

# Enabling Effective Curation of Cancer Biomarker Research Data

Andrew F. Hart, Chris A. Mattmann, John J. Tran,  
Daniel J. Crichton, J. Steven Hughes, Heather Kincaid, Sean Kelly  
Jet Propulsion Laboratory  
California Institute of Technology  
Pasadena, CA 91109, USA  
{ahart,mattmann,jtran,crichton,jshughes,kincaid,kelly}@jpl.nasa.gov

Kristen Anton  
Biostatistics and Epidemiology  
Dartmouth Medical School  
Lebanon, NH 03766, USA  
kristen.anton@dartmouth.edu

Donald Johnsey, Christos Patriotis  
National Cancer Institute  
National Institutes of Health  
Bethesda, MD 20892, USA  
{johnseyd, patriotisc}@mail.nih.gov

## Abstract

*The dramatic increase in data in the area of cancer research has elevated the importance of effectively managing the quality and consistency of research results from multiple providers. The U.S. National Cancer Institute's Early Detection Research Network (EDRN) is a prime example of a virtual organization, sponsoring distributed, collaborative work at dozens of institutions around the country. As part of a comprehensive informatics infrastructure, The NASA Jet Propulsion Laboratory, in collaboration with Dartmouth Medical School, has developed a web application for the curation of cancer biomarker research results. In this paper, we describe and evaluate the application in the context of the EDRN content management process, and detail our experience using the tool in an operational environment to capture and annotate biomarker research data generated by the EDRN.*

## 1. Introduction

Within the domain of cancer biomarker research, collaboration is critical to validating biomarkers as early cancer indicators. By sharing specimen data, researchers can study a marker in multiple contexts and identify common epidemiological characteristics across varying populations. Distributed collaboration between institutions provides an opportunity to access a greater volume of data than would otherwise be available, but can introduce complexity into the task of presenting a unified, coherent picture of the re-

search that has been conducted [6, 8].

Curation, the process of adding value to raw research data through the capture, annotation, and cataloging of relationships among data elements, can improve the caliber and usability of the research data produced. Curation is a process requiring collaborative agreement on standards for research results, and a concerted effort at maintaining those standards over time. Without this effort, distributed research data remains siloed in physically separate repositories, difficult or impossible to cross-reference with data from other silos, and thus fails to achieve its full research potential.

In our experience developing an informatics infrastructure for the U.S. National Cancer Institute's Early Detection Research Network (EDRN) initiative [14], human curation of research results serves as a bridge between data producers. Our integrative approach to biocuration brings together science, policy, and technology to serve the needs of the virtual organization [8]. In addition to injecting an additional level of quality control into the data capture process, curation adds value to the resulting data products by adding *metadata* (data about data) to research components and their relationships. We believe this process of annotating and linking together data elements is key to the construction of a comprehensive knowledge base. We have developed a browser-based web application for biomarker curation as a component of the EDRN Knowledge Environment (EKE) [3] which simplifies the process of preparing results for peer-review and public release.

The rest of this paper is organized as follows: Section 2 provides a context for the effort as well as a brief summary of related work. Section 3 describes the architecture of the

web application, including the choices and constraints that drove the architectural decision-making process. Section 4 describes our experience using the application in an operational environment, and Section 5 summarizes the lessons learned and describes potential directions for future work.

## 2. Background And Related Work

This section provides background on the Early Detection Research Network (EDRN) project and its informatics infrastructure. Additionally, information is provided regarding projects addressing similar challenges in data curation.

### 2.1. Early Detection Research Network

The EDRN is tasked with the discovery and validation of biomarkers as early indicators of cancer. EDRN research is conducted at over two dozen member institutions spread across the United States. To mitigate the technological mismatches between institutions, EDRN initiated an informatics effort [4, 11], with the goal of leveraging recent advances in grid [5] and web technologies to improve the effectiveness and reach of EDRN research. The NASA Jet Propulsion Laboratory (JPL) has worked together with the Fred Hutchinson Cancer Research Center in developing a coordinated informatics infrastructure to connect the activities throughout the EDRN and provide a single, unified location for discovery and analysis of EDRN data. [3].

### 2.2. Bioinformatics Projects

There are a number of projects in the biomedical informatics domain attempting to address similar data curation challenges. We limit our discussion to those projects specifically dealing with the quality of data curation and the accessibility of curated data.

#### 2.2.1 caBIG

The Cancer Biomedical Informatics Grid (caBIG) [15] is an NCI-sponsored initiative geared towards the construction of an all-inclusive bioinformatics grid infrastructure for sharing data among the cancer research community. caBIG recognizes the importance of data curation, and provides several software tools to facilitate curation efforts. However, due to its broad scope, curation in caBIG is a largely decentralized activity, leaving open the potential for quality and consistency issues that can hamper efforts to seamlessly integrate the research.

#### 2.2.2 caTissue

caTissue [1] is an application suite developed as a ‘plugin’ to caBIG which supports the capture of biospecimens.

caTissue’s focus is on gene expression and sequence data related to cancer research. Patient tissue samples are combined with metadata annotations to form a comprehensive specimen bank that can be queried by a caBIG user. Data entry and curation in caTissue is accomplished through either the provided graphical user interface or an application programmer interface that enables users to access the biospecimen data.

EDRN’s specimen capture application, ERNE, was built on a grid middleware package called Object Oriented Data Technology (or OODT [13]), and has successfully plugged into the caTissue suite, enabling EDRN to interoperate with caBIG grids.

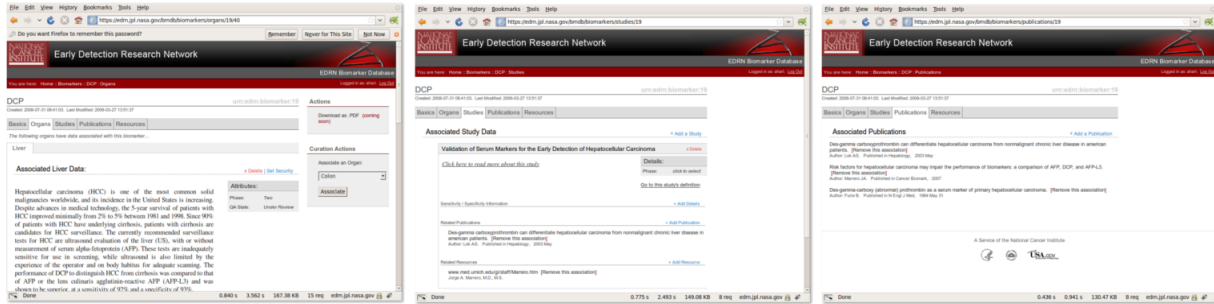
## 3. Architecture

The EDRN Biomarker Database (BMDB) is a relational database containing information linking the various components of EDRN biomarker research. Due to the distributed nature of EDRN research, the BMDB curation web application has been implemented using a client-server architectural model. Several guiding principles informed the design and development of the tool. We discuss these principles in detail below.

### 3.1. Interoperability

Comprehensive integration of EDRN research results requires frequent communication with other components of the EDRN Knowledge Environment (EKE), as well as with services available on the web. The primary method of inter-application communication among the distributed components of EKE is Resource Description Format, or RDF [12]. The curation application both consumes and produces this structured data depending on whether the curator is aggregating or publishing. The application regularly ingests up-to-date RDF data about studies, sites and investigators produced by other EKE components and thus keeps the curator’s view of the data fresh. Following curation, the data, metadata, and captured relationships are packaged as RDF and exported to the EDRN Public Portal where the additional detail is used to provide enhanced search results to the general public.

The scope of the curation task is not limited to data generated by EDRN institutions. Many EDRN studies, for example, reference external publications accessible via resources like PubMed [7]. Additionally, linking external resources of interest (such as laboratory web pages and relevant external registry data) to a particular biomarker or study helps to develop a more complete picture of the current state of research. To accommodate this, the curation application has been designed to incorporate resources from both internal and external sources.



**Figure 1. Various Screenshots of the BMDB Curation Web Application**

As an example, the curation application leverages the PubMed API [9] to provide a seamless integration with the vast repository of medical literature provided by PubMed. Without leaving the application, a curator can query PubMed and import basic information about matching publications. From the point of view of the curator, these remote publications are transparently available for association with EDRN research elements.

### 3.2. Flexibility

EDRN research data is comprised of many data types including studies, sites, investigators, specimens, statistical data, publications, and external resources. Defining and modeling the many relationships between these elements was an iterative process, with increased scientific accuracy achieved over time. In order to minimize the amount of application redesign resulting from data model changes, the application employed a Model-View-Controller architectural pattern [2], logically dividing application code along the lines of its functional contribution to either the data model, application logic, or presentation layer. Because the model code was thus relatively isolated, changes were easier to implement and test since only a limited and well-defined subset of the code was modified.

### 3.3. Usability

Streamlining and simplifying the data curation process meant that architectural decisions affecting usability should minimize complexity and limit ambiguity. Understanding the curation workflow was a critical prerequisite to developing a usable tool. This involved soliciting input from actual curators and provided valuable insight into how the tool could facilitate the data curation process.

One of the key findings was that support for nonlinear editing of annotations and relationships was essential. Due to the distributed nature of the research, it was impossible to guarantee that complete information for a data element

would be available all at once. As a result, the user interface was designed to make it easy to return to a specific relationship or annotation as information became available. Short, simple forms and a hierarchical, tabular display of the data (Figure 1) meant that a curator was free to immediately capture whatever data was on hand at the moment and able to easily identify areas where further curation effort was required.

## 4. Experience

The BMDB curation web application has been operational for several months. At present, we have only one curator, although our design does not impose any strict limits on the number of curators that can be supported. Although it is still early, there are several metrics by which the tool can be viewed as a success.

The first is an improvement in data visibility. Because it was designed with a curator's natural workflow in mind, the tool offers a comfortable interface to enter and review data from myriad original sources, providing a way to track the progress and completeness of the research for a particular biomarker. The interface design, along with integration with EKE and external data sources creates a fluid user experience. Because the tool organizes information hierarchically, a curator can incrementally edit and refine elements. This has contributed to the rapid introduction of new data into the database.

Second, the curated data in the BMDB is of greater value to a researcher than the simple sum of the original data elements. Because of the explicitly codified relationships and annotated metadata, the curator has added semantic value to the database. The additional information, in turn, enables search queries to return a rich array of relevant material spanning multiple data types. The curation web application now plays a critical step in the effort to provide a comprehensive picture of the complete state of research for a given biomarker.

The application's effectiveness in the context of the EDRN is assisted by the fact that the scope of the EDRN

mission is so well defined. Because the EDNRN focuses exclusively on discovery and validation of biomarkers for early detection, the infrastructure and tools could be highly optimized. With the domain constrained by the limited scope of the mandate, a more directed focus has been possible, permitting the capture of comprehensive, meaningful, and scientifically accurate relationships.

## 5. Conclusion

We believe that curation plays a critical role in the data quality management process. We feel this is particularly true in virtual organizations where geographically distributed participants each contribute their efforts to a shared research goal. The curation process adds value to the raw research data through the identification of relationships and annotation of metadata. As an additional benefit, curation adds a level of quality control to the management process, helping to ensure consistency and conformity between data obtained from multiple sites.

In our experience developing and deploying a web-based curation tool as part of an informatics infrastructure for the EDNRN, we found that we were able to improve the content management workflow by providing a curation tool for managing science data and policy. Interoperability of the tool with EKE and external web services was critical to facilitating simple yet comprehensive curation. Furthermore, an uncomplicated architecture reduced the time and effort required to adapt the tool to changing model requirements. Finally, by focusing on the curator's experience, we found that we were able to increase data visibility and completeness, resulting in improvements both the quantity and caliber of data making its way into the public domain.

### 5.1. Future Work

As the volume of science data increases exponentially [10], we anticipate challenges on the horizon for data curation. These challenges include capturing the increasing richness and depth of the research data and scaling to meet the needs of a growing virtual organization as it incorporates research from additional sites and investigators. Our experience with the BMDB to date has been successful in improving the curation workflow as well as both the quantity and quality of the curated data. Our future work involves building on the lessons learned, and extending the BMDB curation approach and methodology to our other EDNRN applications, in particular our science data warehouse, eCAS.

### 5.2. Acknowledgments

This effort was supported by the Jet Propulsion Laboratory, managed by the California Institute of Technology

under a contract with the National Aeronautics and Space Administration. The authors would like to thank Donald Johnsey, Christos Patriotis, and Sudhir Srivastava and the NCI leadership as a whole for their collaborative guidance and support.

## References

- [1] catissue core, <https://cabig.nci.nih.gov/tools/catissuerecore>, 2008.
- [2] P. Clements, D. Garlan, L. Bass, J. Stafford, R. Nord, J. Ivers, and R. Little. *Documenting Software Architectures: Views and Beyond*. Pearson Education, 2002.
- [3] D. Crichton, S. Kelly, C. Mattmann, Q. Xiao, J. S. Hughes, J. Oh, M. Thornquist, D. Johnsey, S. Srivastava, L. Essermann, and W. Bigbee. A distributed information services architecture to support biomarker discovery in early detection of cancer. In *e-Science*, page 44, 2006.
- [4] D. J. Crichton, J. S. Hughes, G. J. Downing, H. Kincaid, and S. Srivastava. An interoperable data architecture for data exchange in a biomedical research network. In *CBMS*, pages 65–72, 2001.
- [5] I. Foster, C. Kesselman, and S. Tuecke. The anatomy of the grid: Enabling scalable virtual organizations. *J. Supercomputing Applications*, pages 1–25, 2001.
- [6] W. H. Frist. Health care in the 21st century. *N. Engl. J. Med.*, 352(3):267–72, 2005.
- [7] T. Greenhalgh. How to read a paper: The medline database. *BMJ*, 315:180–183, 1997.
- [8] A. Hart, J. Tran, D. Crichton, K. Anton, H. Kincaid, S. Kelly, J. Hughes, and C. Mattmann. An extensible biomarker curation approach and software infrastructure for the early detection of cancer. In *Proceedings of the IEEE Intl. Conference on Health Informatics*. IEEE, 2009.
- [9] M. A. Hearst, R. B. Altman, A. S. Schwartz, G. Bhalotia, and D. E. Oliver. Tools for loading medline into a local relational database. *BMC Bioinformatics*, 5:146+, 2004.
- [10] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D. P. Hill, R. Kania, M. Schaeffer, S. St Pierre, S. Twigger, O. White, and S. Yon Rhee. Big data: The future of biocuration. *Nature*, 455(7209):47–50, 09 2008.
- [11] H. Kincaid, S. Kelly, D. J. Crichton, D. Johnsey, M. Winget, and S. Srivastava. A national virtual specimen database for early cancer detection. In *CBMS*, pages 117–123, 2003.
- [12] O. Lassila and R. Swick. Resource description framework (rdf) model and syntax specification. W3c recommendation, World Wide Web Consortium, 2001.
- [13] C. Mattmann, D. J. Crichton, N. Medvidovic, and S. Hughes. A software architecture-based framework for highly distributed and data intensive scientific applications. In *ICSE*, pages 721–730, 2006.
- [14] S. Srivastava and B. Kramer. Early detection cancer research network. *Lab Invest*, 80:11478, 2000.
- [15] A. C. von Eschenbach and K. Buetow. Cancer informatics vision: cabig. *Cancer Informatics*, 2:22–24, 2006.