Preserving Electronic Records: Not The Easiest Task

Fynnette Eaton
National Archives and Records Administration
7th and Pennsylvania Avenue, N.W.
Washington, DC 20408
fez@cu.nih.gov

The National Archives and Records Administration has had a program for accessioning, describing, preserving and providing reference service to the electronic records (machine-readable records) created by Federal agencies for more than twenty years. Although there have been many changes in the name of the office, its basic mission has remained the same: to preserve and make available those records created by Federal agencies that the National Archives has determined to have value beyond the short-term need of the originating agency. A phrase that I once coined for a preservation conference still applies: the National Archives, when it decides to accept the transfer of records into its custody, is committing itself to preserving these records for perpetuity.

Most people think of the National Archives as the keeper of the Constitution and the Declaration of Independence. Even the most experienced researchers are unaware of the growing number of files that have been accessioned by the National Archives in electronic format. Since the creation of the Center for Electronic Records in 1988, the number of files transferred has literally skyrocketed: in 1988 the Archives received 150 files from Federal agencies; in fiscal year 1991, the number was 1500, an order of magnitude increase in two years. This number jumped again this last year to 8700 files. Unfortunately, the number of files that we should be receiving completely overshadows our accomplishments. Beginning in the 1970's, NARA signed agreements with Federal agencies specifying that at certain times or under certain conditions these agencies would transfer files to the National Archives. The Center in 1990/91 developed a database to determine how many of these files have been transferred. As you can see on the graph, we have received very little: less than 10% of what has been anticipated. The second chart is even more daunting. It combines what we knowingly have not received from agencies with projected numbers of new files currently in use at agencies that warrant permanent retention by the National Archives. The National Archives asked the National Academy of Public Administration to perform a study, identifying the most important current databases in use in the Federal government, and to provide s in a study commissioned by the National Archives in 1990 and 1991 identified as candidates for preservation by the National Archives. This second graph has four components. First, almost a blip on the chart are the files received. The second component is what has been projected that we knowingly should have received, based on schedules developed with Federal agencies. The third area is the continuation of series beyond the first file, that should have been received, but have not. Only the fourth component are the files identified by NAPA that NARA should target for transfer.

Consequently, although the Center has successfully increased the number of files being transferred, we have barely made a dent in the ensuring the preservation of Federal records that have clearly been designated as important enough to be transferred to the National Archives.

The successful and almost ubiquitous use of computers for more kinds of record keeping activities in ever increasing quantities poses serious and extensive problems for the National Archives and Records Administration. We share many of these problems with organizations represented in this audience.

The problems posed by electronic record keeping include fragility of the media, rapid obsolescence and incompatibilities between makes and models, and even between different releases of the same product. The magnetic media most commonly used to store electronic records off-line, open reel tape or tape cartridge, are physically fragile and easily erasable and reusable, presenting a serious challenge to the preservation of electronic records.

The lack of standardization coupled with hardware and software dependencies of electronic records means that even if we can identify and physically preserve the records, we may not be able to access the data they contain. And, finally, rapid and unending change in computer technology exacerbates these problems.

The purpose of my paper is to discuss the types of problems we have encountered in trying to preserve the electronic files at the National Archives. But, before I provide specific examples of our problems, I want to quickly acknowledge the fact that my institution's holdings are tiny in comparison to the size of most of this audience's. The amount of information currently in our custody is roughly 1 Terabyte. But NARA makes up for size in the diversity of files that it has accessioned. As of October 1st, 1993, the Center has electronic files from 91 agencies: 19,278 data files in a wide range of formats. For example, although our regulations clearly state that agencies should transfer files that are hardware and software dependent in either ASCII or EBCDIC character code, we have in our holdings files that are BCD binary coded decimal, binary data, EBCDIC and binary data, EBCPARK. EBCZON, Multipunch, NIPS (National Military Command Information Processing System), SAS, and SPSS.

I would characterize our problems as falling into three categories. First there is the very serious problem with metadata. Secondly, the hardware and software dependencies of files, which prohibit us from being able to properly process and preserve the information and thirdly, the medium itself. I would like to spend time on each of these categories.

First, documentation. I would like to use a true anecdote to bring into clearer focus how wide-ranging the problems relating to metadata can be. There was one accession, which we could not properly process, because the documentation was only half-way complete. The agency had made an incomplete copy of the documentation, providing us with only the even-numbered pages, because they had failed to make a proper two-sided copy of the documentation. Fortunately, we were able to secure a full copy of this information, before it was dispersed. Although this appears to be a silly example, it illustrates why it often takes archivists within the Center between six months and a year to secure all the necessary documentation for files that are being accessioned into the National Archives.

Unfortunately, the Center is not always successful in receiving sufficiently complete documentation to ensure a proper understanding of the file. There are numerous examples of files that have been rejected for transfer because documentation was simply too incomplete. We have rejected surveys of taxpayer attitudes, trade statistics, military personnel statistics, and a collection of Vietnam War data, because the archivists could not make sense of the data, due to the lack of documentation.

The likelihood of encountering these problems increases dramatically if the files being transferred are either program files that have not been prepared for distribution or older files that have been in offsite storage for some time. If a file is active, it is easier to find reliable documentation. If the file has been distributed to the public, documentation would have been prepared for those ordering the files.

Yet there is also a middle ground, where NARA has received documentation, but it is incomplete and causes problems when trying to provide reference on these files. One file, the Combat Area Casualty file, which is one of our most frequently requested files, provides a good example of these types of problems in documentation.

The Combat Area Casualty file was maintained by the Department of Defense. The system recorded those people, either military or civilian who were wounded, captured or killed in the conflict in southeast Asia. One of the problems with the documentation appears to have been caused by clerical error. In preparing the list of codes for types of injury, the person creating the documentation skipped one of the letters of the alphabet that represented a code for a type of injury. Thus there were records that, according to the documentation prepared by the office using the file, were invalid. A second problem encountered was the replacement of one set of codes with another, without identifying this change in the documentation. This is a common problem with files that are frequently updated. The documentation must also be updated as codes are modified or replaced. Otherwise the documentation will not accurately reflect the information in the file. This is a serious problem for the National Archives, because we do not actively maintain these files: we simply preserve the information for access by researchers or the agency interested in the file.

The Center continues to confront the problem of hardware and software dependencies when working with agencies to transfer files to the National Archives. Some of our knottiest problems come from records created on systems developed during the Vietnam War. Perhaps the best example is the Filesearch IV system, which was used by the Combined Document Exploitation Center, to film captured Vietnamese documents. In 1977 the National Archives acquired a 106 reels of motion picture film which had 16 mm images superimposed on 35 mm motion picture film, with the sound track used to record a digitized index to the images. There was a major problem. The FileSearch retrieval equipment had ceased to be manufactured in 1969. There have been several efforts to find a way to get access to the index, but there are few systems still in existence (we have one that has never been operational) and we found that there were two incompatible coding structures. Thus, the index to the 3 million images and 1 million documents is inaccessible.

The issue of hardware dependency is still with us, even today. One agency with financial data was interested in transferring the records to us, but this agency had used "Tall Grass" tape drives to create the file and their drives had not been used for many years. When one of our computer systems analysts visited the site, he reported that is was unlikely that the drives would function, and that the data would be inaccessible unless placed on a tall grass tape drive and then outputted to another format.

A third example is provided by a large statistical agency that has permanent files that were created in a proprietary system with both hardware and software dependencies, that they have not been able to successfully convert into a format that could be used outside of that agency. We have been working with this agency for more than 15 years on this issue, yet we have not yet received a file from this system. (Census Input/Output [CENIO]).

Our major problem with software dependencies is found with military files created during the Vietnam War. The military used NIPS software, (National Military Command Information Processing System) to compact files that had numerous repeating fields. Unfortunately, at some point in the 1970's IBM ceased to support this software, and since no one else was a major user of this system, it disappeared. The records, however, are still with us. Currently the Center has approximately 150 files that are in this NIPS format. In this case, we are hopeful that the functionality of relational databases will provide the tools necessary to decode the files recorded in this format.

Some additional examples of software dependent files that we are unable to provide normal access are the National Economic Commission Computer Budget Gameor simulation, which was recorded in a spreadsheet and implemented via macros; the United States Railray Administrations Tracking/Document Management System, which was created in Basis and transferred to us in system backup format. These are problems we currently have. If we did not have the regulations requiring files to be hardware and software independent, this list would be unending.

The last category of problems, fragility of the medium, are the ones that my branch deals with since we have the responsibility for making preservation copies of files transferred by Federal agencies to the National Archives. We often find, particularly with older tapes, that we are unable to read the files because of excessive data checks. We have several accessions of older files that we have not been able to copy because of this problem. Since this is probably the only extant copy of the file, we are in the uncomfortable situation of either rejecting outright these files, or trying to copy as much of the information as possible onto newer media. The problems encountered in trying to copy older tapes is one of the best justifications for working with agencies to provide copies of data that are to be transferred while they are still in use, so if there are problems, the agency can easily supply another copy of the information

Another problem we are finding with older files is the condition that we refer to as "sticky" tapes. On a few occasions we have been notified by the computer center we use that one of our tapes has stuck to their tape drive. So far they have not threatened to kill us, but we must carefully monitor the conditions of the older tapes that we send to the computer center for processing. The National Media Lab has been very helpful in providing us with advice and training in how to screen older tapes.

The fragility of the media has been well documented. Storage conditions are always cited as a major reason for data loss in magnetic recordings. In many cases, the Center has no knowledge as to the storage conditions of the tapes before they are sent to the National Archives for copying. A couple of years ago we received some tapes from Wright Patterson AFB that had files relating to the Vietnam War. The Center has not been able to process most of the records transferred from the Department of Defense on the Vietnam war because the files are in the NIPS format and there is only limited documentation for most of these files. We had hoped to gain additional information about our records from these recently received files, but we discovered that the tapes had not been stored properly: there was the possibility of fungus on the tape and most of these tapes were created before 1975 and exhibited tendencies of sticking.

Again, this is a more dramatic example than what we normally confront, but it helps to convey the wide range of problems facing us as we receive files from the very large universe that is the Federal government.

The Center has techniques for dealing with these problems. For example, to cope with lack of standardization, NARA has issued regulations requiring that any electronic files scheduled for transfer to the National Archives be written in a simple format that is not dependent on any specific hardware or software. We are attempting to find more sophisticated solutions to compatibility by promoting standardization and collaborating with other organizations such as the National Institute of Standards and Technology, the National Environmental Satellite Data and Information Service and the Federal Interagency Coordinating Committee on Digital Cartography. To overcome the fragility of magnetic media, we have implemented controlled storage and testing procedures, and we require other agencies which retain permanently valuable electronic records to do the same. We actively encourage agencies to give us copies of permanently valuable records at the earliest opportunity, and we return the agency's tapes to them for possible reuse. But these are means of coping rather then solving the problems. So the task of preserving electronic records for future generations is not the easiest task: but it is an essential one. That is why we, at the National Archives, look to other agencies and experts in the field of magnetic recording to help us confront the problems of electronic records. By working together we might actually find some solutions to the problems confronting us today. Thank you.

Invited Panel: User Experience with Storage and Distribution Media

MR SAWYER: The panel chairman for today is Jim Berry. He received his BA from the University of Maryland and his master's from American University. He also has an MBA from the University of Southern California. He's held various positions at the National Security Agency, the Department of Agriculture, and the Office of Personnel Management.

Currently, he's the user representative to a processing office which supports one of the major operations groups at NSA. His areas of specialization are massively parallel computing, high speed networking, and mass storage.

Jim Berry.

MR BERRY: This afternoon, each member of the panel is going to make a very short presentation, and then we'll open it up with questions. I invite questions from the floor. Each person is now going to explain to you what they do at their respective organizations.

The first speaker is Lee Bodden. Lee is the Hughes STX manager of the Goddard Space Flight Center Version 0 Distributed Active Archives Center. He has been one of the system engineers for the DAAC since its inception in 1991.

Lee?

MR BODDEN: (Off microphone.)

Thank you. My name is Lee Bodden, as Jim said earlier. I'm with the Goddard Space Flight Center V Zero DAAC. We've been asked to give a short two-minute presentation on what the DAAC is; I understand that there will be other talks on the Goddard DAAC later on.

We're just getting rolling as an on-line, active archive. We -- right now we've got about two terabytes of data on line, but we anticipate growing to twenty terabytes minimum. That's just what we're looking at now. It may be more than that, because there are projects that are coming at us from all directions.

Our computers, we went with SGIs. The mass storage hardware for our archive -- we're going to be talking about Cygnet optical jukeboxes with 12-inch optical platters; and we've got the Metrum RSS600, which is an updated VHS cassette library system, which has been performing very well for us. We've only been using it about six months, but we're very happy with it.

Our storage medium: for the Cygnet, the corresponding media is the optical disk, and the Metrum uses VHS cassettes with a capacity of 14.5 gigabytes per cassette. The distribution that we're currently supporting that we're holding up to is what you see here. There are other types of media available upon special request.

And, as I say, we're just getting going. Our ingest volume right now is 25 gigabytes per month, and our monthly distribution is 125 gigabytes. That's going to increase. By fiscal year 1997 we are looking at perhaps getting up to 60 gigabytes distribution per day. We have a long way to go from where we are and where we're going to be.

Just very quickly, this is an architecture of our system, and these two rectangles, sort of in the middle there, those are our computers with the peripherals surrounding what we call our DADS (Data Archival and Distribution System) systems. So, we've got one computer dedicated to running the archives. Then we have another computer, our second one, which is dedicated to our information management system (IMS). This system is where the users will log into, and users, using our IMS, will be able to browse the summary records, which we call metadata, of all our current data holdings, and select the data.

Once selected, the IMS/DADS interface is automated so that the IMS will take that request from the user, feed it into the DADS and from there on --we are slowly automating the system-- so the users will not see anything beyond that, except for the fact that their request has been filled onto one of these different types of media, sent back to them. We are supporting on-line FTP distribution of the data for limited amounts of data.

So that's the situation that we have at the Goddard DAAC.

MR BERRY: Thank you. The next person on the panel is Richard Davis, who is the data administrator and records officer of the National Oceanographic and Atmospheric Administration's National Climatic Data Center in Asheville, North Carolina. He's been in government service for 47 years -- I didn't know anybody had been in that long -- in the field of meteorology and climatology.

He's responsible for the management of 50,000 cubic feet of manuscript records, 1.3 million microfiche, and over 100,000 reels of magnetic tape, which are depicted on his slides. He's also currently the project manager for the receipt and archiving of data from the new Doppler radars (NEXRAD) that will generate in excess of 88 terabytes of data per year.

Dick?

MR DAVIS: (Off microphone.)

Thank you. Can you hear that all right? At the National Climatic Data Center, we receive climatological and meteorological information from the National Weather Service (NWS), all the Department of Defense agencies, and the FAA. This has been going on for many years.-- We started in 1938 in New Orleans, and then we moved to Asheville, North Carolina, in 1952, and we've been there ever since.

We are an officially designated records center for the Department of Commerce. As such we work very, very closely with the National Archives and Records Administration (NARA). We do have a significant amount of data today, about 140 terabytes now. The NWS modernization will generate about 100 terabytes per year, and then who knows in 1999, what it will be like. So there's a fair amount of data to manage..

So right now this is the way it's looking like in gigabytes, how we've been doing. And from 1986 to 1993, you see here we -- right here is the NEXRAD program. We'll be getting about 33,000 eight millimeter EXABYTE tapes per year -- somewhere in the neighborhood of 88 to 90 terabytes per year from this one project alone.

Now, the good news is that within five years, we would start to migrate these things to 3480 cartridges, and we're going to give Fynnette (Ms Eaton of NARA) about 990,000 3480 cartridges each year.

MR BERRY: The next person is Fynnette Eaton. She is going to get another chance to talk to you about her problem, which she described in her previous talk. So maybe she can say a few more words about it now. She is the chief of the Technical Services Branch of the Center for Electronic Records, and has been an archivist at the National Archives and Records Administration since 1977.

MS EATON: Yes, if it's alright -- is this working?

MR BERRY: Why don't you use the one in front of you?

MS EATON: I do not have any overheads. I am wearing two hats for this meeting. Margaret Adams, who is the chief of the reference services, supplied me with the information about the Center's reference services.

As I said in my talk, we have approximately 19,000 files. They range from military files from Vietnam to the 1990 Decennial census files. The Bureau of the Census constitutes almost 30 percent of our holdings. About 50 percent of our requests are for the census files.

The date span of our files is from the 1960s up to this past year. We do not have information on line. As I was telling my host at lunch, I think NARA is the dinosaur of the group. We do not provide on-line access. People can get documentation about our files and then we will supply copies of the files either on 3480 or 9-track open reel.

The National Archives has a new facility that is very close to this campus. Our office will move into this new building in January 1994. There are plans to provide access to Internet for NARA staff at this building. Once Internet is available to the Center for Electronic Records, we will explore ways for making information available to users through the Internet.

MR BERRY: Thank you. The next panelist is Jordan Gottlieb, who works at the NASA Goddard Space Flight Center. For the past ten years he has participated in the design, development, implementation and maintenance of small and large software projects, project implementation and management, and also had extensive involvement in system integration and archive management of near-line and off-line data.

MR GOTTLIEB: I work at the National Space Science Data Center, and we have an enormous range of media we deal with, most of which are listed. We consider 7-track no longer a current media, but it will show up on the next slide as having significant holdings.

What we do is we act as a twofold division. We actually are chartered with archiving data, scientific data, but we also have taken on a new role in the not too distant history of also providing distribution services. Distribution services get to be rather interesting because we support an on-line distribution service for that data which is electronic, but we also have to support the analog portions, scientific photographs, fiche film of very old history, as well as alternate media for electronic data.

The current estimates of the holdings are listed there, and right now the holdings are probably somewhere around three and a half terabytes. We're looking to have that continue up to six terabytes in the coming year, and then we're looking at a projected growth of approximately six terabytes every year after that.

It gets to be a rather large problem as the equipment gets more and more sophisticated and the data rates keep going up, and it becomes a problem to manage these issues.

7-track is rather interesting because we are in the process of migrating that to a more modern media, and one of the things we are challenged with at the Data Center is continually finding ways to promote data up into more current media. As was stated in many of the presentations, the media changes so rapidly that by the time we actually can do a study and acquire new media technologies, the next media is out and touted as better and more efficient.

MR BERRY: Thank you. The next panelist is Laura Potler, who has been with Goddard for nine years. She started with compiler debugging on the massively parallel processor and has transitioned to systems analysis of data systems for satellite projects, including ROSAT, SeaWiFS, and TRMM.

MS POTLER: Hello. Well, I'm really surprised to be with such a distinguished group. I'm in a very different category-- I do not work for an archive and distribution center. My job is one step removed. I'm a systems engineer, and I've been working at Goddard on an assortment of projects over the last... close to ten years, actually. I half suspect I was invited because a few years ago at one of these conferences I shot my mouth off about the state of eight millimeter, and I think they're getting even with me.

Anyway, I have worked on a variety of projects. I started out on ROSAT, which is short for Roentgensatellite. This is x-ray astronomy; high-energy astrophysics. It was launched in 1990. Then I worked on the design of SeaWiFS, which stands for the Sea-Viewing, Wide-Field-of-View Sensor, which is the follow-on to CZCS (Coastal Zone Color Scanner). SeaWiFS is scheduled to launch next year. Very recently I switched to TRMM, which is the Tropical Rainfall Measurement Mission. I've been doing system design of the data systems which produce all the data that goes to these folks here.

I wrote up a few notes about the projects themselves, the formats of the data. As you see, the ROSAT data is in FITS format. SeaWiFS and TRMM, HDF. We're hoping to feed both of these data sets to the Goddard DAAC. The size of the holdings, as the years go by, changes dramatically from ROSAT to TRMM. I have it broken up by *proprietary* and *public*. It could be *intermediate* and *final* products or however you want to term it. In terms of proprietary data holdings, ROSAT is currently working towards 300 gigabytes. TRMM is expecting 66 terabytes over the life of the mission. So you can see how dramatically the data holdings size increases.

The archive medium for ROSAT, both the proprietary and the public, is 12-inch WORM. SeaWiFS has an intermediate (or proprietary) archive on 5.25-inch magneto-optical platter and is planning to have the DAAC as the public archive, which would mean a combination of 12-inch WORM and VHS. TRMM is still TBD.

So, as you see, we've got projects that are in really very different stages. ROSAT has been operational for several years, SeaWiFS is about to be operational, and TRMM is still very much in the design stages. So I have a real interest in the discussions going on here.

In terms of volume distributed, I called up ROSAT. I haven't actually been involved with ROSAT for several years. I called up Cynthia Cheung and she told me for the month of August, 334 requests were made. These comprise 6,000 files ranging from 1 to 10 megabytes per file.

For SeaWiFS and TRMM, we haven't started distributing data yet, so we don't really know what we're going to be up against. We have estimates based on their predecessors.

Mode of distribution available: ROSAT is shipping out uncompressed data, mainly electronically. That's the way they prefer to deal with it. Again, the volume is such that it's manageable at this point. They do get some 8-millimeter and 9-track requests, but the demand is minor, really, compared to the network.

SeaWiFS' mode is dependent on the DAAC, and I suppose TRMM's will be, also; and, as well, of course what mode the science communities wish to receive the data by.

Wishes and responses to problems: maybe we'll get into these as we get further discussion. I don't want to ramble on here, but there are a lot of wishes that the various groups have. So I'll wrap it up with that.

MR BERRY: And the last panelist is Darla Werner, who is section manager of integration and technology assessment, affiliated with the Hughes STX Corporation. She is project manager for the Landsat Digital Archive Conversion System and managed EDC's digital archive computer operations and technical support areas for over ten years. She implemented tape baking in April 1993, which I believe is a solution for some of the problems where the lacquer on the tapes is peeling off.

MS WERNER: The EROS Data Center is a U.S. Geological Survey Facility. It was established in 1972 in Sioux Falls, South Dakota, to receive, process and distribute Landsat data. EDC was designated as a national land satellite remote sensing data archive in 1992. EDC archives over 10 million space and aircraft images of the earth's land surfaces. Three million of those images are Landsat.

The national archivists focussed on developing advanced data archiving and retrieval to permit more efficient storage and retrieval of the large amounts of data that we will be receiving in the next 10 to 15 years.

As far as the digital archive, the facility is over 12,000 square feet of environmentally controlled storage space in the lower level of our facility. It is accessible to the computer room via an elevator, and all other accesses are card key. The original 4,400 square feet were constructed in 1978; and 10 years later, we finished off another 8,000 square feet to accommodate the early historical Landsat wide band videotapes and also to create an overflow area for 9-track and 3480.

As far as security controls and environmental controls, we do follow the National Archives' Code of Federal Regulations for Electronic Records Management and also use the Care and Handling of Magnetic Storage Media publication from NIST.

The EROS Data Center has made a major commitment to the long-term preservation of data. We currently have two media conversion projects in process: copying 9-track tape to 3480 and also transcribing Landsat data from the 1-inch high density tape to DCRSi digital cassettes. We began baking sticky tapes in April of 1993, and we have experienced a 100-percent success rate with that venture.

As far as our current storage media, our primary media is still the 9-track tape; however, we are copying to 3480. We did start with about 105,000 9-track tapes 3 years ago, so we have made some significant progress. As far as 3480, we have about 50,000.

The 8-mm cassettes are used as system backups, and they are used basically on all of our major systems throughout the building. QIC tapes are used by the users on their work stations, and I do believe that we have more QIC tapes in the building. This is all that is registered in our digital archive, but users tend to keep them in their desk drawers and wherever.

We currently have at the Data Center about 37,000 1-inch high density tapes. We have another 26,000 tapes that are stored in Alexandria, Virginia, that are being incrementally shipped out to the Data Center. We will be transcribing the 26,000 tapes which hold the TM Landsat data and also these 13,000 high density tapes holding some of the older, or the more recent, Landsat MSS data.

The DCRSi cassettes are the result of a conversion that we are doing, which began in December 1992. By the way, these 200 cassettes hold the data that was on 7,200 high density tapes.

CD-ROM: we have what's reported here just a small number less than 200, but again, this is what's registered in the digital archive. If you look in the offices at the EROS Data Center, I believe that we've got hundreds and hundreds of CD-ROMs. They're quite popular.

We have two robotic systems, an EPOCH file server that is used for browse images and electronic file transfers, and also a newly installed STK silo, which is used for raw data sets to be used later for image processing.

As far as distribution, our primary media is 9-track tape. We put out very few 3480s on a monthly basis. We'd like to see that increase. And 8-mm cassettes have been requested more often, even just within the last six months.

As far as issues, problems, and challenges, the first is the rapidly changing technology and the challenges of technology obsolescence. I believe we need to put more emphasis on retrieval. Due to the changing technology, the question I often ask myself is: are we going to be able to play back the data 10 or 15 years from now that we've recorded today?

Also, our distribution data sets are getting larger. Because of the large size, the distribution media options are more limited. We are starting to put out more one gigabyte-sized data sets, and naturally, that's a lot of 9-track tapes, which most universities are still using. There are some universities that have been requesting 8 mm, but because we like to verify our products before they go out the door, that's a very slow process for product generation.

We wish there were more 3480 users. We are watching advances in the 3480 technologies. Media management and maintenance are expensive. As our digital archives grow, there's more and more tasks that are associated with maintaining an archive. We can't just put a tape on the shelf or in the archive and then call it done. Media maintenance and management require people and specialized equipment, and that costs money.

As our digital archives get older, we have to convert to newer media or advanced recording technologies, and conversions cost money and take a lot of time. I also believe that conversions are a never- ending process.

Data management is a science. It involves many tasks and considerations. It's not just archiving and it's not just data handling. It's defining metadata and knowing the environmental conditions and specifications for good archiving practices; making sure that data is going to be available and useable by future users.

I believe that data management is a very complex system of processes that depend on one another. For many data facilities, I believe that data management is a number one problem as far as having funds allocated. I think that more budgets need to have data management as a line item versus just a category under a project. It seems as though when funds are allocated for data management, that the moneys tend to go into the systems that are used to record the data; and there's very little resources that are available for archiving and the tasks that are associated afterwards.

Thank you.

MR BERRY: Thank you.

Now you know a little bit about the panel. First of all, we would like to entertain questions from the floor or, alternatively, we'll discuss a series of points. This is your opportunity to ask questions of the panel, to get their opinions or find out more about what they've been doing.

I'm going to start it off with a question. What I'd like to know from the panel is the following: I notice you are using things like 3480, 7-track, 9-track. Where do you anticipate going in the future, let's say two years from now, three years from now? Are you still going to be in the same place or will you be in a different place?

MR DAVIS: We'll be still 3480 or perhaps 3490, if we get that capability.

MR BODDEN: For the Version 0 DAAC, which is part of the EOS project, we anticipate staying with the Metrum, which is giving us a lot of good use. And also with the Cygnet jukebox. But the problem for Version 1 becomes a lot more complex, and they will be receiving up to one terabyte of data per day to process. So the EOS project itself has to still be looking at what kind of options and alternatives are out there that can handle this kind of load. So there is somewhat of an open page here as to where we're going to go in the long run.

MS WERNER: For our long-term preservation of data in our lower density archives, we have made a commitment to go with 3480s, but we are also looking at the advances in that technology. The 3490 looks to be a promising substitute for that.

In our higher density archives, we will be using the DCRSi cassettes, but longer term, we are keeping our options open.

MR BERRY: Could each of you speak to the question?

MS EATON: Most of our files are much smaller, and what we are interested in doing is actually downsizing it. What we would like to do for most of our users, who are not -- they're more interested in specific information from the files as to try to move to floppies to actually send it out, so that it can be more exactly what they want.

We currently use 9-track and 3480. Most of our users are from universities, so they have the mainframes. But if we could move to other modes so the PC could be used, we could get a much wider distribution.

MS POTLER: Again, in my situation I'm not so much looking at archiving and distribution of the final products but finding ways to keep data, large amounts of data, near-line so that I can do reprocessing in order to get the final product and give it to these folks.

We are keeping our eyes on the market, trying to figure out what the best solution is. I haven't seen it yet, but we're looking. We're trying different things.

MR GOTTLIEB: The Data Center has an idea of what to do in long term. We currently are supporting 9-track, 3480/90, 8 mm, 4 mm, 12-inch optical, CD- ROM, and adhering to the ISO 9960, and we're looking to progress into other areas. We are currently looking into D2, D1, D3, when and if it becomes available, quad density platters, optical platters, blue laser CD-ROM, which will be a 2 gig CD-ROM.

But we also have to continue to support the user base, which may be regressive in the current technology trends, so that right now we can't plan on migrating everything to a forward technology and lose the capability of being able to provide data to our user base.

MR BERRY: Okay. Thank you.

VOICE: (Off microphone.) The issue of archive came up with six panelists; they probably identified about eight different long-term archive media. The first question is, I guess: Is anybody working the issue to say the United States shall go to -- in some kind of synchronous fashion to D2 or whatever? And if that were to ever occur, what would be the impact?

MR GOTTLIEB: The answer is that the decision of how an archive actually manages and stores its data is an internal question. And yes, there are standard activities being processed by the National Institute of Standards and Technology. The question is whether or not excluding certain media out of the marketplace would present problems in the U.S. and whether or not there could be more than one sanctioned archival media.

So what the data centers tend to do is look at those media which are currently providing adequate storage and recovery, as well durability for the long-term archiving, with philosophies of future promotion into more sophisticated media.

MR DAVIS: I would take perhaps a little exception to the internal situation with the archiving. In our case, we must go ahead and keep our archives in a format that will be transferable to NARA in the future. Therefore, it does not become a totally internal situation of how we're going to keep those.

Right now, they accept the 9-track or 3480 in ASCII or EBCDIC. So we attempt to go ahead and do that for those files that we know may well be transferred to NARA at some time in the future. So there's always a problem with the term archive and long-term retention.

Archive is for permanency, real, in perpetuity, where long-term retention *is temporary*.-- NARA says that temporary can be up to 75 years. Well, we're in the process of trying to keep our data at least for 75 years before we turn it over to the Archives in many instances. But we must follow their dictates.

MR BODDEN: Let me add just one more thing. I'm not so sure that it makes sense to go to just one standard media for archives, and I just want to quickly point out that in the Goddard DAAC we have selected two different types of media for two very different reasons. We went with the optical platters for what we consider our most highest priority data, most important data, and the data that we could spend money on.

For the VHS system, our Metrum system, what we selected there was a media that provided us a very economical amount of storage per terabyte or per gigabyte, whatever you want to call it. So there are two different types of media that are being selected for very different reasons. So that's some of our justification.

MS WERNER: In 1988 we went through a lengthy process of reviewing the types of media that were available at the time for both our low-density and our high-density archives. At that time we elected to go with the 3480 to replace the 9-track tapes. But, with the large amounts of data that we have at the Data Center, we have to be more conservative with our choices rather than choosing maybe the newest technology at the time. So we have made a commitment to go with 3480.

MS EATON: If I can second with the last two comments that speakers have said and then add an additional thought. Because of the diversity of formats used by agencies, the National Archives uses standards to ensure that we will be able to process those files deemed to have long-term value and that is why our current regulations cite 9-track and 3480 cartridge. We can find drives that can read the data, so unless there is a problem with the tape itself, we know we will be able to process the file. There is also, though, the issue of what do you need the information for. If the information is scheduled to be transferred to the National Archives, then it must be in a format that we can process. But if it is current information that your agency will need for five or ten years, and it has been scheduled as temporary, then you should use whatever format is best for your institution. We don't feel that we have a right to impose standards on temporary information. We can give suggestions, but it is really up to the individual institution as to what they use.

MS POTLER: Well, I keep coming to these things hoping that I will hear from various committees the answer we are all seeking. There is no one good answer. I agree with everything that's been said so far. It's interesting that every time I come to these, I hear about more and more technologies. It's diverging instead of converging. It's exciting, it's interesting. I'd like to see more work done in terms of committee work or various organizations getting together to try and do some more standardization of what's already in existence, because once we do commit to something, we have to stick with it and make it work. And I'd like to see more elegant software to support a lot of the hardware technology and so forth. But I agree with you that standards is a big issue right now.

VOICE: My question deals with the use of compression and what experience either using the industry standard 3480 -- I don't know the name of the compression algorithm *(Editor: IBM calls this IDRC, Improved Data Recording Characteristic, and has licensed it to other manufacturers under names such as Improved Character Recording Characteristic, or ICRC)*-- for some of the more specialized algorithms which might be applied, especially considering the cost that CPU power is decreasing at a significant rate now. Has anybody had any experience with applying compression techniques to the data? And does it impact your error rates?

MR DAVIS: We are just experimenting now with compression techniques on the 8-mm tapes on the EXABYTE drives. In some recent tests we found a compression rate of about 8 to 1 for the NEXRAD radar data, which *meant* we got about 38 gigabytes on a tape, which sounds great. But then when we read it back, it took a little over 20 hours to read it (*Laughter*). We didn't have any error rate problems, but our concern is, of course, that if you get into the middle of that thing and get some sort of little burp or something like that, you've got a lot of time invested in that one tape.

We'll probably use that compression and go to something less than the 8 to 1, maybe 5 tapes to 1 or something like that, where it would really be beneficial to us.

MR KOBLER (NASA): If I could just interrupt for one quick second. I know Sandra Woolley is in the audience, and she will be doing a paper the last day. I invite you to listen to that paper. That might address some of your questions, unless she wants to respond to that now perhaps.

MS WOOLLEY(Manchester University, England): Thank you. Yes. Data compression does impact your error rates. If you have, say, a single uncorrected bit pass through the system, it can scramble all data to follow, and that's the main theme of my talk. Robust error control is absolutely essential to preserve data integrity. Thank you.

MR BODDEN: Continuing to talk about compression, at Goddard DAAC we've also just started looking at compression as our on-line system is just really getting going now over the last few months. We are looking to compress at three different points. We transfer data over the network from different data projects, and we were looking to compress the data at these points. We've had difficulty getting that started so far.

The second point that we're looking to compress is, as we receive the data, process it, then we put it to the archive, we were looking to compress it at that point. And that has yielded some very good results. For one class of data, like AVHRR (Advanced Very High Resolution Radar) data, we're getting a 70- to 80-percent compression rate. Each file is about 240 megabytes in size, and we're able to compress that down quite nicely. So that has worked there.

We've also been trying to compress the data as we write it out to media to send out to researchers and scientists, and we're just getting started with this. In all of these first three compression techniques, we're using just a very standard UNIX compress. We're not going into any fancy compression algorithms, but we have looked into them. And we've found that the UNIX compress doesn't give us as much as some of these other algorithms, but, given the difficulty that the researchers would have out there in handling the different kinds of compression, we stuck with just a pure UNIX compress.

MS WERNER: We've attempted to apply IDRC on our STK 3480 rack mountable tape drives, interfacing with SGs and DGs, and have not been successful at doing so.

PANELIST: We have not tried to compress our data yet. There hasn't been the need.

MS POTLER: ROSAT doesn't compress. SeaWIFS is just starting to look at compression algorithms, so I don't have anything to say about that. TRMM will definitely have to compress at 60 gigabytes a day, but I don't have results yet.

MR GOTTLIEB: The Data Center actually uses compression but not on most of the scientific holdings. The places we've encountered compression and had it successfully implemented and then extracted again was on the CD-ROMs that the Data Center distributes. There is a concern that taking compression techniques and applying them to data, there is a possibility that you will find some data loss. And when dealing with pure bit streams where every bit is either meaningful or unmeaningful and having that change and become meaningful is a real concern. So there is a danger in that compression will not yield 100-percent accuracy all the time.

DR HARIHARAN (Systems Engineering and Security): What was the problem in writing out the data in compressed form on distribution media?

VOICE: Right. The problem wasn't so much in trying to compress the data. It was just -- it's a new function for us and we haven't got it working yet. That's all it is. We will get it working.

MR BERRY: The question was: what are the problems that he has experienced in using compression on his data?

VOICE: A question for the National Archives. Why is it necessary to centralize all the data at National Archives? Why not network the data and have each agency who owns the systems that are unique to their data archive them in place with you maintaining some index?

MS EATON: That's an interesting possibility. We have not considered that because we don't have access to a network. There is also the concern about documentation. It is only by working with agencies when they transfer files to us, that we're able to determine what the problems are with the documentation. Further, it is only when we can compare the documentation to the file that we know everything is complete. Too often, the documentation is incomplete and additional work is required. So, there would be a problem with the agency maintaining the file, unless they ensured that the documentation was complete, which has not been the case to this point.

As I alluded to in my talk, when the National Archives commissioned the study about current federal data bases, they specifically excluded scientific data bases. And actually, my friend on the far right has been dealing with the National Academy of Sciences' study in which they're looking at what should be done with the scientific records. My personal view is let the agencies keep them, since the agencies have the expertise. I don't know what the study will recommend, but perhaps it will be close to what you recommend.

There is also, with the National Information Infrastructure initiative, the idea of creating a government information locator system. So that might be a way of going about it as well. There's a lot going on with these issues, and I'm not real sure what the direction for all of this will be in five years. But at least for now, the National Archives accepts custody of files in order to maintain the integrity of these records.

DR ANDREW OGIELSKI (Bellcore): Our panelists represent publicly funded archives. What projects are under way in your institutions to improve access over the data networks?

MR DAVIS: Right now we are working on and have several files on line through Internet that are free to the scientific community over Internet. These are both metadata files, inventories, where you can browse and in some cases you can go ahead and go a step further and actually order off-line data that way, but then we also have actual data files out there for, in some cases, the most recent period, like the last month or two, that you can access and use. That effort is expanding fairly rapidly at our center.

MR BODDEN: For the Goddard DAAC, part of the mission for EOS is to try to bring the Earth science data that NASA and the affiliated agencies, such as the U.S. Geological Survey, hold, to bring this data on line. So, that's our mission. As we bring it on line, we are also going to provide network access for the world to log into the Goddard DAAC and browse through our data holdings, which will be represented by metadata records, summary records, of the data.

The researcher will then be able to select, during this session, some samples of data up to a certain limit. That limit we haven't really set yet. And that data, if it's small enough, the amount that the user has selected, can be FTP'd back to the user during that same session. So we are trying to set up a system where you can research your data, you can access it, and retrieve it, all during the same session.

MS WERNER: The EROS Data Center has developed a system called a Global Land Information System (GLIS), and it is available to users on the network for access to US and foreign Landsat data, AVHRR, and other miscellaneous earth sciences datasets. The GLIS, as we call it, has the capability to allow research; and some browse files are available, so that the scientist or user can actually see what their area of interest might be like.

If you would like information on GLIS, please talk to me afterwards and we can give you the address for that.

MR BERRY: Will the panelists make sure you speak into the mikes so everybody can hear you?

MS EATON: What I'm going to say next really applies to the entire National Archives. The National Archives has begun to put some of its selective guides on the Internet so people can get a taste of what is available. We have our title list that is available on the Internet, but that is the only thing that we can provide across networks at this point.

We are hoping in the next couple years to determine if there are some files that possibly should be included so that people could access it that way.

MS POTLER: We certainly have been working on the browse capabilities in conjunction with the DAAC. The hardware itself is not the problem. SeaWIFS has both Ethernet and FDDI connectivity, which is more than adequate to handle the load right now. The problem is the users don't have the high-speed connections, and they have so many hops to go across and so forth. So it's more in terms of what we have to deal with our audience, our user base, and what they have to deal with.

We do support anonymous FTP and so forth to get the data and the browse capabilities. We're trying to improve the software so that they can work with it more easily.

MR GOTTLIEB: Well, it turns out that the Data Center actually has an on-line system that is actually -- to the users it is a mail interface system. So by simply sending a mail message to the system, it...

MR BERRY: Could you speak into the mike a little bit more?

MR GOTTLIEB: -- will actually stage your data into an anonymous FTP area and you can come and get it. As stated before, the Goddard net is a T1, and, as we go outside -- actually it's a hypernet -- as we go outside, we find that 9.6 may become very painful to transfer, you know, three or four megabyte files to a 9.6 station.

There is a selective process that a committee meets on as to which data files are put into the on-line system; but also as a distribution center, the entire holdings that are cataloged are available through other means and media requests. So you can actually write to the Data Center, the User Support Office, which I can give you more information on, and acquire any of the catalog holdings through the National Space Science Data Center.

MS POTLER: I'd like to add one more thing. Even though the users don't have the FDDI or the even higher speed networks, by putting some of the Goddard systems on the FDDI, we can send data to the DAAC via FDDI, and therefore limit the contention on the Goddard Ethernet, which is a dramatic improvement. So there is that advantage.

VOICE: I have a two-part question. The first part is for Laura. If you had to make a decision now as to what type of storage technology you were going to use, do you think you would go for the 3480s? Or would you go with perhaps the Metrum, or another optical?

MS POTLER: Can you be more specific? Are we talking about TRMM launching in '97?

VOICE: Yes.

MS POTLER: If I had to make a decision now for archiving, for distribution?

VOICE: For archiving.

MS POTLER: For archiving. I -- you're really putting me on the spot here in front of all of these people. I would -- if I had to make a decision now, I would go with WORM.

VOICE: Okay. I'm afraid I'm going to put somebody else on the spot. For Darla and Dick, you both have a commitment to the 3480s, and it sounds like you have an awful lot of data to store.

Doesn't that become kind of a management nightmare for all those cartridges that you're storing?

MS WERNER: Actually, copying the 9-track tapes to 3480, is a lot less of a nightmare using the 3480. We do have a lot of data, but we have seen about a four time decrease in space requirements. The advantages of the 3480, as far as speed, reliability and just the easy handling make that an easy choice over what we have right now.

MR DAVIS: And my answer is yes. (*Laughter*)

VOICE: Lee has gone on to the Metrum and I think -- it sounds like he's satisfied with it, where you get a lot higher density per cartridge. Have you all thought about that? It probably wasn't out when you made your decisions.

MR GOTTLIEB: Is this question to me?

VOICE: (Off microphone.)

MS WERNER: Excuse me. Could you please repeat the question?

VOICE: Well, I guess the Metrum technology -- I'm asking a lot of questions about it because I'm looking at it. It probably wasn't out when you all were making your decisions. I guess, if I understand correctly, it would probably decrease the number of cartridges even further. Have you all thought about that at all? Or are you all firmly committed to the 3480s now, so it can't really be an issue?

MS WERNER: Right now we are committed to 3480s for our lower density data. We are always keeping our eye open and watching for current technology and future advances. As I stated before, we tend to be a little bit more conservative because of the large datasets and large data volumes that we deal with.

MR BERRY: Let me give a corollary question to that, because you all have picked a technology that's relatively expensive from a per bit standpoint. There's sort of an implication here that cost is not a driver. If you look at the relative cost of storing something in one media versus another -- that's some of the dramatic differences between media -- is the cost of storing or even sending data to someone.

And it also tends to preclude -- most workstations, for example, don't have that capability. Certainly almost no PCs do. So are those kinds of considerations having any impact in your systems? I mean, you're using historically what have been sort of the mainframe kind of approach.

MR DAVIS: From our standpoint, of course, cost is always a concern. But the initial cost of the new systems, when you're talking *big bucks, is important*. Well, if you take a CREO system, a dual drive CREO system, for example, which we've been looking at, and we'd like the optical tape option. We'd like the permanency for our function, which is primarily archive, and has good recall. But you're talking about a situation there where you've got about $750,000 to $800,000 initial price in order to buy a two-station system, and that's a pretty good chunk of change, particularly for fairly new technology, even though I think it's very viable technology.

The cost of media for 3480s is down about $5 a cartridge. We're paying more than that even for the 8- mm stuff right now. We, too, are constantly looking at new technology. We're always open to new ideas, but we do have to take the more conservative approach for our basic archives.

Now for the NEXRAD system where we're using 8 mm, we have no choice because the National Weather Service has installed and are installing those recorders in the field, and we had no option on that. That's strictly an economical situation there. The drives are reasonably priced. The media is reasonably priced for the amount of data that you can get on one of those. So we're forced into that. We really are not going to convert those to 3480 and give them to NARA.

DR MARIA ZEMANKOVA (MITRE): My question is on data base management; that is, are you satisfied with the current state of the art of data base management systems so that given that we can store all this stuff out there, we can actually get the information from it that we need?

MR BODDEN: I'll start that response for the Goddard DAAC. We anticipate that, as we approach our 20 terabyte goal, that our DBMS problems are going to become very severe. Right now our data base is performing quite well, and we don't see that changing in the very near future. But we have started talking with vendors, and I'm not going to tell you which ones, but we have started talking with vendors who have new and innovative approaches to data bases where they will distribute one data base over several platforms and control that data base through several other CPUs. So there are new ideas out there, and we are exploring them for the not too distant future.

MR BERRY: Anybody else?

MR GOTTLIEB: There are two aspects of the information you're talking about. I think maybe what you were addressing was both the pointer to retrieve the actual information, but also what, at the Data Center, we might call metadata, which is information about the file you might want to retrieve. And the answer is that I don't think you have to position yourself to use a single data base to answer that question. So that would be the first approach I would say, if somebody were to say "my data base can't handle it anymore" -- perhaps you have to reengineer it into several data bases.

But also with medium proof -- there's also a host of other technologies that are improving, perhaps a little more slowly or even more quickly, and data bases are progressing in some fashion. I know object bases have become available, and object bases, I think, start handling the metadata questions where a traditional RDBMS could actually point into the archive and direct you to retrieve a file.

So I think that the answer to your underlying question is that we will have to rethink how the traditional data base is implemented as the archives grow.

MS POTLER: Well, I think the DAAC, in particular, that I'm familiar with is doing a lot of work in terms of metadata and organizing with the user in mind. I know they're holding the projects to the line on these things.

My concern is more with how to get the data out of the hardware, the media, to locate it and get it out rather than what particularly you're looking for within the data base and how the data is organized. I would like to see more elegant handling of the data as it's stored and as it's retrieved, that sort of thing, is what I'm seeing. I see a lot of jukebox-type mechanisms, but not necessarily elegant software to use it.

MS EATON: We have developed a data base for capturing the metadata so that we can do automated validated files that are transferred to us. That project has been ongoing for about a year and a half. It's still not fully operational, but we're trying to capture the metadata in that way.

And we are also considering building a system that we call X-WEF so that we can get specific information about files, across the files, to do a reference system, and we're just beginning looking into that now.

MS WERNER: We have several metadata data bases to handle the different data sets that we have at the Data Center. It doesn't seem to be an issue for us right now.

MR BERRY: We have had a request from the audience. If all of you would be willing to provide the Internet addresses where some of this data could be reached, then we could make it available to the participants. What we will do is make it available on a sheet and have it out front so that you can pick it up later on during the conference. We will do that for you for those people that can supply the information.

VOICE: There have been a lot of questions as to whether in ten years will there be the drives to read particular types of media. What about the flip side of the question, the software format of the data on the media? Are any of you facing the issue of supporting either computers or software packages simply for backwards compatibility to be able to access data that you would just as soon retire?

MS EATON: We cheat a little bit. When certain agencies transfer files to the National Archives, they give us two copies of the file, one in ASCII, the other in SAS or SPSS format. We will copy both formats. We will make the dependent format available to researchers for the first ten years, because we will feel that that version will still be supported by the software that's out there. When we recopy the file after ten years, we do not recopy the dependent file. We just keep the software-independent file at that point. So that is how we cheat.

MS POTLER: I'm not facing that problem; that is up to the DAAC.

MR GOTTLIEB: Actually, we are facing that problem, and the way we're dealing with it is we currently have another committee to attempt to describe and conclude what the best formatting techniques will be. One of the ways we are thinking about solving that is basically internalizing a standard format for the archive so that we could create an *in* filter and an *out* filter, which could either ingest or export any of the required formatting necessary.

MS WERNER: This has been a big problem for us. We have tried likewise to develop an internal archive format, but many of our datasets are in a native format. So that's what we've elected to go with so we don't have to rewrite documentation and such. But we are trying to go with an internal archive format that would include an ANSI standard label, and therefore, to reduce the amount of software required to use the data.

MR DAVIS: We're basically doing the same thing. Of course, we're obligated to provide data to customers, as well as have data for use in our own center for as far as we can see into the future. So it's important for *users* to be able to go ahead and read these data.

We have basically an internal format for our standard records that we use; but we also get a lot of stranger tapes that we get from other organizations, and we have to be able to read those and define what's on there and that sort of thing. So it's a pretty good maintenance programming job to keep up with all that.

One hopes that the industry will go ahead and, as it progresses, will allow you to progress in some reasonable fashion rather than a fruitbasket turnover-type thing.

MR BODDEN: For the Goddard DAAC in dealing with this issue, we have to look at it in two different parts. For the data that is supported by commercial software, that we access through commercial software, that's sort of setting a standard for the data, and, by that, I mean that we are now asking projects that transfer data to us to put the data in a standard format, such as HDF, which is supported by NCSA. Or there are some other formats that we would be willing to take but are not sanctioned by the EOS project, such as CDF.

One concern that we have is that our archive is being placed under a file management system called UniTree, which so far we've been doing okay with. But our concern is: where is UniTree going down the road? And is our archive going to be able to evolve with UniTree? Or are we going to have to at some point move into some other kind of file management system?

So that's a concern for us. And I might say that the people who we buy UniTree from, Titan, have been very responsive in trying to meet our goals with Unitree.

And then one last aspect of this question is the application software that is used to access certain types of data files, and that is a real concern. That's something we want to avoid in the DAAC in the future. We have a lot of old data that is being accessed by programs that have been written 10 or more years ago. So that, we want to try to shy away from, and want to try to move towards standard formats for the data that we're receiving.

DR KING (National Space Science Data Center): Our data centers have a dual responsibility to provide convenient access today to data in the data centers, as well as to preserve the data for the long term so they will be available 50 years or whatever downstream. In your discussions of data media choices, I've not heard that particular distinction brought out. In fact, might the optimal scenario be one where one type of media are in our jukeboxes, providing that convenient, current access, and perhaps the same data on either the same type or a different type media in our deep archives?

MR DAVIS: Well, we keep statistics on what our customers want. Of course, there are some customers who want everything and they want everything on-line, and we're not able to go ahead and do that. We just don't have that capability from hard disk or jukeboxes or anything else at this point.

However, we have noticed some trends. For example, we no longer get any requests for 7-track tapes, and we are getting fewer and fewer requests for 9-track, 1600 BPI tapes. We are seeing an increase in the requests for 3480s, and 2 months ago for the first time, we discovered now the most popular, most frequently requested format is on floppy disk. We are also seeing a great increase in the request by customers for data on 8-mm tape. We do have some CD-ROMs. We don't produce those at will, but we have about seven of those that we distribute.

So we have seen a definite trend towards the users of particularly floppy and 8 mm, and our obligation is, and what we do is, we take data from our 3480 files, we will go ahead and put out data files in any of those formats that the customer wants.

MR BODDEN: Yes. The same thing for the Goddard DAAC. Internally, we store the data on optical platters and VHS tapes. We used two different types of media because of the importance of the data. We don't want to put all the data in just one type of media and have a single point of failure. We view the data as very important, number one, and it is very hard to replace. So a lot of this data is coming down from satellites that once they're -- it's just irreplaceable.

As far as distributing the data, we distribute on popular media, such as 4 mm, 8 mm, 3480s and the 9-tracks. Let me cancel the 3480s. That is available but only through special request. And the same is true with optical disks. That is available through special requests. So we handle quite a full range of media.

MS WERNER: Our Landsat data will all be transcribed to DCRSs is, so we are being consistent there with one media choice. And there will be about 50 terabyte of data.

Our lower density, as I've mentioned, we're using 3480. We've got a commitment there, and we've got about 35 terabyte of data that will be going to 3480. However, our distribution media, we, too -- it's driven by user request, user demand, and somewhat by what we offer. We are still putting out a lot of 9-track tapes. We would like to see that move toward more 3480 and 8 mm. In the last 6 months, we have seen a big increase in the use and demand of 8 mm for file transfers.

MS EATON: We have had to use the computer center for doing both our preservation work and reference work, so we've been very limited in what we could offer. We are trying to build an in-house preservation system. If that works, we would then build a reference system, and we would probably try to hang at least floppy drives from this system, as well as other types of drives, if there are enough requests for another format. It's one of the issues we are looking at.

MS POTLER: Obviously, there has to be different criteria for distribution and archiving and backup, which is something we really haven't talked about here and is a major concern, as well as internal reprocessing and so forth. I spoke with ROSAT project recently to see how they're doing, now that they're in operation, and how they feel about it. One of the things they said that they're looking into is that they have the proprietary and the public archive on WORM, because they have the most confidence in WORM. But at the same time, they have all their eggs in one basket, and if something goes wrong or the company goes out of business or whatever, they're stuck. It's certainly a major concern.

SeaWIFs took the opposite tact. They took everything. They have 8 mm and 4 mm and 9-track and MO. As long as the budget allowed, we got a little bit of everything to be covered. But it would be nice if there were some way to narrow that down a little bit.

MS WERNER: Good enough.

MR GOTTLIEB: I guess I was kind of set up by the questioner, but at the Data Center we use dual media philosophy and that is we actually take in a media and as best we can immediately produce the second media and even go one step further: try to get it to a secondary storage site so that we have an original and a safe copy which is termed off site.

Unfortunately, the world is not utopian, so we are struggling with how to back up some of the data that has arrived electronically, and we did not have an original media choice.

The other dilemma that we're faced with is, how to back up half a terabyte of data in a reasonable fashion without impacting our requesting community's throughput. So the Data Center is currently supporting a dual media philosophy and are coming up to speed on getting the dual media throughout the entire archive and also supporting off-site storage.

MR BERRY: I think we have time for one more question. You'll notice that even the panel has started to leave, so one more. (*Laughter*)

VOICE: I have a question for the panel. How important is backwards compatibility in your decision-making process? If you take as an example 3480, 3490, 3490E, you've got a clear migration path for where you are going to head with your capital investment over quite a number of years. Or, are you concerned about that you downsize your archives to the extent you can forego that migration path and go into what we call leapfrogging technologies?

MR DAVIS: Backward compatibility is critical to us because of our customer base. We've got -- although we get about 90,000 requests a year at the Center, that's not all for digital data. We have somewhere in the neighborhood of 2,800 requests or so per year for digital data. All you've got to do is to stop being able to provide one type of media or one format, and you hear about it very, very quickly.

So, it's important to us, the backward compatibility. But, on the other hand, that means that we've got to be able to do that to the customers. The only thing we've stopped in the last 38 years that I've been there is punched cards and 7-track tape. We can still do everything else, and I think we will continue to have to do that for some time to come.

MR BODDEN: Backward compatibility is also important for the Goddard DAAC, but we are using the EOS project as a breakpoint where there are certain data sets that we will no longer support the backward compatibility or the old version of these data sets. And we're actually migrating them forward to a new technology. An example of that is the coastal zone color scanner system data, CZCS, some of you may know. This data was produced and was available through a VAX VMS system.

We are now in the process of moving this data over, going through all the bit and byte conversions to move it over to a Unix system. That is our intention, really, for all of the old data to little by little bring it on line in our new Unix system.

MS WERNER: Backward compatibility is very important to the EROS Data Center. Even though we are upgrading to new technologies, the data migration, data conversions are a lengthy process. So, for several years, we need to make sure that we can use the data on the older technology to allow our users to access what they need for their project.

MS EATON: Since we often get the older technologies, it is very important to us. As I was telling someone, we still receive 7-track tapes from agencies, so we have to have that functionality.

MR GOTTLIEB: The issue of backward compatibility, I think, can really be addressed as to whether or not it is an archive concern or a distribution concern. And if you manage your archive as progressive and even conservative, I think the entire issue of backward compatibility goes out to a distribution concern.

At the Data Center we do live and breath with that concern every day. I don't think we've had a recent request for a 7-track tape, but it wasn't too far back where somebody actually did request a 7-track tape. Fortunately, we still had a functional drive with which we could fulfill that request.

So I think you have to look at the whole scenario, and the question you need to ask is: is the backward compatibility an issue for the archive? I think the answer is no, if you manage your archive progressively. But it always remains an issue for your community support, depending on how flexible you wish to be in supporting that community.

VOICE: (Off microphone.)

MS EATON: Ten years.

MR BERRY: Let me repeat the question so that other people can hear it. The question is: "Would you consider a system that did not have a clearly defined migration path from where it is today to where it would go out into the future?"

MS EATON: No. We always look at things that we know we will be able to access in 10 years.

MR BODDEN: The answer for Goddard DAAC is also no, we would not look at a system that did not have a clear migration and evolutionary path.

MR DAVIS: I believe that we would (*laughter*) if there's such an animal out there. Yes, we'd definitely look at that and actually have been looking at it. I would say, for example, a jump from 3480 to a CREO system would be the kind of thing you're talking about as a leapfrog, and I'd have no objection to something like that at all.

MS WERNER: The EROS Data Center would need to have a technology with a clear migration path. However, in our long-term archive for the Landsat data, we did go to a different technology. So I don't know if that would be a leapfrog or not, going from high density to DCRSi.

MR GOTTLIEB: I guess the Data Center would answer -- I don't -- it would have to answer: do you mean a media migration path or a migration strategy?

VOICE: Migration strategy.

MR GOTTLIEB: Then the answer is without a migration strategy, no. But without a clear media migration, it's possible.

MR BERRY: Okay. I'd like to thank the panel. I'd like to thank the audience for your participation, and I think we have a poster session scheduled for 6:00. Do you have any announcements?