S.E.N.S.O.R. – Applying Modern Software and Data Management Practices to Climate Research

Michael J. McMahon Jr.
Department of Computer Science & Engineering
University of Nevada, Reno
Reno, NV, 89557, USA
mcmahon@cse.unr.edu

Frederick C. Harris, Jr.
Department of Computer Science & Engineering
University of Nevada, Reno
Reno, NV, 89557, USA
Fred.Harris@cse.unr.edu

Sergiu M. Dascalu
Department of Computer Science & Engineering
University of Nevada, Reno
Reno, NV, 89557, USA
dascalus@cse.unr.edu

Scotty Strachan
Department of Geography
University of Nevada, Reno
Reno, NV, 89557, USA
scotty@dayhike.net

Abstract

Climate research projects traditionally handle data collection and management activities in an isolated manner, targeting and satisfying only the needs of project researcher(s). While this model successfully allows researchers to meet their immediate data needs for a given project, it has the unfortunate effect of creating data sets that exist in a variety of formats, implement a widely varying number of undocumented data collection and management "standards," and are often designed to be consumed by a limited collection of highly-specialized software. The end result is the creation of massive numbers of data sets across numerous research projects, the majority of which cannot be directly compared / correlated or easily consumed by modern, powerful, industry-standard data processing applications, rendering this body of past research largely unusable by current or future researchers.

Advances in modern research hardware and software have provided opportunities to address the need to create consistent, reusable, managed data sets from research projects. These managed data sets both rely upon and support the use of high-performance data processing technologies, such as relational databases and high-speed TCP/IP networks, to facilitate the dynamic creation of quality, accurate, variable-format, highly-selectable data sets across any number of research projects for researchers across all disciplines. These systems result in data sets across projects that implement standardized data curation, collection, and management processes, allowing comparison and utilization by current and future researchers with minimal manipulative effort.

The migration from the traditional, isolated approach of data management to the modern implementation of data standards and practices is extremely complicated, requiring the application of new hardware, software, programming, and cross-project policies to resolve a variety of issues. This paper explores some of the challenges involved in migrating climate research to this new paradigm, using the experiences gained implementing the NSF-funded Nevada Climate Change Project to illustrate how these problems can be overcome, and, in the end, provide features and facilities to researchers that were difficult – if not impossible – to obtain under the old paradigm.

***Keywords*: Climate research, climate monitoring, data standards, high-speed network, data management, data curation.**

1 INTRODUCTION

Climate research involves the utilization of several common components and / or stages to link climate measurements to data consumers (i.e. researchers, modelers, agencies, etc.). The important, high-level components involved in this process can be categorized as follows (illustrated in **Figure 1**): data collection, transmission, storage, processing, dissemination, and utilization.

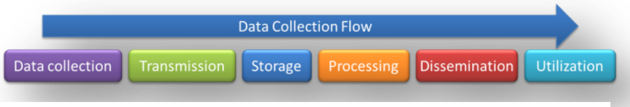

Figure 1. The general process involved in connecting researchers with climate measurement data.

The data collection component addresses the basic need to collect climate measurements from sensors. In a common climate monitoring deployment, numerous sensors are deployed within a geospatial region. These sensors are each connected to a data logger or other aggregation device that transforms the varying electrical signals received into numerical data. The logger itself is configured to take measurements of varying types (e.g. average, maximum, minimum, etc.) over various intervals (e.g. 1-minute, 10-minute, etc.) in specific units (e.g. m/s, meters, etc.) from the sensors, as determined by its program. The logger, for all intents and purposes, is the point-of-origin for raw climate measurements, often acting as local storage for data collected over any period of time (depending upon storage capacity and / or constraints).

The transmission component encompasses the movement of data collected by the logger to a permanent storage location or archive. Abstractly, transmission may include anything from high-speed two-way communication to one-way satellite communication to the physical act of an on-site hardware technician downloading data from the logger and placing it on a server when they return to their office. The precise transmission method employed can affect data accuracy, availability, and long-term verifiability, meaning that this component cannot necessarily be considered transparent to the process.

The storage component represents the repository to which raw climate measurements are transported for later processing and / or archival. Ideally, this component represents storage on a server using fault-tolerant technologies, but it may equally well represent a simple desktop computer or flash drive.

The processing component encompasses the application of various algorithms, steps, or tests on raw data to determine various properties of that data and / or prepare it for use. This may include activities such as quality assurance (QA) / quality control (QC) processing to determine whether transmission errors occurred or whether sensor measurements are within reasonable bounds, data file verification, and / or the import of data into a data store (e.g. database) for later use.

Dissemination specifically addresses the means and mechanisms by which data are made available to researchers and consumers of data. These mechanisms may vary widely depending upon the technologies available to the project members, the requirements of the project, and even the preferences of the project members. Common dissemination methods include: e-mail, local file shares, data files posted to FTP and / or web servers, and web services. Features such as data selection, projection, and format are typically strongly associated with this component (and often supported by the processing component), though they can usually be performed by the end-user.

Finally, utilization of the data is the terminal state of the process, at which a system / organization / individual obtains the data they seek. The mechanism by which they obtained the data is dictated by the dissemination component; further operations, such as selection, projection, and formatting, may occur in this stage if they are not available as a part of the dissemination component. The use of data from this point forward is at the discretion of the consumer.

It is the flow of information through these various components that characterizes the process of collecting and utilizing climate data, thus enabling climate research. While the sequence of components itself has not changed significantly over several decades, the technologies available to implement and connect each of these components has changed dramatically. However, the incorporation and implementation of a particular technology or practice is not easily isolated to changing one component. In fact, implementing most new policies or technologies requires changes to the technologies and policies of most (if not all) components, especially when attempting to implement a modern feature or policy within a traditional data collection model.

For example, the implementation of data curation and management practices [1] – which, in essence, involve the maintenance and collection of quality climate data – involves more than regular data backups or system administration. Maintaining quality data requires that information regarding the sensors used to make measurements be stored, along with maintenance information about monitoring hardware, the personnel involved with the research and their roles, and even documentation relating to logger program changes and any processes that were run on the data. The collection of such error-free data also requires the use of new hardware, namely the use of error-correcting TCP/IP-based networks to transfer data. A change that initially seemed relegated to the storage or processing components has quickly spread to nearly every component. This pattern of cascading changes is also seen when attempting to implement features / policies such as unit conversions, data correlation / comparison, providing multiple output formats, and providing long-term data to multiple audiences both within and outside of the originating research project.

Various advances in modern hardware and software are poised to enable researchers to meet the new data management requirements imposed upon them, both by funding agencies [2] and good, transparent scientific practices [3]. With the introduction of the concept of explicit data management to climate research has come the overarching goal of providing quality project data to current and future researchers in a long-term, sustainable manner.

The NSF-funded Nevada Climate Change Project, being created with a mandate to both answer specific science questions and create a reusable research infrastructure (encompassing hardware, software, and networking) for long-term scientific data collection, has been directly challenged with architecting, designing, and implementing a modern data collection model. This new model makes use of new networking systems, hardware, software, and policies to address both the short- and long-term data management needs of the project. As such, a great deal of effort has been expended migrating from the traditional data collection and management model to the modern one implemented to meet these goals. The result has been the creation of the Spatial Engine for Nevada Scientific Observational Results (SENSOR) system to support advanced data collection activities for the Nevada Climate Change Project [4] and future research.

Section 2 of this paper describes the traditional data management and collection model used by climate researchers, as well as some of its limitations. Section 3 details the implementation of modern software to address the short- and long-term data management needs of the Nevada Climate Change Project, focusing on the challenges faced, resultant policies, and best-practices that have resolved those challenges. The paper concludes with a summary of those policies and best-practices, including their advantages.

2 TRADITIONAL CLIMATE RESEARCH

Figure 2 illustrates the traditional interaction of climate research personnel in the data collection process:

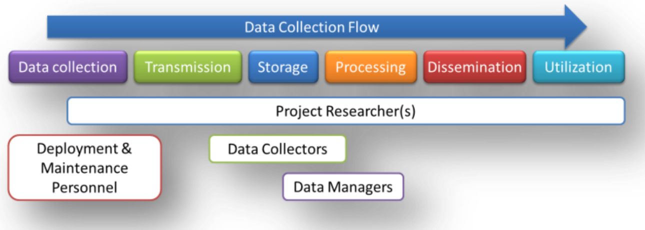

Figure 2. Personnel roles in the traditional data collection model.

As can be seen, the data collection process is researcher-centric. That is, the researcher is involved with and controlling nearly every aspect of the data collection process, with few others involved. Of the few that are, none are directly concerned with managing data over any significant part of the process or period of time, so much as simply collecting that data and ensuring that the hardware generating the data continues to operate correctly.

This model has proven extremely effective at enabling researchers to collect the data they require to answer their specific research questions, thus fulfilling the requirements of their funded project. However, this model has several shortcomings:

- There is no explicit short- or long-term data management concern, only concern for answering the research questions of one project [5] [6].
- The data collected conforms to the arbitrary desires of the researchers (e.g. units, collection intervals, formats, etc.). These parameters may not be explicitly stated in output files, and may not suit the needs of future researchers.
- Implicit information about the data is not recorded, such as QA / QC processes performed on data, filtering, how data was collected, or whether any data corrections were made.
- The future availability of the data (beyond the timeframe of the project) is uncertain.
- The use of the data by researchers other than those configuring the data collection parameters will likely require filtering and unit conversion.
- There is frequently no explicit clock synchronization with data loggers, making the correlation of data between loggers or other projects inaccurate. Further, the time zone in which measurements were made is not documented or recorded, rendering long-term use difficult.
- The data sets generated are frequently in a single format, and consist of a collection of flat files.
- The verifiability of the data is limited, largely due to the lack of explicitly recorded processing and collection practices [7].
- Data systems (e.g. databases) cannot easily be built around the inconsistent data sets generated under this model.

In essence, many of the long-term issues associated with this model revolve around the fact that a great deal of information about the data (e.g. temporal information, maintenance, corrections, etc.) is simply “known” by the researchers within a project. This information is lost over time, rendering the project data less useful to future researchers that wish to utilize the project data.

3 MODERN CLIMATE RESEARCH

Data management is a broad term that includes a variety of data-related activities, such as data acquisition, curation, QA / QC processing, security, and the creation of policies and procedures that track and guide the evolution of that data. Data management is meant to formally establish and codify the manner in which data is collected, transferred, stored, processed, and made available to a broad variety of users, such that the quality and meaningfulness of the data is best retained at all times. In essence, applying data management policies to a project means that a documented plan for efficiently and securely collecting, maintaining, and delivering data to any interested entity is developed, implemented, and enforced.

The key difference between the traditional climate data collection model and the modern model implemented by the SENSOR system is the inclusion of explicit data management activities designed to ensure that data is available for both short- and long-term use by all researchers, not just those of the project. The personnel involved in the data collection process under this model are illustrated in **Figure 3**.

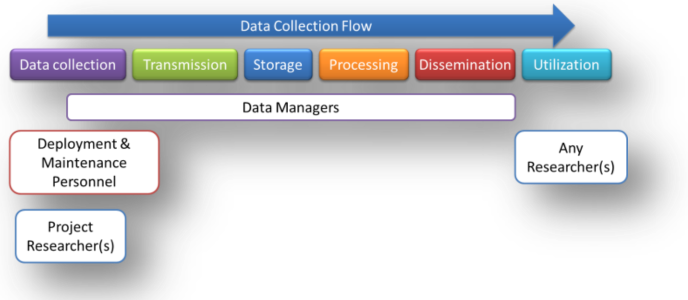

Figure 3. Personnel involved in the modern data collection model.

As shown, the data collection process is no longer centered upon researchers. In this new model, project researchers influence hardware configuration to ensure their research needs are met, while the collected data is made available to all researchers, both within and outside the project. The process of collecting, storing, processing, and disseminating data becomes the responsibility of the data management personnel. Further, the influence of data management is extended beyond a small component to nearly all components in the data collection model in order to ensure data quality throughout the process.

With the focus of the data collection model shifting from providing data to a select group of researchers to answer science questions for one project, to providing data to many researchers across various projects to solve any number of science questions, several high-level requirements appear:

- Support the selection, projection, and formatting of data in any manner required by a data consumer.
- Incorporate data curation practices into the data collection process, requiring the explicit collection of additional hardware information and metadata.
- Minimize or eliminate the potential for data corruption.
- Support the temporal correlation of data.
- Support the retrieval of data in arbitrary units.

Fulfillment of these requirements necessitates a significant number of changes across all components of the data collection model. Because these changes transcend any one component, they have been organized into these categories: data standards, hardware & networking, software, and policies. The following sub-sections describe the data management challenges associated with – and addressed by – each category, as well as the specific implementation recommendations developed as a part of the implementation of the SENSOR system.

3.1 Software

Supporting the data-intensive requirements of data management and curation practices requires the ability to efficiently search, select, and project potentially trillions of collected data points. The traditional approach of collecting flat files for each project is clearly inappropriate and not amenable to this kind of search. A relational, geospatially-enabled database is therefore employed to provide high-performance, advanced data search and organization features. For the SENSOR system, a Microsoft SQL Server 2008 R2 database engine is employed, which natively supports complex OGC-standard geospatial constructs.

The implementation of a database allows the storage of climate measurement values, sensor information, maintenance information, and virtually all additional pieces of information relating to how the climate data was collected. In the case of the SENSOR system, this metadata for the measurements provides an extensive knowledge base upon which long-term questions relating to the data can be answered, and which help verify the authenticity and accuracy of the data collected. For example, tracking the maintenance history of every piece of hardware deployed helps verify that the data collected during a given period were accurate or potentially erroneous, i.e. whether they were part of a trend of erroneous values or associated with a sensor that was soon after replaced. By tracking this type of extended information, the long-term value of the data to future and current researchers is retained, allowing a very comprehensive analysis at any current or future time.

The selection and implementation of a database schema is a thorny issue that can affect the data and functionality of a system. As an example, if one implements a schema that is modeled after a metadata standard (e.g. FGDC [8]), they run the risk of collecting an insufficient amount of information to satisfy the requirements of subsequent standards (e.g. ISO 19115:2003 [9]), thus restricting future extensibility. In implementing the SENSOR system, we found that a standards-neutral database schema that was designed around the data curation philosophy of collecting all possibly relevant data was the most appropriate. Following this methodology, we collect all potentially-relevant data and metadata with the knowledge that we can remove extraneous data later, but we cannot generate missing data. Further, by implementing a standards-neutral schema, we were free to optimize the performance of the database without data structure constraints, yet we retained the ability to reorganize data for output into any format.

3.2 Hardware & Networking

Under the traditional data collection approach, most monitoring stations are accessed via remote one-way links (i.e. GOES [10]), other low-bandwidth telemetry, or direct on-site download. While these methods are less complicated in terms of infrastructure and often the most inexpensive, they introduce the potential for transmission errors (one-way transmissions lack error detection and correction) and ultimately constrain the amount of data that can be collected and transmitted (i.e. via low-bandwidth). Ideally, high-speed, two-way communication links would be established between central data systems and data loggers / sites to support current and future research needs, while eliminating transmission errors.

The SENSOR network utilizes the secure virtual private network (VPN) of the Nevada Seismological Laboratory (NSL) [11] to provide connectivity to deployed data loggers. This network uses TCP/IP communication to provide multi-megabit connectivity to all sites, while simultaneously eliminating transmission errors (TCP/IP networks detect and correct transmission errors) and securing the communication link. As a large number of the sites were not near NSL monitoring stations, a number of high-speed TCP/IP radios (AFAR Communications Pulsar Long Range Industrial Wireless Ethernet Bridge) were deployed at the sites and on several peaks to extend the high-speed communication network that NSL provides to our stations. This network is capable of supporting high volumes of climate data collection (1-minute or less intervals), as well as live web camera streams. Not only do such modern site deployments meet data quality requirements, but they provide unique opportunities for future research.

Monitoring deployments for the Nevada Climate Change Project consist of multiple Campbell Scientific data loggers (CR1000 and CR3000) and more-than 40 physical sensors, recording over 100 data values at any given moment. The uniform use of Campbell Scientific data loggers is not a requirement of the SENSOR system, as it is designed to be extended to interface with any system, but a choice of consistency that has simplified the initial development and deployment efforts of the project.

To utilize the high-speed communication at each monitoring deployments, the data loggers have been outfitted with TCP/IP networking adapters (Campbell Scientific NL100). Combining these devices, the high-speed VPN connectivity provided by NSL is extended to each monitoring deployment via the TCP/IP radios, whose bandwidth is shared amongst site devices – including the data loggers via their standard TCP/IP interface – using an off-the-shelf networking switch. This arrangement (depicted in **Figure 4**) facilitates two-way communication with the loggers for programming and maintenance changes, prompt and frequent collection of data from the loggers, and use and monitoring of any other TCP/IP-connected device deployed at the site (e.g. web cameras).

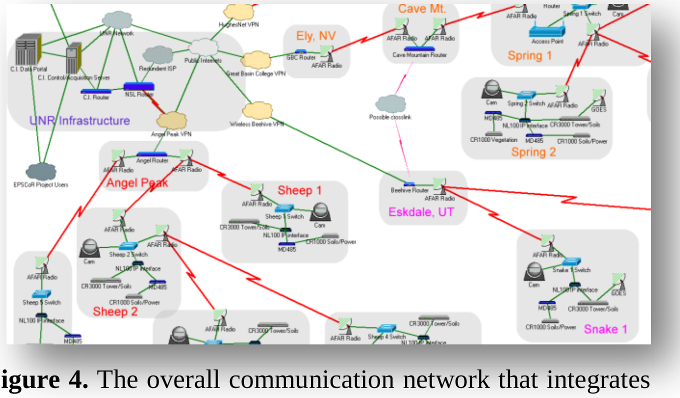

Figure 4. The overall communication network that integrates remote data loggers with the SENSOR system for the Nevada Climate Change Project.

As has been stated previously, the SENSOR system communicates with the data loggers at each deployment. The logger is the effective source of measurements, as it is programmed to measure any set of sensors installed at a monitoring deployment, interpreting their electrical signals and recording numerical values. The SENSOR system is, therefore, capable of supporting any device that the logger can support (i.e. interpret electrical signals into discrete values). This flexibility allows researchers to configure their site deployments to suit their research needs, largely independent of the fact that the SENSOR system is collecting those values. Whenever a logger is programmed or updated, the SENSOR system import settings are updated to properly identify and handle the data being collected; the system simultaneously employs various detection mechanisms that identify changes to the logger program, preventing the import of data from a logger that has an unexpected or changed program. In this way, the import of potentially corrupt or invalid logger data is averted, prompting a data manager to review the change and adjust the import settings as necessary for the new program (and, thus, sensors) and document the change information.

3.3 Data Standards

The efficient and meaningful long-term sustainability of multi-project data requires that all data ultimately conform to certain data standards. With respect to climate data measurements, two long-term concerns present themselves: time and measurement units.

Tackling the issue of measurements is deceptively simple: one would expect to collect data and store it in the standard units for climate research. However, during implementation, one would discover that the “standard” units of measurement for climate research vary between both disciplines and researchers. For example, whether rainfall is measured in millimeters or inches, or distance / length in meters, kilometers, feet, or miles is a decision that varies between atmospheric scientists and hydrologists, and even with the preference of each researcher within that field. Implementing the SENSOR system, it was quickly apparent that there were no actual standard units of measurement; there were “common” units of measurement that were accepted by a body of each field, but such an arbitrary “standard” is clearly not suitable for long-term data storage and management.

As a result, our implementation relies on the use of standard SI units for data storage within the database, omitting scale prefixes. This means that when we store data measured in, for example, millimeters, we store it as meters, losslessly converting the value appropriately; a value measured in ounces is similarly converted to kilograms. This decision has several advantages. Firstly, it standardizes all units within the database, making conversions to other units significantly easier and faster (ultimately satisfying the unit needs of *all* researchers, not a select few). Secondly, by standardizing units within the database instead of the loggers, maintenance personnel can verify logger functionality in the units they’re most comfortable seeing, while data is losslessly converted during import to the database. The configuration of the loggers for standardized output is, indeed, a logical step; however, in practice, technicians must travel a significant distance in remote terrain to access and maintain the loggers, their fatigue making them more apt to configuration errors. The least error-prone solution – which we implemented – is to automate the conversion of logger values to standards for database storage.

Standards for representing and synchronizing time become exceptionally important for supporting the temporal correlation of data collected. Under the traditional data collection approach, logger clock synchronization is done infrequently, allowing the logger clocks to drift significantly throughout any given interval. Given that deployments often have limited connectivity, synchronization opportunities are obviously rare. Further, loggers are often set to the local, standard time in which they make measurements – a fact that is often undocumented and assumed to be “obvious.”

The collection and curation of high-quality climate measurements requires that these shortcomings be addressed. Ideally, logger clocks should be synchronized with a source several times a day to ensure accurate measurements and prevent clock drift. Protocols such as NTP [12] are ideal, as they inherently correct for network latency and other common potential transmission issues, and should be implemented whenever possible. Further, logger clocks should ideally be set to use UTC time, which is not subject to annual corrections (i.e. Daylight Savings Time) and is a globally-recognized standard for time.

In practice, the use of UTC on loggers poses a problem similar to that of using SI units for measurement on the logger – namely that maintenance personnel are more apt to make configuration mistakes when seeing unfamiliar time entries. Again, the most reasonable method of minimizing errors – and the one utilized by the SENSOR system – is the implementation of standard local time (not Daylight Savings Time-sensitive) on the loggers, with a conversion to UTC during import into the database. This standardization allows the database to uniformly search data based on a standard time designation, allowing optimal search efficiency and output conversions to any specified time zone.

With regard to clock synchronization, the network implemented by the SENSOR system supports NTP time synchronization to remote loggers and systems. In fact, the web cameras are synchronized to the central data servers several times a day. However, due to hardware limitations, the loggers do not currently support NTP synchronization and are limited to a vendor-specific 1-second drift. While hardware can be upgraded to support NTP in the future and, in fact, future deployments will be required to support and utilize NTP clock synchronization, the implementation of some kind of clock synchronization is a significant improvement over the traditional approach, in terms of data quality.

3.4 Policies

In implementing the SENSOR system, we discovered that several system-wide policies needed to be implemented to guide and enforce data quality standards, and to provide additional metadata consistently. The most important policy developed requires that changes to the hardware and / or software configuration of loggers and site hardware be documented within the database. This information is vital to later evaluations about the quality of data measurements, as it may record a sensor fault or other malfunction that casts doubt on measurements.

The next most important policy states that data collected from loggers be in a particular format (Campbell Scientific XML [13]), as this format provides data not present in other formats which is vital in describing and identifying the logger, both for import and long-term data archival of raw data files. The selection and use of such a robust and detailed file format also addresses disaster recovery planning requirements. The mobility of the complete dataset is then possible using either the database or collected XML files, both of which contain the complete set of data and, (at minimum) much of the metadata. This type of data safety and robustness is a cornerstone of any long-term data acquisition and curation system.

Various other policies have been developed to reiterate the use of data standards, as described earlier.

4 CONCLUSIONS

The introduction of data management and curation activities as a primary concern for climate research data collection has a significant impact upon the data collection process. Addressing the quality and additional information requirements of these activities requires the use of new, modern software, hardware, and policies across multiple projects.

The implementation of the SENSOR system, which supports very robust data management and curation, illustrates some of the changes required to the data collection model, and also provides insights into potential best-practices for implementers. Chief amongst those changes is the use of a high-performance geospatial database to organize climate measurements and metadata in a standards-independent manner, providing long-term data extensibility and variable output format options. Similarly important is the use of high-speed, error-free communications networks, such as those employed by the SENSOR system, that eliminate transmission errors and provide ample bandwidth for monitoring and troubleshooting activities.

The implementation of data standards that codify the use of global standards for measurement and timekeeping play a central role in maintaining long-term data usability – specifically, the use of SI units and UTC clock settings, respectively. Similarly, the incorporation of system-wide policies that assist in the collection of maintenance data and enforce configuration standards can provide a variety of long-term benefits.

In summary, by applying standards and modern software and hardware systems to the traditional, silo-like data collection approach of individual research projects, data can be collected and maintained in such a manner as to make it available, relevant, and verifiable for current and future researchers in any field. The extensibility of the SENSOR system is such that it can provide these features and options to smaller-scale projects that would otherwise have no budget to construct capable data management infrastructure. Moreover, the application of these modern systems can provide features and options that were previously extremely difficult to offer, if not impossible.

5 ACKNOWLEDGEMENTS

The material presented in this paper is based upon work supported by the National Science Foundation under Grant No. 0814372.

6 REFERENCES

[1] Digital Curation Centre, http://www.dcc.ac.uk/, 2011.

[2] National Science Foundation: Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans. Press release 10-077. http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928, 2011.

[3] JISC, Managing Research Data (JISCMRD), http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx, 2011.

[4] McMahon, M., Jr., Dascalu, S.M., Harris, F.C., Jr., Strachan, S., and F. Biondi (2011). Architecting Climate Change Data Infrastructure for Nevada. In Salinesi, C. and Pastor, O. (eds.), Advanced Information Systems Engineering Workshops CAISE-2011, Lecture Notes in Business Information Processing, LNBIP-83, June 2011, Springer, pp. 354-365.

[5] National Oceanic and Atmospheric Administration, http://www.noaa.gov, 2011.

[6] National Climatic Data Center, http://www.ncdc.noaa.gov, 2011.

[7] JISC, ACRID: Advanced Climate Research Infrastructure for Data, http://www.jisc.ac.uk/whatwedo/programmes/mrd/clip/acrid.aspx, 2011.

[8] Federal Geographic Data Committee, Content Standard for Digital Geospatial Metadata (version 2.0), FGDC-STD-001-1998. http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf, 2011.

[9] International Standards Organization, ISO 19115:2003. http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020, 2011.

[10] Geostationary Operational Environmental Satellites. http://www.oso.noaa.gov/goes/, 2011.

[11] The Nevada Seismological Laboratory. http://www.seismo.unr.edu/, 2011.

[12] Mills, D.L., Network Time Protocol Version 4 Reference and Implementation Guide. Electrical and Computer Engineering Technical Report 06-06-1, University of Delaware, June 2006, pp 83-. http://www.eecis.udel.edu/~mills/database/reports/ntp4/ntp4.pdf, 2011.

[13] Campbell Scientific, LoggerNet 4.1 Manual, pp. B-6 – B-15. http://www.campbellsci.com/documents/manuals/loggernet.pdf, 2011.