A Virtual Bioinformatics Knowledge Environment for Early Cancer Detection

Daniel Crichton
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California 91109, USA

Sudhir Srivastava
National Cancer Institute
National Institutes of Health
Bethesda, MD 20892, USA

Donald Johnsey
National Cancer Institute
National Institutes of Health
Bethesda, MD 20892, USA

ABSTRACT

Discovery of disease biomarkers for cancer is a leading focus of early detection. The National Cancer Institute created a network of collaborating institutions focused on the discovery and validation of cancer biomarkers called the Early Detection Research Network (EDRN). Informatics plays a key role in enabling a virtual knowledge environment that provides scientists real time access to distributed data sets located at research institutions across the nation. The distributed and heterogeneous nature of the collaboration makes data sharing across institutions very difficult. EDRN has developed a comprehensive informatics effort focused on developing a national infrastructure enabling seamless access, sharing and discovery of science data resources across all EDRN sites. This paper will discuss the EDRN knowledge system architecture, its objectives and its accomplishments.

Keywords: Knowledge system, distributed systems, data management, bioinformatics

1. INTRODUCTION

Eradication of death from Cancer is a major focus of the National Cancer Institute (NCI). The discovery and validation of biomarkers as "indicators of early cancer, surrogate endpoints, prognostic factors of preinvasive cancer, or markers of risk" [5] will provide a significant scientific leap forward for cancer treatment and prevention. The Early Detection Research Network (EDRN) was formed by NCI to conduct research in the discovery of cancer biomarkers through early detection. EDRN is a national asset consisting of 31 research laboratories focused on the research of cancer biomarkers in multi-discipline environments.

Informatics plays a key role in supporting the scientific discovery process associated by interconnecting 31 research institutions together into a virtual knowledge space. The EDRN has established an informatics project focused on building a knowledge system that adapts to the heterogeneous and geographical nature of the EDRN. The knowledge system makes it possible for multiple organizations to access, share and analyze data regardless of its location or format. The concept behind the knowledge system is that each of the 31 institutions is a potential "peer" of the knowledge system, conceptually contributing data using a common ontological model[1] for early cancer detection. This aggregate of data across all institutions creates the knowledge space for all EDRN data.

The EDRN knowledge system has been broken into three critical pieces: science processes, system architecture and data architecture. Each of these have been critical to developing an informatics knowledge infrastructure supporting the scientific discovery process. The system architecture is based loosely on a REpresentational State Transfer (REST) style architecture leveraging a distributed-based network infrastructure. The system architecture is "data-driven" enabling the system to communicate using a standard data architecture. The EDRN data architecture implements the ontological model and provides a common language by which all disparate, heterogeneous peers in the knowledge space can share data.

[1] An ontological model defines concepts and their relationships

2. SCIENCE PROCESSES

A key to establishing a scientific knowledge system is defining the associated scientific processes that support the discovery and validation of cancer biomarkers. The EDRN established working groups across the institutions in order to flush out the associated discovery and validation processes that would be used across the network to perform the data collection and analysis. These cross-disciplinary working groups were represented by basic scientists, clinicians, epidemiologists, biostatisticians and computer scientists. The key processes that were identified enabled the computer scientists to generate "use-cases", an important element necessary in constructing an information system.

EDRN's first efforts in building a unified infrastructure was the delivery of tools to aide scientists in locating biospecimens across participating EDRN institutions. Informatics experts and scientists worked together to understand how scientists would search and use biospecimen data. This approach was critical to developing a tool that would support the scientists in their discovery process. Further detail on the deployment of this tool will be discussed in section 5.

EDRN also initiated working groups with scientists to standardize the content collected in EDRN studies using a technique called "Common Data Elements." This provided a common language that future studies could use that would be enable consistency across institutions collecting data.

Working groups were also formed that defined the science processes associated with various multi-center validation studies. This allowed for use-cases to be developed that would support the collection, transmission and analysis of the data used for a validation study. These use-cases were used to drive the knowledge system implementation for the validation study. EDRN is currently completing implementation of two key validation studies: Surface-Enhanced Laser Desorption/Ionization (SELDI) and Microsatellite Instability (MSI). SELDI will allow investigators to "see how the molecular map of proteins changes between healthy tissues, precancerous tissues, and cancer" [5]. MSI will support early detection for bladder cancer through an "assay that detects instability in microsatellites [which are] repetitive stretches of short sequences of DNA" [5]. The informatics teams are working with the SELDI and MSI studies to ensure that data collected, analyzed and stored in these studies will be interoperable with the knowledge system and future studies within EDRN and NCI.

Informatics experts have also been careful to ensure that any data shared is compliant with federal privacy and security regulations including the Health Insurance Portability and Accountability Act (HIPPA). This requires that certain identifiers be removed from to protect the confidentiality of the patients described by the research data. In addition, the system architecture discussed in section 3 provides secure transmission of data between all distributed institutions.

3. SYSTEM ARCHITECTURE

The system architecture for the EDRN knowledge system is based on a distributed framework developed at the National Aeronautics and Space Administration's (NASA) Jet Propulsion Laboratory (JPL) called the Object Oriented Data Technology (OODT) framework [2]. The framework, used as the infrastructure for NASA's planetary missions, provides a middleware infrastructure enabling access to heterogeneous distributed data resources across geographically distributed data systems.

OODT is a set of software components that create a framework for constructing interoperable data systems. It combines a metadata model with a software technology middleware to enable the creation of virtual databases from data sets across the nation. Scientists and other researchers can make discoveries using different data sets produced by different organizations with different meanings, as if they are a single, large repository of knowledge. This means that it can be configured to fit into several different domains that are critical to scientific research such as biomedicine and space science. A key benefit is that it connects disparate databases and systems together over the Internet without requiring those systems to be re-implemented or modified.

The framework significantly reduces the level of effort required by a project to connect distributed data systems together by providing components with well known interfaces that can be used to construct an information system. It follows the principles of a REpresentational State Transfer (REST) [3] style architecture that enables a client/server implementation in an n-tier or layered configuration. Communication between distributed nodes uses the Extensible Markup Language (XML), a leading industry standard, for describing data interchange.

The components provide the base software objects necessary to archive, discover and share data resources. The components are able to communicate using a variety of methods provided by various industry initiatives in distributed computing including the Common Object Request Broker Architecture (CORBA), Java's Remote Method Invocation (RMI), and emerging standards in Web Services along with research in the area of Peer-to-Peer. The components are implemented using the Java programming language and provide mechanisms to be

easily extended to fit specialized requirements for the domain. Each component of OODT has an "implementation class" that provides the domain specific interface. This allows for simple software reuse while still providing the necessary framework to plug the components together. Client applications use common application program interfaces (APIs) to access OODT compliant data sources. Client APIs are provided using Java, C and HTTP.

OODT generalized the problem of finding and accessing distributed data systems into two key steps. The first step is to discover what resources[2] exist. The second is to access those resources. Discovery of resources means being able to query distributed data catalogs about data resources despite their unique implementations. OODT developed the concept of a data "profile" in XML to capture that attributes that describe a data resource. One of OODT's components, the profile server, enables the unification of heterogeneous data catalogs by querying local data catalogs and returning profile descriptions of data resources using a common schema defined in XML. OODT was able to use the Resource Description Framework (RDF) developed by W3C and the Semantic Web community as a means for describing the common XML profile schema. Particular attention has been devoted towards ensuring scalability as the number of integrated data systems and data sets increase.

The OODT "profiles" use a project's data dictionary specification as a means for describing the distributed data resources. This enables applications to search on any data element associated with a particular data resource. In constructing the OODT schema, several standards have been used. OODT uses ISO/IEC 11179, Dublin Core and the Resource Description Framework (RDF) for describing its data resources. Each play a critical role in developing a common interoperable framework for managing data. Further detail on the data architecture will be discussed in section 4.

Another OODT component is called a "product server". It provides access to a physical resource (such as a database) once it is discovered. The product server provides a gateway to local systems enabling access and transformation of the data product. For example, accessing a specimen database at the University of Colorado requires connecting to a product server running on a local computer at that institution. The product server is capable of transforming and packaging the product or set of products into a form that is digestible by the requesting client. The design of the product server component allows it to be specialized for accessing any backend data source. Several OODT product servers that are deployed use the Java Database Connectivity API (JDBC) to connect to SQL-compliant databases such as Oracle, Sybase, SQL Server, MS Access, MySQL and Filemaker Pro.

The interfaces between components use XML. OODT has developed a common XML schema to ensure uniform communication between all OODT components. OODT enables the data sources to have common interfaces despite their local heterogeneity at distributed sites. This means that new applications can be built that access heterogeneous data sources using predictable data interfaces. A conceptual web-based architecture for OODT is shown in figure 1.

OODT has been used on Linux, Windows 2000/XP/NT/98, Solaris 2.x, and Mac OS X. Due to its implementation in Java, it has been successful in delivering one version of OODT for all of these environments.

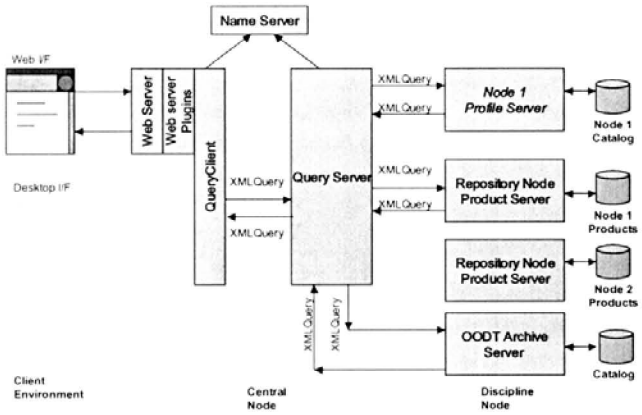

Figure 1: Conceptual OODT Component Architecture

4. DATA ARCHITECTURE

The data architecture provides the ontological model necessary to construct the knowledge system. A data architecture is critical to effectively search heterogeneous distributed data systems as well as enabling correlative science. It defines the common data elements and their relationships within the knowledge space. It enables the interoperability between distributed institutions by providing a common semantic language for communication. Several standards have been developed that support the definition of a data architecture. ISO/IEC 11179 [4] provides a standard definition for describing data elements. This enables consistency when developing data dictionaries. The ISO/IEC standard recommends that a data element consist of attributes for four key categories: identification, definitional, representational and administrative. We use ISO/IEC 11179 in conjunction with Dublin Core as a mechanism for developing a minimal set of data elements that must be provided in any data architecture.

[2] A data resource refers to any electronic data object or too such as gene image, a specimen catalog or a website.

The EDRN developed a data architecture for its knowledge system. An over-arching ontological model for cancer was created that identified the relationships between key objects within the knowledge space. A cross-disciplinary working group was formed that focused on defining the common data elements (CDEs) for each of the objects of the ontological model. Specific objects, for example, include organ specimen models such as lung, breast and prostate. A data dictionary based on ISO/IEC 11179 was developed that provided a common representation of each of the elements. The following fragment defined using the Extensible Markup Language (XML) illustrates an example common data element used by the EDRN:

```
<dataElement>
        <name>ANOTOMIC_SITE</name>
        <version>1.0</version>
        <registration_authority>NCI.EDRN
        </registration_authority>
        <definition>Anatomical site
        </definition>
        <dataType>Integer</dataType>
        <unit>Integer</Integer>
</dataElement>
```

Permissible values for the data element "ANOTOMIC_SITE" could include "bladder (1)", "Lymph node (25)", etc. One of the key benefits of establishing a data dictionary is that it provides a mechanism to validate data objects against the data dictionary. In the above example, the value "334" might not be in the valid set of permissible values. Having an online data dictionary would enable one to validate the data elements.

Once the common data elements are established, each of the participating institutions within the knowledge system are able to map their local data models to the knowledge system model in order to provide semantic consistency across the system. Specific mapping tools were developed that allowed the EDRN informatics experts to capture the mapping of local site data models to the EDRN knowledge system model. Attributes of the data element including permissible values, units, format, and data type were captured and mapped to one another in order to provide the mapping at the informatics level. This enabled the deployed OODT software described in section 3 to run a translation function as part of the process of querying and retrieving data from the distributed EDRN institutions.

5. DEPLOYMENT

EDRN has already deployed the knowledge system to ten institutions providing a common web-based client interface called "ERNE" or the EDRN Network Exchange system (see Figure 2 below) [1]. ERNE unifies search and retrieval of biospecimen data from all institutions regardless of where it is located, how it is stored, or the differences in the underlying data models. This enables a scientist, for example, to locate tissue specimens for breast cancer by searching data catalogs at participating institutions across the country.

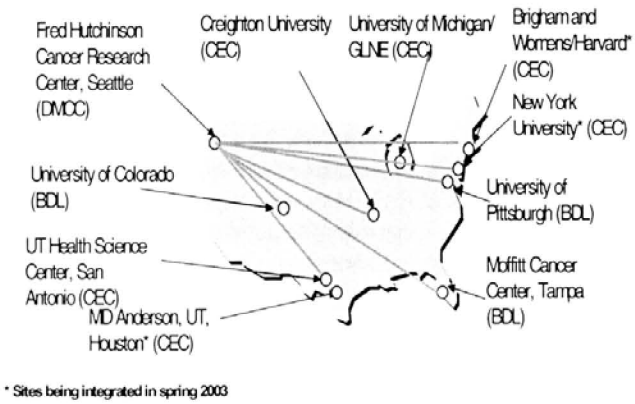

Figure 2: EDRN Knowledge System Deployment

As the knowledge system evolves, the governing cancer ontological model and the use-cases derived in the working groups will be used to drive the relationships between the data sets enabling discovery through data mining. Scientists, for example, will be able to query an assay result from a validation study and then find the associated specimens that were collected as part of that assay.

6. CONCLUSION

The EDRN is leading the agency in developing a research informatics infrastructure for scientists. This scalable infrastructure will enable EDRN to expand its data and tools providing a long term platform for cancer research. New tools that enable scientists to mine and correlate data across multiple data sets and studies will be created that will aide the discovery process. This will include the introduction of data understanding software and algorithms that are capable of using the existing knowledge system infrastructure to construct knowledge bases of metadata using automatic feature detection. This metadata will augment existing metadata used to describe EDRN data products. This will enhance the informatics infrastructure enabling more sophisticated search and correlation capabilities.

The EDRN knowledge system has proven that informatics is a critical component of a collaborative research

network. As the discovery process progresses, it is clear that informatics will continue to play an even greater role. The EDRN knowledge system will be a core tool that enables scientists to seamlessly access, share and correlate data as new approaches and paradigms to the discovery process are defined. It will enable not only interoperability among PI institutions focused on early detection, but will become a major informatics infrastructure providing data to cancer researchers at all stages supporting the eradication of death and suffering from one of mankind's most deadly diseases.

7. REFERENCES

[1] D. Crichton, G. Downing, .H. Kincaid, S. Hughes, S. Srivastava, "An Interoperable Data Architecture for Data Exchange in a Biomedical Research Network", **The 14th IEEE Conference on Computer Based Medical Systems**, Institute of Electrical Engineers, July 26, 2001.

[2] D.J. Crichton, J.S. Hughes, J.J. Hyon, S.C. Kelly, "Science Search and Retrieval using XML", **The Second National Conference on Scientific and Technical Data**, U.S. National Committee for CODATA, National Research Council, March 13-14, 2000, http://oodt.jpl.nasa.gov/doc/papers/codata/paper.pdf.

[3] **ISO/IEC 11179 - Specification and Standardization of Data Elements**, Parts 1-6, ISO/IEC specification, http://www.iso.ch/iso.

[4] R. Fielding, **Architectural Styles and the Design of Network-based Software Architectures**, Dissertation for Doctor of Philosophy, University of California Irvine, June 2000. http://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf

[5] S. Srivastava, J Kagan, M. Verma, **Early Detection Cancer Research Network 2nd Annual Report**, October 2002.

8. ACKNOWLEDGEMENTS

The authors wish to recognize the contributions of the EDRN Data Management and Coordinating Center (DMCC) at the Fred Hutchinson Cancer Research Center who are jointly working with the Jet Propulsion Laboratory and the National Cancer Institute on developing the knowledge system infrastructure.

We also wish to thank the following partner institutions for their involvement in this project: H. Lee Moffit Cancer Research Center, University of Texas San Antonio, Creighton University, University of Colorado, University of Pittsburgh, Dartmouth Medical School, Brigham and Women's Hospital, University of Texas MD Anderson, and New York University.

The work described was partially performed by the Jet Propulsion Laboratory, California Institute of Technology which is under contract to the National Aeronautics and Space Administration.