

The eCAS Model for Scientific Data Warehousing of Biomarker Data

Chris A. Mattmann, PhD, Jet Propulsion Laboratory

Authors

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

Abstract

Cataloging and archiving, and distributing large, complex datasets is a core competency as modern scientific datasets continue to scale dramatically in both size and complexity. To support these functions, we have designed and developed the EDRN Catalog and Archive Service (eCAS) for the U.S. National Cancer Institute's Early Detection Research Network (EDRN) program. The EDRN consists of several dozen organizations and institutions spread across the U.S. all researching into the detection and validation of cancer biomarkers. The eCAS system is part of the EDRN informatics infrastructure, developed by the NASA Jet Propulsion Laboratory (JPL), in collaboration with the Fred Hutchinson Cancer Research Center (FHCRC), whose goal is capturing results from EDRN biomarker studies.

eCAS is a distributed, metadata-driven system that supports ingestion, curation, storage, and retrieval of scientific datasets. eCAS builds upon NASA's Object Oriented Data Technology (OODT) software framework, which has been used on NASA's robotic missions. eCAS provides an integrated, web-based management interface to support biocuration and management of the data holdings, directly capturing EDRN research results.

eCAS shares data seamlessly with the other components of the EDRN informatics infrastructure by adhering to the EDRN Ontology, a shared semantic description of biomarker research elements and the relationships between them. This allows data in eCAS to be accessed and shared through the EDRN scientific portal known as the EDRN Knowledge Environment (EKE) using state-of-the-art search mechanisms. eCAS also employs advanced security policies so sensitive data is only viewed by specific individuals or research groups.

eCAS is expected to play a critical role in the future of EDRN as scientific collaboration continues to expand and the associated data sets continue to increase. We anticipate expansion of capabilities to support both curation and distribution of the data across both the EDRN enterprise and the broader research community. We will continue to work with the EDRN scientific community to ensure eCAS plays a critical role in preserving the scientific results and supporting new discoveries.