Lawrence Berkeley National Laboratory
LBL Publications

## Title

Towards a Data-Centric Research and Development Roadmap for Large-Scale Science User Facilities

## Permalink

https://escholarship.org/uc/item/0jh6b99x

## Author

Bethel, E Wes

## Publication Date

2017-10-01

## DOI

10.1109/escience.2017.72

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

# Towards a Data-Centric Research and Development Roadmap for Large-scale Science User Facilities

E. Wes Bethel
Lawrence Berkeley National Laboratory
Berkeley, CA, USA, 94720

***Abstract*—The U.S. Department of Energy (DOE) Office of Science (SC) operates approximately four dozen large-scale science user facilities (SUFs), each of which generates a tremendous amount of scientific data from experiments, observations and computations. To better understand the data needs and challenges, DOE has run many workshops in recent years to identify and articulate data-centric challenges and opportunities at varying resolution, from facility to community scale. Building on those workshop reports, as well as others from elsewhere in the community, this article goes beyond the findings-recommendations typical of workshop reports to consider how one might structure a broad, technology- and data-centric, coordinated research effort that would realize progress towards solutions that address the well documented challenges and opportunities. We focus on identifying practical issues of strategic relevance, along with offering a view about the focal points for a coordinated research and development effort that would target meeting data-centric needs of a broad set of science users and SUFs. These focal points would, by their nature, engage a spectrum of researchers from computer science, computational and experimental sciences, and data science in a coordinated fashion.**

## I. Background and Introduction

Nearly all fields of science are quickly evolving into a regime where data is the central focal point, where new opportunities for discovery await and where new impediments and obstacles abound. There has been significant activity in recent years to document those opportunities and challenges, with results appearing as individual workshop reports ranging from those focused on a particular regime of science like High Energy Physics [7] as well as more broadly across multiple areas of science [2], [3]. From these reports, others have generated documents describing cross-cutting issues that focus on data [8] and workflows [4]. Those cross-cutting issues are viewed through different lenses in strategic planning documents [1], [10].

This work takes these previous works a step further by focusing how one might structure a large-scale research and development effort that targets cross-cutting challenges and opportunities. One way to approach thinking about structuring such an effort, which is coordinated across a diverse set of technology and science areas, is to identify broad strategic objectives (§II), to define focal points for interdisciplinary projects that serve as the basis for implementation and deployment (§III), and to enumerate key technology areas where R&D is needed (§IV) to that are put into use for data-intensive science.

## II. Strategic Objectives

By *strategic objectives*, we mean desired outcomes, or a long-term goal. Such an outcome might be something relatively tangible, such as a new capability or some generally applicable technology, or it may be more ethereal, such as developing and retaining a scientific and data-savvy workforce. The following set of strategic objectives are themes that are present in many different workshop reports and strategic plans.

*Data archival, dissemination, preservation.* Science is generating increasing amounts of data, and also becoming increasingly oriented around hypothesis formation and testing using data [9]. A critical dependency for data-intensive science, where there are both producers and consumers of data, is the ability to store, share, find, and preserve data. The complexity of this objective quickly deepens when considering the fact that sharing/preserving data is not enough: one must also preserve the software used to work with the data. When considering the preservation of software, one must also consider the entire runtime environment and third-party tools needed by the software that makes the data usable.

*"Google for science"*. Many reports have identified the desire to have the same kind of straightforward access to data they enjoy in finding other types of information on the Internet. One key element of "finding things" is having a basis for search; for web pages, text-based search is the staple approach in broad use due to its relative simplicity and low barrier to entry. Scientifically meaningful search quickly becomes complicated: locating all datasets obtained by a given instrument during a particular period of time requires a significant amount of additional effort and technology. The provenance of such data – who created it, on what date, at what location – needs to be recorded as metadata, which in turn becomes the basis of search. The mechanics and lexicography of scientific search can be significantly more complex than the text-based search that has enabled the rapid growth of public internet in the past two decades.

*Key software infrastructure, tools, libraries for data-centric use.* It is reasonably well accepted that software plays a key role in both operating facilities and making use of data. A common theme concerns the brittle and unmaintainable nature of key software, which is often written by graduate students or postdoctoral researchers working in relative isolation, and that becomes unsupportable once they leave. Advances in the computational ecosystem — increasing cores/processor,

increasing imbalance between compute and I/O rates, evolution in APIs for accessing system and network services — catalyze software obsolescence. Many reports point to the need for a stable, sustainable collection of core software tools that are easily customizable and extensible for use at a particular facility, much in the same way that a web server core is customizable and extensible, and provides broad service and value. From an economic standpoint, this kind of approach helps to increase the lifespan of key, long-term investments. Equally important is the value of best practices in modern software engineering: revision control, bug tracking, multiple platform testing, and rigorous QA.

*Exemplar science use cases and key design and execution patterns.* These serve to capture the most important requirements and characteristics of the scientific data lifecycle. They provide the blueprints upon which software tools and data-centric services/facilities can be designed, architected, implemented, and operated. They also provide the basis for defining clear metrics for evaluating the success and value of data-centric infrastructure and programs. One aspect of these use cases that merits special attention is *data etymology*, or how data will be used now and in the future. There is a deep interplay between the concepts of data etymology, data lifecycle, and the definition of exemplar use cases. There is significant complexity lurking in these use cases when taking into account objectives like facilitating the linking of publications to data (and the software used to generate analysis of data), and vice versa, establishing the links/dependencies between a given datum and all derived publications.

*Fostering Growth and Communication.* Promoting the growth of robust and vibrant data-intensive science research activities requires thinking about and acting on the previously mentioned strategic objectives: success in one area will serve as an example to others, who in turn may be more likely to adopt new methods and approaches to achive similar successes. Communicating these successes to the public and funding agencies, in terms of new science or economics, will help to promote a deeper understanding of and appreciation for these activities.

*Better integration of facilities across science, network, computing.* The increasing complexity of the computational and technological landscape is itself an impediment to the use of such resources. A desired objective is to simplify and streamline use of increasingly complex and distributed compute and storage resources.

## III. Program Focal Points

This set of focal points is by no means complete, but does provide traction on a place to get started in a meaningful way.

*Data "galaxies" and "hamlets".* These are community or topically centered places where code and data live. They serve as repositories and archives for both data and the software tools/environment needed to work with data, and are designed and engineered to meet the needs of exemplar use cases. Examples of successful operational data galaxies include the Earth System Grid Federation [6] and the HEPData repository for high energy physics data [11]. In both these cases, the effort focuses on meeting the data archival and dissemination needs for a specific scientific community, each of which is global in nature.

*Pilot projects.* These efforts, which are framed by thinking about the data lifecycle, target specific types of capabilities that can be applied in a focused, and scope-limited way. One such example is the Neurodata Without Borders project [12], which aims to make databases of neurological data usable and accessible through a unified API for access. Another is the Center for Advanced Mathematics for Energy Research Applications (CAMERA) [5], which focuses on developing mathematical methods for data understanding problems across several light source projects.

## IV. Data Thrust Areas – Basis Vectors

While the problem space is complex, with many different interdependencies, the following set of basis vectors represents an orthogonalizion that is intended to cultivate core R&D, to assemble the right teams needed to create and operationalize working software for data-intensive science, and to be somewhat scalable with available funding.

*Computational data understanding.* Design, use, application of computationally based methods to derive insight from data. Includes algorithms/methods/tools for analysis, visualization, mathematics, machine learning, and so forth.

*Data management.* Methods, systems, approaches for managing (storing, finding, sharing) data. Includes archival, search, metadata, provenance, curation, digital libraries, along design and implementation of data models/format architectures and software tools.

*Computational Data Systems.* The customizable, extensible software infrastructure needed to build data pipelines. Includes things like workflow systems; the design/engineering of "boundary infrastructure" at facilities (computing and SUF) to accommodate data movement, resource discovery/provisioning/marshaling. A desired target is for these systems to make use of methods and tools that emerge from the data understanding and management thrusts, as well as related technologies that might emerge from elsewhere in the computational ecosystem.

*Data "carpentry" and "engineering".* Construction and implementation of user-facing tools, applications, data portals, and so forth. Includes integration of methods/tools from above categories, as well as a plan for ongoing support/maintenance. These may be project-focused to expand/shrink with available budget.

*Outreach, etymology, workforce development/training.* Spectrum of efforts that aim to cultivate a systematic understanding of how data is used now, and how it ought to be used in the future; proactive engagement with potential stakeholders and collaborators. Data bootcamps and "hack-a-thons", for making rapid progress in both training and in quickly standing up working, operational systems. Interactions with industry/other agencies/other offices.

## ACKNOWLEDGMENT

This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, through the grant "Towards Exascale: High Performance Visualization and Analytics," program manager Dr. Lucy Nowell.

## REFERENCES

[1] F. Berman, R. Rutenbar, H. Christensen, S. Davidson, D. Estrin, M. Franklin, B. Hailpern, M. Martonosi, P. Raghaven, V. Stodden, and A. Szalay. Realizing the Potential of Data Science, Dec. 2016. https://www.nsf.gov/cise/ac-data-science-report/CISEACDataScienceReport1.19.17.pdf, last accessed July 2017.

[2] E. W. Bethel and M. G. (eds.). Report of the DOE Workshop on Management, Analysis, and Visualization of Experimental and Observational data – The Convergence of Data and Computing. Berkeley, CA, USA, 94720, May 2016. http://science.energy.gov/~/media/ascr/pdf/programdocuments/docs/ascr-eod-workshop-2015-report_160524.pdf.

[3] E. W. Bethel, M. Greenwald, K. K. van Dam, M. Parashar, S. M. Wild, and H. S. Wiley. Management, Analysis, and Visualization of Experimental and Observational Data – The Convergence of Data and Computing. In *Proceedings of the 2016 IEEE 12th International Conference on eScience*, Baltimore, MD, USA, Oct. 2016. LBNL-1006061.

[4] E. Deelman and T. P. (editors). The Future of Scientific Workflows, Apr. 2015. http://science.energy.gov/~/media/ascr/pdf/programdocuments/docs/workflows_final_report.pdf.

[5] J. Donatelli, M. Haranczyk, A. Hexemer, H. Krishnan, X. Li, L. Lin, F. Maia, S. Marchesini, D. Parkinson, T. Perciano, et al. CAMERA: the center for advanced mathematics for energy research applications. *Synchrotron Radiation News*, 28(2):4–9, 2015.

[6] S. Fiore, M. Płciennik, C. Doutriaux, C. Palazzo, J. Boutte, T. Zok, D. Elia, M. Owsiak, A. D'Anca, Z. Shaheen, R. Bruno, M. Fargetta, M. Caballer, G. Moltó, I. Blanquer, R. Barbera, M. David, G. Donvito, D. N. Williams, V. Anantharaj, D. Salomoni, and G. Aloisio. Distributed and Cloud-Based Multi-Model Analytics Experiments on Large Volumes of Climate Change Data in the Earth System Grid Federation Ecosystem. In *IEEE International Conference Proceedings on Big Data*, Washington D.C., USA, Dec. 2016.

[7] S. Habib and R. R. (eds.). High energy physics exascale requirements review. Bethesda, MD, USA, June 2015. https://science.energy.gov/~/media/ascr/pdf/programdocuments/docs/DOE-ExascaleReport-HEP.pdf.

[8] B. Hendrickson and A. S. (editors). Data Crosscutting Requirements Review. Technical report, U. S. Department of Energy, Office of Science, Apr. 2013.

[9] T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, Washington, 2009.

[10] J. Kurose, K. Marzullo, and et al. The Federal Big Data Research and Development Strategic Plan, May 2016. https://www.nitrd.gov/PUBS/bigdatardstrategicplan.pdf, last accessed July 2017.

[11] E. Maguire, L. Heinrich, and G. Watt. HEPData: a repository for high energy physics data. In *22nd International Conference on Computing in High Energy and Nuclear Physics (CHEP 2016) San Francisco, CA, October 14-16, 2016*, 2017.

[12] J. Teeters, K. Godfrey, R. Young, C. Dang, C. Friedsam, B. Wark, H. Asari, S. Peron, N. Li, A. Peyrache, G. Denisov, J. Siegle, S. Olsen, C. Martin, M. Chun, S. Tripathy, T. Blanche, K. Harris, G. Buzsáki, C. Koch, M. Meister, K. Svoboda, and F. Sommer. Neurodata Without Borders: Creating a Common Data Format for Neurophysiology. *Neuron*, 88(4):629 – 634, 2015.