Comments on EOSDIS Plans, October 1997
(Roy Jenne, 17 October, 1997)

This text has a discussion of factors that I think should be considered when trying to formulate a present and future strategy for NASA EOSDIS. Other issues are covered in the draft ERG report of October 1997.

1. Projects that can handle data, independently of ECS.

   a. PI teams such as for Total Ozone, Sage, Altimeter, Scatterometer, etc. have a good track record of preparing products from their instrument data.
      — It seems to me that it is likely to be cheaper and better to handle these projects outside of the ECS project.
      — The downside is that this would possibly diminish support for the ECS project. Is this a downside? I think that the ECS project wants to keep these within ECS.
      — Bruce Barkstrom told me that the Sage people have argued strongly to be outside of ECS. They can easily do it on a workstation. They may win. Why are we trying to complicate life for all of these PI groups?

   b. Landsat-7 data and Aster.

      EDC tells me that they could handle Landsat data outside of ECS. It would probably be less risky, but time is short. They have a long history of handling Landsat, so it is not a big deal for them.

      EDC also says they can handle Aster outside of the project. I doubt if there is much downside to letting EDC handle Landsat and Aster outside of the project.

      EDC tells me that they would need about $2 million per year to independently do Landsat and Aster.

   c. CERES.

      There is one CERES instrument on TRMM (Launch November 19, 1997) and two on AM1. The CERES team will obtain 108 GB/day of imager data from TRMM (sent from Goddard on tapes). They (Langley, etc.) have spent $2 million on the initial development to handle CERES by themselves for TRMM.

      The plan is for them to handle CERES on TRMM. And ECS do it on AM1. They want to do it for AM1 too. But they would need a lot of Modis inputs too.

   d. Moppitt (on AM1), John Gille and Paul Bailey
They want to make products independently of ECS. They have proposed a very attractive price. We should let them do it. With the present ERG document, I guess that the project will say “no.”

2. TRMM.

The Goddard DAAC is handling this separate from ECS. They are proud that 5 FTE working for 10 months are now ready for it. So about $500K for people and $1.5 million for hardware to get ready. They plan a total 5 year, $10 million budget, which I think may be higher than necessary.

3. PI teams who want to split off.

Our ERG panel has not been informed about the number of PI teams who want to prepare their products independently of ECS. This information has come to me from other sources. It was only on October 15 that I learned that the Sage team has expended a very large effort to try to become free of ECS, and then I also learned that Moppitt wanted to be free. Why wasn’t I told about this in our ERG October 7-9 meeting? It seems strange to me that the PI teams have to put a lot of work into requesting permission to do their work in a less cumbersome and more efficient way—and then their requests will probably be turned down. Does our panel want to be used to enforce this type of policy?

4. Making an EOSDIS that is more dependent on ECS.

Almost everyone realizes that we still need backup activity within EOSDIS. And this is in the panel report (emergency plans).

NASA Headquarters says that we will go toward a less centralized approach after AM1. There is a clear worry by dominant forces on our review panel that talking too much about backup approaches or emergency work can undermine support for ECS now. The arguments to keep strong support for ECS now are:

1) There is no way to handle Modis outside of ECS (if we started two years ago, there would be ways).
2) Moshe doesn’t believe that the teams could really handle Aster and CERES in AM1. I’m not convinced he is right. The panel has not had the right people there for a debate.
3) People do not want to upset the AM1 work at this late date—but why are they also so eager to keep all the easy stuff under ECS too?
4) They want ECS to be the dominant force for a long time (quite often it is hard to escape this conclusion).
5) At this late date, we can’t do much else (but we really could easily separate a number of projects).

5. Is the current ECS software almost done?

The August 1997 test was encouraging in that it demonstrated that a number of software functions existed. The plans to deliver one or two versions of the software during the next few months are also encouraging. However, the project will be working as hard as possible for more months to
include necessary capability in the software. And there still are performance and scaling issues. Thus, there are still many software uncertainty issues, which is why the ERG report advises that the emergency backup work should continue.

6. Locking in ECS.

By including all projects but TRMM under ECS, we will tend to lock in ECS. If the software more or less works, we will become very dependent on the huge amount of very complex software. If it does not work, we will also be in a lot of trouble.

Bruce Barksdale tells me that the annual cost to keep ECS going later on will be about $100 million per year. I think that may be a bit high. But it seems to me that we are not building in enough hedges against high software maintenance costs or failure.

7. Strategies for EOS data.

I think that the best strategy would be to take advantage of opportunities to become less dependent on ECS. This is not happening now and it is not happening in the ERG report. In many cases, this approach would result in less cost, 100% certain of results, and certainly less top heavy. The general sense I get now is that PIs do not like to work with ECS. This is the strategy of making ECS less necessary. This is the option that I favor now. Our ERG panel is not discussing or supporting this option. The next harder option (item b below) should also be discussed.

a. Some comments.

The project says it is very short on funds. So it is hard to finance the alternative methods. But note how cheap a lot of them are.

Barkstrom tells me that one view is that the real last date for an alternative approach to ECS is about December 1997. Probably about right.

ECS people always say that an alternative approach would cost more, not less. I simply do not believe this. But it is true if we keep ECS at full strength, and spend any increment of money on a different method. Look at the rather modest costs that PI teams now talk about. I have collected other examples of good work in the US.

b. Harder option.

Let the PI projects do it. Radically cut ECS. Accept the fact that Modis will take some time to find its way (but this may be true anyway). There would be lots of political flack with this. But PIs would generally cheer. It would save money. This option should be debated more than it has been.

8. Budget strategy.

If we are trying to keep the ECS budget as high as it is, then it seems to me that the best strategy may be to try to lock in ECS as much as possible. However, this probably leaves us more vulnerable to failure and to possible budget cuts. If we are looking for ways to bring the budget
down gracefully on the one to two year time scales, we need to find methods that do not require an ECS team with 500 or 600 staff.

9. Where are the costs in EOSDIS?

Bruce Barkstrom notes that the hardware costs in EOSDIS are only about 10% of the budget. It may take another 4 to 10% to handle such functions as buying equipment, testing it, keeping it going, software for hardware, operating it, etc. These budget facts put a limit on how much money can be saved by changes in data volume, etc.

However, changes in volume do affect complexity. In ERG we have largely ignored the problems involved in user access to huge datasets. If the volume gets too high, it tends to scare users off, and it prevents them from doing necessary research.

It is impossible to change the costs by very much in EOSDIS unless ways are found to decrease the staffing that is needed.


I saw these numbers in October 1997. The EOSDIS budgets are projected to be fairly constant at about $230 million per year for 1997 through the year 2003.


Robert Harriss, Texas A&M (was NASA HQ), has a letter in Science, 3 October 1997. The funding for MTPE science grants has been cut from $178 million in 1994 to $125 million in 1997. The function of building satellites, algorithms for products, and data management is costing over 80% of MTPE budget.


The station overruns mean that NASA will need $330 million more in the 1998 budget to keep the program on track. In addition, NASA needs to spend $100 million to be ready in case Russia can not meet their promises.

The present total NASA budget is $13.7 billion. It is unlikely that congress will increase this, so NASA will have to absorb most of the station overruns. In February 1997, Goldin warned that the total NASA budget could drop to $13.4 billion in 1999, as part of the “balance the budget” national plans. (See Science, 227, 26 Sep 1997, p. 1920)


The House and Senate agreed that NASA’s budget for FY1998 will be $13.65 billion, which is $148 million more than the administration request. The committees denied a request by NASA to take money from the science, etc. funds to help cover the space station shortfall. Senator Barbara Mikulski supports the station, but supports MTPE even more (Goddard is in her home state).
In August 1997, the White House OMB budget office told NASA to only request $12.6 billion for FY1999. People agree that it will really end up being more than this (Nature, Vol. 389, 9 Oct 97, p. 530).

My comment: with NASA budget getting tighter, NASA HQ is likely to try hard to find savings in infrastructure budgets.


We need better tools to estimate costs for handling data from different missions. We especially need to know what the costs would be if the work is done using the simplest, most direct method. To do this, we need to gather several types of information about projects, methods, and costs.

a. Compile methods and costs for handling data from similar instruments in the past (altimeter, ozone, etc.).

b. Examples of the cost of handling several types of data across the US (factors such as diversity of data types, volume, and cost would be given). I have some of this information.

c. Summary of the changes in the cost and power of computing equipment and storage subsystems over time. These estimates can be compared with the data volume from instruments and with the PI estimates of Mflop rates needed for EOS. We can gather information from several computer centers (NCAR, NCEP, GFDL, Goddard, and ECMWF). How much data is stored? What is the data flow rate? What is the total compute power? What is the total cost? How does this compare with EOSDIS? These comparisons make EOSDIS appear a little less hard than it was. In cost analyses, Barkstrom finds that the hardware and maintenance is only 10 to 15% of EOSDIS costs. This makes sense.

These examples and analyses lead me to think that EOSDIS is overpriced. The problem is that the methods are more complicated than necessary or desirable. This also makes it less likely to work well. PIs have usually done a very good job of producing products in the environment of a general purpose computing system, not using a specialized data system. Examples are ISCCP clouds, climate model runs, reanalysis. Reanalysis has huge amounts of automation of job streams, and it has automatic checking of results to trigger alarm bells. Thousands of computer tasks are automatically initiated every day. This did not cost a lot to set up. And one person at NCEP is running production (plus other tasks). There is a lot more checking of results (about 80% science work) that involves more people in reanalysis at NCEP.
EOSDIS
Comments by Roy Jenne

1. Some cost markers are:
   a. The hardware cost over a 3 year period to store 275 Tbytes of data in a silo (robot retrieval) is about $1.4 million ($500K per year).
   b. NCDC (Asheville) is now copying 100 Tbytes of weather radar data per year at a cost of $800,000 per year.
   c. The University of Wisconsin hopes to copy about 210 Tbytes of GOES satellite data in a 5 year project that would cost $3 or 4 million total. And do other calculations too.
   d. The amount of NASA EOS data coming down will be under 100 Tbytes per year.
   e. These cost markers aren’t all fair, but they also are not all wrong.
   f. We have to be careful about people blaming all of their huge costs on big data volume.

2. We need ECS help to move the data to DAACs.

3. Steps needed to have a working EOSDIS whether or not the ECS software to make products and data services work.
   a. Put the big datasets Landsat and Aster into separate projects that we know will work (this has been done).
   b. Put TRMM and 2 others into projects that will work (this was done Fall 1996).
   c. Use known DAAC and PI methods to handle other smaller datasets (this can be done at any time we want to).
   d. This leaves only AM1 data.
      • Scale out a backup to process AM1 that we know will work. I am told that this is being done.
      • Centralize some of the AM1 data flows in a better way.
      • Have only a few big DAACs. The rest can be small

Note: Recommendation 3 says we need EOSDIS for Landsat, AM1 and Sage. Landsat is mainly separate from ECS. We can surely easily do Sage again. That leaves AM1. Bretherton told me that two groups are working on backup methods to do it. So yes, we need an EOSDIS, but the real question is whether we need an ECS.

4. Scale costs by using dataflow methods.

Think of the most direct approach of saving AM1 data and making products.
This is like the many PI programs that are run on general computing systems.
I can’t even imagine how a direct approach would need a team of 700 or 800 people (that is where the money is).

5. Phase out ECS.

R. Quayle and J. London made this suggestion. My approach is to start in this direction by making ECS less necessary. It is really surprising to what extent this is already happening. This is happening in spite of the fact that people say that EOSDIS can not be changed before about 2002.
Option 3.1 (in our text) talks about reducing data centers. It seems to me that we should separate the question of how many small DAACs there are from the question of how to process AM1 data in the best way. I sense that ECS would like to simplify AM1 and I am almost certain that the backup procedures being scaled out will simplify it.


I believe that it is a big mistake to keep pushing NASA toward either a fully centralized or fully distributed approach. We will continue to need a mix. NASA PIs and others will need to look at 10 year data series that are easy to access. For the huge datasets, we should at least make them accessible on a computer such as Milt Halem’s at Goddard. We have not even done this.

Note 1: In the last week, I made some other EOSDIS notes, and will send them in a day or two.

Note 2: My big problem is that it is hard to understand how it was possible to make EOSDIS this difficult and this expensive.

7. The EOSDIS system.

It seems to me that a lot of the problem has been the focus on a big system that handles everything. Especially for large datasets, we need to focus on designing simpler systems that will work. These methods can still be powerful.

I am not convinced that ECS would be able to design a simpler system.

But the backup teams probably will.
Some Thoughts on Data Systems

Roy Jenne  
Sep 1998

An agency is periodically faced with the question of how to organize its data activities. This text has some ideas to consider when making these decisions.

1. There is reason to think the following about data systems within a large agency:
   a. A completely centralized data system is not a good idea
      - Not good for quality
      - Not good for costs
   b. A completely distributed (atomized) data system is not a good idea
      - Not good from standpoint of quality and ease of use
      - Not good for costs
      - Not good for long term survival of data.
      - However, some functions can be widely distributed such as:
        • Making products
        • Gathering certain types of data

2. To control costs, have to get the constraints right:
   a. Have to limit the desire to build the “big” software system.
   b. Have to limit monopolies – the ability of any one unit to be a sole source and to name its own price.
   c. Have to develop better tools to evaluate operating units in terms of both quality and cost. The costs are often left out of this equation.

3. We can’t put pressure on a discipline area to control its data costs if we force them to do the following:
   a. Make them use a high cost data service
   b. Or force them to use high cost methods, when simpler and faster methods are available.

4. Does a big, complex data system give a more functional data system?
   a. The big one can be 2 to 20 times more costly than alternatives, but it usually is not better.
   b. Is the big one easier for users? So far, experience says “No.”
   c. The common adage that “You pay for what you get” still has some validity, but it is largely not true in the big picture.

5. More automation and more software in data systems are usually sold on the basis of more functionality and saving cost. Is this sales pitch true?
   a. If we do not get the data strategy right, and the cost constraints right, more automation can easily lead to systems that are more costly and less functional.
   b. But sensible use of automation has the potential of actually giving us these good goals.

6. Make good data systems that have cost control.
   If our goal is to obtain good scientific data systems, and to also have cost control, what factors should we look at?
   a. We want to see a focus on “getting the job done” and not a focus on the big system.
   b. We want to see that there is proper knowledge and a focus on getting the big data flows right, with proper cost constraints. Then there is a good chance of getting both the timing right and the costs right.
Note: In the "big system approach" it is common not to see the above. All the talk is about the big system and what it can (in principle) do. That hides a lot of problems, and makes it impossible to make a good judgement on whether it is likely to work well.

7. Specs for a new data system:
The most common way to decide on specs for a new big data system is as follows:
   a. Hold a big meeting and ask everyone to say what they want the system to do.
   b. Compile a long list of these data system functions.
   c. Don't tell these users much, if any, about trade-offs and costs.
   d. Another group of people then designs the system and they also get to tell what it will cost.
Note: This is a common strategy, but it often leads to big systems and high costs.

8. Get many bits from satellite; there are trade-offs.
Do people make sensible trade-offs between how much data they plan to bring down from a satellite, and whether appropriate ground data systems are set up?
   a. With big data flows, we see cases where people just bring down as many bits as possible, without adequately thinking about how much extra science we get by flooding the system with data volume. The constraints need to be:
      -- You can't just dump costs over the wall to data systems.
      -- You don't get your mission if the costs are too high, including data.
   b. When we, in fact, really need a lot of bits, we need a data strategy that focuses directly on handling the big data flows, and controlling costs. If the focus is on a big software system, it is almost certain that we will not achieve the combined goals of quality and controlled cost.

9. The status of the present EOSDIS
   a. During about 1996-97, the common idea was that the ECS software would handle all of the making of data products in the DAACS, and it would also handle the delivery of data services for all of the AM-1 era of EOS satellites.
   b. But what has actually happened? Because of delays, backup plans to generate data products have been developed. Most PI teams have strongly desired (and argued for) a less cumbersome working environment for making products than they have when working with a "big data project."

We observe that, by Sep 1998, most data processing beyond the initial stages is in the hands of the PI teams and the DAACS, and not under ECS software. There is one big exception and that is Modis. It is complicated, and it is a large data flow. But in July 1998, I learned that the Modis team also wants to do their own processing.
Some Thoughts on Data Systems

An agency is periodically faced with the question of how to organize its data activities. This text has some ideas to consider when making these decisions.

There is reason to think the following about data systems within a large agency:

a. A completely centralized data system is not a good idea
   - Not good for quality
   - Not good for costs
b. A completely distributed (atomized) data system is not a good idea
   - Not good from standpoint of quality and ease of use
   - Not good for costs
   - Not good for long term survival of data.
   - However, some functions can be widely distributed such as:
     - Making products
     - Gathering certain types of data

2. To control costs, have to get the constraints right:
   a. Have to limit the desire to build the “big” software system.
   b. Have to limit monopolies – the ability of any one unit to be a sole source and to name its own price.
   c. Have to develop better tools to evaluate operating units in terms of both quality and cost. The costs are often left out of this equation.

We can’t put pressure on a discipline area to control its data costs if we force them to do the following:
   a. Make them use a high cost data service
   b. Or force them to use high cost methods, when simpler and faster methods are available.

4. Does a big, complex data system give a more functional data system?
   a. The big one can be 2 to 20 times more costly than alternatives, but it usually is not better.
   b. Is the big one easier for users? So far, experience says “No.”
   c. The common adage that “You pay for what you get” still has some validity, but it is largely not true in the big picture.

5. More automation and more software in data systems are usually sold on the basis of more functionality and saving costs. Is this sales pitch true?
   a. If we do not get the data strategy right, and the cost constraints right, more automation can easily lead to systems that are more costly and less functional.
   b. Sensible use of automation has the potential of actually giving us these good goals.

6. Make good data systems that have cost control.

If our goal is to obtain good scientific data systems, and to also have cost control, what factors should we look at?
   a. We want to see a focus on “getting the job done” and not a focus on the big system.
   b. We want to see that there is proper knowledge and a focus on getting the big data flows right, with proper cost constraints. Then there is a good chance of getting both the timing right and the costs right.
Note: In the "big system approach" it is common not to see the above. All the talk is about the big system and what it can (in principle) do. That hides a lot of problems, and makes it impossible to make a good judgement on whether it is likely to work well.

7. Specs for a new data system:
The most common way to decide on specs for a new big data system is as follows:
   a. Hold a big meeting and ask everyone to say what they want the system to do.
   b. Compile a long list of these data system functions.
   c. Don’t tell these users much, if any, about trade-offs and costs.
   d. Another group of people then designs the system and they also get to tell what it will cost.
   Note: This is a common strategy, but it often leads to big systems and high costs.

8. Get many bits from satellite; there are trade-offs.
   Do people make sensible trade-offs between how much data they plan to bring down from a satellite, and whether appropriate ground data systems are set up?
   a. With big data flows, we see cases where people just bring down as many bits as possible, without adequately thinking about how much extra science we get by flooding the system with data volume. The constraints need to be:
      - You can’t just dump costs over the wall to data systems.
      - You don’t get your mission if the costs are too high, including data.
   b. When we, in fact, really need a lot of bits, we need a data strategy that focuses directly on handling the big data flows, and controlling costs. If the focus is on a big software system, it is almost certain that we will not achieve the combined goals of quality and controlled cost.

9. The status of the present EOSDIS
   a. During about 1996-97, the common idea was that the ECS software would handle all of the making of data products in the DAACS, and it would also handle the delivery of data services for all of the AM-1 era of EOS satellites.
   b. But what has actually happened? Because of delays, backup plans to generate data products have been developed. Most PI teams have strongly desired (and argued for) a less cumbersome working environment for making products than they have when working with a "big data project."

We observe that, by Sep 1998, most data processing beyond the initial stages is in the hands of the PI teams and the DAACS, and not under ECS software. There is one big exception and that is Modis. It is complicated, and it is a large data flow. But in July 1998, I learned that the Modis team also wants to do their own processing.
1. THE PROBLEM OF SETTING REQUIREMENTS
There still seems to be a dominant idea about the way that setting requirements for a task and then doing the job of meeting the specs should work. One set of committees makes a list of all the requirements. Another group tries to figure out what the requirements really mean, they determine the methods to do the work, and they determine the costs.

We can’t completely avoid this model of doing the work, but it has the problem of not encouraging the right trades between benefits, costs, and methods. One needs to try to set up methods of doing business where these trade-offs are encouraged. When people had to do a whole satellite mission under a more or less fixed budget, these trade-offs were made.

When individual people buy a car or a house with their own money, these trade-offs are made, unless they are very rich. We should also note another factor about buying a car. There are few times when a salesperson will encourage us to buy a car that costs 70% as much, yet has 95% of the benefit. It is up to us to get facts to determine these tradeoffs.

2. MAKING MANAGEMENT DECISIONS WITHOUT KNOWING THE RANGE OF OPTIONS
Management often gets many inputs that are more like sales pitches than a real summary of the main options and costs. To help control these effects, management needs to develop a range of other inputs that help lead to better decisions. These can be examples of different ways of doing business and the results. The other inputs can lead to questions like “If Ford can build a car in X units of time, why does it take us 25% more time?”

There is IBM’s famous frustration (about 1976) with Seymour Cray’s ability to build a better supercomputer even though IBM had invested a lot more time and money in a similar project.

The same problems happen in data management in spades. NASA needs some better tools to sort out these issues. A good start has been made in some of the cost comparisons of different data activities that NASA has made.

3. MARCH 1998: “INTO THE BATTLE WITH NASA’s CHIEF”
This is a 6-page story about Goldin’s life on the job in NASA. He had been a senior manager in TRW where he set up innovative technology programs and then won a number of small satellite contracts with the USAF. He ran into a wall when he suggested that NASA build smaller earth observation satellites. NASA forced him to withdraw a paper at an international conference laying out the case for smaller and cheaper satellites. “I was told if you don’t withdraw that paper, you’ll never get any more NASA business.”

Goldin has been head of NASA since 1992, and he advocates smaller, cheaper, and faster missions. Naturally, a lot of people did not like these ideas. A young fan of Goldin’s at JPL said it was demoralizing to come out of grad school and find that planetary science was no longer a national priority. But now, according to this fan, with faster and smaller missions, it is a thrill a minute—something is always “hopping.” (Astronomy, March 1998, pp 40-46) We are getting many missions now, because the cost of each one is not as high as it once was.
Considerations for Data System Planning  
(by Roy Jenne, NCAR, 28 Sep 1998)

INTRODUCTION
EOSDIS has been evolving so that PIs will have much more control of making data products in both the eras of AM and PM satellites. This reduces the need for software from the central ECS project. Then the question arises about how the user services will be provided. The DAACs have built up a rather good capability for providing data services. I think that some DAACs will not need any help from ECS in this regard, but a few will. We need to hear the present arguments about user services. What about the future form of digital data systems in NASA? We need to invent a system that does not burden NASA with high cost methods, or with monopolies that are hard to work with. There are special needs for long term archives which require more infrastructure that the typical small PI group has. This means that we need central archives and DAACs. But there needs to be enough subtle competition (cost and quality), and enough data overlap that NASA has the right tools to maintain a good system that is also cost effective.

The cost of technology to handle data and make computations has come down a lot. However, the cost of many data services has gone up a lot. Only a fraction of this puzzle is explained by the larger amounts of data.

1. ONE VISION FOR US DATA SYSTEMS.
One observation of a few committees was that some government data systems in the US, such as NASA EOSDIS, had become far too centralized, cumbersome, and costly. When a data system becomes too big, there often are problems with the working relationships with PIs; and no one has the ability to manage it. There may be questions about whether it will work, and it may cost too much. The suggested solution was a radical decentralization of the data activities with small groups (maybe 200 of them) doing most of the data work. The long-term archives would be "digital libraries" that might often be in universities. These ideas were pushed by at least 2 NRC reports. I like the spirit behind some of these ideas, but they scare me. The data activities need enough long-term vision and enough institutional glue so that the data will survive, and the data services will exist. This much decentralization would very likely lead to data loss and high costs. Nevertheless, some of the best and most cost-effective work tends to come from small groups. How do we blend the different ideas to get a working solution, control monopolies, and control costs?

The bottom line is simple. There will probably be budget pressures. If good service and cost control can not be achieved with one type of organization for these services, we will have to move to another.

2. SUMMARY OF THE TASKS EOSDIS NEEDS TO ACCOMPLISH.
There are several tasks that EOSDIS must accomplish.
   a. Provide command and control of the satellites.
   b. Move the raw satellite data to selected archive points on the ground.
   c. Data needed for real-time operations must be delivered within about two or three hours. Most of the data do not have to be moved this fast.
   d. Prepare useful products from the satellite data and possibly prepare them again in 3 or 4 years when the methods have been improved.
   e. Put the primary data and the products into archives and help users obtain data.
   f. Develop a strategy for long-term data archival that focuses on providing basic data access (that is relatively easy for users), and has strongly controlled costs.

3. THE STATUS OF THE PRESENT EOSDIS
During about 1995-97, the common idea was that the ECS software would handle all of the making of data products in the DAACs, and it would also handle the delivery of data services for all of the AM-1 era of EOS satellites.
But what has actually happened? Because of software delays, backup plans to generate data products have been developed. Most PI teams have strongly desired (and argued for) a less cumbersome working environment for making products than they have when working with a “big data project.”

We observe that, by Sep 1998, most data processing beyond the initial stages is in the hands of the PI teams and the DAACs, and not under the ECS software. There is one big exception and that is Modis. It is complicated, and it is a large data flow. But in July 1998, I learned that the Modis team also has plans to do their own processing.

4. NEWS SEP 1998: LOCKHEED LANDS HUGE NASA DEAL.
The Lockheed aerospace firm beat Boeing to win the $3.44b, 10-year contract. Lockheed Martin will manage all of NASA’s data collection, measurement, and communications operations supporting the earth-orbiting satellites, planetary exploration and human space flight. Goldin said that this contract is projected to save taxpayers about $1.4 billion over 10 years (Boulder Camera, Sep 26, 1998).

Note: All the NASA data systems were going to be included in this contract, but Goldin was talked out of that over a year ago.

5. UP FRONT SUMMARY
We will discuss several issues such as:
   a. Issues about organizing data activities and controlling cost
      i. Not too centralized, not too dispersed
   b. Issues about making data products
      i. We are getting more PI control and that is good
   c. Big data flows and subsets
      In many cases, a carefully chosen 10% data subset will provide for 90% of the science needs.
   d. Providing data services
      No need to completely centralize this, and no need to standardize it.
   e. Long-term archives
      How do we store data for 50 years and give service at a controlled cost?
   f. Technology and cost comparison issues
      We can learn a lot from a selection of projects that handle a lot of data, and the costs seem quite low.
      Changes in technology continue to give us opportunities to do more for less money.
   g. General
      i. More of the total work of EOSDIS can now be done by the DAACs.
      ii. And there is a lot of work that can be done by PIs.
      iii. There are opportunities to give good service and save money. Any suggested solution has to be able to do this.

6. ISSUES ABOUT THE ORGANIZATION OF DATA ACTIVITIES. SEVERAL OF THESE ISSUES FOLLOW:

7. DATA SYSTEMS; HOW CENTRALIZED AND HOW DISTRIBUTED IS WISE?
   An agency is periodically faced with the question of how to organize its data activities. One question is how centralized the activities should be. There is reason to think the following about data systems within a large agency:
   a. A completely centralized data system is not a good idea
      i. Not good for quality
      ii. Not good for costs
   b. A completely distributed (atomized) data system is not a good idea
      i. Not good from standpoint of quality and ease of use
      ii. Not good for costs
iii. Not good for long term survival of data.

c. However, some functions can be widely distributed such as:
   i. Making products
   ii. Gathering certain types of data
   iii. Serving the education and policy communities

Note: I have saved a little information about what is happening in the commercial world. Some big companies are saving money by centralizing. Others take different strategies. It is interesting and frustrating to read about all of the problems in some of these large systems.

8. A STRATEGY TO CONTROL COSTS IN DATA SYSTEMS
We need a strategy that will enable our data systems to preserve the data, give good service, and control costs. If we are going to achieve these goals, we have to get the constraints right, and we need to avoid certain methods that can lead to excess complexity and high costs:
   a. Have to limit the desire to build the "big" software system.
   b. Have to limit monopolies – the ability of any one unit to be a sole source and to name its own price.
   c. Have to limit the tendency to become unnecessarily large.
   d. Have to develop better tools to evaluate operating units in terms of both quality and cost. The costs are often left out of this equation.

9. FACTORS OF CULTURE THAT AID IN COST CONTROL
Goldin has been the head of NASA for several years, and he has been a strong advocate of excellent projects and strong cost control (smaller, better, faster, cheaper). People are urged to develop technology and practices that will achieve these goals. As many articles point out, Goldin really has been changing the culture of NASA in this way.

In the area of data, I have not yet seen a real application of the skills that exist to achieve these goals. This does not say that nobody has tried. But we should all feel challenged to sort out these problems.

10. MAKE GOOD DATA SYSTEMS THAT HAVE COST CONTROL.
If our goal is to obtain good scientific data systems, and to also have cost control, what factors should we look at?
   a. We want to see a focus on “getting the job done” and not a focus on preparing the big computer software system.
   b. We want to see that there is proper knowledge and a focus on getting the big data flows right, with proper cost constraints. Then there is a good chance of getting both the timing right and the costs right.

Note: In the “big system approach” it is common that we do not see the above. All the talk is about the big software system and what it can (in principle) do. That hides a lot of problems, and makes it impossible to make a good judgement on whether it is likely to work well. It also costs more.

11. MAKING SPECIFICATIONS FOR A NEW DATA SYSTEM:
The most common way to decide on specs for a new big data system is as follows:
   a. Hold a big meeting and ask everyone to say what they want the system to do.
   b. Compile a long list of these data system functions.
   c. Don’t tell these users at higher management very much, if any, about trade-offs and costs.
   d. Another group of people then designs the system and they also get to tell what it will cost.

Note: This is a common strategy, but it has few constraints and it often leads to big systems and high costs. Compare this with a mission PI who must have an excellent mission plan and still control costs.
12. A VISION FOR STAFFING OF EOSDIS
There should be some vision for the needed staffing levels of EOSDIS. The history and the projections of staff size should be readily available along with budget information. Since most of the costs are people-costs, we can obtain an overall idea of costs by getting a feeling for the number of people in the work force. There should be separate sets of information for the DAACs, the ECS core project, the PI teams that develop products, and other activities.

In 1998 it appears to me that the plan is that the ECS staff will remain constant for 3 to 5 years, and the staffing in the DAACs will increase. Also, we will add many ESIPs. This vision does not lead to level costs or to reduced costs.

Information about staffing and budgets should be available each year, with a projection and a 5-year history.

13. DOES A BIG, COMPLEX DATA SYSTEM GIVE MORE USEFUL FUNCTIONS?
   a. The big one can be 2 to 20 times more costly than alternatives, but it usually is not better.
   b. Is the big one easier for users? So far, experience says “No.”
   c. The common adage that “You pay for what you get” still has some validity. However, when systems become unnecessarily big and complex, we can often pay more and get less in return.

14. SELECTED ISSUES ABOUT MAKING DATA PRODUCTS
NASA has moved to a system for making data products in which the PI teams have more control. This still leaves issues about where the computing is done, and the forms of automation that are used.

15. WHERE CAN PRODUCTS FOR BIG DATASETS BE MADE?
Sometimes each PI group that needs to prepare data products will want their own very large computer installation, paid for by NASA. A large computer setup will need staff experienced in fast computers, fast data flows, fast networks, storage devices, etc. It can become expensive if NASA has to pay for very many high-end computer facilities.

Suppose that PIs are fully in charge of preparing products from a large dataset. NASA still has the option of either buying them their own computer, or giving them time on another NASA computer. In either case, they will need some special systems support to help solve problems that they will encounter. Also, they will need access to a fast, high-capacity mass storage system that has a simple file management system.

The location of the computing resources might be one of the following:
   a. At the location of the main PIs (This is most likely for smaller datasets)
   b. At a NASA DAAC
   c. At a NASA computing facility that has a good design for fast data flows, simple file management, error control, and controlled costs.

Note: a center needs procedures to make data storage and data flow easy and fast. They need a track record of handling user needs. If they try to force excess data rules, complexity, or costs onto users, then they are not a viable place to handle large data flows or to make products.

16. MAKE DATA PRODUCTS: WE NEED COMPUTE SERVICES, NOT A “DATA SYSTEM.”
What is the working environment that PI teams, DAACs, or other groups need in order to develop products and give good data services? They do not need a full-blown “data system” developed by some other group.

What are the compute services that they need? A mass storage system is needed that can accept data files, keep them under error control, and deliver them back when needed. It can have both on-line and off-line components if that helps to control costs. Enough computing power is needed to accomplish the tasks. Good
data channels are needed from the mass store to the computer systems. Standard software libraries are needed to aid the scientific calculations.

17. ISSUES ABOUT BIG DATA FLOWS AND SUBSETS.

18. CONSTRAINTS ON DATA FLOWS AND DATA PRODUCTS FROM A SENSOR.
Suppose that a group of PIs wants to propose a mission. They want to define and build a sensor that will measure some quantities from space. They will often want as much scientific return as possible from the mission, which is good.

We need to ask ourselves whether the PI team is working under a set of constraints so that they will have to make reasonable trade-offs between the data flows, the amount of data captured, the products to be derived and the costs.

In the days of the old NASA project, the PI team often did everything from managing the satellite to making data products. They had to do all of this within a budget. This generally worked well, but perhaps 5% of the time the proper data did not get into long-term archives.

In the EOS era, NASA lost most of the budget constraints that were associated with the previous NASA projects. The data costs were thrown over the wall into the EOSDIS activity.

I was on the NRC CES committee for several years (June 1993-May 1997). This group handles issues for earth-looking satellites, mostly for NASA and NOAA. Kennel (head of NASA MTP) usually briefed the group twice a year. I think it was Oct of 1995 when he said that PI teams would be required to estimate the data costs of their proposed mission before approval. It is a good idea for the PIs to estimate the data costs, so that there is a constraint on total mission costs.

However, if we ask the mission PIs to estimate their data costs, they must have some freedom to make sensible choices about how to accomplish the data tasks. We must not force them to do the following:
  a. Make them use a high cost data service
  b. Or force them to use high cost methods when simpler and faster methods are available.
  c. Or force them to work in an environment with lots of rules and a high bureaucratic overhead.

19. DATA FOR LONG PERIOD, GLOBAL SCALE PROBLEMS (NEED 15 TO 30 YEARS OF DATA).
Under climate change research it is expected that many problems should be studied that have a global view, and 15 or 20 years of data. Such projects can be stopped by too much data volume, as well as by the lack of data. Therefore, PI teams need to think about what data they will offer to help such projects. If the data volume is very high, it may be necessary to produce a sub-sampled dataset.

Data volume is not a problem for many satellite instruments (altimeter, Sage, ozone, etc.). In a few cases, it is a problem. Some teams already have a good strategy to manage the problem of high volume:
  a. Ocean color radiances at 4 or 8 Km resolution as well as at 1 Km
  b. An ISCCP project for global clouds uses a sampled data set for calculations (either 8 or 24 Km resolution, not 1 Km). Data for 1983-1998 (16 years) are being processed. The original satellite data has a volume of about 8 to 10 Tbytes per year. The ISCCP archive of basic radiances is less than 1% of the original volume yet it permits PIs to calculate many products. It is also easy to recalculate the products when the volume is small.
  c. There are archives of global reanalysis data at full resolution. There are also global 2.5 degree resolution grids in pressure coordinates. The latter grids are lower in volume and they are used by 95% of the users.

Status of this issue in EOSDIS: for small datasets, and some midsize datasets, it is not an issue because new data subsets can be quickly generated when problems arise. For big datasets, this issue is likely to be a big
problem. It appears that PI teams have not given it sufficient attention to head off problems in advance. One of the goals of climate change science is to look at issues of climate change and variability. If we do not have datasets that will support these studies (at reasonable cost) we are missing the boat.

20. PROVIDING DATA SERVICES

21. THE TASK OF PROVIDING BASIC DATA SERVICES IN EOSDIS
The DAACs have gained a lot of experience in providing data services, using the Version 0 software that was developed.

I think that the plan was to use new ECS software in the DAACs to provide data services for AM-1 era products, and beyond. That plan was developed when the ECS software was also being used to generate the data products. But now the PI teams will usually make the products.

It appears to me that the ECS project has planned to bring the data products back into the fold of ECS software during the next few years. But that may be unnecessary if the DAACs are able to provide good data services using an extension of present methods. I want to hear more about the plans and arguments in this regard. The following subjects should be covered:

a. What are the cost implications of using Version 0 methods?
b. Are Version 0 methods simpler, and would maintenance cost less and be more reliable?
c. What are the arguments in favor of using ECS software for providing data services?
d. What do we see as the role of ECS software to provide data services on the 3 to 10 year time frame? At present, it does not make sense to me to use it for 3 to 5 years and then change to something else.

22. TAKE ADVANTAGE OF SELECTED SUPER COMPUTER CENTERS.
A supercomputer center, such as the one at Goddard, should be seen as one of the effective ways that users can access large quantities of EOS data (without moving it long distances). Such a center already has user services so that people scattered all over the USA can use the computers there. By setting up a high-speed local link, the Goddard supercomputers could ingest data from the mass store at the Goddard DAAC.

23. DON'T TRY TO SERVE 10,000 USERS, ONLY SERVE 500. NOT GOOD.
I have been in several data meetings were people believe that NASA data systems have grown too big and costly. One common argument to reduce costs is to say, "Don't try to serve 10,000 people; focus on serving the main 500 NASA PI's." We share the motivation behind these statements, but think that it is an unnecessary restriction. We can point to several examples where an effort of 0.3 FTE to 1.5 FTE can accomplish a large amount of work toward making big amounts of data available to a wide community in a manner that is very user friendly.

We can imagine a plan for data systems where the focus is on serving the core science missions, but where a little extra work opens many datasets to a much wider community. This is happening in the DAAC system.

Conclusion: keep a focus on the science mission, but do a little extra work to serve a broad community. People doing this extra work should have to demonstrate that they are doing it at low cost. We also know that there are high cost methods to achieve the added goals, often with small results.

24. DATA TASKS TO HELP EDUCATION AND POLICY PEOPLE.
I believe that NASA should always support some data work to accomplish goals in education and policy. NASA has a choice to strongly control costs and still accomplish a lot of these goals. One model to use for doing this work is to grant a number of ESIP-3 proposals at a low price per grant.

I have noticed that the NASA USRA activity (NASA plus universities) has done a lot of work to enhance education at a low cost. This is another model that can be used. More information is available about this
subject. I have been to too many meetings where it seems to be assumed that NASA either needs to spend a huge amount of money on these goals or do nothing. There are other good options as noted above.

25. DATA DISTRIBUTION (CENTRAL SERVER OR BLOCK DISTRIBUTION)
The vision of some people is that almost all data delivery would be from a central server. That way, data selections would be made and users would not have to move a high volume of data over the Internet. Some of this idea needs to be implemented, but it has drawbacks. The central data system may lose perspective on what is a reasonable amount of money for hardware and software to do this work. Also, people will have timing problems if most of the large archive is on tape (it takes 70-100 seconds to mount a tape and find a random dataset).

The other strategy is to distribute a lot of the data in bulk. This is commonly done on CD-ROMs and people often do a good job of providing software with the CD-ROM. The CD-ROMs are popular.

The idea of bulk data transfer can be extended further by using modern small tapes. In this way, various organizations could have large archives at home. In this future, different groups would develop simple software that others could use to operate on the data. This solution has some good properties:

a. There are backup copies of the data in many locations
b. Small groups are enabled to use their skills to develop software for data access and manipulation.
c. The structure of the basic data is not changed unless necessary.
d. The need for large amounts of software in a central site is curbed.
e. This gives NASA management some tools to control monopolies and to use other options to distribute data as they develop.

Most data archives have datasets that can be classified by size as follows:

a. Many small datasets
b. Some middle-sized datasets
c. A few huge ones

The huge datasets dominate the total volume in the archive, and they dominate storage costs. They are hard to fully distribute in bulk. There are strategies to cope with the special issues of huge datasets.

26. ISSUES ABOUT DAACs AND OTHER LONG TERM ARCHIVES

27. LONG TERM ARCHIVES FOR NASA DATA.
Until about 1996, I had just assumed that NASA would take care of its own data for many years although I heard a few contrary arguments. Now it appears that NOAA wants to do this job, and NASA is happy if it does not have to budget for this work.

There has been talk about turning over the NASA data to NOAA after it has been in the DAACs for 4 or 5 years. But the nation needs to support global change work and that often requires PIs to use data from 15-30 year periods, or longer. What is the vision for how a PI obtains 30 years of ozone data? Do they obtain the first 25 years from NOAA and the last 5 years from NASA? This seems overly cumbersome and costly. Does the data have to be moved from a NASA archive to a NOAA archive, or does it still sit in one place as it ages? Aren’t these ideas just removing functions from the DAACs that they could easily handle? What strategy leads to adequate services and good cost control for the USA, not just for any one agency? The arguments need to be put onto the table.

28. A DESIGN PHILOSOPHY FOR LONG TERM ARCHIVES AND DATA ACCESS
The system design that is used for National archives and data access should always keep the basic files of archive data in a form that is mostly independent of any software that may operate on these archives. It is important that the basic data should be in simple and efficient file structures. The reason that the data should
be separate from the software is that the various computer software systems will keep changing, and we should not put ourselves in the position of having to move the data or change the data every time these changes occur. A different philosophy is likely to lead to unsustainable costs, data corruption, and data loss.

We also note that even small groups, of size 2 people or 0.5 people, have come up with clever ideas to implement methods to make displays or to calculate products that are commonly used by meteorological modelers, as an example. We can expect that similar groups can develop small software tools to simplify the ingest of the basic data sets on a variety of computing platforms that exist during any one period of years. Data access routines that have a small amount of software, and which have little to learn are an example of this practice now in operation. But improvements can be made.

29. THE DAACs FOR EOSDIS

The distributed active archive centers (DAACs) for EOSDIS have been successful. They have proven that they can handle data from projects, archive it, and help users obtain data. It is less clear that there is enough budget control in the DAACs. Several DAACs that started with staff sizes of 20 or 30 are now up to 75. If the data system is distributed to 8 or 9 places, and if many of them become quite large in staff, then the total budget of the DAACs is going to become quite large. (We also note that the really large data flows from the EOS era have not yet started to arrive at the DAACs.)

The idea of having a distributed data system is usually that people will be in smaller groups, with a better focus on core discipline problems and better user interfaces. The bigger systems run into problems of focus, priority setting, and lack of knowledge of the science. Also, there are management problems associated with size. The big groups are more likely to focus on building software systems rather than on solving data problems.

There are two ways that the idea of a distributed data system can run into problems:

a. It is distributed across so many groups that it becomes (1) costly, (2) the essential functions for archiving are not done, and (3) it becomes harder for users to locate the data (the ESIP idea can get into these problems).

b. It is distributed to ten or so sites (such as DAACs) and each of these starts becoming a large organization (the DAAC system can get into this problem).

There is no reason that DAACs should all be the same size. It is likely that only 2 or 3 have a need to be of significant size. However, there are often social forces that try to bring the size of smaller units up to that of larger units. After all, they are all like members of a family; shouldn’t they be treated alike in staffing? The answer is “No.”

A DAAC often needs 3 years or more to really learn how to do its job. This means that we can’t replace DAACs on a rapid time cycle. However, NASA should feel free to assume that scientific discipline boundaries will not be too rigid for particular DAACs. This would give NASA some freedom about where to locate a new dataset. This would send a signal that a DAAC must compete on both quality and cost in order to get new work. And if another non-DAAC entity can do a certain job well, we should keep open the possibility that they would be used for the new work.

30. QUALITIES NEEDED FOR A DAAC OR LONG ARCHIVE

Data archives usually need to last 10 to 50 years, or longer. If an organization is going to take on this job, we should look for several qualities:

a. They have the skills and knowledge to set up and operate long-term archives.

b. Their main goal is to take care of the data and provide services, rather than to build systems. The latter is an enabling technology.

c. There is an institutional setting that is likely to permit a long-term home for the data.
d. The group feels a mission to handle data and provide services, and is not doing it because “that is where the money is.”

Note: A single professor or other PI often has the right skills to carry out many functions such as:
a. Run a network (such as the Pacific tide gage net)
b. Gather data (such as the set daily rainfall from Pacific atolls)
c. Gather many datasets of observations from a field experiment. There are several excellent examples of this work.
d. Prepare an atlas and dataset (such as tropical winds and clouds)
e. Prepare data from a satellite instrument (such as ocean heights and waves from an altimeter)

Once the data have been prepared by the PIs, there should also be a copy in longer-term archives.

It seems to me that the present role of the ESIP-2 groups is fuzzy. NASA has a lot of focused work it needs to do to develop data products, handle long-term archives, and give data services. Different PI groups can help with these tasks, but the methods of selection have to be focused on the core needs of NASA to get the “real work” done.

31. LONG TERM ARCHIVES
The four most important objectives for long term data storage are as follows:
a. Cost, cost, cost (it must be controlled)
b. Give service, but keep it simple and limited, but still good. Do not design a “big system.”
c. A design must be used that can achieve a 50-year archive. It must be able to accommodate changes in hardware and software, without falling into the traps of high costs and high complexity.
d. The bits have to be under error control, so that any changes can be detected.

Discussion:
Consider the problem of book libraries. There are thousands of new books each year. Society cannot afford to spend unlimited funds on libraries. Therefore, the new books and all of the old books have to be accommodated under constrained budgets.

There will also be a huge increase in the data volume in digital libraries. If the funding to run the libraries is allowed to increase in an uncontrolled manner, the funds will have to come out of current missions and science. It is not acceptable to build a “data empire” that prevents us from accomplishing all of our different goals. It is still true that taking care of the data will require some effort, and it will have costs.

The new data will be used more than the old data, but most of the old data will have to be carried along. The National Archives notes that the typical old book is used only once each 200 years, if it is not useful for genealogical studies. There should be data services for all data, but there should not be a high level of service for data that are seldom used.

31. WILL LARGE DATASETS STAY STUCK IN ARCHIVES?
Users have four problems in gaining access to large datasets:
♦ They often cost so much that they are not affordable.
♦ The chore of handling many, many tapes may become overwhelming.
♦ The data formats and access software may be excessively complicated and unnecessarily bulky.
♦ There may still be a lot of “warts” in the dataset that make it hard to use (bad date/times on some records, some data out of order, garbage scan lines, etc.)

We need multiple strategies to give users access to large datasets. First, we can give copies of the bits by making a copy from the mass store. Second, for popular datasets, we could prepare a set of master tapes that could be copied off-line and sent to users. Third, in some cases, it is better to bring the user to the data; users
of the main Goddard computers should be able to directly read data from the archives into their computer programs.

32. THE HUGE DATASETS; DO NOT MOVE THEM OFTEN.
When plans are made for archiving the huge datasets, and making products from them, we should keep in mind the cost factors that are involved.
   a. The computer that generates the products should be located close to the archive.
   b. The archive of basic data and high-volume products should not have to be physically moved in 3-5 years to another archive for long period storage.
   c. The data system needs to be designed for rapid data flows and low cost archival. Large amounts of slow software should never be in the paths of main data flows.
   d. The data system needs to be optimized for access to files of data of significant size. It is still possible to have other software that operates on the data in a more detailed way, but that should be judged on the basis of real needs and costs.

Note: There are many small and middle-sized datasets that are easy enough to move that we do not have to worry about these rules.

Examples: Landsat is a large dataset. The initial archive is located at the EROS DAAC. The products will be made there. And EROS will be the permanent archive. This makes sense.

33. SOME TECHNOLOGY AND COST COMPARISON ISSUES

34. SCALING COSTS BY ANALYSIS AND BY COMPARING DATA PROJECTS.
For a new data project, NASA needs to have several methods to help scale how much it might cost to handle the data. An analysis of the data flows and the volume of the archives helps to determine some of the costs. It also helps to know the costs a variety of past and existing data projects so that comparisons can be made. Therefore, we need to collect brief information about a variety of projects to help us make these comparisons.

The radical drop in the cost of technology to accomplish selected data tasks opens up new possibilities for handling the data. Summary information about the history of these changes, and projections for the next five years should always be available.

An initial set of some of the above types of information is available now, but it needs more work.

35. MORE AUTOMATION; WILL IT SAVE MONEY?
More automation and more software in data systems are usually sold on the basis of more functionality and saving cost. Is this sales pitch true?
   a. If we do not get the data strategy right, and the cost constraints right, more automation can easily lead to systems that are more costly and less functional.
   b. But a good use of automation is essential, and a sensible use of it has the potential of actually letting us achieve these good goals.

36. THERE ARE METHODS TO AUTOMATIC REPETITIVE TASKS TO REDUCE STAFF TIME.
In large programs there is often a need to automate the submission of many computer programs, and do this thousands of times on different sets of observed data. This is necessary to make products for EOS and it is necessary for the big reanalysis programs. There is existing experience in how to do this so that very little staff time is taken. In several large programs one person sets up the software and runs production (day and night). We need to make certain that different groups are aware of these methods so that experience can be shared when necessary.
THE CAPACITY OF HARD DISKS, CD-ROMS AND TAPES WILL IMPROVE.

Hard disks now seem very cheap compared with a few years back. A small disk only 3.5 inches in diameter now holds 10-18 Gbytes and only costs about $500 or $600. During the next 5 years, there will likely be another factor of 10 increase in the data volume held on a small cheap hard disk.

Tapes can be used to move large blocks of data that contain many small datasets and middle-sized datasets. The capacity of one tape has been about as follows on different years:

<table>
<thead>
<tr>
<th>Year</th>
<th>Tape Type</th>
<th>Data Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982</td>
<td>A 6250 BPI tape</td>
<td>125 MB</td>
</tr>
<tr>
<td>1986</td>
<td>A new IBM 3480 tape</td>
<td>200 MB</td>
</tr>
<tr>
<td>1995</td>
<td>New version of Exabyte</td>
<td>5 GB</td>
</tr>
<tr>
<td>1998</td>
<td>Some small tapes</td>
<td>20 or 30 GB</td>
</tr>
<tr>
<td>2000</td>
<td>Some small tapes</td>
<td>est. 100 GB</td>
</tr>
</tbody>
</table>

Therefore, our ability to move large blocks of data to users will increase a lot. We need to have strategies to stake advantages of these opportunities and that will decrease costs and enhance services.

There will also be improvements in data speeds on networks.

SELECTED READINGS ABOUT DATA ISSUES.

a. "Digital Libraries: Assembling the World’s Biggest Library on your desktop." (Science, 18 Sep 98, p 1784-86)

I was at the NRC Review of US-GCRP and EOS and EOSDIS in July 1995 at Scripps. I was bothered by some of the outcomes. I wrote this to collect my thoughts and to document some of the things going on in US data systems. It includes the following and more:
♦ How distributed should data systems be?
♦ Internal software for data systems; Standardize it?
♦ Lower cost methods to move large datasets.
♦ Big datasets are different.
♦ Can we simplify product generation in EOSDIS?
♦ Some small data activities that accomplish a lot.

c. "Readings about data formats," June 1994, by Jenne at NCAR. About 1.5 cm thick.
For small datasets, some experience shows that about 95% of the users will want the data in simple ASCII formats. This often leads to problems because most planners of big data systems do not think this way. This text has the following sections:
♦ Introduction to formats.
♦ Brief information about some common formats.
♦ Uniform data handling without common formats.
♦ Data delivery to users and to display systems.
♦ Attributes of desirable data structures.
♦ Etc.
Some Information about NASA ESDIS
(Roy Jenne, August 1998)

The purpose of this paper is to present some talking points about ESDIS. The main purpose has been to update my own thinking on the subject. The name used to be EOSDIS. Now they want to emphasize the idea of earth science (ES) rather than earth observing system (EOS).

1. Summary of the tasks ESDIS needs to accomplish.

There are several tasks that ESDIS must accomplish.

a. Provide command and control of the satellites.
b. Move the raw satellite data to selected archive points on the ground.
c. Data needed for real-time operations must be delivered within about two or three hours. Most of the data do not have this fast requirement.
d. Prepare useful products from the satellite data and possibly prepare them again in 3 or 4 years when the methods have been improved.
e. Put the primary data and the products into archives that can help users obtain data.
f. Develop a strategy for long-term data archival that focuses on providing basic data access (and relatively easy for users), and has strongly controlled costs.

2. An overall view of making products in NASA ESDIS

In 1996 the view of making products in ESDIS was that the Instrument PI teams would develop the program software to process the data streams (which they had not yet seen). The ESDIS project would imbed this PI software into other big software releases that would be run operationally at the NASA DAACS. The idea was that the preparation of products for later satellites (after 2000) might have more PI control, but for AM1, Landsat, etc. the current ESDIS plans would be used. Since that time, there has been a slow evolution toward letting PIs and the DAACS directly prepare the products.

We will now give a very brief overview of the various instruments for which data products are needed. Nearly all of these can be handled by methods that the community has used before (more distributed, more local control, more PI control).

a. Lots of smaller instruments that provide data. In many cases, there is a history in which PI’s have done a good job of calculating products from one or two earlier satellites that had similar instruments. And the costs were often modest. The science is complicated, but experts know how to make data products (atmospheric temperature, sea ice coverage, ocean winds, etc.).

- Ocean altimeter
- SAGE (aerosols, etc)
- BUV (ozone)
- ERB (radiation)
- Microwave (ice, etc.)
- Scatterometer (ocean wind)
- Atmospheric sounders
- Etc.

Note: we should look up the costs of some of the previous work that handled these data streams.
b. Landsat and Aster. The EROS data center and their PI teams have a long history of handling Landsat data. From March 1997-on, they have been telling me that they can definitely do Landsat & Aster by themselves. And their projected costs seem rather low and reasonable.

c. The TRMM (tropical rain mission). In 1996 the Goddard DAAC was given this mission (data from several instruments) to handle. The bird is up. I think things are working well.

d. Handle data from the CERES and Moppitt instruments. In fall 1997, these PI teams obtained permission to handle their own data streams. It was a relief for these teams to obtain permission, because they needed a more streamlined working environment to do their work.

e. Modis: after PI and DAAC methods are used to make products, we are mainly left with the Modis instrument, and its large data flows.

In July 1998, the MODIS PI team requested that they be allowed to process the data from their instrument. I do not know any details.

Note: we see a situation where the processing of data from AM1, Landsat, etc is gradually moving toward a mode of more PI control. I think that this is good, but it does bring up some other questions.

Summary: the making of products in ESDIS is much more distributed than 18 months ago. We endorse this change. But the changes in the making of products will require some new thinking about providing data services and maintaining archives.

3. Local control for making products in ESDIS.

A project in ESDIS needed to develop data products from one of the instruments. They wanted to develop and run the software themselves rather than work through the large ESDIS project to get this done. They knew there would be surprises and problems in the data stream to cope with. The needed an environment where they could rapidly prepare and implement changes in the way that they processed the data. This would not be possible in a large centralized approach.

There has been a trend toward giving PIs more local control in making products. Part of this was driven by necessity: the big project could not have enough of the necessary software ready in time. But a lot of it came from the instrument PI teams. They wanted a more streamlined approach and more local control.

4. The effect of data volume on complexity and cost.

Suppose that we had defined an EOS with essentially the same satellite sensors, but where the data volume was only 1% of the present EOS.

a. It would take just as much knowledge to prepare the algorithms to make the same number of products as now.

b. With the lower data volume, it is likely that the costs would never have been allowed to approach present ESDIS costs. There would have been a sense that there is a history where PIs have accomplished complex calculations at rather modest costs. Since the data volume would be similar to past experience, there would be an expectation that costs should be similar.

c. The big data volumes have scared people, and it is possible to use these partly unknown factors to slip into high cost plans.

d. If the ESDIS costs were mainly driven by data volume, we would expect that the hardware costs would be a considerable fraction of the total costs of the big ESDIS. Instead, they are only about 10 or 15% of the costs.
e. The total costs are mostly explained by the 600 or 700 staff that have been working on the big ESDIS core project. Then we have to ask whether there is a way to accomplish the necessary work, using methods where the plans can be more focused and where a smaller human effort is needed.

5. The new millenium program for satellite technology.

NASA started a program to develop technology that would help foster better and lower cost satellite sensors and satellite missions. During the technical development process, it was interesting to see that the advanced thinking was applied in the following ways:

- Design a sensor that would use less power (then the solar arrays could be smaller & cheaper)
- Struggle to reduce the weight and size of a sensor. This will help to reduce launch costs.
- Struggle to reduce the sensor cost, while keeping performance high.
- Struggle to reduce overall mission costs.

Note: for the ESDIS data systems, where is the equivalent of this type of thinking in the data system process?

6. Factors needed to reduce costs.

In cases where it is possible to both achieve good results and reduce costs, several factors are needed to make it happen.

- Need the will to try to reduce costs.
- Need the knowledge to make it happen (and the knowledge necessary is often distributed across a number of people).
- Need a process that will enable the lower cost approach to be planned.

7. Who gets to name the price of a data task?

If the selection of who will do data tasks is essentially a sole source activity that can name the price, then there is little incentive to control costs.

- Is there any competition? For a given task, one could get a bid from 2 or 3 centers. The lowest bid would not necessarily get the work; it depends on cost, quality, knowledge, and track record.
- Is there any way to scrub down costs for the higher priced activities?
- Is there a history of selected previous costs to permit comparisons of costs and activities?

8. Command and control of a satellite.

One function of ESDIS is to control the satellites. In mid-1998, there were problems found in the new software to do this. The AM1 launch was delayed.

In about 1996 a small group considered ways to achieve good results in controlling a satellite while also lowering the cost. There are experts in this area that could help in designing new systems and in reviewing existing practices. One person told me that it is often possible to find ways to save 20 or 30% of the cost. By making changes in the overall approach, it is sometimes possible to achieve much more radical decreases in cost (even a factor of ten, they said).

A white paper should be available that discusses the factors that influence quality and cost. Comparisons of staffing and costs for several missions should be included.
9. The projects and costs in the Pathfinder program (during ~1993-1998?).

The idea of the NOAA/NASA Pathfinder program was to calculate a range of satellite data products, based on satellite data collected from about 1979-on. But state-of-the-art algorithms were developed and applied. This has been a very successful project, funded by NASA. More information about the Pathfinder projects and the costs is available. One part of the program had to take steps to make the satellite data really available for processing. Otherwise, the cost and complexity of data access was often too high for PI’s to readily handle.

10. Is a big staff needed to produce high-volume products?

In the next sections we will review several examples of large complex projects, where the work to develop and operate the necessary controlling software is handled by a very small staff.

Therefore, if the staff members are carefully chosen, and if the methods that are used follow the spirit of the examples, the staff does not have to be big.

11. Automated software methods to produce products in big projects.

One task of ESDIS is to manage many large streams of data and to process them through computers to make products. It is useful to review some examples of how people have handled these tasks in some large, complicated projects.

There are a number of very large projects that have developed experience in preparing automated sequences of complex software programs which automate the running of thousands of programs every day, over a span of years. Considering the complexity of the tasks, it is surprising that the staffing needs can be as low as they often are. Each big job is going to be different from other problems, but we can still use the experience that has been developed to help us define methods and scale the staff requirements. Consider two examples.

a. Three large reanalysis projects.

The world has three large reanalysis projects (NCEP/NCAR, ECMWF, and NASA/DAO). There is an associated large project to gather and prepare the global daily observations that are needed (raobs, aircraft, land surface, ocean surface, satellite sounders, and satellite cloud winds). The reanalysis projects are complex efforts that are based on the cutting edge of research.

All of these projects produce global analyses, each 6-hours, for many levels in the atmosphere, and for many years. Each 6-hour period means using one of the world’s best analysis methods (data assimilation), together with a “state-of-the-art” forecast model. Many steps are needed: prepare the observations for the 6-hour period, do QC, calculate many, many different variables in several forms for the users, prepare “alarm bells” to help warn operations people if problems have occurred.

Each real day, this production is able to accomplish 10 to 40 days of new analyses, done for each 6-hours. Every day (and night) hundreds of complex program processes are automatically triggered to do the production.

Usually one or two people handle this production, and write other programs to understand the nasty problems that sometimes arise. The actual programs for analysis methods and forecasts have usually been developed over a number of years, using research inputs from a group of people. The core group to handle production, data flow, problems, and documentation in reanalysis is often 2 to 5 people, with periodic help from a wider set of experts.
b. The COADS surface marine project.

This project prepares the world's best set of surface marine observations. Some years ago, we realized that during a 7-month production period, we would have to run about 5000 different computer jobs. There is no way to set up and execute this many jobs without making mistakes and taking a lot of people time. So we set up an automation procedure that is placed on top of normal computer software systems. This did not take very long. It gives the project quick control to handle problems, because the necessary software is not buried in a larger amount of complex software. The project worked well, but it certainly kept us busy!

c. Conclusions.

PIs usually have their own methods to develop data products for smaller datasets and these methods have been very effective. There is typically no need to try to centralize all of this work. When the work is centralized too much, the experience is that it leads to high costs, and too much time spent in interactions. There is also a lack of sufficient PI control necessary to respond to real-life problems and still get the job done. People should be aware of methods that have been used to help automate many functions in data processing.

12. The project director for ESDIS is changing jobs.

The director of NASA Goddard announced (3 August 1998) that Rick Obenschain (Project Director of ESDIS) has agreed to become the first Center Chief of the Electrical Systems Center (Code 560). This is one of the largest engineering centers within AETD. Dorothy (Dolly) Perkins will take over Rick's present ESDIS management tasks.

Rick has been in the ESDIS job about two years. He is a very active manager and very oriented toward results. It is hard to know what effect these changes will have on the ESDIS project.

13. Understand useful methods and costs for big data flows.

To help understand the costs of computing and storage for large data flows, it helps to have selected information about the costs of projects that have done similar work. Some of the comparisons that can be developed are as follows:

a. Compare experience with other big data flows
   - Landsat data (handled by EROS and Goddard)
   - The SAR data handled by Fairbanks (high resolution ice information)
   - The long experience at University of Wisconsin in handling GOES data
   - Projects in Europe to handle data from ERS-1, ERS-2, and related satellites

b. Big computer centers
   - Get a brief history of computing power, data flow, and total archives at selected computer centers
   - Have data for a few of the main production centers (GFDL, NCAR, NCEP, Goddard, and ECMWF).

14. Where can products for big datasets be made?

Sometimes each PI group that needs to prepare data products will want their own very large computer installation, paid for by NASA. A large computer setup will need staff experienced in fast computers, fast data flows, fast networks, storage devices, etc. It can become expensive if NASA has to pay for very many high-end computer facilities.
Suppose that PIs are fully in charge of preparing products from a large dataset. NASA still has the option of either buying them their own computer, or giving them time on another NASA computer. In either case, they will need some special systems support to help solve problems that they will encounter. Also, they will need access to a fast, high-capacity mass storage system that has a simple file management system.

The location of the computing resources might be one of the following:

a. At the location of the main PIs
b. At a NASA DAAC
c. At a NASA computing facility that has a good design for fast data flows, simple file management, error control, and controlled costs.

Note: a center needs procedures to make data storage and data flow easy and fast. They need a track record of handling user needs. If they try to force excess data rules, complexity, or costs onto users, then they are not a viable place to handle large data flows or to make products.

15. The huge datasets; do not move them often.

When plans are made for archiving the huge datasets, and making products from them, we should keep in mind the cost factors that are involved.

a. The computer that generates the products should be located close to the archive.
b. The archive of basic data and high-volume products should not have to be physically moved in 3-5 years to another archive for long period storage.
c. The data system needs to be designed for rapid data flows and low cost archival. Large amounts of slow software should never be in the paths of main data flows.
d. The data system needs to be optimized for access to files of data of significant size. It is still possible to have other software that operates on the data in a more detailed way, but that should be judged on the basis of real needs and costs.

Note: There are many small and middle-sized datasets that are easy enough to move that we do not have to worry about these rules.

Examples: Landsat is a large dataset. The initial archive is located at the EROS DAAC. The products will be made there. And EROS will be the permanent archive. This makes sense.

16. Will data be ready on time? Will data be lost?

It is important to summarize a little of the 20-year history of archive strategy questions within NASA. For many years the data from a sensor was sent to the appropriate PI groups for archival, and to make products. Then the products and the data were gathered back into the data center that was organized within Milt Halem’s computer center at Goddard. As late as 1990, this center had a comprehensive set of archives for data from both the earth science disciplines (from NASA satellites), and from space sciences.

In some cases there were problems. A PI might hold onto the data too long before delivering products. In few cases, archives were lost. When I reviewed the archives in 1988, it appeared that nearly all of the archive of basic radiance data and products were there. There were problems in lack of access to huge amounts of early GOES data at Goddard, and the THIR data was not being properly cared for (still true).

About 1987, NASA had an opportunity to take care of the THIR data (Copy 30,000 tapes to 3000 new tapes). The NASA center only needed help to buy new media ($30,000). The chance to do this job for $30,000 when it would usually cost $750 K to $2100 K was lost. It is unknown why NASA was not able to respond to an opportunity like this. In any case, this problem had nothing to do with poor planning by a single PI team.
We note that the problems of timely access to data products can be addressed by appropriate agreements with PI's. Also, no PI team should have monopoly rights to any primary data stream. There also should be a copy of data in a central archive so that the data is secure, and so that is can not be used for “monopoly bargaining.”

Conclusion: we see no reason why the data processing for many small and mid-sized datasets should not be handled by the PI's. But the results must get into other archives in a timely way.

17. The DAACs for ESDIS.

The distributed active archive centers (DAACs) for ESDIS have been successful. They have proven that they can handle data from projects, archive it, and help users obtain data. It is less clear that there is enough budget control in the DAACs. Several DAACs that started with staff sizes of 20 or 30 are now up to 75. If the data system is distributed to 8 or 9 places, and if many of them become quite large in staff, then the total budget of the DAACs is going to become quite large.

The idea of having a distributed data system is usually that people will be in smaller groups, with a better focus on core discipline problems and better users interfaces. The bigger systems run into problems of focus, priority setting, and lack of knowledge of the science. Also, there are management problems associated with size. The big groups are more likely to focus on building software systems rather than on solving data problems.

There are two ways that the idea of a distributed data system can run into problems:

a. It is distributed across so many groups that it becomes (1) costly, (2) the essential functions for archiving are not done, and (3) it becomes harder for users to locate the data (the ESIP idea can get into this problem).

b. It is distributed to ten or so sites (such as DAACs) and each of these starts becoming a large organization (the DAAC system can get into this problem).

There is no reason that DAACs should all be the same size. It is likely that only 2 or 3 have a need to be of significant size. However, there are often social forces that try to bring the size of smaller units up to that of larger units.

A DAAC often needs 3 years or more to really learn how to do its job. This means that we can’t replace DAACs on a rapid time cycle. However, NASA should feel free to assume that discipline boundaries will be too rigid for particular DAACs. This would send a signal that a DAAC must compete on both quality and cost in order to get new work. And if another non-DAAC entity can do a certain job well, we could keep open the possibility that they would be used.

18. Don’t try to serve 10,000 users, only serve 500.

I have been in several data meetings were people believe that NASA data systems have grown too big and costly. One common argument to reduce costs is to say, “Don’t try to serve 10,000 people; focus on serving the main 500 NASA PI’s.” We share the motivation behind these statements, but think that it is an unnecessary restriction. We can point to several examples where an effort of 0.3 FTE to 1.5 FTE can accomplish a large amount of work toward making large amounts of data available to a wide community in a manner that is very user friendly.

We can imagine a plan for data systems where the focus is on serving the core science missions, but where a little extra work opens many datasets to a much wider community. This is happening in the DAAC system.

Conclusion: keep a focus on the science mission, but do a little extra work to serve a broad community. People doing this extra work should have to demonstrate that they are doing it at low cost. We also know that there are high cost methods to achieve the added goals, often with low results.
19. EOS data for K-12 education and policy makers.

The need for data and information for K-12 education for policy people has come up in several meetings. There is often a fear that a large amount of money will be spent in a diffuse way on these topics. The fear is justified, but there is a way to handle the need for this activity in something other than an “all-or-nothing” approach.

First, the spending can be rather modest. For example, people might argue about whether the pot should be $1 million or $3 million per year in size. There are many small groups in universities, etc. that can accomplish a lot of work using grants that only range from about $30K to $70K per year in size. A group of universities working within the NASA-USRA program has been effective. In 1997, NASA gave contracts to some small groups working within the ESIP, type-3 program to accomplish this sort of work. In addition, it should be expected that other PI groups and data centers will use a small (but important) fraction of their time to help education.

Conclusion: by establishing the limited pools of money for grants, and by using mechanisms that are now working, the education community should have a range of data and information products for use by teachers and students.

By about 1990 NASA had decided to form a distributed DAAC system. They also decided that there was not enough in-house experience to handle all of the EOS data, so they looked to a central contract – awarded in about 1992 to Hughes.

In the early 1990’s, the DAAC’s were becoming more experienced so that they could handle many tasks once intended for the central contract. The PI teams were also able to handle more jobs, either by themselves or by working with a DAAC. For a given cost, the amount of computing and data handling that can be accomplished has increased by many factors over the past 10 and 20 years. This improvement in technical capability means that PI teams and DAACs can handle bigger tasks than was previous possible.

These changes give the opportunity to do more work at less cost, but only if the strategy is modified.

20. How much data can users handle?

Possible users of data can be pushed away by any of three factors:
   a. Problems of handling the data
      • Too many pieces of media
      • High local costs to handle and store that much data.
   b. High cost to obtain a copy of the data
   c. Complexity of formats and data access software

Most users can readily handle small files in ASCII format. Many of these files will be obtained by Internet or CD-ROM.

Scientific users can readily handle costs up to $800-$1000 for significant amounts of data. Costs of $3000-$5000 or more would often have to be built into a contract or a cooperative plan to get data. People are also limited by too much fuss to handle many pieces of media, or by the cost of tape drives to read the media. In the mid-1980’s an appropriate lower cost tape drive often cost $18,000. Now there are options with much better capability that cost $1000 to $6000. Most PI’s will readily handle 10-30 pieces of media and some realize that they must be willing to handle more than that.

Conclusion: any plans that assume that the typical PI will be able and willing to handle 1000 pieces of media are probably due for a surprise.

21. How to run a satellite project to obtain quality and cost control.

To be useful, a satellite mission has to accomplish several things:
   a. Bring down data from the satellite.
b. Solve (or work around) problems that will occur in the satellite system or in the data stream.

c. Prepare sensible data products. Give reasonable data access for users.

d. Save much of the data in long period archives.

Most satellite missions were in the position where they had to accomplish all of these on a limited budget. They made the tradeoffs necessary to give results that were generally good. But there were complaints that when there were hardware problems, then the money was used to fix these, and the data budget was shortchanged too much. Some of those complaints were true, but if one looks at the archives, it appears that most of the necessary products were somehow created.

If the data activity is organized completely separately from the flight projects (and from PI's), it is much easier to get budgets that have run away from sanity. One can then see several problems, which include:

(1) Lack of focus
(2) Unreasonable requirements
(3) Big systems rather than streamlined processes
(4) Sometimes a focus on software systems rather than on streamlined processes
(5) Lack of interest in cost control
(6) Too often there is too much empire building.

Some of the problems do, in fact, result from bad motives. But in very large projects, there seems to be an effect of things becoming worse than necessary for no very good reason.

22. Data for long period, global scale problems.

Under climate change research it is expected that many problems should be studied that have a global view, and 15 or twenty years of data. Such projects can be stopped by too much data volume, as well as by the lack of data. Therefore, PI teams need to think about what data they will offer to help such projects. If the data volume is very high, it may be necessary to produce a sub-sampled dataset.

Data volume is not a problem for many satellite instruments (altimeter, Sage, ozone, etc.). In a few cases, it is a problem. Some teams already have a good strategy to manage the problem of high volume:

a. Ocean color radiances at 4 or 8 Km resolution as well as at 1 Km
b. An ISCCP project for global clouds uses a sampled data set for calculations (either 8 or 24 Km resolution, not 1 Km). Data for 1983-on are being processed.

Status of this issue in ESDIS: for small datasets, and some midsize datasets, it is not an issue because some new data subsets can be quickly generated when problems arise. For big datasets, this issue is likely to be a big problem. It appears that PI teams have not given it sufficient attention to head off problems in advance. One of the goals of climate change science is to look at issues of climate change and variability. If we do not have datasets that will support these studies (at reasonable cost) we are missing the boat.
LEVELS OF DATA SUPPORT

We will consider different types of support that are offered to users of data. Depending on the main data needs to accomplish different scientific or applied goals, we often have to shift our priorities on what computing functions to develop. For given amounts of resources, we often have to decide whether it is more important to make more of the basic data available (and clean up its problems), or to provide more ways to display the data. Different levels of support for data use are:

1. The User Has Access to a Dataset on Tape
   - The user obtains the basic data on tape or other media;
   - Format information is given; other help is limited;
   - No access program to unpack data;
   - Perhaps wait hours for tape to be mounted;
   - Even to provide this amount of support, a considerable effort needs to be made to gather the data and organize it into reasonable datasets.

2. Dataset on Tape or Mass Storage System:
   This is like #1, only an access program is sent with the data. The user also has more variety of software, such as general graphics routines, to use.
   - Good format and other info;
   - Have an access program to unpack arrays for computing;
   - Have a selection of good inventories;
   - Data can be on-line within 2 to 4 minutes;
   - User computes with the data;
   - User has general display packages and other tools to use (data blocking, bit manipulation, etc);
   - User has many math and statistical routines to use;
   - Some data-specific software routines are available:
     - code to make common grid transformations;
     - code for common display routines; These may even include complicated routines such as: take weather data from 6 arrays (one for each hr) and show me a movie of how winds and weather have changed in the last 6 hours.
   - Project to gather display software;

Many universities and laboratories develop selected display routines. It may be useful to have a small activity that gathers together these packages so that they can be generally used. Usually a laboratory gradually prepares
display functions for a given type of data. A modest proposal to NSF from a lab. to gather such software and make it available to a larger community, often should be supported.

Advantages: The user has full flexibility to do anything that he/she wants. He is able to have plots the way he wants them, not just accept what a system offers. He can also use "canned" displays when they are available. Most science will be accomplished in this mode. But, the user has to know how to program.

3. Like #2, Only the User is Remote From the Archive

- For complicated formats, a portable access program is made available;
- The existence of selected software is made known for data transformations, etc;
- User receives data on tape, cartridge, CDrom, etc; Small datasets or selected subsets may be obtained via communications.

4. Interactive Look at the Data, Using a System:

There are some well-defined cases for which many users need about the same type of display products. Also, instead of calling a subroutine, the user may just want to "hit a button" to cause the following to happen:

- Show me the present surface pressure map for my region;
- Put a dashed 500 mb map on top of the surface pressure;
- Show me hourly satellite pictures in motion for the last 24 hours;
- Plot a upper-air rawinsonde sounding;
- Draw an upper-air cross section through the reports from five cities that I specify;
- Show me the plot of 24-hour rainfall on top of the present 700 mb map; use different colors;
- Show me plotted surface winds and type of weather for stations in my area. Do the same, but show me each hour for the last 8 hours in a motion sequence.

For other types of data, examples might be:

- Show me energy use by the US, UK, Germany and Japan, plotted on the same time scale;
- Display a plot of geographic structures through the Appalachian mountains;
- Display fish catches by continents, for the past 40 years;
- Display the depth of the ocean.

These systems are very powerful and useful when the goals are rather well focused. They also tend to be expensive to create. The Univ. of Wisconsin has been a pioneer in developing such systems, with their McIdas system which is
being sold worldwide.

The systems also may be very limited. A user sees a display and says "That is very nice, but I would really like to see ... (another related display)." But, usually the system organization doesn't permit the user to use basic computer data and graphics routines to create his own displays. Thus, the user gives up, uses the methods of #2 above or tries to talk the systems people into adding his display.

5. A Workstation Environment that Permits Some Mix of #2 and #4

- The user can use programming and software tools to create his own displays. Thus, he is not restricted by the limitations of the system.

- If some canned displays (as in #4) have been developed, a person can use them by "hitting a button."

- A PC has programs to do spreadsheet analysis, word processing, play games, etc. It would be desirable to have similar options to make data displays:
  - load a program to do a certain type of weather display;
  - use a program to show paleoclimate changes over the last 200,000 years;
  - a translation routine often will be needed to take the archive data and put it into a structure needed by a given display system.

- Inputs will be tape, CDrom, communications, etc.

6. Discussion

The approach used should take account of how much a given type of data will be used, whether there is a common set of displays, and what the costs are. Many sets are rarely used; for these, level #1 (probably adding an access routine) is what makes sense.

The canned system approach (#4) makes sense only when the objectives can be focused rather well and where the displays are used a lot. This happens in McIdas. It also happens in the types of weather displays and overlays needed in a forecast center. It happens for public TV weather forecasts.

The environment noted by #5 is the one that we hope will slowly evolve.

One should not try to force all archive formats to be the same as the structure that a particular display system wants. We don't want to restructure all archive data every time a new display-level idea comes along. There is no necessity to change the archive formats in order to achieve these goals of data access.
Beyond EOSDIS

These are some comments by Roy Jenne. They refer to the draft by Elaine Hansen (Feb 28, 1994).

1. The ability of a PI to obtain data will improve greatly. The ability to receive data by communications will grow as most people become connected to a network and the data speeds become faster. There will be a variety of evolving networks, but they will be linked together (like Internet). Also, the ability to use media to distribute large amounts of data to local archives will improve dramatically (present examples are CD-ROM, DAT tapes, and Exabyte tapes). Both methods (communications and media) will continue. The method used by a PI to obtain some data will depend on convenience, cost, and the software aids that are available.

2. Some PIs will analyze in situ and satellite data from recent observations. Other users will analyze data series that are 5 to 100 years in length or longer. It must be possible to access appropriate archives so that all of the main research and applications needs are supported.

3. All users do not need all of the original data (often high volume). There will be sampled data subsets and smaller sets of data products that allow users to find a data input that is appropriate for their data needs, processing capability, and budget.

How will costs of data services be managed? Many of the data services will be provided by smaller groups where costs and performance can be compared between groups. The activities of the groups will be determined more by competition (the way grocery stores compete) than by many rules and committee meetings. If government pays for the cost of preparing data services, and if users pay nothing, there may still be problems in managing costs.
Interface Standards

Probably should not be supported until q. unless:

1) The idea is clarified as to what it actually means
2) Whether it is necessary
3) How much extra work it adds
4) How much complexity it adds
5) It is connected to a vision of how it could have prevented some of the problems of the past, and whether it would lead to a better future

Roy Tenne
Finding a Smooth Future for EOSDIS

1. When the data comes down
   - We need to handle it
   - Some normal troubles, no more

2. Inventing good systems
   - That work better and cost less
   - Using our best talent to help

3. A strategy for all the functions
   - Data products
   - Data services
   - Long-term archives

Slides for NASA plans

Roy Jenne
NCAR
Oct 1998
EOSDIS Has Evolved A Lot

1. From 1994 to 1998
   - DAACs much more capable in ’98
   - DAACs need less help from ECS

2. Making products
   - In 1996-97, DAACs got the job of doing products for some data
   - The 1998 trend: PI teams have main focus for products
     - This has been the plan for 2 years for future
     - The plan has been moved up
   - So PIs plus DAACs will do products
     - Software for products from ECS
     - There are simpler ways to automate production

3. For a big software system
   - Hard to check it out
   - Hard to know if it will work
   - We need reliable, more limited software

4. Questions on data services and long-term archives
The 10-Year Trends of Data Activities

- Cost of computing (gone down a lot)

- Cost of storage (gone down a lot)

- Cost of data management
  - Gone up a lot
  - Bigger data volume explains only part of this problem
What are Staff Estimates for EOSDIS?

♦ These are my impressions.
♦ What are the estimates?
♦ Note: An increasing staff does not lead to either stable or decreasing budgets.

Staff in ECS Core Project

Staff in DAACs

Plus
♦ Staff in PI Teams (They will do much more of the work to prepare products.)
♦ Staff in ESIPs, etc. (The buzz has been that there would be many more of these.)
ESIP – 2’s
(Earth Science Information Partners)

- How many tasks like ESIP –2’s
  - Perhaps 500 to 1000
- Costs at $1M each
  - $500M+ per year
  - This would really break the bank

Need:

1) A few places that group many tasks (like DAACs, other options)
   And have institutional longevity and cost control.

2) Places only 2 to 5 staff in size
   - but still do lots of tasks

3) PI operations
   - often small (1 to 3 in staff)
We Need Ways To Get Cost Trade-Offs Right

1. For a proposed project by PIs
   • NASA needs a cost estimate up front
   • This forces people to do necessary cost trades
   • If it costs too much; maybe say no to projects
     (I think NASA started doing this.)

2. Unnecessary requirements on data providers
   • They increase costs a lot
   • Usually, they are not needed

3. NASA needs second sources for much work
   • Otherwise you have monopolies
   • And hard to measure costs

4. Bulk data transfer methods
   • Ways to ship data in bulk
   • And others can help with access tools
Cost to Prepare Data from PI Instrument
(Information from a PI Team)

Consider the Effects of Excess Requirements

1. PI team worked up costs a year ago

2. They recently did costs again
   • After lots of requirements and standards from the big project
   • The costs were almost doubled
Data Services; Help People Get Data
(Include Long-Term Archives)

Case I

- ECS
- DAACs


Case II

- DAACs
- ECS

ESIPs 4
Uni. Libraries
NOAA
daacs

Case III

- DAACs
- ECS

The lowest cost option

Distinct efforts
NOAA
PROVIDING DATA SERVICES

Help 10,000 more PIs
5%

Help the 500 PIs of EOS Core
92%

Help Education

Info for policy people
1.0%

A small part of the budget can accomplish a lot.
Aspects of Data Services

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Percent of Datasets</th>
<th>Percent of Volume</th>
<th>Service Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Huge size</td>
<td>1%</td>
<td>90%</td>
<td>Basic access, reduce cost</td>
</tr>
<tr>
<td>Mid size</td>
<td>5%</td>
<td>8%</td>
<td>Basic access, bulk distribution</td>
</tr>
<tr>
<td>Small size</td>
<td>94%</td>
<td>2%</td>
<td>Mix (CD-ROMs, bulk, web)</td>
</tr>
</tbody>
</table>

Some observations:

♦ Some groups (size 0.5 to 2 people) make clever access methods

♦ So if we give them bulk data, they can do good work

♦ Many groups want to build big systems
  - Any number of groups could be funded
  - Any amount of money could be spent
Long-Term Archives (15 to 50 Years)

1. Where does the data sit?
   ♦ In NASA for 5 years?
   ♦ Then in NOAA?
   ♦ Does the data have to be moved?

2. Give two plans
   a. Do basic archives
      ♦ Give only basic service
      ♦ Be ready for bulk delivery of blocks of data
      ♦ Staff and costs?
   b. Do basic archives, plus more
      ♦ With much on-line access, etc.
      ♦ And much software
      ♦ Staff and costs?

3. Questions
   a. Plan for error protection?
   b. Methods to determine costs?
   c. Methods to scrub costs?
   d. Plan to farm out part of work?

Roy Jenne
Oct 1998
Compare Costs of NASA Data Systems  
(Data from NASA, Oct 98)

1. Data Systems for all Space Sciences  
a. Planetary, Astronomy, and Space Physics  
b. They cost $39m each year, total.

2. The EOSDIS for Earth Sciences  
a. Costs $230m a year  
b. Trouble getting it to work.

3. Comparison  
a. $230m vs. $39m  
b. Volume of data much higher for EOSDIS  
c. But volume only explains a fraction of the cost difference  
d. The Space Science example is a rather good one.

Roy Jenne  
Oct 1998
Guide for Future Strategy

1. Handle data from many smaller projects.
   This includes many projects like ozone, altimeter, Sage, etc. In many cases, we have a history of good data preparation by PIs in the past. There is no reason that similar data effort for new data should cost a lot more than in the past.
   ➢ Find what some past projects cost
   ➢ Compare costs with similar data activities in Space Sciences

2. Data such as Landsat
   The main task is to store and retrieve scenes of data.
   ➢ There is experience with previous Landsat data
   ➢ Compare costs with some commercial places (Spot, new USA activities)
   ➢ Do a data flow analysis to help determine costs; look at technology costs and trends.
   ➢ Landsat-6 failed. The planned data costs were very high; a mistake. We should record the history.

3. Some huge datasets (e.g. Modis)
   ➢ Preparing the algorithms is probably similar in effort to small size datasets
   ➢ The cost of automating the steps handling the operations is either small or large, depending on methods.
   ➢ Estimate the costs of computing and storage gear to handle the big data flows and store data.
     a. Compare with what is known about hardware costs, and the history and trends.
     b. Compare with other projects that handle a lot of data.

4. Examine the constraints on different groups.
   Is the system encouraging people to manage costs or to increase costs?
   Are there any penalties for increasing the costs?

Ray Pomer
22 Oct 1998
The DAACs of EOSDIS

Greg Hunolt, April 28, 1997

2. Products and Volume Delivered

The number of products delivered in CY 1996 was 2,077,942, which accounted for a total volume of 44.9 TB of data. These can be broken down by type, average product size, and total volume of data delivered for the year.

<table>
<thead>
<tr>
<th>Product Type</th>
<th>Number of Products</th>
<th>Average Size</th>
<th>Total Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staged FTP Retrieval</td>
<td>79,857</td>
<td>7.6 MB</td>
<td>0.61 TB</td>
</tr>
<tr>
<td>Anon FTP Retrievals</td>
<td>1,133,638</td>
<td>0.6 MB</td>
<td>0.63 TB</td>
</tr>
<tr>
<td>WWW Retrievals</td>
<td>375,594</td>
<td>0.2 MB</td>
<td>0.07 TB</td>
</tr>
<tr>
<td>8mm Tape</td>
<td>380,975</td>
<td>29.4 MB</td>
<td>11.2 TB</td>
</tr>
<tr>
<td>4mm Tape</td>
<td>39,050</td>
<td>23.3 MB</td>
<td>0.91 TB</td>
</tr>
<tr>
<td>9 Track tape</td>
<td>65</td>
<td>151.4 MB</td>
<td>0.01 TB</td>
</tr>
<tr>
<td>CD-ROMs (includes sets)</td>
<td>41,901</td>
<td>748.2 MB</td>
<td>31.4 TB</td>
</tr>
<tr>
<td>CD-Recordable</td>
<td>45</td>
<td>307.5 MB</td>
<td>0.014 TB</td>
</tr>
<tr>
<td>Floppy Disk (incl. sets)</td>
<td>1,508</td>
<td>15.1 MB</td>
<td>0.022 TB</td>
</tr>
<tr>
<td>Other</td>
<td>25,309</td>
<td>3.2 MB</td>
<td>0.08 TB</td>
</tr>
</tbody>
</table>

Total                                           44.9 TB
Several Technology Issues

Roy Jenne
23 Oct 1998

We have given some information about the volume of data in several large archives in the USA. We have noted that there are only a small number of huge datasets, and yet they have most of the volume of data. If we want sensible data management, we must focus on how the large datasets are handled. Have we considered whether there are better ways of handling them, and probably at a lower cost? In too many cases, they are just part of a big system without much attention given to a thorough analysis or to cost.

The next few slides show a few technology issues and technology trends. Compare some of these numbers with the volume of the datasets. In 1960 it required 83,000 tapes to hold one terabyte of data. By 1986 this was down to 5000 tapes, and it was 30 tapes in 1997. By year 2004, it may require only about 2.3 tapes to hold one terabyte. There has also been a large decrease in the cost of media to hold one terabyte.

A large tape silo with a robot to mount tapes will hold about 5000 tapes and it costs about $200,000 not counting the necessary tape drives. In 1986, one silo with 5000 tapes held about 0.87 Tbytes. By 1997, the main technologies (10 or 50 Gbytes per tape plus compression) permitted a silo to hold 60 to 275 Tbytes per silo. By 2004, a similar silo should hold about 2200 Tbytes, which is an enormous amount of data.

How many tapes can a silo mount in an hour? NCAR first got a Storage Tek silo in Dec 1989. It could mount a maximum of 180 tapes in an hour. Around 1993 it was upgraded to a system that could mount 350 tapes an hour. In 1998, NCAR still has a peak of 350 tapes/hour, but with a microcode change, we could do 450. The vendors project about 800 per hour in the future, still using one robot in a silo. This is fast enough that it is scary (fetch a tape, deliver it, and move back in under 5 seconds!) The big silos (5000 tapes) have permitted a maximum of 16 tape drives. Each extra drive reduces the number of tapes that can be stored. By around year 2000, a silo will hold up to 40 drives and they are smaller (40 will fit in the same space as 16 present drives).

But why should a person want more tape drives? After a tape is inserted in a drive, it takes 60 to 100 seconds to find a random dataset on a tape. With more drives, there can be more simultaneous tape searches and tape reads that are happening. But more drives also means more cost, so one needs to make sensible trade-offs depending on the application.


**Table 1: Tapes needed to hold one Tbyte of data**
The amount of data held by tapes and the data rate is for no compression. Devices with built-in compression have a (c) by them.

<table>
<thead>
<tr>
<th>Date of Technology</th>
<th>Device</th>
<th>Holds</th>
<th>Data rate</th>
<th>Drive cost</th>
<th>Tapes per Tbyte</th>
</tr>
</thead>
<tbody>
<tr>
<td>1960</td>
<td>1/2 in</td>
<td>12 MB</td>
<td></td>
<td></td>
<td>83,300</td>
</tr>
<tr>
<td>1972</td>
<td>1/2 in, 1600 BPI</td>
<td>40 MB</td>
<td></td>
<td></td>
<td>25,000</td>
</tr>
<tr>
<td>1980-1996</td>
<td>9 track, 6250</td>
<td>125 MB</td>
<td>0.9 MB/sec</td>
<td></td>
<td>8000</td>
</tr>
<tr>
<td>1986</td>
<td>IBM 3480</td>
<td>200 MB</td>
<td>3 MB/sec</td>
<td></td>
<td>5000</td>
</tr>
<tr>
<td>Sep 1991</td>
<td>IBM 3480 (c)</td>
<td>420 MB</td>
<td>3 MB/sec</td>
<td></td>
<td>2380</td>
</tr>
<tr>
<td>Sep 1996</td>
<td>Exabyte 8900 (c)</td>
<td>20 GB</td>
<td>3 MB/sec</td>
<td>$4500</td>
<td>50</td>
</tr>
<tr>
<td>1993</td>
<td>DLT 2000 (c)</td>
<td>15 GB</td>
<td>1.25 MB/sec</td>
<td>$2200</td>
<td>67</td>
</tr>
<tr>
<td>Apr 1995</td>
<td>DLT 4000 (c)</td>
<td>20 GB</td>
<td>1.5 MB/sec</td>
<td>$4200</td>
<td>50</td>
</tr>
<tr>
<td>Jan 1997</td>
<td>DLT 7000 (c)</td>
<td>35 GB</td>
<td>5.0 MB/sec</td>
<td>$7000</td>
<td>29</td>
</tr>
</tbody>
</table>

Note: The Exabyte 8900 is also called the Mammoth drive.
# COST OF MEDIA PER TBYTE
*(FOR ONE COPY OF DATA)*

<table>
<thead>
<tr>
<th>Year</th>
<th>Type</th>
<th>Hold</th>
<th>Media</th>
<th>Media Cost/TB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1986</td>
<td>IBM 3480</td>
<td>200 MB</td>
<td>$7</td>
<td>$35,000</td>
</tr>
<tr>
<td>1996</td>
<td>DLT or Exb</td>
<td>20 GB</td>
<td>$90</td>
<td>$4,500</td>
</tr>
<tr>
<td>2000</td>
<td>DLT or Exb</td>
<td>100 GB</td>
<td>est. $90</td>
<td>$900</td>
</tr>
</tbody>
</table>

## MEDIA FOR NEWER BIG SYSTEMS

<table>
<thead>
<tr>
<th>Year</th>
<th>Type</th>
<th>Hold</th>
<th>Media</th>
<th>Media Cost/TB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>IBM</td>
<td>20 GB*</td>
<td>$50</td>
<td>$2,500</td>
</tr>
<tr>
<td>1997</td>
<td>STC**</td>
<td>50 GB</td>
<td>$50</td>
<td>$1,000</td>
</tr>
<tr>
<td>e2004</td>
<td>Same</td>
<td>e400 GB</td>
<td>$60</td>
<td>$150</td>
</tr>
</tbody>
</table>

*available ~Nov 98  
DATA HELD BY A SILO  
(A SILO HOLDS 5000 TAPES)

<table>
<thead>
<tr>
<th>Year</th>
<th>Each Tape</th>
<th>Total Silo</th>
</tr>
</thead>
<tbody>
<tr>
<td>1986</td>
<td>175 MB</td>
<td>0.87 TB</td>
</tr>
<tr>
<td>1990</td>
<td>~400 MB</td>
<td>2.0 TB</td>
</tr>
<tr>
<td>~1995</td>
<td>~900 MB</td>
<td>5 TB</td>
</tr>
<tr>
<td>1997</td>
<td>50 GB</td>
<td>275 TB</td>
</tr>
<tr>
<td>e2004</td>
<td>400 GB</td>
<td>2200 TB</td>
</tr>
</tbody>
</table>

Cost of STK silo that holds 5000 tapes:

- Actual price for NCAR is $175,000
- Each Redwood drive about $100,000
- Each drive for 800MB tapes about $25,000 plus about $10K each for controllers
- Each drive for IBM 10 or 20 GB tapes – guess about $45,000

So: A silo with 8 drives costs $600,000 to $1M.
Our Data Support Unit at NCAR

1. A special feature at NCAR
   - Fast access to read bulk data into programs
   - 400 unique on-line users each year

2. A large archive, several discipline areas
   - Over 500 datasets
   - Over 10 Tbytes

3. Total staff of 11 FTE
   - About 3.5 on data delivery
   - About 2 on updates and archive maintenance
   - About 4 on new projects
     - Data from mesoscale models
     - Data from climate models
     - Prepare observations for reanalysis
   - One for administrative support
Figure 1: Data from Data Support archives that are read into user programs (not DSS), which are run on main computers at NCAR. The Gbytes read during each year are shown. A large portion of the use is by the universities. Most of the "other users" are NCAR users.
Use of NCAR Data Support Archives

1. Internet use
   - About 30 GB/year

2. Number of CD-ROMs sent
   - 30 CD-ROMs sent in each of 1995, 96
   - 1200 in 1997
   - est. 1300 in 1998

3. Summary of volume of data used

<table>
<thead>
<tr>
<th>NCAR DSS DATA USED (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Tapes</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1995</td>
</tr>
<tr>
<td>1996</td>
</tr>
<tr>
<td>1997</td>
</tr>
<tr>
<td>1998</td>
</tr>
</tbody>
</table>

4. On-line users of DSS data at NCAR

<table>
<thead>
<tr>
<th>Unique Users</th>
<th># Reads</th>
<th>GBytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>391</td>
<td>81,476</td>
</tr>
<tr>
<td>1996</td>
<td>399</td>
<td>88,994</td>
</tr>
<tr>
<td>1997</td>
<td>414</td>
<td>95,044</td>
</tr>
</tbody>
</table>
The Volume of Data

1. World observations for reanalysis

   a. The rawinsondes, pibals  
   b. Surface synop (3 hr)  
   c. Other aircraft  
   d. Surface ocean  
   e. Satellite sounder (2.5°)

<table>
<thead>
<tr>
<th>Period</th>
<th>Years</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>1946-97</td>
<td>52</td>
<td>40 GB</td>
</tr>
<tr>
<td>1946-97</td>
<td>52</td>
<td>90 GB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 GB</td>
</tr>
<tr>
<td>1946-97</td>
<td>52</td>
<td>6 GB</td>
</tr>
<tr>
<td>1969-96</td>
<td>28</td>
<td>27 GB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>165 GB</td>
</tr>
</tbody>
</table>

2. Archives from global reanalysis (204 Km resolution)
   - All archives (54 GB/yr), 50 years, 2600 GB
   - Most used archives, 50 years, 500 GB

3. Data from US GOES satellites.
   1978-1996 (19 yr) 200 TB/19 yr
   - Capture/archive cost $150K/year
   - Cost to copy & improve archive $2 million

4. Data from US weather radars
   - Copy, mailing, archive cost $800K/yr 100 TB/yr

5. Data from NASA EOS satellites
   - Launch mid-1998
   - Volume primary data
   - Cost for Archive, make products ($230 m/yr) 80 TB/yr

Roy Jenne
Dec 1997
Some Satellite Data, Global Coverage

   ♦ 20 years of great data!

2. TOVS sounder, primary data
   ♦ Oct 78 – Mar 92 (13.4 years) 631 GB
   ♦ About 48 GB per year
   ♦ IR, VIS, microwave (MSU)

3. Cloud cleared TOVS, and soundings (2.5°)
   ♦ 15.1 years 24 GB

4. VTPR sounder, primary data and soundings
   ♦ Nov 72 – Feb 79 (6.4 years) 1 GB

5. Global 4Km NOAA scanner (GAC)
   ♦ 4 Km, 5 chan (2x/day world coverage)
   ♦ About 232 GB/yr for 1.1 satl
   ♦ Total archives for 7.8 years 1805 GB
   ♦ Gives SST, clouds, vegetation index, etc.

Ray Kemme
Sep 1998
University of Wisconsin

- Doing good work since before 1970
- They collect huge amounts of GOES data
- National leaders in handling large volume data
  - at a very low cost
- Keen knowledge of science areas

But

They do not qualify to be an ESIP
- This is very strange
- They would have seemed better if they had emphasized more research on computer methods (very strange).
- The rules on being an ESIP seem focused on “leading edge systems,” rather than on appropriate systems to achieve “smaller, better, faster, cheaper.” Why?

Roy Jenne
Oct 1998
Some Information about NASA ESDIS
(Roy Jenne, August 1998)

The purpose of this paper is to present some talking points about ESDIS. The main purpose has been to update my own thinking on the subject. The name used to be EOSDIS. Now they want to emphasize the idea of earth science (ES) rather than earth observing system (EOS).

1. Summary of the tasks ESDIS needs to accomplish.

   There are several tasks that ESDIS must accomplish.

   a. Provide command and control of the satellites.
   b. Move the raw satellite data to selected archive points on the ground.
   c. Data needed for real-time operations must be delivered within about two or three hours. Most of the data do not have this fast requirement.
   d. Prepare useful products from the satellite data and possibly prepare them again in 3 or 4 years when the methods have been improved.
   e. Put the primary data and the products into archives that can help users obtain data.
   f. Develop a strategy for long-term data archival that focuses on providing basic data access (and relatively easy for users), and has strongly controlled costs.

2. An overall view of making products in NASA ESDIS

   In 1996 the view of making products in ESDIS was that the Instrument PI teams would develop the program software to process the data streams (which they had not yet seen). The ESDIS project would imbed this PI software into other big software releases that would be run operationally at the NASA DAACS. The idea was that the preparation of products for later satellites (after 2000) might have more PI control, but for AM1, Landsat, etc. the current ESDIS plans would be used. Since that time, there has been a slow evolution toward letting PIs and the DAACS directly prepare the products.

   We will now give a very brief overview of the various instruments for which data products are needed. Nearly all of these can be handled by methods that the community has used before (more distributed, more local control, more PI control).

   a. Lots of smaller instruments that provide data. In many cases, there is a history in which PI’s have done a good job of calculating products from one or two earlier satellites that had similar instruments. And the costs were often modest. The science is complicated, but experts know how to make data products (atmospheric temperature, sea ice coverage, ocean winds, etc.).
   - Ocean altimeter
   - SAGE (aerosols, etc)
   - BUV (ozone)
   - ERB (radiation)
   - Microwave (ice, etc.)
   - Scatterometer (ocean wind)
• Atmospheric sounders
• Etc.

NOTE: We should look up the costs of some of the previous work that handled these data streams.

b. Landsat and Aster. The EROS data center and their PI teams have a long history of handling Landsat data. From March 1997-on, they have been telling me that they can definitely do Landsat & Aster by themselves. And their projected costs seem rather low and reasonable.

c. The TRMM (tropical rain mission). In 1996 the Goddard DAAC was given this mission (data from several instruments) to handle. The bird is up. I think things are working well.

d. Handle data from the CERES and Moppitt instruments. In fall 1997, these PI teams obtained permission to handle their own data streams. It was a relief for these teams to obtain permission, because they needed a more streamlined working environment to do their work.

e. MODIS: After PI and DAAC methods are used to make products, we are mainly left with the Modis instrument, and its large data flows.

In July 1998, the MODIS PI team requested that they be allowed to process the data from their instrument. I do not know any details.

NOTE: We see a situation where the processing of data from AM1, Landsat, etc is gradually moving toward a mode of more PI control. I think that this is good, but it does bring up some other questions.

SUMMARY: The making of products in ESDIS is much more distributed than 18 months ago. We endorse this change. But the changes in the making of products will require some new thinking about providing data services and maintaining archives.

3. Local control for making products in ESDIS.

A project in ESDIS needed to develop data products from one of the instruments. They wanted to develop and run the software themselves rather than work through the large ESDIS project to get this done. They knew there would be surprises and problems in the data stream to cope with. They needed an environment where they could rapidly prepare and implement changes in the way that they processed the data. This would not be possible in a large centralized approach.

There has been a trend toward giving PIs more local control in making products. Part of this was driven by necessity: the big project could not have enough of the necessary software ready in time. But a lot of it came from the instrument PI teams. They wanted a more streamlined approach and more local control.

4. The effect of data volume on complexity and cost.

Suppose that we had defined an EOS with essentially the same satellite sensors, but where the data volume was only 1% of the present EOS.
a. It would take just as much knowledge to prepare the algorithms to make the same number of products as now.

b. With the lower data volume, it is likely that the costs would never have been allowed to approach present ESDIS costs. There would have been a sense that there is a history where PIs have accomplished complex calculations at rather modest costs. Since the data volume would be similar to past experience, there would be an expectation that costs should be similar.

c. The big data volumes have scared people, and it is possible to use these partly unknown factors to slip into high cost plans.

d. If the ESDIS costs were mainly driven by data volume, we would expect that the hardware costs would be a considerable fraction of the total costs of the big ESDIS. Instead, they are only about 10 or 15% of the costs.

e. The total costs are mostly explained by the 600 or 700 staff that have been working on the big ESDIS core project. Then we have to ask whether there is a way to accomplish the necessary work, using methods where the plans can be more focused and where a smaller human effort is needed.

5. The new millenium program for satellite technology.

NASA started a program to develop technology that would help foster better and lower cost satellite sensors and satellite missions. During the technical development process, it was interesting to see that the advanced thinking was applied in the following ways:

a. Design a sensor that would use less power (then the solar arrays could be smaller & cheaper)
b. Struggle to reduce the weight and size of a sensor. This will help to reduce launch costs.
c. Struggle to reduce the sensor cost, while keeping performance high.
d. Struggle to reduce overall mission costs.

NOTE: For the ESDIS data systems, where is the equivalent of this type of thinking in the data system process?

6. Factors needed to reduce costs.

In cases where it is possible to both achieve good results and reduce costs, several factors are needed to make it happen.

a. Need the will to try to reduce costs.
b. Need the knowledge to make it happen (and the knowledge necessary is often distributed across a number of people).
c. Need a process that will enable the lower cost approach to be planned.
7. Who gets to name the price of a data task?

If the selection of who will do data tasks is essentially a sole source activity that can name the price, then there is little incentive to control costs.

a. Is there any competition? For a given task, one could get a bid from 2 or 3 centers. The lowest bid would not necessarily get the work; it depends on cost, quality, knowledge, and track record.
b. Is there any way to scrub down costs for the higher priced activities?
c. Is there a history of selected previous costs to permit comparisons of costs and activities?

8. Command and control of a satellite.

One function of ESDIS is to control the satellites. In mid-1998, there were problems found in the new software to do this. The AM1 launch was delayed.

In about 1996 a small group considered ways to achieve good results in controlling a satellite while also lowering the cost. There are experts in this area that could help in designing new systems and in reviewing existing practices. One person told me that it is often possible to find ways to save 20 or 30% of the cost. By making changes in the overall approach, it is sometimes possible to achieve much more radical decreases in cost (even a factor of ten, they said).

A white paper should be available that discusses the factors that influence quality and cost. Comparisons of staffing and costs for several missions should be included.

9. The projects and costs in the Pathfinder program (during ~1993-1998?).

The idea of the NOAA/NASA Pathfinder program was to calculate a range of satellite data products, based on satellite data collected from about 1979-on. But state-of-the-art algorithms were developed and applied. This has been a very successful project, funded by NASA. More information about the Pathfinder projects and the costs is available. One part of the program had to take steps to make the satellite data really available for processing. Otherwise, the cost and complexity of data access was often too high for PI's to readily handle.

10. Is a big staff needed to produce high-volume products?

In the next sections we will review several examples of large complex projects, where the work to develop and operate the necessary controlling software is handled by a very small staff.

Therefore, if the staff members are carefully chosen, and if the methods that are used follow the spirit of the examples, the staff does not have to be big.

11. Automated software methods to produce products in big projects.

One task of ESDIS is to manage many large streams of data and to process them through computers to make products. It is useful to review some examples of how people have handled these tasks in some large, complicated projects.
There are a number of very large projects that have developed experience in preparing automated sequences of complex software programs which automate the running of thousands of programs every day, over a span of years. Considering the complexity of the tasks, it is surprising that the staffing needs can be as low as they often are. Each big job is going to be different from other problems, but we can still use the experience that has been developed to help us define methods and scale the staff requirements. Consider two examples.

a. Three large reanalysis projects.

The world has three large reanalysis projects (NCEP/NCAR, ECMWF, and NASA/DAO). There is an associated large project to gather and prepare the global daily observations that are needed (raobs, aircraft, land surface, ocean surface, satellite sounders, and satellite cloud winds). The reanalysis projects are complex efforts that are based on the cutting edge of research.

All of these projects produce global analyses, each 6-hours, for many levels in the atmosphere, and for many years. Each 6-hour period means using one of the world’s best analysis methods (data assimilation), together with a “state-of-the-art” forecast model. Many steps are needed: prepare the observations for the 6-hour period, do QC, calculate many, many different variables in several forms for the users, prepare “alarm bells” to help warn operations people if problems have occurred.

Each real day, this production is able to accomplish 10 to 40 days of new analyses, done for each 6-hours. Every day (and night) hundreds of complex program processes are automatically triggered to do the production.

Usually one or two people handle this production, and write other programs to understand the nasty problems that sometimes arise. The actual programs for analysis methods and forecasts have usually been developed over a number of years, using research inputs from a group of people. The core group to handle production, data flow, problems, and documentation in reanalysis is often 2 to 5 people, with periodic help from a wider set of experts.

b. The COADS surface marine project.

This project prepares the world’s best set of surface marine observations. Some years ago, we realized that during a 7-month production period, we would have to run about 5000 different computer jobs. There is no way to set up and execute this many jobs without making mistakes and taking a lot of people time. So we set up an automation procedure that is placed on top of normal computer software systems. This did not take very long. It gives the project quick control to handle problems, because the necessary software is not buried in a larger amount of complex software. The project worked well, but it certainly kept us busy!

c. Conclusions.

PI's usually have their own methods to develop data products for smaller datasets and these methods have been very effective. There is typically no need to try to centralize all of this work. When the work is centralized too much, the experience is that it leads to high costs, and too much time spent in interactions. There is also a lack of sufficient PI control necessary
to respond to real-life problems and still get the job done. People should be aware of methods that have been used to help automate many functions in data processing.

12. The project director for ESDIS is changing jobs.

The director of NASA Goddard announced (3 August 1998) that Rick Obenschain (Project Director of ESDIS) has agreed to become the first Center Chief of the Electrical Systems Center (Code 560). This is one of the largest engineering centers within AETD. Dorothy (Dolly) Perkins will take over Rick’s present ESDIS management tasks.

Rick has been in the ESDIS job about two years. He is a very active manager and very oriented toward results. It is hard to know what effect these changes will have on the ESDIS project.

13. Understand useful methods and costs for big data flows.

To help understand the costs of computing and storage for large data flows, it helps to have selected information about the costs of projects that have done similar work. Some of the comparisons that can be developed are as follows:

a. Compare experience with other big data flows
   - Landsat data (handled by EROS and Goddard)
   - The SAR data handled by Fairbanks (high resolution ice information)
   - The long experience at University of Wisconsin in handling GOES data
   - Projects in Europe to handle data from ERS-1, ERS-2, and related satellites

b. Big computer centers
   - Get a brief history of computing power, data flow, and total archives at selected computer centers
   - Have data for a few of the main production centers (GFDL, NCAR, NCEP, Goddard, and ECMWF).

14. Where can products for big datasets be made?

Sometimes each PI group that needs to prepare data products will want their own very large computer installation, paid for by NASA. A large computer setup will need staff experienced in fast computers, fast data flows, fast networks, storage devices, etc. It can become expensive if NASA has to pay for very many high-end computer facilities.

Suppose that PIs are fully in charge of preparing products from a large dataset. NASA still has the option of either buying them their own computer, or giving them time on another NASA computer. In either case, they will need some special systems support to help solve problems that they will encounter. Also, they will need access to a fast, high-capacity mass storage system that has a simple file management system.

The location of the computing resources might be one of the following:

a. At the location of the main PIs
b. At a NASA DAAC
c. At a NASA computing facility that has a good design for fast data flows, simple file management, error control, and controlled costs.

NOTE: A center needs procedures to make data storage and data flow easy and fast. They need a track record of handling user needs. If they try to force excess data rules, complexity, or costs onto users, then they are not a viable place to handle large data flows or to make products.

15. The huge datasets; do not move them often.

When plans are made for archiving the huge datasets, and making products from them, we should keep in mind the cost factors that are involved.

a. The computer that generates the products should be located close to the archive.
b. The archive of basic data and high-volume products should not have to be physically moved in 3-5 years to another archive for long period storage.
c. The data system needs to be designed for rapid data flows and low cost archival. Large amounts of slow software should never be in the paths of main data flows.
d. The data system needs to be optimized for access to files of data of significant size. It is still possible to have other software that operates on the data in a more detailed way, but that should be judged on the basis of real needs and costs.

NOTE: There are many small and middle-sized datasets that are easy enough to move that we do not have to worry about these rules.

EXAMPLES: Landsat is a large dataset. The initial archive is located at the EROS DAAC. The products will be made there. And EROS will be the permanent archive. This makes sense.

16. Will data be ready on time? Will data be lost?

It is important to summarize a little of the 20-year history of archive strategy questions within NASA. For many years the data from a sensor was sent to the appropriate PI groups for archival, and to make products. Then the products and the data were gathered back into the data center that was organized within Milt Halem’s computer center at Goddard. As late as 1990, this center had a comprehensive set of archives for data from both the earth science disciplines (from NASA satellites), and from space sciences.

In some cases there were problems. A PI might hold onto the data too long before delivering products. In few cases, archives were lost. When I reviewed the archives in 1988, it appeared that nearly all of the archive of basic radiance data and products were there. There were problems in lack of access to huge amounts of early GOES data at Goddard, and the THIR data was not being properly cared for (still true).

About 1987, NASA had an opportunity to take care of the THIR data (Copy 30,000 tapes to 3000 new tapes). The NASA center only needed help to buy new media ($30,000). The chance to do this job for $30,000 when it would usually cost $750 K to $2100 K was lost. It is unknown why NASA was not able to respond to an opportunity like this. In any case, this problem had nothing to do with poor planning by a single PI team.
We note that the problems of timely access to data products can be addressed by appropriate agreements with PI's. Also, no PI team should have monopoly rights to any primary data stream. There also should be a copy of data in a central archive so that the data is secure, and so that is cannot be used for “monopoly bargaining.”

CONCLUSION: We see no reason why the data processing for many small and mid-sized datasets should not be handled by the PI's. But the results must get into other archives in a timely way.

17. The DAACs for ESDIS.

The distributed active archive centers (DAACs) for ESDIS have been successful. They have proven that they can handle data from projects, archive it, and help users obtain data. It is less clear that there is enough budget control in the DAACs. Several DAACs that started with staff sizes of 20 or 30 are now up to 75. If the data system is distributed to 8 or 9 places, and if many of them become quite large in staff, then the total budget of the DAACs is going to become quite large. We also note that the really large data flows from the EOS era have not yet started to arrive at the DAACs.

The idea of having a distributed data system is usually that people will be in smaller groups, with a better focus on core discipline problems and better users interfaces. The bigger systems run into problems of focus, priority setting, and lack of knowledge of the science. Also, there are management problems associated with size. The big groups are more likely to focus on building software systems rather than on solving data problems.

There are two ways that the idea of a distributed data system can run into problems:

a. It is distributed across so many groups that it becomes (1) costly, (2) the essential functions for archiving are not done, and (3) it becomes harder for users to locate the data (the ESIP idea can get into these problems).

b. It is distributed to ten or so sites (such as DAACs) and each of these starts becoming a large organization (the DAAC system can get into this problem).

There is no reason that DAACs should all be the same size. It is likely that only 2 or 3 have a need to be of significant size. However, there are often social forces that try to bring the size of smaller units up to that of larger units. After all they are all like members of a family; shouldn’t they be treated alike in staffing? The answer is “No.”

A DAAC often needs 3 years or more to really learn how to do its job. This means that we can’t replace DAACs on a rapid time cycle. However, NASA should feel free to assume that discipline boundaries will not be too rigid for particular DAACs. This would give NASA some freedom about where to locate a new dataset. This would send a signal that a DAAC must compete on both quality and cost in order to get new work. And if another non-DAAC entity can do a certain job well, we should keep open the possibility that they would be used for the new work.

18. Don’t try to serve 10,000 users, only serve 500.

I have been in several data meetings were people believe that NASA data systems have grown too big and costly. One common argument to reduce costs is to say, “Don’t try to serve 10,000
people; focus on serving the main 500 NASA PI’s.” We share the motivation behind these statements, but think that it is an unnecessary restriction. We can point to several examples where an effort of 0.3 FTE to 1.5 FTE can accomplish a large amount of work toward making large amounts of data available to a wide community in a manner that is very user friendly.

We can imagine a plan for data systems where the focus is on serving the core science missions, but where a little extra work opens many datasets to a much wider community. This is happening in the DAAC system.

CONCLUSION: Keep a focus on the science mission, but do a little extra work to serve a broad community. People doing this extra work should have to demonstrate that they are doing it at low cost. We also know that there are high cost methods to achieve the added goals, often with low results.

19. EOS data for K-12 education and policy makers.

The need for data and information for K-12 education for policy people has come up in several meetings. There is often a fear that a large amount of money will be spent in a diffuse way on these topics. The fear is justified, but there is a way to handle the need for this activity in something other than an “all-or-nothing” approach.

First, the spending can be rather modest. For example, people might argue about whether the pot should be $1 million or $3 million per year in size. There are many small groups in universities, etc. that can accomplish a lot of work using grants that only range from about $30K to $70K per year in size. A group of universities working within the NASA-USRA program has been effective. In 1997, NASA gave contracts to some small groups working within the ESIP, type-3 program to accomplish this sort of work. In addition, it should be expected that other PI groups and data centers will use a small (but important) fraction of their time to help education.

CONCLUSION: By establishing the limited pools of money for grants, and by using mechanisms that are now working, the education community should have a range of data and information products for use by teachers and students.

By about 1990 NASA had decided to form a distributed DAAC system. They also decided that there was not enough in-house experience to handle all of the EOS data, so they looked to a central contract – awarded in about 1992 to Hughes.

In the early 1990’s, the DAAC’s were becoming more experienced so that they could handle many tasks once intended for the central contract. The PI teams were also able to handle more jobs, either by themselves or by working with a DAAC. For a given cost, the amount of computing and data handling that can be accomplished has increased by many factors over the past 10 and 20 years. This improvement in technical capability means that PI teams and DAACs can handle bigger tasks than was previous possible.

These changes give the opportunity to do more work at less cost, but only if the strategy is modified.
20. How much data can users handle?

Possible users of data can be pushed away by any of three factors:

a. Problems of handling the data
   • Too many pieces of media
   • High local costs to handle and store that much data.

b. High cost to obtain a copy of the data

c. Complexity of formats and data access software

Most users can readily handle small files in ASCII format. Many of these files will be obtained by Internet or CD-ROM.

Scientific users can readily handle costs up to $800-$1000 for significant amounts of data. Costs of $3000-$5000 or more would often have to be built into a contract or a cooperative plan to get data. People are also limited by too much fuss to handle many pieces of media, or by the cost of tape drives to read the media. In the mid-1980's an appropriate lower cost tape drive often cost $18,000. Now there are options with much better capability that cost $1000 to $6000. Most PI's will readily handle 10-30 pieces of media and some realize that they must be willing to handle more than that.

CONCLUSION: Any plans that assume that the typical PI will be able and willing to handle 1000 pieces of media are probably due for a surprise.

21. How to run a satellite project to obtain quality and cost control.

To be useful, a satellite mission has to accomplish several things:

a. Bring down data from the satellite.

b. Solve (or work around) problems that will occur in the satellite system or in the data stream.

c. Prepare sensible data products. Give reasonable data access for users.

d. Save much of the data in long period archives.

Most satellite missions were in the position where they had to accomplish all of these on a limited budget. They made the tradeoffs necessary to give results that were generally good. But there were complaints that when there were hardware problems, then the money was used to fix these, and the data budget was shortchanged too much. Some of those complaints were true, but if one looks at the archives, it appears that most of the necessary products were somehow created.

If the data activity is organized completely separately from the flight projects (and from PI's), it is much easier to get budgets that have run away from sanity. One can then see several problems, which include:

1. Lack of focus
2. Unreasonable requirements
3. Big systems rather than streamlined processes
4. Sometimes a focus on software systems rather than on streamlined processes
5. Lack of interest in cost control
6. Too often there is too much empire building.
Some of the problems do, in fact, result from bad motives. But in very large projects, there seems to be an effect of things becoming worse than necessary for no very good reason.

22. Data for long period, global scale problems (need 15 to 30 years of data).

Under climate change research it is expected that many problems should be studied that have a global view, and 15 or twenty years of data. Such projects can be stopped by too much data volume, as well as by the lack of data. Therefore, PI teams need to think about what data they will offer to help such projects. If the data volume is very high, it may be necessary to produce a sub-sampled dataset.

Data volume is not a problem for many satellite instruments (altimeter, Sage, ozone, etc.). In a few cases, it is a problem. Some teams already have a good strategy to manage the problem of high volume:

a. Ocean color radiances at 4 or 8 Km resolution as well as at 1 Km
b. An ISCCP project for global clouds uses a sampled data set for calculations (either 8 or 24 Km resolution, not 1 Km). Data for 1983-on are being processed.

Status of this issue in ESDIS: for small datasets, and some midsize datasets, it is not an issue because some new data subsets can be quickly generated when problems arise. For big datasets, this issue is likely to be a big problem. It appears that PI teams have not given it sufficient attention to head off problems in advance. One of the goals of climate change science is to look at issues of climate change and variability. If we do not have datasets that will support these studies (at reasonable cost) we are missing the boat.