomarker Development for Early Detection of Cancer," *Journal of the National Cancer Institute*, vol. 93, pp. 1054-1061, 2001.

[18] J. J. Garrett, "Ajax: A New Approach to Web Applications, http://www.adaptivepath.com/publications/essays/archives/000385.php," in *Adaptive Path*, 2005.

[19] H. Gomaa, et al., "A Software Architectural Design Method for Large-Scale Distributed Information Systems," *J. of Distributed Systems Engineering*, pp. 162-172, 1996.

[20] A. Finkelstein, et al., "Relating Requirements and Architectures: A Study of Data Grids," *J. of Grid Computing*, vol. 2, pp. 207-222, 2004.

[21] C. Mattmann, et al., "Unlocking the Grid," in *Proc. 8th ACM SIGSOFT International Symposium on Component-based Software Engineering*, pp. 322-336, 2005.

[22] A. Finkelstein, et al., "Computational Challenges of Systems Biology," *IEEE Computer*, vol. 37, pp. 26-33, 2004.

[23] R. Begent, et al., "Challenges of Ultra Large Scale Integration of Biomedical Computing Systems," in *Proc. CBMS*, pp. 64-69, 2005.

[24] "National Cancer Institute. http://cabig.nci.nih.gov/, February, 2004."

[25] S. Kelly, "ERNE Interface to caTissue, http://oodt.jpl.nasa.gov/wiki/display/edrn/ERNE+Interface+to+caTissue," 2006.