J4.6 NOAA'S SCIENTIFIC DATA STEWARDSHIP PROGRAM: CONNECTING ESSENTIAL CLIMATE VARIABLES TO SCIENCE QUESTIONS AND SOCIETAL IMPACTS FOR LONG-TERM INFORMATION ACCESS

Bruce R. Barkstrom, John J. Bates *, and Jeffrey Privette
NOAA National Climatic Data Center, Asheville, North Carolina

1. INTRODUCTION

The goal of NOAA's Scientific Data Stewardship (SDS) Program is to provide high quality Climate Data Records (CDRs) for data from the atmosphere, oceans, and land surface. The data in these CDRs will have been identified as Essential Climate Variables (ECVs) within the Global Climate Observing System (GCOS) plans. During an initial development phase, the SDS program expects to place its emphasis on ensuring the quality and survivability of satellite observations with high levels of scientific and preservation maturity, as well as high societal benefit. As experience with such data sets expands, the SDS program expects to extend the data with these attributes so that it will produce CDRs routinely on an operational basis.

In support of this work and in support of the Global Earth Observation System of Systems (GEOSS), we have designed a web site that provides

1. A unified view of the GCOS ECVs, criteria for prioritizing measurements, and a view of potential instruments and in-situ data sets that can contribute in measuring each ECV.
2. An ability to interactively navigate from summary information (that can assist managers in prioritizing measurement technologies) to detailed inventories of data set versions (to assist data set users in finding and assessing data sets of use in research and decision making).
3. An interactive input approach that can help develop community consensus regarding priorities for measurements or data set production schedules.

This web site is intended to assist agencies and organizations in prioritizing measurement strategies and for managing data in a way that fosters efficient, wide participation in the prioritization process. It may also help sophisticated data users discover data for particular uses involving highly technical assessments of measurement uncertainty.

* *Corresponding author address*: John J. Bates, NOAA National Climatic Data Center, 151 Patton Avenue, Asheville, NC 28804; e-mail: John.J.Bates@noaa.gov.

2. SCIENTIFIC DATA STEWARDSHIP

Scientific Data Stewardship is a subset of data management, emphasizing data quality, quantification of uncertainty, and long-term data access. These features make NOAA's SDS program particularly well-suited to production of Climate Data Records. A recent National Research Council report (2004) defines a CDR as "a time series of measurements of sufficient length, consistency, and continuity to determine climate variability and change."

That report also differentiates fundamental climate data records (FCDRs), which are calibrated and quality-controlled sensor data that have been improved over time, and thematic climate data records (TCDRs), which are geophysical variables derived from the FCDRs, such as sea surface temperature and cloud fraction. In most senses, it is appropriate to associate FCDRs with Level 1 data (i.e. geolocated and calibrated) and TCDRs with higher level data. Thus, TCDRs may contain instantaneous geophysical fields or spatially gridded and temporally averaged fields. The instantaneous TCDRs usually have the full temporal and spatial resolution of the original data and are thus often described as Level 2 data. The gridded and averaged TCDRs would often be described as Level 3 data.

In the material that follows, we first describe some characteristics of CDRs. Then, we formalize those characteristics in terms of a maturity model. Finally, we tie the maturity model to a Data Submission Agreement between the CDR data provider and the archive that retains the data.

2.1 *Characteristics of CDRs*

There are several characteristics that distinguish CDRs from other kinds of data products:

- *Record Length* – Because CDRs are intended to provide reliable data for climate research, it is important that the data record be as long as possible. This characteristic implies that a CDR will need to include data from several different sources, such as similar instruments on a succession of spacecraft.
- *Error Structure Homogeneity* – Climate investigations usually seek to detect and measure small signals embedded in a highly variable record. Instrument or algorithm artifacts add uncertainty and can substantially reduce the usefulness of data sets for investigations that seek to measure these small climate signals. To achieve error structure homogeneity, a CDR data provider

will usually need to carefully check (and often reprocess) data for consistency.

- *Assurance of Data Provenance* – Because of the importance attached to being able to draw reliable conclusions regarding climate change from these data, CDRs need to ensure that users can retrieve and understand the heritage and chain of custody of the data they want to use.
- *Ability to Retain Usefulness over Long Time Periods* – Because CDRs are dealing with long data records with high attention to data quality, they may take longer to produce than weather forecast data as they need to be dependable over very long periods of time. This characteristic means that CDRs need documentation of their context and assurance of long-term survivability in the presence of huge changes in their Information Technology (IT) environment and in user access needs.

These characteristics of CDRs place rather stringent requirements on the production configuration management. Figure 1 shows a generic Data Flow Diagram for a family of CDRs. Level 0 data appear on the left of this figure. The processes of calibrating and geolocating the raw data are those that produce the Level 1 data that constitute a FCDR. The instantaneous Level 2 data form one kind of TCDR; the gridded Level 3 data form a second.

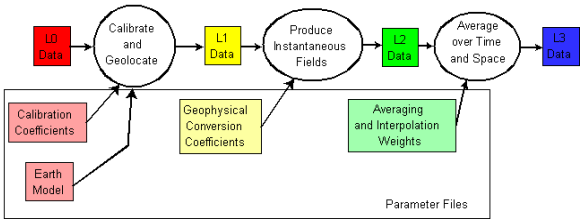

Figure 1. Generic Data Flow Diagram for Climate Data Records.

In addition to illustrating the relationship between FCDRs and TCDRs, figure 1 illustrates how errors propagate between data products and what needs to be tracked in order to maintain a record of data provenance. If we consider the FCDR that appears as Level 1 (or L1) data, for example, it is clear that there are four potential sources of error:

- Level 0 data, which may have gaps or errors introduced by the instrument telemetry or ground receiving systems
- Source code (and algorithms) for the calibration and geolocation process
- Calibration coefficients and Earth model used in the data reduction
- System configuration (computers, operating system, compilers, production scripts) used to actually run the jobs that produce the data

Errors any of these potential sources will propagate into the FCDR. Thus, quality assurance in forming the FCDRs is of the utmost importance.

The same emphasis on production quality also applies for the higher level data products. Figure 1 makes clear that errors or gaps in Level 1 data propagate into Level 2 data. This figure also shows that errors in the Geophysical Conversion Coefficients or in the Level 2 source code (or algorithms) will affect the instantaneous TCDRs. The system configuration may also contribute to error propagation. For gridded and time averaged TCDRs, this figure illustrates the same connections.

Figure 1 illustrates the difficulty of maintaining version control. To produce a single version of the Level 1 data, the source code and coefficients that convert the Level 0 data to Level 1 must be kept constant for the time span of the data in the version. To produce a single version of the Level 2 data, there needs to be a homogeneous version of Level 1 data, as well as a single version of the Level 1 to Level 2 source code and a single version of the Geophysical Conversion Coefficients. A homogeneous Level 3 TCDR version adds additional requirements on the homogeneity of the algorithms and coefficients for the time and space averaging process.

In practical terms, then, the production of CDRs must adhere to the highest standards of configuration management. To avoid introducing inhomogeneities into CDRs, the production system must ensure that the input data sources are as homogeneous as possible and that source code and input coefficient changes are as small as possible, but still support generation of homogeneous time series. As our field matures, it appears probable that interdisciplinary CDRs will become more common. Such data records are particularly challenging for climate research because they require maintenance of homogeneous input versions. Thus, attempts to produce CDRs with input data sets whose source code changed in an irregular fashion are likely to be less valuable than CDRs whose input data has homogeneous versions.

2.2 *A Model for Maturity of Data Sets*

The CDR characteristics just mentioned seem daunting. Furthermore, it appears necessary to allow different degrees of stringency for different kinds of data – some applications are more demanding than others. It seems useful to draw on the experience of the software development community, in which a Capability Maturity Model, Chrissis (2003) has proven useful. In this model, maturity is quantified according to the degree of reproducibility a team exhibits in estimating and producing a software product. At a low level of maturity, the results are scattered; the team does not have the ability to estimate and produce results on time and within budget. At a high level of maturity, the process becomes consistent.

Adopting this suggestion, the SDS program is pursuing an approach to quantifying the maturity of a data set based on three axes:

- *Scientific Maturity* – a set of measures of the scientific quality of a data product

- *Preservation Maturity* – a set of measures of the long-term sustainability of the information in a data set
- *Societal Benefit and Impact* – a set of metrics that assess the potential value of a data set

The intent of these axes of maturity is to provide a systematic measurement approach to ensuring that data sets are valuable and ready for long-term preservation.

In quantifying the level of maturity for each axis, we have found it helpful to develop a hierarchical breakdown of the appropriate attributes. At the most detailed level, each attribute can be ranked non-dimensionally, avoiding the troubles that arise with different units for different attributes. The lack of units makes it easier to quantify maturity and rank the attributes of various properties.

Scientific Maturity attributes break down into the following categories:

- Physical understanding of the measurement process, including
  - Measurement of spectral sensitivity
  - Measurement of Point Spread Function (PSF) (spatial sensitivity)
  - Pre-launch calibration
- Capability to detect important changes in calibration
  - Changes in spectral sensitivity
  - Changes in PSF
  - Changes in calibration parameters
- Public accessibility of data production processes
  - Documentation of data flow diagram
  - Algorithm Theoretical Basis Documents
  - Documentation of data editing algorithms
  - Availability of source code for modification
- Rigorous validation
  - Documentation of a validation plan
  - Documentation of validation data and results
  - Understandable uncertainty analysis

Preservation Maturity attributes break down into the following categories:

- Low Total Cost of Operation, including automated operations
- Highly Reliable Operations, including
  - Reliable and automated configuration management
  - Balanced approach to redundancy and dispersed site storage
  - Robust and graceful exception handling
- Evolvability
  - Documentation of designer intent and design evolution
  - Modularity of architecture while being traceable to user needs
  - Formalization of Operations and of operational procedure evolution
  - Outside participation in design, development, and evolution
- Integrity Maintenance
  - Intellectual Property Rights Considered in design
  - Permanent file and data naming registration
  - Ability to track provenance
  - Transactional Basis for system operation and auditing

Societal Benefit and Impact attributes are more difficult to quantify. This difficulty arises because of the diversity of the communities that produce and use CDRs. The diversity is perhaps easiest to describe by thinking of both producers and users as members of tribes. Each tribe has a distinctive dialect, a distinctive set of data world views (that include data structures and formats that are "easiest" or "most natural" for a particular kind of data), and a distinctive set of tribal "customs". The latter include data search strategies and use of particular visualization tools.

A second aspect of the difficulty of quantifying societal benefit and CDR impact lies in the fact that there are at least two measures of data value.

One measure of impact or benefit arises from the ability of data to confirm or negate cause-and-effect hypotheses. In other words, data in a CDR may be sufficiently accurate to confirm that a particular cause-and-effect relationship holds – or it may not be that accurate. Thus, this approach to valuing data is perhaps best quantified in terms of the uncertainty in the data or in terms of its ability to answer important scientific questions.

A second measure arises from the flow of economic impact a CDR can produce. While our attention is often drawn to the ability of data to assist emergency first responders, we should also keep in mind that CDRs may help quantify the probability of long-term extreme conditions and can assist with planning for mitigating the effect of those conditions. With this kind of metric, data value may be measured by cost avoidance for particular conditions. Thus, to the extent that we could quantify savings associated with better water or energy management, CDRs might be evaluated in terms of the net present value of future cost savings for expenditures.

The Intergovernmental Panel on Climate Change has noted that scientific value and monetary value are not measurable on the same scale – and therefore we need to develop methods that appropriately balance the valuations that these two different systems use. It seems useful to think of the monetary valuation as emphasizing near-term data uses, while the scientific valuation emphasizes the long-term use.

2.3 *Using the Maturity Model for Setting up Data Submission Agreements*

One of the most significant frameworks for systematically dealing with long-term preservation of the information in CDRs lies in the Open Archive Information System (OAIS) Reference Model, CCSDS

(2003). This model is an international standard that formalized a systematic description of the functions an archive needs to perform. This framework combines with the maturity model to provide a way of ensuring that the scientific community and its archives can agree on policies regarding which data are valuable enough to archive and what needs to be done to ensure that the collection will continue to be valuable after being placed in the archive. The maturity model provides the attributes for evaluating data sets that are candidates for CDRs.

In the SDS program, the process of using the maturity model involves a systematic evaluation of the candidate CDRs that follows a procedure such as the following:

1. Establishing expectations for the process of accepting a CDR
2. Obtaining a description of the data collection to be included in the CDR
3. Conducting a collection valuation and risk assessment that identifies why the candidate CDR would be of long-term value and what steps need to be taken to assure continued value
4. Quantifying a user access model that will allow the archive to plan for resources needed to provide adequate user access over the foreseeable future
5. Conducting an engineering analysis of the Data Provider Submissions, allowing the archive to estimate the cost of accepting the CDR and maintaining it
6. Developing a service delivery agreement, if that is necessary
7. Defining the Data Producer Logical Namespace and Metadata Model, balancing the unique nature of each CDR and the need for interoperable data discovery
8. Designing (or modifying) the archive's metadata databases and database access interface to accommodate the proposed CDR
9. Designing the Submission Schedule that will govern the timelines and milestones under which the data provider will submit data and when the archive can provide user access to the CDR
10. Finalizing the submission agreement

2.4 *Risk Assessment and Mitigation for CDR Information Survivability*

While many of these steps need formal documentation, the point of this procedure is to ensure that information accepted into an archive for long-term preservation maximizes the probability of long-term survival. In many ways, archives need to take advantage of the standard approach to identifying and managing risk. This approach is spelled out nicely in Boehm, (1981).

Item 3 in the list above is particularly important in the context for survival. Suppose the probability of loss per year is p. The probability of having information survive N years is $(1 - p)^N$. For archival work, a reasonable value for N is 200 years. In order to have a 99% probability of survival, information will need $p < 5 \times 10^{-5}$, a rather stringent requirement.

Given the fact that new hardware models for reading and writing media emerge every five years, with a similar value for software upgrades or new environments, it is clear that the probability of loss per year due to obsolescence is about 20%. For archives, there is little choice but to migrate data from old storage media to new media at least once every five years. Equally important is the migration of both data and metadata from old software versions to new ones. This kind of migration places a heavy burden on both data providers and archives because it requires them to consider a strategy for minimizing the probability of loss in the face of continual evolution of the Information Technology environment.

It is also important for data providers and archives to ensure that when data are accepted for long-term preservation that there is sufficient context so that after the providers have moved on the community preserving the data record can interpret the results. The most obvious part of this problem lies in documenting data formats and ensuring that the documentation is kept up to date. However, the preservation of information context for earth science data also requires ensuring that the data production software be available, understandable, and useable – even if the IT environment has changed out from under the original production environment. This requirement may also place stringent demands on configuration management, as represented by the question "which version of the input data do we need to have available if we want to reconstruct the original data record?"

While obsolescence is almost certainly the largest risk to preserving the long-term data record, there are two other risks that appear to be almost as large: the probability of loss of data integrity or provenance due to IT incidents and the probability of loss due to operator error or hardware and software failures. While it is difficult to obtain documentation regarding both of these risks, it appears that they are each on the order of 5% to 10% per year. While a discussion of the methods required to systematically reduce their probability to acceptable levels, a first look at the mechanisms for risk reduction involves keeping at least one copy of data completely off-line and developing highly reliable, automated data migration and handling procedures.

A corollary to these issues lies in the policy related to data product preservation. Some members of the Earth science community have suggested that all that is needed for long-term preservation is the raw data stream and the software to produce higher-level data products on demand. Proponents of this policy are very concerned over the projected increases in stored data volume. The alternative policy is to keep all of the produced higher-level data products. This policy would appear to reduce the probability of data loss owing to operator error in what are often highly complex data production processes, as well as the difficulties associated with migrating processing software.

Resolution of the conflict between these two alternative policies will require a careful assessment of the risk and cost associated with each alternative.

3. PRIORITIZING MEASUREMENTS AND CDRS

The GEOSS is an ambitious consortium of U.S. and international agencies. There are several entry points. The one for NOAA (http://www.noaa.gov/eos.html) states that "More than 60 countries, the European Commission and more than 40 international organizations are supporting the development of a global Earth Observation System that, over the next decade, will revolutionize the understanding of Earth and how it works. The U.S.-led initiative promises to make people and economies around the globe healthier, safer and better equipped to manage basic daily needs. The aim is to make 21st century technology as interrelated as the planet it observes, predicts and protects, providing the science on which sound policy and decision-making must be built.

3.1 *GEOSS Societal Benefits*

GEOSS has created a strategic plan (http://usgeo.gov/docs/EOCStrategic_Plan.pdf). The plan identifies nine societal benefit areas that provide focus for the U.S. effort and links to international activities. These areas are

1. *Weather*: Improve weather forecasting
2. *Disasters*: Reduce loss of life and property from disasters
3. *Oceans*: Protect and monitor our ocean resource
4. *Climate*: Understand, assess, predict, mitigate, and adapt to climate variability and change
5. *Agriculture*: Support sustainable agriculture and forestry, and combat land degradation
6. *Human Health*: Understand the effect of environmental factors on human health and well-being
7. *Ecology*: Develop the capacity to make ecological forecasts
8. *Water*: Protect and monitor water resources
9. *Energy*: Monitor and manage energy resources

The GEOSS plan includes the phrase "System Of Systems" because, as a practical matter, there is a recognition that building a single source system and deploying it would be prohibitively expensive and politically infeasible, particularly given the need for international cooperation. Thus, the architectural approach expects that "this system will be built upon existing and planned systems and will identify and document observation gaps and needs in the societal benefits areas. Currently, most of the data and information related to Earth observations are encompassed within the U.S. National Spatial Data Infrastructure and integration of Earth observations will be implemented within that legal, policy, and institutional framework. The technical implementation component will establish the standards, protocols, and metadata for observation systems. The strategy will also recommend the optimum operating environment and support developing the associated human and institutional capacity.

The U.S. Integrated Earth Observation System builds upon current cooperation efforts among existing observation systems (including but not limited to the physical integration of observation systems on the same platform or at the same ground site, and by sharing space platforms and observation towers on the ground for various observations), processing systems, and networks, while encouraging and accommodating new components. Across the processing cycle from data collection to information production, participating systems maintain their mandates, their national, regional and/or intergovernmental responsibilities, including scientific activities, technical operations and ownership." (GEOSS Strategic Plan, pp. 38-39.)

3.2 *Key Science Issues*

In addition to providing societal benefits, Climate Data Records also assist in answering key science issues. Dr. Francis Bretherton formulated such a list in 1994 for the U.S. Global Change Research Program. Perhaps remarkably, this list still appears to have relevance.

"Each product results from ongoing activities, with anticipated periodic upgrades and involving improvements focused on the items in italics. Items are listed in approximately the order in which significant improvements can be expected, given the present state of the science.

a. Dection, causes, and i*mpacts* of significant changes in the stratospheric ozone layer, and analysis of the environmental effects of response strategies.
b. Regularly scheduled ongoing predictions one year ahead of interannual climate fluctuations associated with El Nino, together with regional *impact* and *forecast utilization* studies.
c. Plausible scenarious for regional climate and ecosystem change in a *form suitable for various impact models*.
d. Estimates of the relative global warming potential of various gases and *aerosols*, including *interactions* and the *indirect effects* of other chemical species.
e. Ability to determine *national* sources and sinks for atmospheric Carbon Dioxide and other greenhouse gases and *aerosols*, for application as part of the *monitoring system* for the Framework Convention on Climate Change.
f. *Reduction in the range* of predictions of the rate of global climate change over the next century through improved models and understanding of the effects of clouds, ocean heat and carbon storage, and land surface processes.
g. Predictions of anthropogenic inter-decadal changes in *regional* climate in the context of

statistics for natural, unpredictable, interannual and interdecadal *variability*.

h. Detection *beyond reasonable doubt* of greenhouse gas-induced global warming, and *documentation of other significant changes* in the global environment.

i. *Understanding* of the major interactions of human societies with the global environment, *enabling quantitative analyses* of existing and anticipated patterns of change."

3.3 *Clarifying Prioritization Issues*

GEOSS places high value on making the measurement system relevant to meeting society's needs. The "Bretherton Issues" emphasize the importance of dealing with key scientific issues for understanding the Earth as a system. The importance of developing improved measurement technologies provides a third perspective. These three perspectives flavor many of the discussions regarding priorities for flying particular instruments, for developing candidate CDRs into operational CDRs, and for connecting CDRs with decision support applications. It is also clear that different scientific communities have different priorities that can make it difficult to reach a community consensus on what future measurements need to be made and which CDRs produced next.

It is not easy to resolve the apparent conflict between these different prioritization approaches. Economists [e.g. Bruce, et al., 1996; Stern, 2006] identify these as "multi-valued" prioritization frameworks. The authors had examined several different approaches of different complexity, including social learning by rational Bayesian agents [Chamley, 2006], analysis of variance with clustering, Delphi approaches, and economic analysis [Stern, 2006]. In the end, a simple approach that identifies the contribution of each Essential Climate Variable (ECV) to the societal benefits and to the Bretherton issues appears to be a useful start. In this approach, members of the measurement community should be able to contribute a simple 0 (low) to 3 (high) ranking of the contribution. After a round of prioritization input, software can sum each ECV's contribution to produce a weighting of the variable. Because these weightings are not precise, it seems sensible to group them into a small number of categories. For example, if an ECV contributes strongly (with a ranking of high) to three societal benefits and four Bretherton issues, then its average priority is high and this ECV is in the top category. On the other hand, if this variable were ranked as moderately or weakly tied (with a ranking of medium or low, respectively) on only one societal benefit or one Bretherton issue, it would be placed in a low priority category.

This approach is easy to understand and is quite amenable to visualization. It is also easy to place into a web site that members of the community can use to provide input and to visualize the priority rankings. In addition to ranking the importance of measurements of ECVs, experience in the ocean community suggests that in deciding which measurements need to be made and which CDRs need to be produced immediately and which may need further development, it is necessary to incorporate a view of the "feasibility" of producing the measurement or the CDR. The same approach of simple categorization based on a few of the maturity attributes we have listed, as well as the probable cost appears applicable to the feasibility ranking. In Section 4 we describe how this approach facilitates development of a community consensus on priorities.

3.4 *Engineering Data Flows and Data Production in a "System of Systems"*

It is clear that the scientific community is concerned with provision of adequate measurement capability in the future. However, it is also clear that the community can ill afford to abandon the data record already accumulated. A significant effort is required to ensure that this record is useful in the future. Knapp, et al (2006) provide a number of concrete examples of some of the difficulties, including ensuring that the data set documentation is as carefully preserved as the data itself – and that the data, metadata, and documentation are kept "in sync".

In practice, data users do not directly connect with instruments – they use data products and data services for this purpose. The first step in creating and managing CDRs is to inventory data sources and data sets to ascertain their condition. This step moves beyond several current databases that contain information on platforms and instruments.

There appear to be several key ingredients to such an inventory:

- *Essential Climate Variables* – and the deeper ties to individual parameters and data structures contained in the files that usually contain archived data
- *Justification and Prioritization* – which provide the ties between ECVs and the societal benefits as well as scientific issues
- *Data Sources* – meaning the platforms and instruments that provide the Level 0 data, including both satellite-based systems and in-situ data sources
- *Data Set Versions* – by which we mean data collections that have common contents, common time intervals of data collection, common data sources, and homogeneous error structure (insofar as the data sources allow that)
- *Data Flow Diagrams* – and the extensions that provide the connectivity between different data sources and data set versions
- *Uncertainty Assessments* – which we treat like error budgets, with the added assessment of systematic biases and probability distributions
- *Production Schedules* – based on both data sources and (re)processing expectations

The details of the connections between these ingredients are beyond the scope of this paper, and will eventually be available in Unified Modeling Language (UML). For now, we note that these ingredients form a

highly complex network of concepts. The dimensionality of connections is sufficiently high that it does not fit well within a normal report structure. Rather, it has been much easier to conceive of what we need as a web site, in which users gain insight by navigating between concepts.

4. WEB SITE DEVELOPMENT

The genesis for developing a web site to present this information occurred in several meetings, where the authors found that prioritization discussions often diverged from their agendas when the participants needed to incorporate holistic views of complex prioritization issues. The GCOS Implementation Plan, for example, has a format that makes it difficult to compare and contrast features of measurement technologies within a common framework. Participants in a recent Climate Change Science Program Prioritization Workshop found it necessary to incorporate in-situ measurements with satellite measurement systems. It was almost as though the participants were trapped one and one-half dimensions formed by the format of the traditional report structure and the representation of complex webs of relationships.

A web site (or, more theoretically, a hypertext approach) offers a useful extension of capability because such a site can hide "details" while it keeps them available. Even more important is the fact that a user's navigation through the site can enhance his or her capacity to use "spatial navigation" as an organizing tool for holistically viewing complex relationships between objects. Thus, we suggested developing a web site that could assist in the prioritization of measurements, as well as managing data for GEOSS.

In the subsections that follow, we identify several sources of complexity in creating this kind of web site. These include the fact that there are a number of sources of variable lists, different approaches to creating data sets, and different standards for metadata.

4.1 *Complexity 1: Parameter Nomenclature*

One of the complexities of dealing with data products and measurements lies in the diversity of nomenclature. While the GEOSS Strategic Plan provides a list of twenty-six "Earth Observations," there are a number of alternative lists, such as a list of "Essential Climate Variables," the Global Change Master Directory (GCMD) "Topics and Terms (see http://gcmd.nasa.gov/User/difguide/difman.html)," the list of parameter names appearing in the Climate Forecasting (CF) profile (see http://www.unidata.ucar.edu/projects/THREDDS/GALEON/netCDFprofile-short.htm), as well as the lists of parameters that appear in existing and planned data documentation.

We can distinguish three basic levels of specificity in these lists:

- *High-Level Category Lists*, such as the "Earth Observations" or "Essential Climate Variable" lists. These have between twenty-five and one hundred categories and provide top-level classes of observation types
- *Controlled Vocabularies of Key Phrases,* such as the "GCMD Topics and Terms", the CF Profile or the list of parameters in the future spacecraft requirements. These sources typically range from one or two hundred phrases to about one thousand. They provide an intermediate level of specificity and are often used as metadata in key phrase approaches to data discovery.
- *Inventories of Parameters in Data Products,* which are usually available in documentation for particular data products. There are probably several thousand distinct data products identified in various web sites. An informal examination of such lists suggests that the number of parameters in a file is typically between ten and one hundred. In many cases, the parameter identifiers extend the concepts used in either controlled vocabulary lists or high-level category lists.

It is perhaps expecting too much to think that there will ever be a single vocabulary for data products. While the concept of the "Semantic Web" has received wide publicity in the IT domain, the concept spaces of different communities are still not susceptible to automated translation. At the same time, we do not need to remain within a list as the only way of dealing with translations between different ontologies and vocabularies. A web site allows several alternative approaches, including tables that show the correspondence between terms, presentation of alternative search terms when a user types in a key phrase for searching, and the use of concept maps.

4.2 *Complexity 2: Production Description*

A second source of complexity in dealing with data sets and data set versions lies in the diversity of ways teams have taken to describe their approaches to produce data products. It appears that most Earth science data are produced by discrete, batch processing. In this case, the fundamental description of production takes the form of a graph, in which the input and output files form nodes and the executable that ingests input files and creates output files is an arc, Barkstrom, (2003). Appropriate graph algorithms can traverse the graph from an output file back to the fundamental data sources – provided the graph has been recorded in a usable form.

In addition, because the source code for producing CDRs will almost certainly be the same across all of the production instances that create a data set version, a production graph for a single instance may be regarded as an instantiation of a data flow diagram template. This means that a data flow diagram for producing CDRs is a key component for understanding how different data sources that might create somewhat similar data relate to each other. Specifically, a data flow diagram (DFD) is an essential element of

documentation that can assist in identifying sources of error and for comparing algorithmic structures.

Oddly, there is little agreement on how to document this key to understanding. At one end of description lie the Earth radiation data products, where the data flow diagram is built into the CERES Data Products Catalog, Caldwell, (2006) and into the web site for accessing these products (http://eosweb.larc.nasa.gov/PRODOCS/ceres/ceres_dataflow.html). The Earth Observing System Data Products Handbook, King, et al., (2004) is inconsistent in its approach to describing production data flow. Some instrument teams appear to have DFDs. Others do not. Descriptions of production data flows for other agencies are also lacking consistency.

One of the items we expect to include in the CDR web site is an explicit database representation of the data flow diagrams that govern production. With these diagrams, we can make sense of the production scheduling needed for reprocessing. These diagrams also provide help in comparing alternate approaches to producing similar CDRs.

4.3 *Complexity 3: Metadata Structures*

A third source of complexity lies in the diversity of metadata structures used to keep track of data and organize searches by users. In part this complexity arises because it has taken a long time to pull together this part of the infrastructure. There are three fundamental schemas currently in use:

1. *ECS* – the EOSDIS Core System schema, which has been used by NASA's EOSDIS system. This schema is currently in use on more than five petabytes of Earth science data. The schema version available at: (http://edhs1.gsfc.nasa.gov/waisdata/rel6/html/tp4202301.html) comes in a 265 page pdf document that uses entity-element diagrams to show relationships. This schema ties in with both the Federal Committee on Geospatial Data (FCGD) standards and with the GCMD Data Interchange Format for data set documentation.
2. *FGDC* – the schema proposed by the the Federal Geospatial Data Committee in a version that appeared after ECS was widely adopted in the NASA EOS community and before the new ISO standard. The standard, known as the Content Standard for Digital Geospatial Metadata (CSDGM), Vers. 2 (FGDC-STD-001-1998) can be obtained from (http://www.fgdc.gov/metadata/geospatial-metadata-standards).
3. *ISO 19115* – a formalization of the FCGD standard in XML, with revisions and extensions proposed by various interested parties, particularly those interested in applying Geographic Information Systems to data management. This standard is still emerging.

These metadata standards are generally regarded as critical elements of data management. At the same time, they are developed by rather small communities of experts who may not have contact with the full range of data types and data services required by users. Using these schemas requires attention to detail and a great deal of patience. Furthermore, for application to CDR collections, it is important to provide automated ties between the production processes and the metadata associated with the files. Given the decadal time intervals of interest in CDRs, there will be hundreds of thousands of individual files in a single data set version. Data producers should not be expected to supply three hundred parameters per file by hand.

We also note that there has been considerable interest in using databases for managing earth science data. If this approach becomes practical, it has the advantage of making it easier to provide data to users in customized packages. Because Earth science data files can be quite large, this approach would allow data users who want very particular subsets of files to have exactly what they wanted. On the other hand, databases appear unlikely to provide the performance necessary to deal with reprocessing several decades of level 2 imagery to create a new CDR or find all features of interest in such as data set using a new algorithm. At a more detailed level, databases produce nearly continuous versions of data, making it much more complex to ensure error structure homogeneity across a multi-decadal record.

5. SUMMARY COMMENTS

In this paper, we have defined CDRs, providing a list of attributes for these data records that is intended to enchance their usefulness over the long-term. We also described some of the critical components of a web site designed to assist in prioritizing measurements and CDRs. This site is also expected to be helpful in managing data for the Science Data Stewardship program and perhaps other data connected with the GEOSS program.

6. REFERENCES

Barkstrom, B. R., 2003: Data product configuration management and versioning in large-scale production of satellite scientific data, in B. Westfechtel, A. van der Hoek (eds.): *SCM 2001/2003* Lecture Notes in Computer Science 2649, pp. 118-133.

Boehm, B. W., 1981: Software risk management: principles and practices, *IEEE Software*, 8, 32-41.

Bruce, J. P., Lee, H., and Haites, E. F., 1996: *Climate Change 1995: Economic and Social Dimensions of Climate Change*, Cambridge University Press, Cambridge, UK.

Caldwell, T. E., L. H. Coleman, D. L. Cooper, J. Escuadra, A. Fan, C. B. Franklin, J. A. Halvorson, P. C. Hess, E. A. Kizer, N. C. McKoy, T. D. Murray, L. T. Nguyen, S. K. Nolan, R. Raju, J. L. Robbins, J. C. Stassi, S. Sun-Mack, C. J. Tolson, P. K. Costulis, E. B.

Geier, J. F. Kibler, and M. V. Mitchum, 2006: *Clouds and the Earth's Radiant Energy System (CERES) Data Products Catalog*, **Release 4, Version 14**, available at http://eosweb.larc.nasa.gov/PRODOCS/ceres/DPC/dpc_R4V14.pdf.

Chamley, C. P., 2006: *Rational Herds: Economic models of social learning*, Cambridge University Press, Cambridge, UK.

Chrissis, M. B., Konrad, M., and Shrum, S., 2003: *CMMI: Guidelines for Process Integration and Product Improvement*, Addison-Wesley, Reading, MA.

Consultative Committee on Space Data Systems (CCSDS), 2003: *Reference Model for an Open Archival Information System (OAIS)*, **CCSDS 650.0-B-1**, adopted as ISO 14721.2003.

GEOSS Strategic Plan, pp. 38-39, available at: http://usgeo.gov/docs/EOCStrategic_Plan.pdf.

King, M. D., J. Closs, S. Spangler, R. Greenstone, S. Wharton, and M. Myers, 2004: *EOS Data Products Handbook*, **Volume 1**, available at http://eospso.gsfc.nasa.gov/ftp_docs/data_products_1.pdf.

Knapp, K., Bates, J. J., Barkstrom, B. R., 2006: Scientific Data Stewardship: Lessons learned from a satellite data rescue effort, *Bull. Amer. Meteo. Soc.*, in press.

National Research Council, 2004: *Climate Data Records from Environmental Satellites*, National Research Council, Washington, DC.

Stern, 2006: *Stern Review on the economics of climate change*, available in pre-publication form from http://www.hm-treasury.gov.uk/independent_reviews/stern_review_economics_climate_change/stern_review_report.cfm.