

## **Curation of EDRN Cancer Biomarker Research Data**

Andrew F. Hart, MS, Jet Propulsion Laboratory

### **Authors**

Heather Kincaid (JPL), Andrew Hart (JPL), Kristen Anton (Dartmouth), Christos Patriotis (NCI), Mark Thornquist (FHCRC), Jackie Dahlgren (FHCRC), Chris Mattmann (JPL), Dan Crichton (JPL), and John Tran (JPL)

### **Abstract**

The EDRN conducts research on cancer biomarkers from laboratories and universities spread throughout the United States. Due to the distributed nature of the research, data curation is a critical component of the EDRN content management process. Curation provides a quality-control gate through which research metadata (data about the data) must pass before entering the review process and ultimately being made available to the public via the EDRN Public Portal. Furthermore, the curation process allows for linking various EDRN data together to improve access and discovery of biomarker research data from EDRN studies.

To support the EDRN's data curation needs, the NASA Jet Propulsion Laboratory (JPL), the Fred Hutchinson Cancer Research Center (FHCRC), and the National Cancer Institute (NCI) have collaborated on the design and development of two web applications: the Biomarker Database (BMDB) Curation Web Application and the EDRN Catalog and Archive Service (eCAS) Curation Web Application. The applications are part of a suite of software components known collectively as the EDRN Informatics Infrastructure.

The BMDB Curation Web Application provides a hierarchical, faceted view of the EDRN research data stored in the EDRN Biomarker Database. Information about biomarkers, studies, publications, resources, sites, investigators and more can be linked together and annotated by a curator to provide a comprehensive picture of the current state of research for a given biomarker. This data, and the associated metadata, is then exported to the EDRN Public Portal, where it enables the return of a rich array of relevant data in response to user queries.

The eCAS Curation Web Application focuses more specifically on the data sets that are generated by EDRN researchers. It provides a web-based interface for the analysis and annotation of the metadata associated with a data set and its individual data elements, providing a mechanism to ensure the completeness and consistency of the descriptions prior to ingestion of the data into the archive.

These applications facilitate the preparation of EDRN research data for peer-review and public release, providing a flexible environment for the rapid processing of data as it becomes available. The tools follow the practices present throughout the EDRN Informatics infrastructure, including using open standards for communication, and enterprise security mechanisms to ensure appropriate access. These components form the backbone of the EDRN content management process, helping to ensure that consistent, high-quality research results are made available to the public.