Communicating Scientific Data from the Present to the Future*

Position paper, NSF Funded Workshop "Research Data Lifecycle Management"
July 18 - July 20, 2011 Princeton University, Princeton, NJ

Herbert J. Bernstein[a], Michael J. Folk[b], Werner Benger[cg], Matthew T. Dougherty[d], Kevin W. Eliceiri[e], and Erik Schnetter[fc]

[a]Dept. of Mathematics and Computer Science, Dowling College, Oakdale, NY
[b]The HDF Group, Champaign, IL
[c]Center for Computation & Technology, Louisiana State University, Baton Rouge, LA
[d]National Center for Macromolecular Imaging, Baylor College of Medicine, Houston, TX
[e]Laboratory for Optical & Computational Instrumentation, University of Wisconsin, Madison, WI
[f]Perimeter Institute for Theoretical Physics, Waterloo, ON, Canada
[g]Department for Astro- and Particle Physics, University of Innsbruck, Innsbruck, Austria

"The goal [of the "Research Data Lifecycle Management" workshop] is to develop a combination of policy and financial frameworks that ensures maintenance of important data over time scales longer than the career of any individual investigator." We propose viewing this problem as one of designing a communication system from the present to the future. Thibodeau[1] focuses on cyberspace as the communications medium and the implications for the use cases: "the perspective of the consumers who wish to use digitally preserved assets." We propose to look at some of the implications of this model for the choices of the data frameworks to be used.

There is no simple way to reach the goal of maintaining high volumes of important data in heterogeneous environments over many decades. Not only must a cross-generational communication system reach from the present to the future half a century from now, but from many points in time in the near future to many points in time in the distant future. No single current medium, no single current file format specification is likely to be sufficiently robust and adaptable to survive without major changes over half a century. Science will advance and provide us with unanticipated data requirements. System errors, human errors, environmental insults, natural disasters, accidental and intentional disruptions will degrade or destroy data

*Work at Dowling College supported in part by grants from the DOE Office of Science and NIH NIGMS. Work at Louisiana State University supported in part by a grant from BP/The Gulf of Mexico Research Initiative.

[1]K. Thibodeau, "Digital Preservation, Communicating Across Cyberspace and Time" 1st International DPIF Symposium, Dresden, Germany, April 21-23, 2010

sets we had thought to be safely preserved, or irretrievably separate data from its metadata leaving the recipient of the communication with meaningless bits or interesting commentary on unknown data. Data and metadata must be sent coupled or, better, together.

As with any communication system, we need a design that tries for low error and loss rates but that accepts the reality of imperfect communications and provides sufficient error detection and redundancy to reduce the inevitable data losses to acceptable levels. No mechanism exists for the future recipient to send us a NAK (negative acknowledgement) requesting retransmission. We must plan on forward error correction, with multiple independent repositories in widely distributed locations with mechanisms for cross validation among those repositories.

We must encode the sender's version of the data as we see it now, send it through the system and provide the mechanisms for the receiver to decode the information and extract not just the original bit patterns but the original meaning as well. The more difficult or lossy or more poorly documented we make the initial encoding step, and the wider the variety of one-off experiment-specific data formats we use, the higher the probability that the original meaning will not be recoverable.

Even when a unique format is appropriate for short-term use in an experimental effort, long-term maintenance of the resulting data is unlikely to be successful without the use of well-documented interoperable data/metadata formats working within a common, extensible data/metadata framework with robust high performance software support. A highly promising candidate framework is HDF5[2]. There are other possibilities that need to be considered, and HDF5 itself can and should be further improved, refined and extended, but the widespread and increasing use of HDF5 argues for it as a unifying "hub" framework among whatever interoperable formats are adopted. As noted by Dougherty *et al.*[3]: "Hierarchical Data Format Version 5 (HDF5) is a generic scientific data format with supporting software. Introduced in 1998, it is the successor to the 1988 version, HDF4. NCSA (National Center for Supercomputing Applications) developed both formats for high-performance management of large heterogeneous scientific [datasets]. Designed to move data efficiently between secondary storage and memory, HDF5 translates across a variety of computing architectures. Through support from NASA (National Aeronautics and Space Administration), NSF (National Science Foundation), DOE (Department of Energy), and others, HDF5 continues to support international research. The HDF Group, a nonprofit spin-off from the University of Illinois, manages HDF5, reinforcing the long-term business commitment to maintain the format for purposes of archiving and performance. Because an HDF5 file can contain almost any collection of data entities in a single file, it has become the format of choice for organizing heterogeneous collections consisting of very large and complex datasets. HDF5 is used for some of the largest scientific data collections, such as the NASA Earth Observation Systems petabyte repository of earth science data. In 2008, netCDF (network Common Data Form)[4] began using HDF5, bringing in the atmospheric and

[2]http://www.hdfgroup.org

[3]M. T. Dougherty, M. J. Folk, E. Zadok, H. J Bernstein, F. C. Bernstein, K. W. Eliceiri, W. Benger, C. Best, "Unifying biological image formats with HDF5," CACM **52**:10, 2009, pp. 42 – 47.

[4]http://www.unidata.ucar.edu/software/netcdf/.

climate communities. HDF5 also supports the neutron and X-ray communities for instrument data acquisition. Recently, MATLAB implemented HDF5 as its primary storage format. Soon HDF5 will formally be adopted by the International Organization for Standardization (ISO) as part of specification 10303 (STEP, Standard for the Exchange of Product model data). Also of note is the creation of BioHDF[5] for organizing rapidly growing genomics data volumes."

Since the publication of this paper, we have witnessed continued adoption of HDF5 as a replacement for or alternative to legacy data formats. As data management policy makers recognize the challenges of preserving data for the long term, especially challenges involving complexity, scalability, and the proliferation of one-off formats, HDF5 offers an increasingly attractive option. Just as PDF provides a universal exchange standard for page images, and XML offers an powerful exchange format for text markup, HDF5 offers a comprehensive, general, binary and meta-data friendly framework for scientific data exchange, while being open source.

If we focus on bringing as many existing data formats as possible under the HDF framework, and carefully organize and record the metadata ontologies, then we can use and extend HDF5 as the necessary sender's encoding for our communication system to the future. Without adopting such a coherent framework, we face the problems noted by Millard *et al.*: "data from imaging, multiplex biochemistry, flow cytometry and other cell- and tissue-based assays usually reside in loosely organized files of poorly documented provenance"[6]. The approach they chose was "an adaptive approach to managing experimental data based on semantically typed data hypercubes (SDCubes) that combine hierarchical data format 5 (HDF5) and extensible markup language (XML) file types."

Rising data volumes and the continual addition of new data and metadata types combined with short time-horizon funding policies make it very difficult to sustain the effort needed to ensure maintenance of such data, data/metadata frameworks and supporting software on long time scales. Neither individual research groups nor even major experimental facilities are likely to have sufficient resources, inclination or ambitions, and most importantly the time, to solve these problems in isolation. The subtle differences between inexpensive solutions for recording data for the short term and the more expensive robust solutions needed to communicate reliably with the future combined with tight budgets for doing "real science" create the temptation in funding decisions to settle for the inexpensive solutions. That would be a mistake. Gathering all the data and all the metadata in a coherent, self-contained message to the future by use of HDF5 greatly increases the chances that future recipients will be able to understand and use what we send. If data is to survive reliably into the future, to arrive at the end of multi-decade communication systems reliably, we need to encourage use of a common data framework such as HDF5 as early in the life of the data as possible and we need to create the matching multi-decade funding model to keep the system not just marginally operational but growing and healthy.

[5] www.geospiza.com/research/biohdf/

[6] B. L. Millard, M. Niepel, M., M. P. Menden, J. L. Muhlich, P. K. Sorger, "Adaptive informatics for multifactorial and high-content biological data," Nature Methods, 2011, doi:10.1038/nmeth.1600