

# A National Virtual Specimen Database for Early Cancer Detection

Daniel Crichton  
NASA Jet Propulsion Laboratory  
California Institute of Technology  
dan.crichton@jpl.nasa.gov

Sean Kelly  
NASA Jet Propulsion Laboratory  
California Institute of Technology  
sean.kelly@jpl.nasa.gov

Donald Johnsey  
National Cancer Institute  
National Institutes of Health  
johnseyd@mail.nih.gov

Heather Kincaid  
Fred Hutchinson Cancer Research Center  
hkincaid@fhcrc.org

Mark Thornquist  
Fred Hutchinson Cancer Research Center  
mthornqu@fhcrc.org

Marcy Winget  
Fred Hutchinson Cancer Research Center  
mwinget@fhcrc.org

## **Abstract**

*Access to biospecimens is essential for enabling cancer biomarker discovery. The National Cancer Institute's (NCI) Early Detection Research Network (EDRN) comprises and integrates a large number of laboratories into a network in order to establish a collaborative scientific environment to discover and validate disease markers. The diversity of both the institutions and the collaborative focus has created the need for establishing cross-disciplinary teams focused on integrating expertise in biomedical research, computational and biostatistics, and computer science.*

*Given the collaborative design of the network, the EDRN needed an informatics infrastructure. The Fred Hutchinson Cancer Research Center, the National Cancer Institute, and NASA's Jet Propulsion Laboratory (JPL) teamed up to build an informatics infrastructure creating a collaborative, science-driven research environment despite the geographic and morphology differences of the information systems that existed within the diverse network.*

*EDRN investigators identified the need to share biospecimen data captured across the country managed in disparate databases. As a result, the informatics team initiated an effort to create a virtual tissue database whereby scientists could search and locate details about specimens located at collaborating laboratories. Each database, however, was locally implemented and integrated into collection processes and methods unique to each institution. This meant that efforts to integrate databases needed to be done in a manner that did not require redesign or re-implementation of existing systems.*

## **1. Introduction**

The Early Detection Research Network (EDRN) created and supported by the National Cancer Institute (NCI) is a 5 year, collaborative, multi-institutional scientific

consortium[reference Sudhir's paper]. The Networks' goal is to identify, evaluate, and validate promising biomarkers to support the early detection of cancer. Access to biospecimens is essential for enabling the Network to obtain this goal. This paper is an update of an informatics infrastructure, ERNE, described previously by Crichton, et al. [1].

ERNE, the EDRN Resource Network Exchange, was developed to enable investigators to easily identify the availability of biospecimens and associated epidemiological information needed for their research. The system provides scientists access to biospecimen information regardless of where it is located across the country. ERNE's specific goal is to provide transparent access to existing specimen repositories providing EDRN a virtual knowledge environment despite the distributed nature of the collaboration. An overall informatics architecture and infrastructure was created for EDRN, plugging in databases which are managed locally by each institution.

The project focused on development of several key aspects including a common semantic architecture, a distributed informatics technology infrastructure that leveraged the semantic architecture, a dynamic portal, and a common study protocol for achieving compliance from each institution's Institutional Review Board (IRB). The project team took special care to minimize the impact of change and informatics skills for each institution and to ensure that all data shared would be compliant with federal regulations.

Scientists use the system to search for and retrieve details about specimens located at collaborating sites. Specimen curators must be freed from the task of updating a central repository. Therefore, each database to become a part of the ERNE system must be integrated without impacting collection processes and methods unique to each institution. There must be no redesign or re-implementation of existing systems.

The project team consists of the following institutions: the Data Management and Coordinating Center (DMCC) located at Fred Hutchinson Cancer Research Center serves as the project management and coordinating mechanism providing the central access point for the data management architecture; JPL provides the expertise and distributed software component infrastructure; and the NCI provides overall guidance. As of July 2003, nine EDRN sites are participating in this project to provide biospecimen repositories for querying. The ultimate goal is for all thirty EDRN institutions to be involved.

The software framework called the Object Oriented Data Technology (OODT) Framework [2] was provided by the Jet Propulsion Laboratory. OODT has been used by the National Aeronautics and Space Administration (NASA) for a wide variety of science disciplines including Planetary, Earth and Space Physics, as well as the informatics infrastructure for the ground data component of some of NASA's key missions. Most recently, OODT was used to provide the infrastructure for release of the planetary products from the 2001 Mars Odyssey mission. OODT provides a distributed component architecture that uses metadata as a means for integrating geographically distributed data resources.

The following were identified as key foci for this project: semantic architecture, informatics infrastructure, security and confidentiality, data model mapping, and a dynamic portal interface. **Error! Reference source not found.** illustrates the multi-phase plan and accomplishments to date for this project.

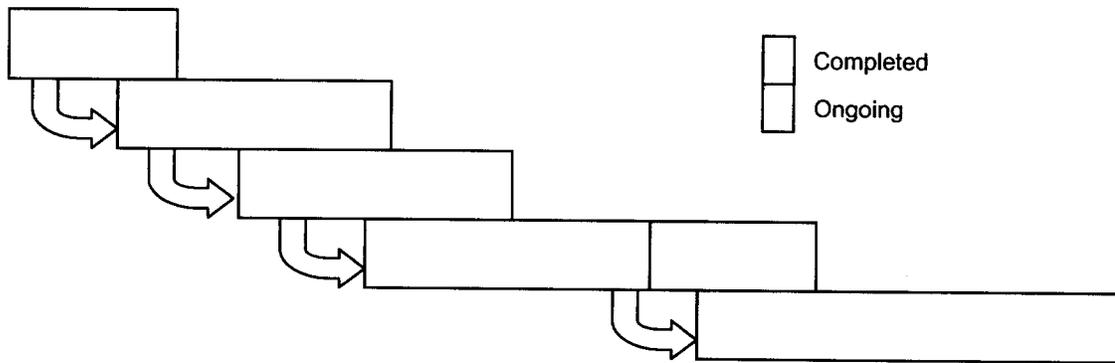


Figure 1. Multiphase plan and accomplishments to date for ERNE.

## 2. Semantic architecture

The underlying data models, both relationships and in many cases, terminology, of each system were locally defined making interoperability difficult. Many of the participating institutions had their own methods of representing data in their databases, presenting a real challenge for creating a virtual database. The EDRN developed a common ontology model for specimens that was useful in generating a set of common data elements (CDEs) for describing specimens and their associated attributes. CDEs are data elements that have been agreed upon by the EDRN investigators as critical data that must be collected by all EDRN sites that describe study participants and specimens [reference MW paper]. One of the key findings early on was that development of a common set of CDEs was essential for enabling interoperability across disparate databases. EDRN adopted the ISO/IEC 11179 [5e] standard. This standard has provided a critical meta model framework for describing the CDEs in a consistent manner.

In developing the CDEs, the DMCC created working groups that combined discipline science experts with computer scientists in an effort to establish the common language. Given that several sites had preexisting implementations using a different semantic architecture than EDRN, it became necessary to establish a mapping process that mapped the local data model at the institution to the common CDEs. The DMCC continues to develop EDRN Common Data Elements (CDEs) based on the ISO/IEC 11179 standard. This common model will continue to define a standard language for EDRN that will be used in all data sharing, data collection and informatics efforts.

## 3. Informatics infrastructure

In addition to creating a common model for describing biospecimens, the project team created an informatics software infrastructure, deployed via the Internet at all collaborating institutions in order to create-find and access information about specimens located in each institution's database. The system employed a metadata-based distributed framework as a synchronous communications infrastructure that tied databases together using the Common Data Elements (CDEs). Developing the common middleware allowed for data, normally tightly coupled to applications, to be de-coupled and integrated as a set of virtual repositories. In middleware, a request broker manages service requests from top tier client applications to server applications. Two server applications, the profile and product servers, provide search

and retrieval functionalities and interface to catalogs and data repositories in the bottom tier of the architecture. This distributed framework makes it possible to query multiple institutional databases concurrently, compiling the results into a unified view of the available specimens [1].

Message-driven processing software (middleware) uses a request broker to handle service requests from clients to server applications. The message-driven paradigm addresses both interface as well as scalability issues since the number of component interconnections increases linearly as new components are added. The Object Oriented Data Technology (OODT) framework [2] is the foundation for the EDRN informatics infrastructure and provides the messaging mechanism, product and profile servers, distributed server management, and plug-in capabilities for user tools. The EDRN ERNE middleware is configured as a single downloadable package and was installed at every site participating in the EDRN Informatics Project.

The software leverages Java's Remote Method Invocation (RMI) to support the distributed object implementation. This enabled a common messaging layer within the system that all distributed servers would use to communicate. The OODT distributed framework is designed in such a way as to support various distributed messaging implementations including Java RMI, Common Object Request Broker Architecture (CORBA), and Sun's new Peer-to-Peer implementation called JXTA. These distributed messaging implementations provide services for distributed communication as well as object naming in order to locate distributed objects.

Product servers provide a common system interface to differing data repositories for data product access and retrieval. Each EDRN site downloaded the software package and installed the product server component. Each product server runs a dynamically loaded Java object called a query handler that negotiates the interface between the EDRN enterprise environment and the local biospecimen repository. The query handler converts an EDRN query into a local query for the database. In general, this is a conversion to a SQL-compliant database (although any transformation is possible) that translates the query from EDRN CDEs to entity and attribute definitions defined by the local database model. Results from the query are mapped to the EDRN CDEs and then formatted with an agreed-upon representation. Any user application can request service from a product server through a standard HTTP or Java API.

Profile servers provide a common mechanism for describing distributed resources. The profile server manages profiles—sets of resource definitions [2]—about distributed data systems and their products. A profile is a metadata description of the resources known by a node in the distributed framework. These resources are interfaces, data products, or other profile servers available in the integrated enterprise. Profiles may be grouped and served by more than one profile server. The query component ties this architecture together by providing and managing the traversal of the integrated digraph node architecture. It also interprets profile definitions that provide mappings between data system nomenclature. The query component also provides the facility to manage concurrent queries across multiple servers to improve performance.

For this implementation, one profile server was used to reference each of the product servers that were ~~distributed across~~ located at the ~~distributed~~ -sites across the country. Figure 2 ~~Figure 2~~ below shows the deployment of the product servers at each institution, while Figure 3 ~~Figure 4~~ provides an overview of the geographic distribution planned for spring 2003. Profile and product server instantiations are uniquely identified by name so they can be located within the distributed name server. These names are used as part of the metadata header encoded to identify the distributed EDRN services that can support queries for products.

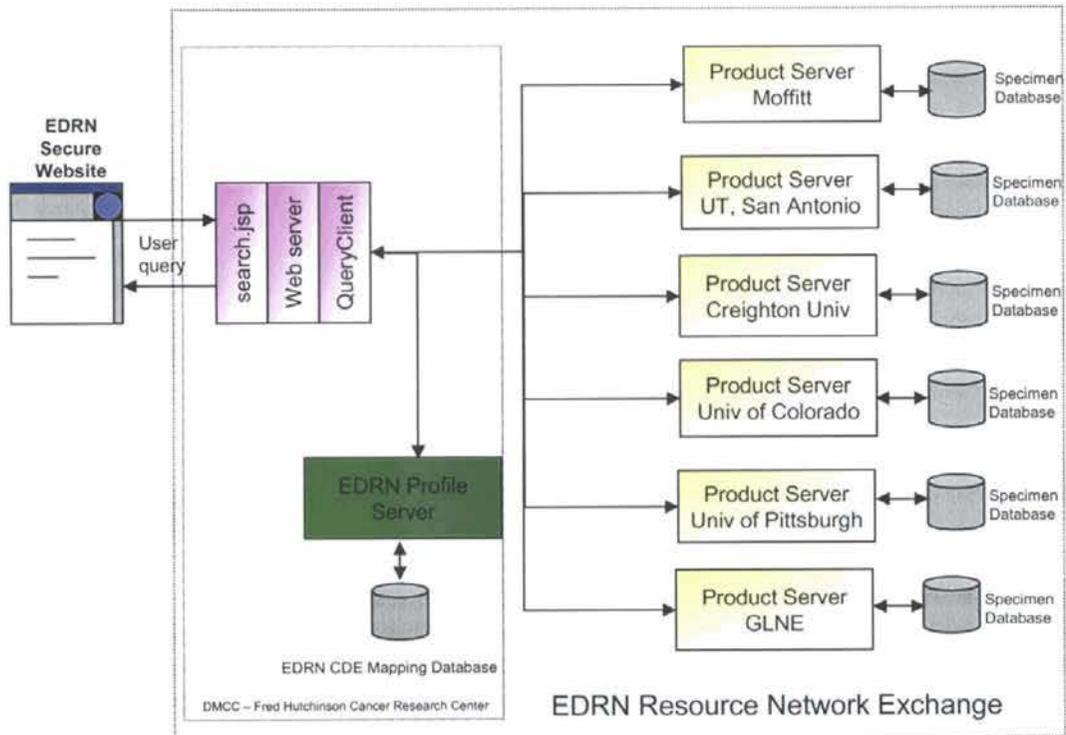
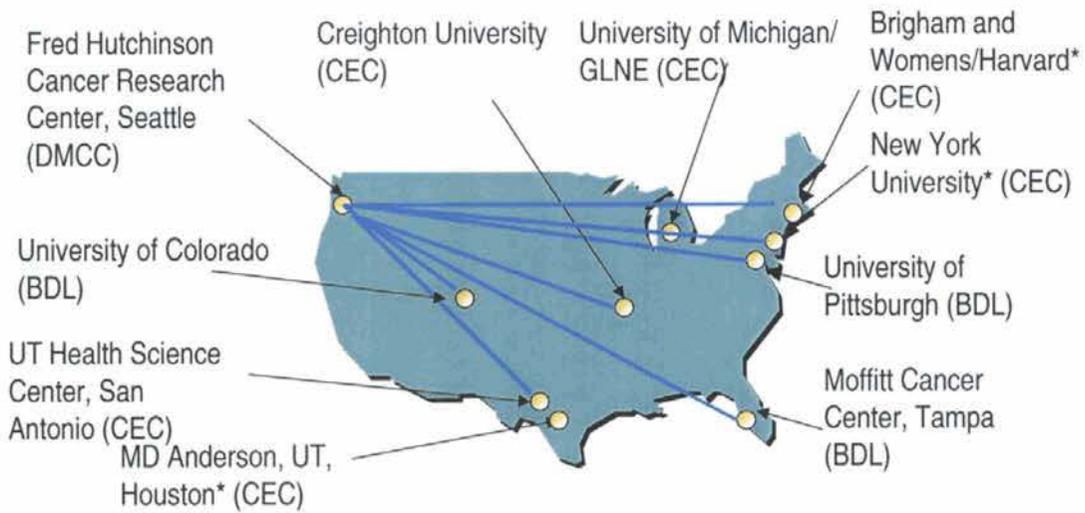


Figure 2. Software component deployment.

The team chose XML since it provides a rich environment for defining and managing metadata. In addition, XML serves as an interface specification on top of the distributed messaging layer between each of the nodes of the system. The query definition is implemented independent of any one database, functional, or programming language and is intended to provide an abstract view of both the query expression and the results. The query definition allows for each data system to be encapsulated. This allows various implementations, ranging from the use of relational and object database management systems to the use of flat file and home-grown databases for cataloging and storing data products to exchange information by plugging into a generic query definition.



\* Sites being integrated in spring 2003

Figure 14. ERNE software installations.

One of the goals of this architecture is to provide a standard application program interface (API) that will allow for generic science analysis tools to be written that can plug into the architecture to retrieve and correlate data from multiple data sources. For this pilot, the team developed a web-based interface using Java Server Pages (JSP). The interface served as a client of the data architecture and allowed for researchers to query distributed databases from a single point.

The team discovered that architectural goals for space science and biomedical research were very similar, and in fact, the components developed for the space science could be directly infused into the EDRN knowledge environment. By focusing on a framework for supporting basic system interoperability, the architecture was able to provide solutions that not only solve problems within a single discipline, but also support integration of cross-disciplinary databases.

#### 4. Security and confidentiality

Security and confidentiality were important considerations for this project due to the sensitive nature of the data and the use of the Internet to link each institution. Security requirements needed to be met for both the participating institutions and the federal Health Insurance Portability and Accountability Act (HIPAA) [add reference]. The system accommodated the local network configurations at various institutions in order to link the them together via the Internet protecting all network traffic using 128-bit strong encryption. Sites opened specific ports in local firewalls allowing traffic between product servers located at participating institutions and the DMCC. In addition, a template protocol was written to describe the project including information about each participating institutions technical environment, security infrastructure and database. This protocol was completed by each participating institution and approved by their Institutional Review Board (IRB). All Personal

Health Information (PHI) identifiers have been removed from these data to make them compliant with HIPAA regulations for the data that are being shared.

## 5. Data model mapping

Sharing of disparate representations of specimen banks could not have been possible without a high degree of communication. The team found it necessary to communicate the EDRN CDEs that had been established and agreed upon for the virtual repository to each participating site. It was also essential to document and understand each site's local data dictionaries and data models. The DMCC created a CDE mapping tool to facilitate this process.[need to include a reference to Moffitt somehow] This online tool resides on the EDRN secure website located at the DMCC. It allows each participating site to view the 39 EDRN CDEs that are being shared for this project as they are defined in the EDRN online metadata repository, and document how each of the data elements in their database maps or matches to a specific CDE. The following information is displayed for each CDE and the institution enters corresponding information about each of their data elements: **[list fields shown]** The tool also allows users to generate automatic emails to ease communication between groups. Each institution underwent this mapping process using the tool described above to document the transformations between the local database model and the EDRN CDE model.

This tool has allowed us to overcome several challenges in understanding the data contained in these repositories and their data models. It has allowed us to have a communication and documentation mechanism when data collected at a center differ due to data collection prior to the establishment of the EDRN. For example, in one site more detailed data may be collected for \_\_\_\_\_ - **give example. Easy example is race/ethnicity. [Show screen shot!]**

## 6. Dynamic portal interface

The team developed a unified portal using Java Server Pages which connects the middleware to the distributed repositories. This portal allows scientists to specify search criteria using the project specimen and epidemiological CDEs. The software translates the selections into an XML-based query definition, and two queries are sent. The first query verifies candidate sites that have a repository that has potential specimens that meet a minimum criteria. The second query is then sent concurrently to the sites that may return results.

The portal is dynamic and will limit choices that appear based on specified user selections. When a user logs in they will see all sites and their operational status. If a site's server is down the user cannot select this site as part of the query. The portal also narrows choices based upon sites selected to show only what specimens are available at those sites and what specimen types are stored at those sites. Results are summarized and returned showing what is available at each institution. Drilling down on a site shows detailed de-identified results.

**[Update Screen shot – possibly add query results if time's room.]**

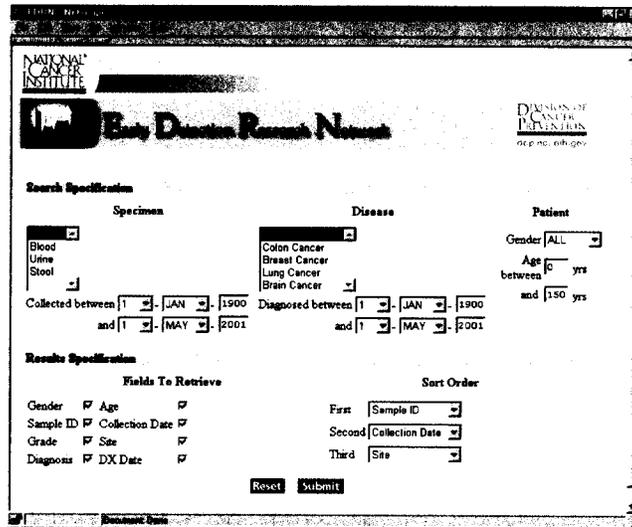


Figure 2. EDRN prototype user interface

## 7. Conclusions

The achievements reflect the current progress of rolling out a national virtual specimen database for early cancer detection. Significant maturity in both the software infrastructure and the experience of the teams has been achieved through this process and we envision that deployment of the architecture to support data access and sharing of similar data sets as specimens will occur. Developing methodologies for deploying software nationally presented challenges on several technical fronts as well as policy and cultural challenges. As we move forward, we continue to change the paradigm by which data is delivered to scientists moving towards the goal of building a national knowledge infrastructure for cancer research. Future phases will target the creation of data profiles that include metadata descriptions of biospecimens, biomarkers including assay sensitivity and specificity, research protocols, and publications that ultimately will yield the biomarker knowledge environment.

Finally, our experience continues to show that the development of common metadata models is critical to development of a data architecture for biomedical research. Combining the data architecture with the Object Oriented Data Technology (OODT) technology architecture provides a platform for leading research institutions around the country to plug into enabling seamless access to specimens regardless of what institution manages them.

## 8. 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 14<sup>th</sup> 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] Biomarkers Knowledge System.  
[http://www1.od.nih.gov/osp/ospp/biomarkers/Biomarkers\\_Knowledge\\_System.pdf](http://www1.od.nih.gov/osp/ospp/biomarkers/Biomarkers_Knowledge_System.pdf)
- [4] Early Detection Research Network. <http://www3.cancer.gov/prevention/cbrg/edrn/>
- [5] ISO/IEC 11179 - Specification and Standardization of Data Elements, Parts 1-6, ISO/IEC specification,  
<http://www.iso.ch/iso>.
- [6] Object Management Group. CORBA/IIOP 2.3.1 Specification. October 1999.
- [7] W3C. Extensible Markup Language (XML), Version 1.0, <http://www.w3.org/TR/REC-xml>
- [8] Data Entity Dictionary Specification Language (DEDSL) - Abstract Syntax, CCSDS 647.0-R-2.0, Draft Recommendation for Space Data System. Standards, Consultative Committee on Space Data Systems, November 1999.
- [9] Special Issue: Planetary Data System, Planetary and Space Science, Pergamon, Volume 44, Number 1, January 1996.
- [10] ISO/IEC11179-1,6, <http://www.iso.ch/infoe/text.htm>.
- [11] NCI Common Data Elements Dictionary, [http://cii-server5.nci.nih.gov:8080/pls/cde\\_public/cde\\_java.show](http://cii-server5.nci.nih.gov:8080/pls/cde_public/cde_java.show).