Scientific Reproducibility via Rapid Algorithm Integration, Intelligent Data Movement and Automated Text and Metadata Extraction

Chris A. Mattmann[1,2]

[1]Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109 USA
chris.a.mattmann@nasa.gov

[2]Computer Science Department
University of Southern California
Los Angeles, CA 90089 USA
mattmann@usc.edu

1. Introduction and Motivation

We propose a framework for scientific reproducibility with three predominant concerns: (1) the rapid, and unobtrusive integration of scientific algorithms; (2) the intelligent and automated selection of data movement technologies and high accuracy and high fidelity approaches for the selection process; and finally (3) automated and rapid extraction of text and metadata from *any* type of file.

Based on our prior work [1, 2] constructing the next generation of Earth science remote sensing data systems for the National Aeronautics and Space Administration (NASA), we have found the above three concerns to be under-represented and not uniformly dealt with in software. Often computer scientists and data system developers ask scientists to rewrite their algorithms in order to integrate them into a data processing system – the request is often something along the lines of "to make this algorithm faster we need to rewrite it as a *blank*" where *blank* has evolved from "middleware service" such as CORBA/Java RMI; to *web service* such as SOAP and/or REST; and most recently "map reduce job" such as a Hadoop job, etc.

Rewriting the algorithm to make it faster or more amenable to modern data system frameworks removes the stewardship of the algorithm from scientists and their lineage (postdocs, research assistants, other professors, etc.) and brings into the realm of computer scientists who may not be properly trained in the underlying discipline necessary to maintain it.

In the area of data movement, there are vast differences among the available data movement technologies which our team has chronicled over the last decade (2005 – 2015) [3, 4] and these differences in scalability, dependability, reliability and efficiency *do* make a difference and it *does* matter what data movement technology you select to satisfy a data distribution scenario. Often these selections are not reproducibility and they are housed in the minds of "gurus" or architects who make the selections of data movement based on "pet technologies" at best backed by anecdotal evidence.

Finally, some estimate that there are between 18,000 to over 51,000 different kinds of file formats on the Internet today[1]. A more fine-grained classification by the Internet Assigned Numbers Authority (IANA) estimates around 1,200 file formats [5], with richly curated metadata including parent/sub-type, digital file signatures and precedence order ("MIME magic"); regular expressions on filename and/or directory structure, and XML root schema detection information for XML-based file types. Scientific data files fall into this hierarchy, and it is our position that there will never be *a single "god" file format* and in reality we need better and more automated approaches to extracting meaning from these files for scientific data. Meaning can be in the form of text and metadata with the text representing the actual "science" or "bits" and the metadata being information about those "bits", e.g., "what instrument took the data"; "when was it captured", etc. Finally,

[1] http://fileext.com

in this distributed collaborative world, the detection of a file's *language* is also necessary as it drives the extraction of text and metadata. Tools and approaches are needed to make these processes faster, more rapid, and repeatable.

2. Towards A Framework for Reproducibility

Our work to date has focused on deriving solutions to the above motivational problems with respect to scientific data processing and it is our position that solutions to the above challenges will make scientific data more reproducible. By integrating science algorithms the way they are written into data systems, we provide a *repeatable* and *reproducible* process for executing the algorithm the same way the scientists did, at scale, simply mirroring or automating the provision of an algorithms input; managing its execution; recording of its outputs for downstream algorithms and future processing and ingestion activities. In turn, by allowing for automated approaches to *reproduce* the selection of data movement technologies made by architects and experts in data systems, we can with confidence and high accuracy continue those decisions, and have scientific data systems make them on their own, without human intervention. Finally, by automatically and rapidly extracting text and metadata and language from *any type of file*, we provide a homogeneous mechanism for arbitrarily large and complex file types (which continue growing), and at the same time, allow for their *repeatable* detection by human or software, and the *repeatable* and *reproducible* extraction of the meaning and information from those files which represent the science as captured in-situ; by airborne instrument, by spaceborne instrument, or by simulation.

What's more is that: *technologies exist today that implement solutions to the above challenges!* Rather than creating new technologies that lay similar plumbing and ground-work even before providing a capability to deal with the reproducibility problem, we are focusing our efforts on expanding the Apache OODT technology [1, 2] which directly deals with issues 1-2 (algorithm integration and intelligent data movement); and Apache Tika [5] which deals with issue 3. These are open source, permissively licensed software products with strong lineage over the last decade, and with strong participation from the software, and science communities within the government, academia, and within private industries. Apache OODT provides a file management component for tracking and recording file locations, their MIME types (derived from Tika), and for cataloging and indexing metadata about files, and finally for data movement. OODT's workflow manager models control flow and data flow (allowing for the recording of provenance, when used in concert with the file management component) and a resource management component for modeling the disk usage, CPU, language and runtime requirements (Python, IDL, Fortran, etc.) for science algorithms. Three client frameworks, one for remote file acquisition ("Push/Pull"), one for automated crawling and metadata extraction and file ingestion ("Crawler"), and finally an algorithm wrapper ("CAS-PGE") for unobtrusively integrating algorithms into the data management system.

In turn, Apache Tika works in concert with Apache OODT. Tika fully implements the IANA MIME registry, it is used in OODT file management to automatically deduce file types; used in ingestion to classify which metadata extractor should be called on a particular file, and used in remote file acquisition to decide what files to pull down, and finally used in algorithm integration to determine what files should be ingested where, and what metadata should be extracted and recorded. The combination of these technologies represents a powerful toolkit for scientific reproducibility and they are currently in use at NASA, NIH, DARPA, NSF, private industry (e.g., Google, Microsoft), academia, and they are used across a number of scientific disciplines including Earth science, Planetary science,

radio astronomy, and bioinformatics to name a few.

3. Future Work and Conclusions

The author's *Vision for Data Science* [6] published in Nature magazine's January 24th, 2013 issue outlined the necessity for the above advancements and pinned much of the hope for progress in the next decade on open source, on people power, and ultimately on the response of government agencies and funding agencies to invest in technologies including OODT, Tika and other efforts.

As a framework for reproducibility in science, we see these technologies as current exemplars, and as mechanisms to prepare for the next generation of systems. It is my sincere hope that progress is made in each of the areas of rapid and unobtrusive algorithm integration, intelligent data movement, and in automated identification of, and extraction of text and metadata from any type of file. Regardless of specific technology choices, these three areas represent foci that require advancement in order to allow for reproducible science.

References

1. C. Mattmann, D. Freeborn, D. Crichton, B. Foster, A. Hart, D. Woollard, S. Hardman, P. Ramirez, S. Kelly, A. Y. Chang, C. E. Miller. A Reusable Process Control System Framework for the Orbiting Carbon Observatory and NPP Sounder PEATE missions. In *Proceedings of the 3rd IEEE Intl Conference on Space Mission Challenges for Information Technology (SMC-IT 2009)*, pp. 165-172, July 19 - 23, 2009.
2. C. Mattmann, D. Crichton, N. Medvidovic and S. Hughes. A Software Architecture-Based Framework for Highly Distributed and Data Intensive Scientific Applications. In *Proceedings of the 28th International Conference on Software Engineering (ICSE06), Software Engineering Achievements Track*, pp. 721-730, Shanghai, China, May 20th-28th, 2006.
3. C. Mattmann, D. Crichton, J. S. Hughes, S. Kelly, S. Hardman, R. Joyner and P. Ramirez. A Classification and Evaluation of Data Movement Technologies for the Delivery of Highly Voluminous Scientific Data Products. In *Proceedings of the NASA/IEEE Conference on Mass Storage Systems and Technologies (MSST2006)*, pp. 131-135, College Park, Maryland, May 15-18, 2006.
4. J. Tran, L. Cinquini, C. Mattmann, P. Zimdars, D. Cuddy, K. Leung, O. Kwoun, D. Crichton and D. Freeborn. Evaluating Cloud Computing in the NASA DESDynI Ground Data System. In *Proceedings of the ICSE 2011 Workshop on Software Engineering for Cloud Computing - SECLOUD*, Honolulu, HI, May 22, 2011.
5. C. Mattmann and J. Zitting. *Tika in Action*. 256 pages. New York: Manning Publications, November 2011. ISBN: 9781935182856.
6. C. Mattmann. A vision for data science. *Nature*, Vol. 493, No. 7433, pp. 473-475, January 24, 2013.