PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

The EDRN knowledge environment: an open source, scalable informatics platform for biological sciences research

Daniel Crichton, Ashish Mahabal, Kristen Anton, Luca Cinquini, Maureen Colbert, et al.

Daniel Crichton, Ashish Mahabal, Kristen Anton, Luca Cinquini, Maureen Colbert, S. George Djorgovski, Heather Kincaid, Sean Kelly, David Liu, "The EDRN knowledge environment: an open source, scalable informatics platform for biological sciences research," Proc. SPIE 10194, Micro- and Nanotechnology Sensors, Systems, and Applications IX, 101942A (18 May 2017); doi: 10.1117/12.2263842

SPIE.

Event: SPIE Defense + Security, 2017, Anaheim, California, United States

Invited Paper

The EDRN Knowledge Environment: An Open Source, Scalable Informatics Platform for Biological Sciences Research

Daniel Crichton[a], Ashish Mahabal[b], Kristen Anton[c], Luca Cinquini[a], Maureen Colbert[c], S. George Djorgovski[b], Heather Kincaid[a], Sean Kelly[a], and David Liu[a]

[a]Jet Propulsion Laboratory, Caltech, Pasadena
[b]Center for Data Driven Discovery, Caltech, Pasadena
[c]Geisel School of Medicine at Dartmouth

ABSTRACT

We describe here the Early Detection Research Network (EDRN) for Cancer's knowledge environment. It is an open source platform built by NASA's Jet Propulsion Laboratory with contributions from the California Institute of Technology, and Giesel School of Medicine at Dartmouth. It uses tools like *Apache OODT*, *Plone*, and *Solr*, and borrows heavily from JPL's Planetary Data System's ontological infrastructure. It has accumulated data on hundreds of thousands of biospecemens and serves over 1300 registered users across the National Cancer Institute (NCI). The scalable computing infrastructure is built such that we are being able to reach out to other agencies, provide homogeneous access, and provide seamless analytics support and bioinformatics tools through community engagement.

Keywords: Data Science, Knowledge Environment, Cancer, Open Source, Apache OODT, Common Data Elements, Biomarker, Ontology

1. INTRODUCTION

NASA's Jet Propulsion Laboratory (JPL), in collaboration with scientific investigators from several biomedical research institutions, is extremely active in research and technology development of scalable biological data networks, leveraging Big Data technologies from JPL's Earth and planetary science programs. In addition to constructing multi-petabyte infrastructures and data archives for Earth, astronomy and planetary science missions, for the past ten years JPL has led the construction of a novel biomedical research environment for the National Institutes of Health, called the Early Detection Research Network (EDRN).[1] The EDRN Informatics Center[22,23] (PI: Daniel Crichton) has developed a comprehensive Knowledge System for the capture, processing, management, distribution and analysis of data to support the discovery and validation of cancer biomarkers across many organ sites dovetailing various open source technologies.[2–5]

The EDRN Knowledge System[22] is a highly modular, scalable infrastructure that supports the capture, generation, management, dissemination, and analysis of data across the network using an open source infrastructure originally developed to capture massive data repositories from planetary science robotic missions. By having a well-defined software and information architecture, the EDRN, through its Informatics Center at NASA Jet Propulsion Laboratory (JPL), has been able to unify biomarker research data and knowledge and make them available to the research community. Rather than providing a centralized database or warehouse, the EDRN chose to build a nationally distributed system, linking disparate resources and providing access through a sophisticated Knowledge Portal built on a secure data infrastructure. The Knowledge System has collected terabytes of data encompassing EDRNs biomarker research including biomarker annotations (approximately 900 from EDRN research), dataset and analysis information (approximately 70 datasets), protocols (more than 220), biospecimens (approximately 200,000), and other information that is integrated and presented to over 1300 registered users.[2,5] Figure 1 represents total publications over time and total datasets by organ. Over 1100 Common Data

Send correspondence to: daniel.j.crichton@jpl.nasa.gov

Micro- and Nanotechnology Sensors, Systems, and Applications IX, edited by Thomas George, Achyut K. Dutta, M. Saif Islam, Proc. of SPIE Vol. 10194, 101942A · © 2017 SPIE
CCC code: 0277-786X/17/$18 · doi: 10.1117/12.2263842

Elements make up a biomarker ontology that provides the foundation for integrating data across the EDRN. The data from the 900 candidate biomarkers encompass 11 organ and tissue sites. EDRN researchers have published more than 230 papers on these biomarkers, with many more papers in preparation. To date, the EDRN program has seen five of its candidate biomarkers receive FDA approval as cancer diagnostics: %[-2]proPSA, urinary PCA3, OVA1, ROMA, and DCP and AFP-L3. In addition, nine CLIA certified diagnostic tests are based on EDRN biomarkers: MiPS (Mi Prostate Score Urine test), IHC and FISH for T2-ERG fusion, GSTP1 methylation, prostate mitochondrial deletion, Somalogic 12-marker panel for lung cancer, lung 80-gene panel, vimentin methylation in stool, galectin-3 ligand for colon cancer, and an 8-gene panel for Barretts Esophagus.

The maturation of the EDRN Knowledge System coupled with the long-standing NASA-NCI Inter-agency Agreement to leverage scalable informatics capabilities from NASA/JPL provide a robust capability and collaboration that can be leveraged to deliver high quality, science-driven informatics to other programs and serve as an integration hub to support the generation, management, distribution and analysis across multiple biological challenges. The models, methodology and design of the EDRN system, derived from implementations in other science disciplines by JPL and adapted for biomarker research, demonstrate the ability of this team to deliver scalable informatics capabilities for other programs. To that end, JPL, NCI, and Caltech held a workshop in May 2013, the Cancer Biomarker Bioinformatics Workshop that included over 80 scientists, statisticians, and experts in computational science to address future data-intensive science needs across the lifecycle from laboratories all the way through to integrated data analysis. The workshop concluded that developing systematic solutions for capturing, generating, managing, and analyzing data, end-to-end, is critical to enabling reproducibility in the era of big data science.[6] Using our cross-disciplinary approach between space science and cancer research to construct scalable, data-driven knowledge environments through reusable software architectures and informatics enables us to build novel, cross-disciplinary scientific solutions for data science.

To date, all of our software is released as open source. The underlying data science framework developed at JPL, Apache OODT,[24] is a top-level project at the Apache Software Foundation (ASF) with over 40 committers all over the world that are using OODT to capture, process, manage, distribute, and analyze data. The EDRN Knowledge System, built on top of Apache OODT, and customized for biomedical research, is released to GitHub for global use.

2. ARCHITECTURAL OVERVIEW

The Knowledge System serves as an integration hub for data, services, and tools to directly provide informatics technology to the NCI research community. This research community is highly diverse and highly distributed, driving a heterogeneous environment. It is critical that this architecture is readily able to adapt and evolve as data from different groups are brought into the Knowledge System. The architecture for the Knowledge System is governed on two levels: first, by adhering to the fundamental principles for implementing highly distributed, diverse science environments—listed in the table below—derived from our experience;[7–10] and second, by specifying components of the architecture that can respond to these principles.

Table 1: Principles of Architecture

S. No.	Principle	Description
P1	Access	Software should provide uniform methods for access to diverse data in the knowledge system .
P2	Location independence	Users should not concern themselves with the physical location of data or services.
P3	Well defined information architecture	An explicit domain information model is critical, independent of the software (e.g., cancer biomarker information model).
P4	Allow plug-in algorithms	Systems should allow for algorithms to be integrated into the software framework that allow for specific data processing workflows and/or automated data discovery techniques from ingestion to data dissemination and analysis.

The collaborative nature of cancer research necessitates bringing together data from multiple institutions, yet far too often systems are not architected as virtual scientific environments that can support the analysis of distributed data. The first principle, access (P1),[11] is targeted at ensuring that data can be made available using secure, uniform software methods, to a large distributed community.

To support the distributed nature of cancer data and information, location independence (P2) has become a fundamental architectural tenet and a valuable feature for the construction of successful informatics systems. Location independence prescribes that the physical location of data and components should be transparent to those accessing them. Specific data and software, whether local or geographically distributed, should be transparent to users and other software applications. The access and interpretation of the data objects should remain consistent despite multiple topologies for the system that may be in place. We allow for location independence of research centers while delivering access to the data through a unified Knowledge Portal. The efficacy of this has been demonstrated by the EDRN Knowledge System, which provides access to distributed biospecimen and other data across funded investigator laboratories.[11–13]

Our experience has shown the importance of defining the information architecture independently from the technical system (P3). This is particularly true for science-driven disciplines where the data and the associations for it are evolving. As a result, we assert that software should respond to the data model defined in the information architecture rather than embedding that model directly into the software. We will expand on this below, where we have developed a comprehensive set of data standards and information models to ensure the Knowledge System can capture diverse data. Similarly, data processing is a critical element, but often needs to be specialized. Algorithms should be plugged in and run as workflows (P4), something that is routinely done in many science-driven informatics systems. We will describe this below as we talk about integration of bioinformatics tools and algorithms.

Finally, we do not view the Knowledge System as a place to deliver data and services hosted and developed only by this team. Rather, we view it as an integration hub allowing programs to incorporate data, services, and tools into the Knowledge System. The architecture, as specified, enables data and services to be managed by other groups. This has been a hallmark capability of the EDRN.[14] For example, the EDRN Resource Network Exchange (ERNE)[15–17] allows for the integration of highly distributed specimen repositories. This allows the EDRN to integrate other systems and tools such as the NCIs caTissue[18] into the Knowledge System to share specimen information.

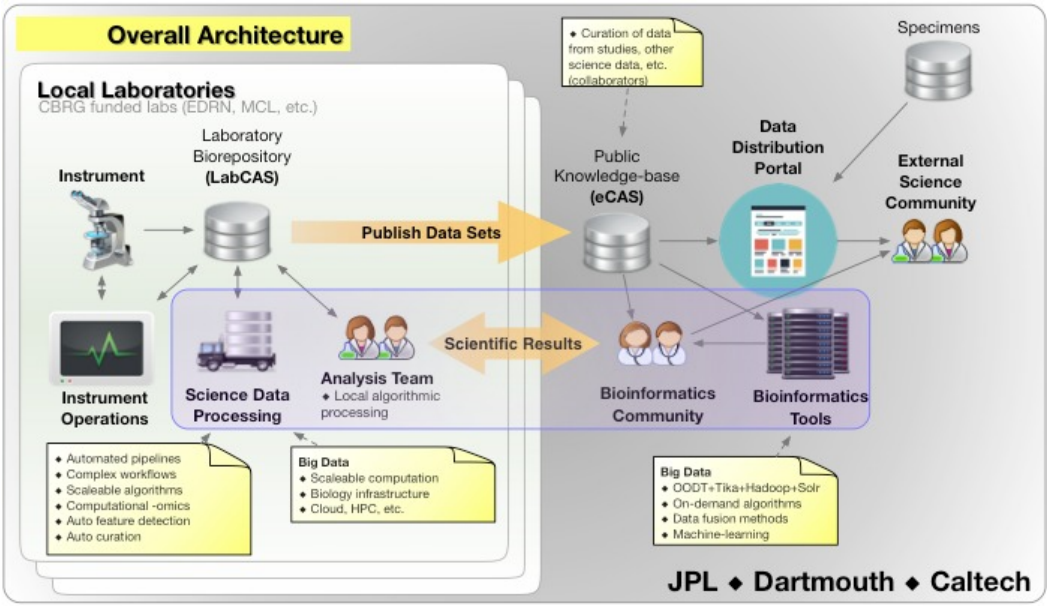

Figure 1: Overall Architecture (see Table 2 for mapping to items of the architecture).

Figure 1 depicts the concept of the Knowledge System including the flow of data and processing, end-to-end, as a distributed system. An overarching core information model defines the domain data elements and their relationships. It is used for describing the scientific data captured in the Knowledge System.

3. SCALABLE COMPUTING INFRASTRUCTURE

To implement the Knowledge System, we have developed and deployed a scalable data and computing infrastructure including identified components and services, as part of the architecture, identified in Table 1, that adhere to our architectural principles and support the capture, processing, management, dissemination, and analysis of data within the system.[2,3,5]

Table 2: Components, Tools and Services in the EDRN infrastructure.

Item	Software Component/Service	Description
3	Information Model	An organized set of Common Data Elements (CDEs) captured as an information model, defined by standards, and expressed as. structures that will be used by metadata, data, and software
2	Infrastructure built on *Apache OODT* and other open source data services	*Apache OODT* and other Apache data services (e.g. *Hadoop*, *Hive*, *Solr*, Mahout, etc.) will be used to form the core data-intensive service infrastructure on which the system is built.
2 and 4	Multi-level Security	Authentication and authorization for managing access to the data and service in the Knowledge System.
4	Science Data Portal	Authenticated and authorized access to the data, services, and tools in the center; search and presentation services
5	Core Systems Databases	Core databases including information about studies, sites, investigators, and publications for linking to research results and driving science portal
5 and 6	Scalable Archive, Processing, and Catalog Services	User tools for data movement, ingestion, data generation and workflow, catalog management, and dissemination provided through the eCAS and LabCAS software systems.
5	Virtual Storage Services	Virtualized storage services to support a growing knowledge-base.
6	Scalable Computation	Virtualized computational services that will be used to support on-demand processing and integrated with the LabCAS system.
6 and 7	External Software Tools and Servcies	Integrate external software tools, applications and data services for interoperability.

The core infrastructure supports a scalable set of data-intensive services based principally on Apache OODT. Apache OODT was developed by JPL over the past decade for distributed data management and processing services, solving problems in data capture, management, processing, and exchange. It now has approximately 40 committers worldwide at the ASF with systems managing 100 terabytes and plans to capture petabytes of data. It is domain independent and follows all architectural principles specified above. A key feature is the separation of data cataloging from data storage, allowing data to scale on massive servers, but still enabling discovery through a registration/cataloging process. It provides APIs in Java and Python and its services can be accessed through a variety of modern software client interfaces including discovery and remote access to both metadata and data. This allows for integration of different types of elastic services such as cloud computing. *OODT* is routinely coupled with *Apache Hadoop* and *Solr* to support scalable data management and search.

Other infrastructure capabilities include multi-level security, scalable storage, and scalable computation. Multi-level security is implemented using the Apache Directory Service,[19] in which all security information is recorded. This information includes principals (users), group memberships, and roles granted to those groups. Principals are identified with a combination of username and password. All data managed within the infrastructure is communicated with encrypted protocols, making use of both AES256 (FIPS 197)[20] with random keys

from secure number generators and TLS1.2.[21] Furthermore, such data is stored on servers using encrypted file systems (with AES256) and the servers themselves reside in secure locations with physical access controls. Such servers have full firewalls and all unnecessary services disabled. Developed applications exist solely within this security framework, validate users, and prevent actions not within the users' roles. For storage and computation, we use shared computing services from other massive JPL data initiatives in climate and space science. We also take advantage of JPLs institutional agreement with Amazon, providing direct access to cloud computing services on demand.

4. DATA ARCHITECTURE

Our approach for data management heavily leverages an information model-driven architecture where the definition of data is captured and managed in an information model, and used to drive the construction of the software and the system. This is employed in our space and biological data system implementations, and allows us to adapt and support many different types of data.

For the EDRN Knowledge System, we capture a variety of information objects that are well defined by our information model including various types of biomarkers, studies, biospecimens, raw and processed study data, investigators, etc. Each of these are interrelated in our information model and used to construct the semantic architecture that enables search and discovery in our knowledge system. While much of this data is being generated and shared across the EDRN, the data architecture is critical for defining how the data can be interrelated. Metadata within our information model is described as Common Data Elements (CDEs). These CDEs are captured using ISO/IEC 11179 as a standard reference model for describing data elements and their associated values—something that has been done for both the EDRN and Planetary Data System. For EDRN, we have worked with science teams to agree on the standards for data collection using the CDEs. All data objects have specific identifiers ensuring their uniqueness as well as their access requirements.

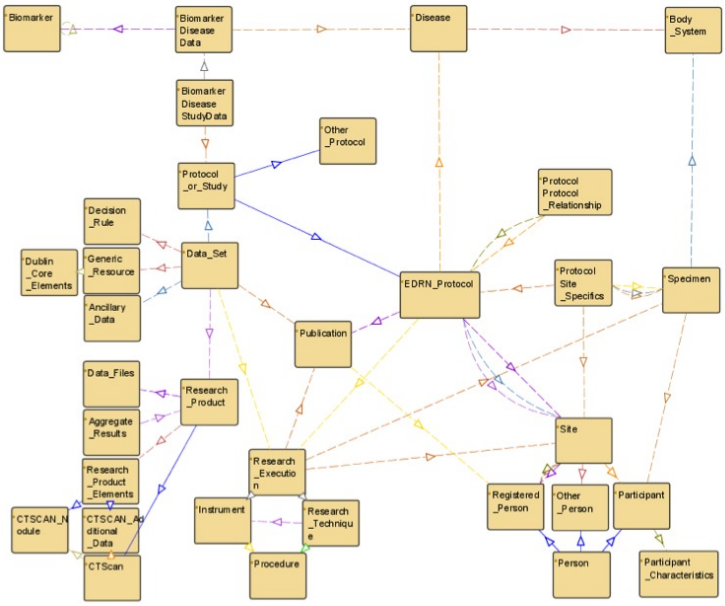

Figure 2: EDRN Biomarker Ontology. Many blocks are tightly interconnected with several other nodes.

The EDRN biomarker information model, shown in Figure 2, is captured formally as a cancer biomarker ontology managed in Protégé. Classes of data include biomarkers, protocols, specimens, people, publications, organs, science data, and other information relevant to the definition of the data in cancer biomarker research. The Common Data Elements (CDEs), derived from the ontology, encompass the parlance through which scientific data may be accurately described, searched, and interchanged. The CDEs are widely used for data collection and for software within EDRN to increase interoperability and bring the Knowledge System together.

Data within our system is captured and validated against the CDEs. Both our scientists and our curation team curate the metadata. Data goes through an ingestion, review, and release process to ensure it is of high quality. This is further described below under the data curation process.

5. DATA CURATION PROCESSES

The EDRN Knowledge System captures clinical and biological information about cancer biomarkers from a biomarker-centric perspective, through the Biomarker Database (BMDB), and makes that data available through the Knowledge Systems Portal. The knowledge system also captured high quality data sets and experimental data sets through its scientific data warehouses that we call eCAS and LabCAS (EDRN Catalog and Archive Service and Laboratory Catalog and Archive Service, respectively).

The cancer biomarkers are identified by an official HUGO Gene Nomenclature Committee (HGNC) name when available and then programmatically uploaded into the database in sets ranging in size from individual biomarkers up to several hundred. Before the upload, a variety of metadata about each set of biomarkers is assembled and incorporated into the upload process. These metadata include biological and clinical data as well as security parameters and links to additional resources, both from within EDRN and from public biological databases, to ensure that system can be semantically linked.

The biological metadata consists of a comprehensive list of synonyms by which the biomarker is known (and may be searched from the Knowledge System Portal) and a general description of the function of the gene or protein. This biological data is extracted from highly reliable informational repositories, including HGNC, UniProt, and NCBI Entrez Gene. Clinical data provided by the EDRN cancer biomarker research laboratories includes organ from which the biomarker was isolated, phase of discovery, and type of biomarker. Additionally, security parameters are added which limit access to the biomarker to authorized groups. Once data is published, the annotated biomarker is made public in the EDRN Knowledge System. Finally, administrative metadata is included which points to resources used to obtain the biological and clinical data. This includes links to the specific records in HGNC, UniProt, PubMed, and NCBI Entrez Gene used to compile the metadata for each biomarker. With inclusion in the Knowledge System, biomarkers are systematically linked to the lab studies, protocols, biospecimens, data sets, and investigators involved in the discovery and validation of the biomarkers.

Once the upload process is complete the cancer biomarkers are curated in a more manual process by domain expert biocurators who read each publication accompanying the biomarkers and add specific information describing performance in assays and experimental procedures. Biocurators work collaboratively with the Informatics Center team to influence optimal database design and processes around biomarker curation.

6. TYPES OF DATA IN EDRN

The types of data captured and managed in our knowledge system are extensive and varied. The research network includes experts in genomics, proteomics, epidemiology, epigenetics, biostatistics, and other disciplines. Figure 4 shows various molecular types of biomarkers in our knowledge system. These biomarkers are associated with various organs, most commonly with prostate, ovary, breast, and lung. Other related information related to biomarkers are also captured Each collaborative group's contributions to panel, biomarker, data, protocol, and member information is shown in Figure 3.

The Informatics Center handles a huge variety of biomedical and clinical data generated and captured around understanding biomarker structure and function—in a molecular sense—and biomarker sensitivity and specificity—in the clinical setting. The Knowledge System framework is particularly appropriate for large-scale research undertakings such as EDRN because it accommodates data from any platform and in any format, requires extensive standardized metadata for data sets and products, readily manages and integrates large data files (including images), and is inherently flexible to accommodate future data whose characteristics are at present not yet known. The field of biomedical science is evolving rapidly, with new technologies expanding the flow and variety of data requiring management and inclusion in the Knowledge System. By leveraging an ontology that is configured into Apache OODT and not hard-coded into software, the framework is both prepared to manage data we cannot yet imagine and available to be re-purposed for managing large distributed data in any domain.

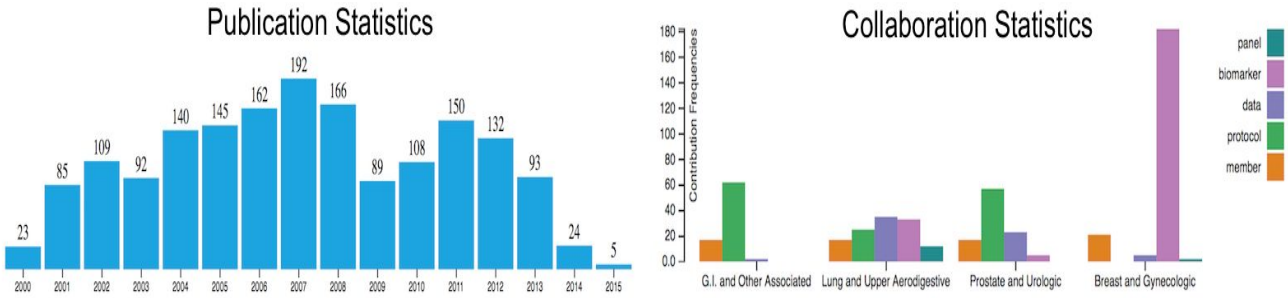

Figure 3: Biomarker and Collaborative Group statistics.

7. DISTRIBUTED ACCESS AND DATA SHARING

The centerpiece of the Knowledge System is the Portal enabling access to analysis tools and dissemination of data captured by the system. The portal serves the diverse needs of the cancer community ensuring that multiple groups (researchers contributing data, applicants, program managers, collaborators, potential future research members, and the broader research community—as well as cancer advocates and the general public) have appropriate access to a variety of information. Motivated by this diverse audience, we ensure that the Portal has sufficient flexibility in its breadth and presentation of information, and that it functions well as the main entry point into the Knowledge System. Portal functionality includes the abilities to search, access, present, and distribute data captured in online repositories and to link to advanced services for executing data analytics, granting access and rights, based on the our security infrastructure.

The Portal supports a publishing paradigm, whereby information about data, services, and tools from the cancer community are published to the Portal as metadata descriptions based on the information model (as identified above) and the W3C Resource Description Framework (RDF) recommendation. This approach allows for the Portal to capture, catalog, organize, semantically link, and index these metadata descriptions into a knowledge base that then can be used to enable access and presentation of the data. This will allow for the linking of scientific data sets to users, studies, publications, instruments, technologies, and other research programs. The indexing of the information will allow for searching and presenting results using a Google-like approach, but implemented with the enterprise-grade *Apache Solr* open source toolkit. Team members at JPL are open source committers at Apache, members of the board of directors, and have contributed the core development of many Apache projects including *OODT*, *Solr* (above), and *Hadoop*.

This integration architecture via publishing of data to the Portal works well for scaling the data and services.

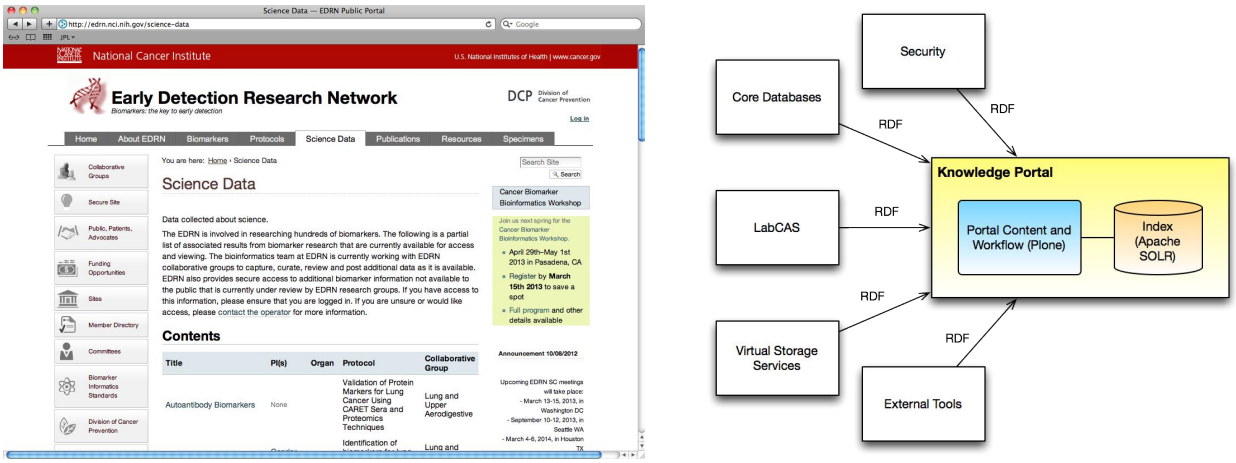

Figure 4: EDRN Knowledge Portal and publishing to the portal.

It provides a platform for linking data captured in science data warehouses, providing a single point of entry for access to large collections of data. It also allows for the addition of external software tools by providing a standard convention for publishing data and software services to the Knowledge Portal so that the community can find, access, and use those services.

JPL has constructed the portal using an open source portal technology called Plone. Plone provides an enterprise-grade content management infrastructure, allowing automated systems as well as content editors to update and publish scientific data, protocols, news, and other information without requiring software changes to a website. Members of JPLs Informatics Center have been some of the leaders in using Plone, contributing to its core development.

Additionally, Plone's robust security features means the Portal can be public-facing on the internet. Researchers may log in using credentials stored in the Apache Directory Service (mentioned above), gaining higher levels of authorization. This allows for access, for example, to biomarkers still being curated, or to collaborative workspaces where groups working on multi-institution protocols can share data and documents.

8. BIOINFORMATICS TOOLS AND ANALYTICS SUPPORT

A major aspect of our architecture is enabling the integration of bioinformatics tools and data processing pipelines to enable analytics support. We have developed LabCAS as an open source, data processing infrastructure that integrates automated ingestion of data into the knowledge system using Apache OODT. The data processing infrastructure provides services for file and metadata management, workflow management, and resource management. All services support APIs and allow for significant flexibility in swapping out the backend software implementation, as well as integrating data from distributed sources. The infrastructure provides client frameworks for remote file acquisition and protocol negotiation, automatic file identification and ingestion, and rapid science algorithm integration.

Apache OODT provides its own scalable workflow services. We are aware that there are many workflow engines in use in the physics and life sciences community (e.g., Pegasus, Wings, Kepler, Galaxy, Taverna etc.). Hence, our architecture allows for the integration of different workflow systems. For example, we have researched the integration of the Berkeley Data Analytics Stack (BDAS) including Spark, Shark, and Mesos to evaluate scalability for our big data science deployments and, in particular, running elastic MapReduce analytic algorithms and jobs.

We have worked with groups at Boston University, Cedars Sinai, M. D. Anderson Centre, NIST, Sloan Kettering, and Vanderbilt, to integrate proteomics, genomics, and imaging data. The LabCAS system provides a flexible infrastructure, based on our Earth and planetary science data pipeline approaches, to build analytic pipelines. As a result, we believe LabCAS provides an excellent framework for the efficient sharing of sequencing, mass spectrometry, and imaging data as well as analyses allowing both the processing and sharing of data as well as the integration of algorithms and tools.

9. COMMUNITY ENGAGEMENT, DISTRIBUTION AND SUPPORT

The programs we have worked with encompass researchers from many different settings, including academia, government and for-profit-organizations. Integrating the data, software, computational methods, and services from highly diverse research groups has required that we develop an extremely flexible architecture.

In the life sciences discipline we have worked with the NCI Early Detection Research Network, serving as the Informatics Center (which contains over 40 different PI sites), with the Helmsley Foundations SHARE network (which contains eight sites), with Consortium for Molecular and Cellular Characterization of Screen-Detected Lesions (MCL; eight sites), and are working with a few commercial entities to integrate the open source framework. This has included working with investigators to set up software services as well as capture their data and support analysis of the data within the system.

The open source framework has been critical to enabling the construction of a national knowledge system. In addition, the relationship to other sectors, including space and Earth Science, has enabled us to significantly leverage mature software as well as to grow an open source community of users and developers.

ACKNOWLEDGMENTS

We acknowledge the EDRN grant from NIH (82-16333), and the support from JPL Center for Data Science and Technology. AAM and SGD also acknowledge support from the Center for Data-Driven Discovery at Caltech, and from the Ajax Foundation.

REFERENCES

[1] S. Srivastava. B. Kramer, "Early Detection Research Network", *Nature*, Lab Invest 80:11471148, August 2000, PMID: 10950105

[2] D. Crichton, C. Mattmann, M. Thornquist, K. Anton, J. S. Hughes, "Bioinformatics: Biomarkers of Early Detection", *Cancer Biomarkers*, IOS Press, Vol. 9, Number 1-6 2011, PMID 22112493

[3] D. Crichton, H. Kincaid, J.S. Hughes, S. Kelly, S. Srivastava, D. Johnsey, "Creating a National Virtual Knowledge Environment for Proteomics and Information Management", *Informatics and Proteomics*, Marcel Dekker Publishers, December 2004

[4] D. Crichton, M. Thornquist, S. Kelly, C. Mattmann, D. Johnsey, J. Dahlgren, D. Steling, G. Warnick, S. Reid, C. Edelstein, A. Hart, H. Kincaid, "A Distributed Informatics Knowledge Environment for Biomarker Research", in Proceedings of the 5th *EDRN Scientific Workshop*, Bethesda, MD, March 17-19, 2008, PMID: 22112493

[5] D. Crichton, S. Kelly, C. Mattmann, Q. Xiao, J. S. Hughes, J. Oh, M. Thornquist, D. Johnsey, S. Srivastava, L. Esserman, and B. Bigbee, "A Distributed Information Services Architecture to Support Biomarker Discovery in Early Detection of Cancer", Proceedings of the 2nd IEEE International Conference on *e-Science and Grid Computing*, pp. 44, Amsterdam, the Netherlands, December 4th-6th, 2006.

[6] M Winget, J Baron, M Spitz, D Brenner, D Warzel, H Kincaid, M Thornquist, Z Feng, "Development of common data elements: the experience of and recommendations from the early detection research network", *International Journal of Medical Informatics*, 70(1):41-8, 2003, PMID:12706181

[7] Cancer Biomarker Bioinformatics Workshop, Pasadena, CA, May 2013, <http://edrn.nci.nih.gov/cancer-bioinformatics-workshop/cancer-biomarker-bioinformatics-workshop-report-may-2013/at_download/file>

[8] D. Crichton, C. Mattmann, J. S. Hughes, S. Kelly, and A. Hart. A Multi-Disciplinary, "Model- Driven, Distributed Science Data System Architecture", *Guide to e-Science: Next Generation Scientific Research and Discovery*, X. Yang, L. L. Wang, W. Jie, eds. Spring Verlag, 2011

[9] C. Mattmann, D. Crichton, A. Hart, C. Goodale, J. S. Hughes, S. Kelly, L. Cinquini, T. H. Painter, J. Lazio, D. Waliser, N. Medvidovic, J. Kim, P. Lean, "Architecting Data-Intensive Systems", *Handbook of Data Intensive Computing*, B. Furht, A. Escalante, eds. 1st Edition. Springer Verlag, 2011

[10] C. Mattmann, D. Crichton, A. Hart, S. Kelly, C. Goodale, R. R. Downs, P. Ramirez, J. S. Hughes, F. Lindsay, "Understanding Open Source Software at NASA", IEEE IT Professional Special Theme on NASA Contributions to IT, Vol. 14, No. 2, pp. 29-35, March/April 2012

[11] National Research Council Committee on the Analysis Massive Data. Frontiers in the Analysis of Massive Data, National Academy Press, 2013

[12] T. J. Fuchs and J. M. Buhmann, "Computational Pathology: Challenges and Promises for Tissue Analysis", *Journal of Computerized Medical Imaging and Graphics*, 35(7):515530, April 2011

[13] T. J. Fuchs, P. J. Wild, H. Moch, and J. M. Buhmann, "Computational Pathology Analysis of Tissue Microarrays Predicts Survival of Renal Clear Cell Carcinoma Patients", Dimitris Metaxas, Leon Axel, Gabor Fichtinger, and Gabor Szekely, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2008*, volume 5242 of Lecture Notes in Computer Science, pages 18. Springer-Verlag, Berlin, Heidelberg, 2008

[14] C. Mattmann, V. Perrone, S. Kelly, D. Crichton, A. Finkelstein, and N. Medvidovic, "A Reference Framework for Requirements and Architecture in Biomedical Grid Systems", Proceedings of the 2007 IEEE International Conference on *Information Reuse and Integration* (IEEE IRI-07), pp. 418-423, Las Vegas, NV, August 13-15, 2007

[15] D. Crichton, G. Downing, J.S. Hughes, H. Kincaid and S. Srivistava, "An Interoperable Data Architecture for Data Exchange in a Biomedical Research Network", proceedings of the International Conference on *Computer-Based Medical Systems*, 2001

[16] D. Crichton, H. Kincaid, S. Kelly, S. Srivastava, D. Johnsey, "A National Data Grid Infrastructure for Sharing Biospecimens in Early Cancer Detection", Proceedings of the *Digital Biology: the Emerging Paradigm*, Bethesda, MD, November 2003

[17] H. Kincaid, D. Crichton, M. Winget, S. Srivastava, D. Johnsey and M. Thornqusit, "A National Virtual Specimen Repository for Early Cancer Detection", Proceedings of the 16th IEEE Symposium on *Computer-Based Medical Systems*, June 2003

[18] J. Klemm, A. Basu, I. Fore, A. Floratos, G. Komatsoulis, "The caBIG Life Sciences Distribution", *Biomedical Informatics for Cancer Research*, Springer, 2010

[19] Welcome to the Apache Directory Apache Directory, The Apache Software Foundation, 2003 2015, May 2015, <http://directory.apache.org/>

[20] Announcing the Advanced Encryption Standard, Federal Information Processing Standards Publication, NIST. November 2001, May 2015, <http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf>

[21] T. Dierks, E. Rescorla, Transport Layer Security (TLS) Protocol Version 1.2, August 2008, <http://tools.ietf.org/html/rfc5246>

[22] Welcome to the EDRN Informatics Center Informatics Center, Jet Propulsion Laboratory, May 2015, May 2015, <http://cancer.jpl.nasa.gov>

[23] JPLs Center for Data Science and Technology Center for Data Science and Technology, Jet Propulsion Laboratory, May 2015, <http://datascience.jpl.nasa.gov>

[24] Apache OODT, The Apache Software Foundation, 2010-2015, May 2015, <http://oodt.apache.org>