



Global Information Industry Center

at the School of International Relations and Pacific Studies

University of California, San Diego

Climate Change at MIT The Scientific Data Flood: A Case Study of “How Much Information?”

Stuart Madnick, John Norris Maguire Professor of Information Technology, MIT Sloan School of Management & Professor of Engineering Systems, MIT School of Engineering

MacKenzie Smith, Associate Director of Technology, MIT Libraries

Kate Clopeck, Masters of Science, Technology and Policy Program, MIT

June 2009

Abstract:

This case study provides an early look into the data growth projections for the embryonic Earth System Initiative (ESI) at MIT. ESI is an umbrella initiative facilitating the development of large scale research efforts in Earth system science and engineering. The case study focuses on the first project initiated under ESI, the Darwin Project. Darwin is focused on the large scale modeling of the physical and biological processes in the oceans. Numerical models produce approximately 90% of the data generated by ESI scientists; the remaining 10% of data is observational data recorded by NASA satellites or NOAA oceanographers. ESI's 25 researchers generated approximately 200 terabytes of data in the last year, primarily from high resolution calculations that model the physical and biological processes occurring in a specific sector of the ocean. For example, one high resolution calculation will occupy from one to two months of machine time and produce 60 terabytes of data. Increases in the Laboratory's computing power and storage capacity have helped drive data to increase by a factor of 100 in five years. At this rate, in the next five years ESI data production could reach 20 petabytes annually. The case concludes with notes on data retention policies and metadata creation and use. Other papers examine other labs at MIT.

- Executive Briefing:** a summary of one or more research projects with preliminary findings for a non-academic audience.
- Research Report:** a completed report drawing on one of more research projects that presents study data, findings and management implications.
- Case Study:** an in-depth description of a firm's approach to an information management issue.
- Research Article:** an academic research paper with sections on hypotheses tested, methods and data, analysis, findings and references.

Contents

Background3

Data Generation3

Metadata4

Data Retention.....4

Data Sharing and Reuse4

Key Trends and Indicators for Data Growth4

About the HMI? Program5

Acknowledgements5

Climate Change at MIT The Scientific Data Flood: A Case Study of “How Much Information?”

Stuart Madnick, MacKenzie Smith, and
Kate Clopeck, MIT

June 2009

Background

In recent years, our society has become more aware of the delicate balance of the Earth system, and has devoted much time and energy to debates over how best to ensure a sustainable future for the planet. The Earth System Initiative (ESI) is predicated on the notion that, to be meaningful, these debates must be informed by reliable scientific data regarding the evolution and current state of our planet. ESI scientists and engineers marshal their efforts around four broad research themes:

- System Characterization
- System Organization
- Evolutionary Processes
- Human Impacts

The Earth System Initiative facilitates the development of large-scale research efforts in key areas of Earth system science and engineering. In December 2006, the Darwin Project, the first example of such an undertaking, was launched.

The Darwin Project is an ESI initiative to advance the development and application of novel models of marine microbes and microbial communities, identifying the relationships of individuals and communities to their environment, connecting cellular-scale processes to global microbial community structure.

For this case study, three scientists in MIT’s Department of Earth, Atmospheric and Planetary Science were interviewed. These research scientists focus on the large scale modeling of the physical and biological processes in the global oceans. To do

this, they build large numerical simulations that are constrained with observational ocean data.

Data Generation

Numerical models produce about 90% of the data generated by scientists in the Earth Science Initiative within the MIT Department of Earth, Atmospheric and Planetary Science, and particularly for the Darwin Project. The remaining 10% is observational data that is recorded by NASA satellites or oceanographers. There are about 20-25 people who work on the Darwin Project at MIT. They post all of their numerical models on the project website for others in the field to download and use. The primary research estimates that there are another two hundred people around the world who download these models to use in a variety of different ways. The group at MIT communicates with these other scientists through the web, and often collaborates on projects.

Over the last year, this project has generated about 200 terabytes of data. The majority of the data is from high resolution calculations that model the processes (both physical and biological) occurring in a certain area of the ocean. One high-resolution calculation will run for about 1-2 months and will produce about 60 terabytes of data. However, the amount of data produced is very dependent on the specific processes that are being modeled.

The amount of data produced by these models has increased over time. However, this increase is largely due to better storage technologies, not changes in the models. Five years ago, the research group was “theoretically capable” of generating just as much data as today, but they did not have enough storage to handle the size of the data files. As the storage hardware continues to improve and become less expensive, the amount of data that is generated by improvement to the resolution of the models will continue to increase. According to the primary research scientist, mathematically speaking, there is “no upper bound.” Since 2003, the data generated by these researchers has increased by a factor of 100. In the next five years they predict that it will increase by another factor of 100 as the computer infrastructure continues to become less expensive and more

widespread. Based on these predictions, the group could produce about 20 petabytes of data in 2014.

Metadata

Metadata for this project's research includes the grid (area of the ocean) that the model is using, the physical fields that are produced by the model configuration, and the biological fields being modeled. Like other areas of scientific research, the amount of metadata is very small in comparison to the experimental or model data produced, but is critical to making use of the primary data.

Data Retention

The data generated by this group of researchers is stored on their computational cluster's file system. The cluster is a collection of 750 hard drives, with a certain amount of redundancy in case a drive fails. Currently, they do not have the capacity to do a redundant back up of the entire cluster. Instead, they use national facilities to back up important data. One facility that they frequently use is the NASA-Ames Lab, which has an archive system to which they can transfer data over the network. It is relatively simple (although sometimes time consuming) to re-run a model, so data can also be reproduced if it is lost.

The storage capacity of his team's cluster is 500 TB. They have had this storage technology for about 18 months. Before purchasing this hardware, the team was storing all of their data the NASA-Ames facility, which had 100 terabytes of storage available for them to use, however the transfer time was very slow (approximately 1 terabyte per day).

This research group saves the source code for all of their model configurations, but has no other data retention policies. The individual scientist running a model will usually decide what data to keep, and what to delete. Five years after a project is completed, only about 10% of the data from that project is still available. The research group employs a computer system administrator to manage the computational cluster, however he only informs the researchers when they are close to their storage capacity limit. He does not make any decisions about

data retention.

Data Sharing and Reuse

As mentioned earlier, this research team posts all of their computational models online and is open to sharing their data with other researchers in this field. They also use the NASA-Ames facility to archive important data. However, the NASA-Ames archive is only shared with immediate collaborators and is not publicly accessible.

In addition to sharing their data with other researchers, the research group is constantly re-using their own data, mainly to re-run models to test for reproducibility of results, or to re-analyze data with different models.

Key Trends and Indicators for Data Growth

Several key trends and indicators for data growth can be identified for the Earth Science Initiative on Climate Change:

1. The amount of data currently generated is more than 200 terabytes per year (based on one large project within the Initiative).
2. Increases in computing power and storage capacity have caused the amount of data generated to increase by a factor of 100 over the past five years. If hardware trends continue, in the next five years data production could reach 20 petabytes annually.
3. Like other scientific areas, although metadata is important to research projects, the amounts generated are not significant.
4. Data retention decisions are up to the individual researchers and retention is not a major concern today. Retention is constrained by the volume of data produced, and is facilitated by use of national data archiving facilities (managed by NASA, in this case).
5. Data sharing and reuse are commonplace in Climate Change research, and would be facilitated by improved data storage and archiving capabilities.

About the HMI? Program

The How Much Information? (HMI?) research program is a multi-discipline, multi-university project, formed to investigate the nature of data and information generated and used by individuals and enterprises. The program is sponsored by seven companies, including AT&T, Cisco, IBM, Intel, LSI, Oracle, and Seagate, and involves multiple research universities. The Principal Investigator is Prof. Roger Bohn and the Research Director is Dr. James Short, at UC San Diego's Global Information Industry Center (<http://giic.ucsd.edu>). Founded in 1960, the University of California, San Diego is one of the nation's most accomplished research universities, widely acknowledged for its local impact, national influence and global reach.

Acknowledgements

This case study is the product of industry and university collaboration in applied research. We are grateful for the support of our industry partners, sponsor liaisons, university research partners, and administrative staff at the University of California, San Diego.

Financial support for HMI? research and the Global Information Industry Center is gratefully acknowledged. Our sponsors are:

AT&T

Cisco Systems

IBM

Intel

LSI

Oracle

Seagate Technology

Additional support was provided by the Alfred P. Sloan Foundation of New York.

Questions about this research may be addressed to the Global Information Industry Center at the School of International Relations and Pacific Studies, UC San Diego:

Roger Bohn, Principal Investigator, rbohn@ucsd.edu

Jim Short, Research Director, jshort@ucsd.edu

Pepper Lane, Program Coordinator, pelane@ucsd.edu