Digital Library Issues Arising from Earth Science Data

Bruce R. Barkstrom
Atmospheric Sciences Data Center
MS 157
NASA Langley Research Center
Hampton, VA 23681-2199
1-757-864-5676
Bruce.R.Barkstrom@nasa.gov

John J. Bates
National Climate Data Center
Asheville, NC
1-828-271-4378
John.J.Bates@noaa.gov

ABSTRACT

NASA, NOAA, and USGS collections of Earth science data are large, federated, and have active user communities and collections. Our experience raises five categories of issues for digital librarianship:

- Organization of the data in the collections is not well-described by text-based categorization principles
- Metadata organization for these data is not well-described by Dublin Core and needs attention to data access and data use patterns
- Digital librarianship requires risk management approaches to dealing with the unique threats to knowledge preservation specific to born-digital information
- Digital librarianship requires careful attention to archival cost management
- Professional reference librarians for these collections my require special training.

This paper suggests three mechanisms for improving the quality of digital librarianship:
- Using a maturity model to assess the readiness of data for accession, for preservation, and for future data usefulness
-Developing a risk management strategy for systematically dealing with threats of data loss
- Developing a life-cycle cost model for continuously evolving the collections and the data centers that house them.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]; H.3.6 [Library Automation]; H.3.7 [Digital Libraries]; H.4.1 [Office Automation] Workflow management; K.3.2 [Computer and Information Science Education] Curriculum; K.6 [Management of Computing and Information Systems].

General Terms

Management, Documentation, Economics, Reliability, Security, Human Factors, Legal Aspects.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
JCDL'06, May 1–2, 2006, Raleigh, NC, USA.
Copyright 2004 ACM 1-58113-000-0/00/0004…$5.00.

1.INTRODUCTION

NASA, NOAA, and the USGS have about a dozen Earth science data centers with some important characteristics that influence their view of digital librarianship:

- The collections in a data center can be very large, with several of these centers having more than a Petabyte (PB) of born-digital data and several tens of millions of files.
- Data centers have some degree of federation, in that they allow interoperable searches that can identify and allow users to order large amounts of data.
- These centers have active and growing user communities who need a diverse and evolving repertoire of services, including data reformatting, fusion, and visualization.
- They have already evolved their systems through at least three major generations of storage technology and a similar number of generations of computer hardware and software.

The Atmospheric Sciences Data Center (ASDC) at NASA's Langley Research Center provides a concrete example of these characteristics. Table 1 shows some of the statistical information available for 2001 and 2004. It is notable that while most of the metrics for work done by this center have increased by a factor of two or more, the number of personnel involved has decreased by about 20% over the time period shown.

Table 1. Properties of ASDC Holdings and Work

Metric	2001	2004
Data Volume of Holdings [TB]	340	1,250
Number of Files	5 M	20 M
Data Volume Distributed [TB]	34	114
No. of Distinct Users	6,000	12,000
Production Jobs Run per Day	2,000	5,000

It should be noted that ASDC both ingests data from external data providers and creates specific data using production software provided by NASA science teams, led by their Principal Investigators. These statistics are not unique – they are similar to several other of the large NASA data centers.

In the body of this paper, we present five categories of issues for digital librarianship:

1. Collection Organization
2. Metadata Organization and Use
3. Threat Mitigation
4. Data Center Cost Effectiveness
5. User Services and Interfaces as Reference Librarians

We suggest three mechanisms for dealing with these issues:

1. Using a Maturity Model for Judging Curation
2. Risk Management
3. Cost Modeling for Continual Evolution

2.DIGITAL LIBRARIANSHIP ISSUES

Born-digital data poses special challenges to data access, data search, and context preservation when compared with data arising from text on tangible media. Our long experience with print has produced conventions about appropriate collection organization, metadata, and other issues. These conventions are not necessarily applicable to digital collections.

2.1Collection Organization

Born-digital data collections have both large-scale and small-scale organization issues. The key factors that impinge on these issues arise from search performance, preservation of knowledge in the presence of technological evolution, and user expectations.

In text libraries, a long historical tradition has led to conventions for organizing collections by subject matter. Both the Library of Congress and the older Dewey Decimal classification give guidance to the spatial organization of collections. Even in multi-building departmental libraries, it is possible for users to consult a spatial map of the collection. In detail, these conventions and their classification labels lead to a quasi-spatial search strategy that allows interpolation search between members of the collection. We identify the issues involved in file collection architecture as issues of large-scale collection organization.

Within text collections, the long history of experience with organizing the physical media has produced conventions for orienting the reader, including pagination, tables of contents, indexes, and conventions for organizing headers and footers on each page. Born-digital data collections do not have this long experience. Thus, there is a natural difficulty in creating conventions for individual data files so that individual elements in a file can be found and cross-referenced. In detail, this kind of issue arises with particular pungency in architecting file formats. We identify these issues as those belonging to small-scale collection organization.

2.1.1Large Scale Collection Organization Issues

In born-digital, Earth science libraries, there are currently no cataloging principles that represent a community consensus. In other words, there are no broadly accepted handles (equivalent to a LoC call number or ISBN identifier) that would help new institutions organize their collections. Of course, in huge data volume collections, such principles do not necessarily govern the physical location or ordering of the files.

Effective engineering suggests that a random organization may be appropriate if files are to be accessed in a random order; sequential organization is probably appropriate if the files are accessed in sequential order. For the Earth sciences, this means that if users are searching for one file at a time, based on a random spatial pattern, then random ordering might lead to good search performance for the users. On the other hand, a global climate investigation may need to access a large number of files in their time sequence. A random organization to the file collection can create a performance penalty of at least a factor of four for such a request.

In addition to performance, data centers need to consider long-term provenance and integrity in their large-scale collection organization. In the Earth sciences, data often play a key role in policy decisions with large economic consequences. Data needs to be trustworthy – meaning that its history of production and preservation should be traceable in a publicly accessible form.

This requirement suggests the need to create a reference classification system that can uniquely identify individual files and identify the original source of the data. It appears that a system similar to that used for generating and registering IP addresses would be useful. Such a system needs both a sound technical basis and an institutionally acceptable mechanism for delegating responsibility for the naming convention. From the standpoint of the authors, one candidate for such a registration scheme would be to organize the classification in a hierarchy based on production history [1].

2.1.2Formats and Small Scale Collection Organization Issues

In the humanities, it is possible to find useful fragments of text in a particular volume by citing “chapter and verse”. We expect books to contain chapters, which may be subdivided into sections, subsections, sub subsections, paragraphs, sentences, and words. While words may be hyphenated across lines, we usually build indexes by referencing the page number on which individual words or phrases lie. This approach allows users to reference and extract useful fragments of text.

Of course, electronic data storage introduces new possibilities for such referencing as we may note in hypertext (or web pages). In our browsers, there are no fixed page numbers. Thus, an index built on pagination is not trustworthy. On the other hand, text fragments are directly referenceable through the use of “anchors”, which allow the browser to switch pages and land the user on the reference.

Because of the binary nature of Earth science data, users have not often encountered the need to reference individual fragments of files. Thus, many members of the Earth science community encounter the detailed layout of data files only through the intermediary of data formats. Such formats play a key role in the usability of the files, since the only way to access the data is through software that can read the file in the format in which it is stored. It is fair to regard the users as broken into “tribal

communities" with distinct dialects, "world views" that include expectations regarding useful data structures, and "tribal customs" that include approaches to citing data and authorship.

However, as Earth science librarianship develops, we should expect that the need for cross-referencing data will increase. The process of identifying hurricanes in a large collection provides a useful example of the need for such detailed referencability. In this example, a science team might create algorithms that identify hurricanes as features in a collection of images. A single hurricane could appear in several files as it follows its path across the Earth. Only a portion of each file in which it appears may belong to the hurricane, so the community of investigators interested in the hurricane may need to be able to reference the individual pixels within the image that belong to the hurricane. Such individual pixels may be important in comparing one algorithm against another or in identifying the same hurricane in other kinds of data, such as microware or radar data.

While it may not be apparent at the outset, there are some special issues arising for Earth science data because of the way in which the data are created. With text collections, there is a relatively stable production approach: the author writes the text, provides it to the publisher, who then creates multiple identical copies. With scientific data, the production process may become more transient and less transparent. Algorithms in executable code produce data products – thus changes in the code, in the parameters that may be contained inside the code or in external files, changes in input data, or even changes in the computers, compilers, operating systems, or hardware storage compression algorithms may all create variant versions of the data files. Furthermore, scientists often string together complex sequences of data transformations that produce final data products that depend on previous products in a way not usually encountered in textual collections. Maintaining a coherent account of provenance may require automated methods of recording complex workflows.

In addition to the issues of tribal customs of data formatting and referencability, data centers also need to deal with technological evolution and intellectual property rights for data formats. In Geographic Information Systems, for example, companies own patents on data formats. In NASA's Earth science data collections, the Hierarchical Data Format (HDF) has been an open source formatting "standard" developed at the National Computational Sciences Alliance (NCSA). A significant library of software that provides efficient access to HDF files and a "self-documenting" capability supports this format. While public funding has supported the development and dissemination of this format, the development group has obtained additional, proprietary funding and may well produce proprietary versions of this system in the future.

To some extent, we can regard the issues of data formatting as being subject to a dynamic equilibrium between the pressures to develop new and improved ways of accessing data and the inertia developed by user communities that must make a significant investment in learning how to use these improvements. This balance is also affected by the economics involved in having data centers maintain expertise and software to deal with multiple formats. There is a tendency to try to develop single standards for all data canters to use that would reduce the cost of maintenance. Of course, developing these standards can require years of negotiation by expensive digital librarians and agency managers. An alternative approach is to develop automated translation tools that can readily create products for users in the format they customarily use. In this case, the expense lies in developing the translation tools. It is the author's opinion that the translators are likely to be easier to implement – and are certainly less susceptible to user complaints. However, they do not expect a single, universally accepted approach to emerge within their lifetimes. Format wars appear to be a permanent feature of this digital library landscape.

2.1.3Summary of Collection Organization Issues

We can summarize the issues involved in collection organization for Earth science data into a modest number of categories:

- User needs for high throughput access with a variety of granularities
- User needs for transformation on input and output, which makes provenance and integrity more difficult issues than they would be if files were static
- Complex production chains influence collection organization and referencing schemes
- Technology evolution, intellectual property rights, and data center costs interact with large-scale and small-scale collection organization

2.2Metadata Organization and Use

Metadata is an important aspect of digital libraries. However, the nature of Earth science data and the way in which users interact with it create some special issues for digital librarianship of these data. To understand these, we need to deal first with the serious differences between metadata based on text collections and metadata for these data. Then, we need to understand how user "tribal" behavior influences the nature and structures of metadata for Earth science data. Finally, we need to consider new models for metadata presentation.

2.2.1Differences Between Metadata for Text Collections and for Earth Science Data Collections

Much of the metadata architecture for digital libraries naturally arises from long experience with text materials. The Dublin Core metadata standard (available at http://dublincore.org/) encapsulates this experience, emphasizing the fundamental categories of

- Author
- Publisher
- Title
- Subject

While the Dublin Core standard includes a richer vocabulary than these, these three categories do not appear to map well to Earth science data collections.

The "authorship" of Earth science data collections is probably best described as collective. Large science teams that include a Principal Investigator, ten to twenty Co-Investigators, and a still larger set of staff, graduate students, and programmers have assembled many of the large NASA Earth science collections. In most cases, the community refers to the collection by the Project title, rather than by the human author(s).

The "publisher" of an Earth science data set also differs from the publisher of a text. Typically, there is no publishing firm for Earth science data. One view would be that the "publisher" is the data set provider. A second is that the "publisher" is the data center that is responsible for long-term curation, particularly when the government is the responsible party.

The "title" of a data set or of a file in a data set version also has a different context than the title of a text work. One way of appreciating this difference is to consider the number of titles a single author might publish. A particularly prolific author might have one hundred books, although usually authors do well to publish ten or fewer over their lives. In contrast, a single Earth science data set version may contain 100,.000 files. The only difference one might encounter in file names for a data set version is likely to be a reference to the date and time period during which the data were collected. Thus, we do not usually expect data center users to search for files based on "title" (or file name) alone, although users might use this attribute to select appropriate files to extract from the collection.

Searches by "subject" also differ between text collections and Earth science data. Probably the closest analogue between the two kinds of collections is identification of parameters contained in the files. Here, it often appears that controlled vocabularies are an important resource. In the Earth science data collections, the Global Change Master Directory is often cited as a key resource. At the same time, users of Earth science data may have an interest in features that are not present in the formal structure of the data within the files. For example, the data collections in NASA and NOAA do not necessarily include an identification of which files contain "hurricanes" or "dust storms", or, for that matter, which data might indicate that local meteorological conditions indicate "winter" or "summer" or "drought" to plants or animals. Identification of such features is usually done on an "after-the-fact" basis and would lead to an extension of the metadata compared with the metadata supplied by the data provider.

2.2.2Dealing With User "Tribes"

In many ways, a major difficulty in preparing metadata lies in the fact that the user communities exhibit "tribal" behavior in at least three ways:

- Distinct dialects
- Distinct data world views
- Distinct customs, particularly as defined by data search patterns and data use profiles

Distinct dialects develop in Earth science data using communities because of differences in the detailed technical knowledge required to work with data of a particular kind. Some of the dialect differences arise because of differences in measurement technology. A passive microwave instrument uses very different technology than does a radiation budget instrument. The microwave community uses frequency in GHz, whereas the radiation budget community expresses the equivalent of frequency in terms of a wavelength interval in say microns. Units are one key dialect feature that distinguishes one community from another. One tribe's "aerosol" is another tribe's "haze" and still another's "smoke".

It should be noted that dialects complicate digital librarianship. For one thing, they create aliases that disrupt the assumptions associated with controlled vocabularies Furnas and his colleagues [5] suggested using adaptive indexing, rather than controlled vocabularies because users have dialects with aliases for the terms in the vocabulary. In terms of searches, this means that users are highly likely to misidentify what they need. In response, controlled vocabulary is often presented in the context of alternatives that help clarify the meaning of the terms. At the same time, this combined presentation of the terms and their context appears to add additional information beyond a simple "key phrase" search.

Distinct data world views develop because the Earth science tribes use different data structures in their work. While one key community relies on images for their work of classifying the Earth's surface by vegetation type, another community thinks of the Earth's atmosphere as having columns, some of which contain clouds and others do not. The "horizontal" community thinks of the Earth's surface as broken into "images", in which each pixel has adjacent pixels in a row or column within the image. The stratospheric aerosol and gas community that infers vertical profiles of aerosol concentration by observing the sun set through the atmosphere thinks of their data as an array of vertically-stacked "pencils" through the stratosphere (with each pencil being about 200 km long and 1 km in diameter).

Distinctive customs, particularly those involving search patterns and data use profiles, arise naturally from the tribal data world views. For Earth science data, we can distinguish at least three search patterns that appear to lie outside the normal bounds of classic metadata:

- Spatial searches
- Temporal searches
- Coincident data searches
- Searches for features and phenomena

Spatial searches arise from a need to find data appropriate for investigation of conditions associated with phenomena affecting a particular location, region, atmospheric layer, or oceanic current. The usual visualization of such searches is for data that might provide clues to the weather affecting a particular city, county, or state. A variant of such "horizontal" spatial search might involve a search for data affecting a particular ecosystem, such as a forest or agricultural region. However, we could also consider searches for data relevant to understanding the stratosphere or troposphere are (at least approximately) "spatial" in nature.

Temporal searches arise from a need to find data associated with a particular time interval. For example, researchers might want

to compare global statistics for the same season in different years. They might be looking for the Earth's responses to particular events, such as solar flares or volcanic eruptions. Often, of course, users need to make searches that combine spatial criteria and temporal ones.

Coincident data searches involve looking for data from multiple sources that cover the same time interval and spatial region. These are particularly valuable for scientific data validation, where a data producer is seeking to find data that can help assess the correctness or uncertainty of a particular data product. An investigator who has created an algorithm to identify hurricanes on the basis of reflected sunlight and emitted thermal radiation may want to identify whether a radar observation of the same hurricane would verify the derived altitude of cloud top. While there is some similarity between the kind of search a humanities researcher might make for text passages in two histories to this kind of search, there are differences, particularly in the number of instances being sought and the lack of human meaning in digital data.

Searches for features and phenomena arise naturally from our need to interpret phenomena and identify cause and effect. Digital data do not speak to our human understanding without interpretation – and, in the large volume data collections, without algorithms that can identify natural features and phenomena. Our televised experience with weather forecasts creates a misleading expectation that a high-pressure air mass has a large, green letter "H" high in the atmosphere, where it could be seen by looking up. This interpretation of conditions requires human intervention. More subtle phenomena, such as ocean currents, drought-afflicted agricultural areas, sand storms, or diseased plant communities, require algorithmic interpretation. However, once the algorithms have operated and the features recorded as metadata not in the original catalog, searches for the identified features and phenomena become very useful.

2.2.3Extensibility of Metadata

In the world of Earth science data, the original data records and the original metadata tend to be relatively bland. The metadata by and large records the conditions of measurement and the provenance provided by the science teams who built the instruments and applied the algorithms to calibrate, geolocate, deduce indirectly measured quantities, and perform spatial and temporal averaging. However, there is an extensive amount of validation effort that the science teams contribute in determining the uncertainty of the data and the limits of usefulness. Typically, this work is iterative, leading to multiple versions of the data.

There are two significant extensions to metadata that appear to be needed beyond what the original data providers contribute:

- Feature and phenomena identification
- Critical commentary on the quality and usefulness of alternative data sets

Feature and phenomena identification is a key area for increasing human understanding of Earth science data. In some cases, we can visualize the required metadata for this purpose as a database table that identifies instances of a particular feature category or phenomenon. Because the original data providers often have uses for the data that did not include such feature identification, the new metadata is external to the original data. Thus, a data center has to obtain additional resources to provide searches based on this metadata.

Critical commentary is becoming important in the Earth sciences.

For example, the record of temperature and humidity profiles obtained by sampling these variables during a balloon ascent is critical for weather forecasting. Unfortunately, the instruments taking the data may have technical difficulties that cause missing data points. The instruments may also have biases. As a result, users of these data have developed editing procedures and have (often strong) opinions on which data are reliable. In creating a climate record, the data assimilation programs that produce weather forecasts ingest the reliable measurements and adjust their physics parameterizations based on the data. Metadata on data quality that the data assimilation programs can interpret is important to reduce the cost of the systematic reprocessing of the data that the atmospheric science community undertakes periodically.

In a similar fashion, data users for other kinds of Earth science data need the equivalent of critical commentaries in the humanities. In the latter disciplines, there is a well-developed critical literature that includes commentaries, glossaries, critiques, variant editions, and so on. The Earth science community is certainly familiar with peer-review as a method of vetting papers. There is a need to develop a similar kind of review for data itself.

We also note that there is a need to provide the equivalent to the searches users can make to published literature and then move to order data of interest as a result of such a search. In this case, the metadata (author, title, publisher, subject) that is maintained by the normal avenues of publication are stored separately. Extending them to cover data search and ordering can create difficulties owing to intellectual property rights. In other words, the normal bibliographic tools usually have strong proprietary intellectual property rights that may be too expensive for data centers to license. Thus, while it may be useful for users to search through a (digital or conventional) library for papers related to a particular problem and then to have that user connect to a data center from which he or she could obtain relevant data, the cost or proprietary nature of the bibliographic materials may preclude having the data center provide the service. We have also encountered similar difficulties in obtaining such metadata as professional society glossaries that could be of considerable help to our users.

2.2.4Summary of Metadata Issues

What we have seen in identifying digital librarianship issues for Earth science data is similar to the issues involving the collection organizations

- The conventional metadata for text libraries needs modification in order to serve the needs of the Earth science data community.

- Users need extensions to the metadata searches so that they can find data using both horizontal and vertical spatial searches, by temporal interval, for coincident data, and by feature or phenomena.
- Metadata for Earth sciences needs to extend its flexibility to adapt some of the features of humanities metadata, including cross-referencing and critical commentaries to guide users to recommended data sets for a particular use.

2.3Data and Context Threats

A key requirement for data centers is avoiding loss of data and of the context needed for its understanding by the "designated user community". We can identify several threats to this mission:

- Inability to capture key tacit knowledge required for context interpretation and data usability
- IT security threats to data integrity and provenance
- Operator error
- Hardware and software reliability
- Institutional instability

In approaching these threats, digital librarians need to develop a formal model for mitigating the risk from these threats.

2.3.1Tacit Knowledge Capture

Compared with texts written by humans, Earth science data is potentially much more difficult to interpret, particularly with the context provided by source code and documentation, read software that works, and ties to the literature needed to understand the concepts being used by the data producer. Additionally, the chain of processes that create Earth science data may be considerably more complex than the processes an author uses to crate a text. Thus, the context for understanding a digital data collection is longer and potentially subject to more loss mechanisms than the chain for text collections.

If we go back into the workflow that leads to an Earth science data collection, we can begin to understand the difficulty in capturing an appropriately detailed context for using or assessing the value of the data. For data obtained from Earth orbiting satellites, the chain of understanding starts with the instrument design and the calibration of the instrument done on the ground. In verifying the physical cause of artifacts in an instrument's data record, a data user may need as-built drawings from the manufacturer, as well as calibration plans and procedures (plans being an intended approach; procedures being step-by-step activities signed off during the calibration), and the data reduction that produces the calibration. Editing processes for the calibration data often escape capture. Because there may be many perturbations to a complex procedure, a researcher seeking to establish the validity of the ground calibration may need to verify that it was correct to throw out three of the five calibration runs. If this tacit knowledge is not captured and recorded by the data center holding the data, it may be impossible to understand the uncertainty in the data.

A second major threat to tacit knowledge capture lies in obtaining an understanding of the algorithms, source code, and algorithm parameters that are required to produce an Earth science data set. In some cases, even with an "Algorithm Theoretical Basis Document" (which has been required on a number of NASA Earth science missions in the last decade), it may be difficult to verify that the delivered algorithms correspond to the Theoretical Basis Document. Science teams may neglect to update the Document, or add additional editing code to deal with artifacts discovered after launch. Even if the ATBD and code are available, the size of the code may make it difficult to understand the full algorithmic behavior. In several cases, the authors have had to deal with science team algorithms containing multiple subsystems consisting of more than one hundred thousand lines of code per subsystem.

A third major threat to tacit knowledge capture lies in recording the complete history of production. Technically, where the science team is using discrete, batch production, each program execution has source code, input data, input algorithm coefficients, and may create several different output streams. The program is also potentially perturbable by the compiler used to convert source code to executable form, by the computer and its operating system, and by the related production environment. Any change to input, code, or configuration has the potential for perturbing the output from the process. In addition, the production batches are often chained together.

Tacit knowledge capture is quite difficult. In many cases, the process of preparing for capture has been less than optimal. This capture may also be expensive and organizationally hindered. Documentation is often regarded as a huge burden – and put off until the end of mission, where the process of capturing knowledge falls to the few remnants of the team who are stranded while highly productive scientists and project managers move on to new opportunities for creating more data.

2.3.2IT Security Threats

While any facility that connects its computers to the WWW has legitimate concerns over IT security, Earth science data centers have particular areas of concern that differ from commercial sites. The concerns for Earth science data center primarily on issues of data integrity and provenance, rather than direct financial threats.

In some cases, the data itself has commercial value. This is particularly the case with high-resolution imagery. The value of the data is likely to be dependent on how rapidly applications can access new observations. In this case, IT security threats due to denial-of-service are particularly important.

In the cases where the authors have direct experience, the data is valuable for research and for its potential impact on our understanding of the Earth's climate and its variability. For such data uses, the concerns shift to issues related to whether the data can be used to reach reliable conclusions regarding change. The economic interests become large, so the concerns shift from denial-of-service attacks by outsiders to carefully crafted, insider attacks that might alter the data or lead to concerns that the chain-of-custody that guarantees provenance had been disrupted.

In our experience, we estimate that the probability of a serious IT incident is in the vicinity of 10% per year or higher. It is difficult to obtain documentation about this probability, since most

institutions are very reluctant to release information about incidents.

We grant that a potential perpetrator would need a very large amount of knowledge about both the system and the data in order to create an incident. On the other hand, the economic and political stakes are perhaps sufficiently large that large firms and national governments might have an interest in such an attack. While we suggest later a systematic approach for dealing with the risk of data loss, two defensive strategies appear worth mentioning here. The first is ensuring that all key data and documentation is replicated in off-line storage. The second is ensuring multi-institutional data replication. Both of these mechanisms operate by geographically dispersing the risk and reducing the probability of success.

2.3.3Operator Error

A second area of concern for data centers lies in losses that may be incurred through operator error. Technically, such losses appear to be of particular concern when data are ingested into a data center and when the data need to be distributed. We have also had difficulties ensuring that metadata are synchronized with the contents of data files.

In our experience, operator error may be similar in its probability to the probability of successful IT incidents. Again, it is difficult to obtain documentation about the actual probability of data loss due to operator error.

The Open Archive Information System (OAIS) Reference Model [4] incorporates functionality for Ingest QA as part of this standard. From the standpoint of data integrity, such Quality Assurance is critical to reduce the probability of errors, such as nearly empty files being placed into the data stored in a data center.

The same care is needed to deal with errors on data distribution, particularly when data are being sent out on media. In this case, operators may send data to incorrect addresses, as well as mislabel the contents of the shipment.

Finally, it is important to reduce the probability of error during periodic migration of data from media type to media type as hardware changes. This process probably needs checks to ensure that no errors crept into copies when these are made. This work is non-trivial, since we expect the entire digital holdings of a data center will need migration about once every five years.

2.3.4Hardware and Software Reliability

Humans are not the only agent by which errors and data loss can be injected into the contents of an Earth science data center. Hardware can inject errors, as can software.

One particularly difficult experience involved errors in a router when it transmitted data. The data were seriously corrupted, but it took a substantial length of time to identify the cause and correct it. Such incidents are difficult to detect and may be quite subtle.

It is clear that software reliability is difficult to achieve. The direct evidence lies in the frequency of patches distributed by major software product vendors.

As with operator error, this problem may be found somewhat earlier if systematic procedures are developed to check for data integrity are implanted in routine operations.

2.3.5Institutional Instability

It is easy to identify hardware as a culprit in data loss ("the tape reader ate my tape"). It is harder to admit that a hacker put a rootkit in the system. It is probably hardest to get institutions of human beings to agree to cooperate. All of the normal forces that make us defend our belongings operate on data center budgets and the institutional components that need to cooperate together to produce long-term reliability. In this realm, the key appears to lie in building self-governance mechanisms that reward cooperative behavior and reduce the cost of lack of trust (not to mention the comment that "the enemies in the office across the hall stole my budget").

2.3.6Summary of Threat Mitigation

We have seen that while many data losses have mechanical causes, human behavior is probably a larger causative factor. It also appears likely that it will be impossible to completely guarantee against data loss. We suggest a more useful and systematic approach below. Even here, we suggest the following

- Careful attention needs to be paid to the workflow required to reconstruct particular data products. In many cases, this analysis will suggest key pieces of documentation and ancillary data that need to be available to reconstruct the original data if it were lost.
- IT security can be enhanced by off-line storage, although this creates a requirement for periodic migration of data as technology changes. Fortunately, this approach should decrease the cost of storing a particular data product, since the cost per bit of storage is decreasing with time. On the other hand, procedures for checking that data has been copied successfully from old media to new must be in place to ensure satisfactory retention of data.
- Operator error needs careful attention. Automation and good design of operator interfaces can help reduce the probability of error. On the other hand, a computer can make errors a lot faster than a human can – and on a lot more data.
- Hardware and software are as suspect as operators. Careful procedure design may help catch errors before they become too serious. However, wisdom suggests that plans will probably be rewritten as error diagnosis and fixing proceed.
- Humans are difficult. Humans in groups are even more difficult. Managerial care in creating self-reinforcing cooperative behavior may mitigate some forms of institutional instability. Luck is also helpful!

2.4Data Center Cost Effectiveness

If avoiding data loss is the first commandment of data center policy, making sure that a data center can survive on its meager budget is the second. One of the real difficulties lies in the fact that most cost discussions focus on the tangibles of storage devices and media, while the operations budget that is much less tangible is usually the dominant factor in costs. In other words, budget reduction discussions tend to focus on reducing the hardware and software costs. Budget discussions that may lead to reductions in staff costs are much more difficult to carry on. Everyone is aware of the human costs of budget reductions that cut staff.

Perhaps the most appropriate motivation for such discussions is to provide an opportunity to change the skill mix of the work being done so that data center staff have an opportunity to improve their productivity on the important work of making the data more accessible and usable. This suggests that we need to design appropriate automation into the system evolution. We need to reduce repetitive, error-prone work with work where the staff can contribute creatively to the evolution of the data center.

As part of this approach, it is important to invest staff time into reducing the long delay between the point where a new idea appears in the IT realm and the point where it can be embedded in an operational system. At present, some in the IT community call this gap "the valley of death." One data center manager has suggested that it is about six years from the initial appearance of a change until it can be incorporated into that center's operations. Given that the normal academic grant period is three years, there is clearly a gap that should be shortened.

One approach to this problem may be to consider arrangements in which grant recipients are required to identify a data center and a user community that would benefit from the services. Grant renewal would then be contingent upon demonstrating that the data center can incorporate the idea into its operational environment at an acceptable cost.

2.5User Services and Interfaces as Reference Librarians

The final set of issues that concern us here is representing the role of Earth science data center user services personnel. Often the current approach seems to identify these key personnel as telephone answering servants at a data center. However, that does not appear to do justice to the role that user services personnel fulfill. In cases with which we are familiar, our user services personnel fulfill the following roles:

- Assisting users who have trouble understanding the user interfaces for data discovery, search, and order
- Actually ordering data and arranging for suitable reformatting for users who have difficulty dealing with the technology (which is roughly equivalent to the mechanism of asking one's children for help programming the video recorder)
- Translating from the user dialect into the data center dialect for finding data – including highly specialized requests for data with qualities or properties that are not well represented in the metadata
- Assisting users with read software
- Assisting users by providing visualizations
- Summarizing user experience to improve the data center's understanding of how the user community's needs are evolving

These roles are essentially those we associate with reference librarians in specialized collections. As with reference librarians, data center user services personnel perform an extension to automated search tools in a niche market.

It seems sensible to expect that user interfaces will continue to evolve as we explore various technologies. We note that the combination of semantic web and semantic grid technologies appear to offer significant advances in helping users. In particular, the combination of RDF, ontologies, and triple stores may be of significant assistance in improving user search success. These technologies may also allow us to automate some work, so that intelligent agents can traverse the automated search mechanisms to answer important, previously unanswerable queries.

One metaphor for this work would be avatar based interfaces, in which talking heads (with animated expressions and gestural languages) provide improved communication possibilities. In this case, the scenes from the recent movie "The Time Machine" provide an impressive expression of an intelligent interface, capable of holding a meaningful and helpful dialog with a user needing search assistance for touring the British Museum.

At the same time, one of the author's recalls a notable experience that encapsulates the most rewarding kind of help experience. In this case, the author was attempting to enter and use a system for mathematical equation manipulation, MACYSMA. At the time, the advanced technology consisted of 80-column terminals that were only capable of handling characters – no graphics at all. After about fifteen minutes of increasingly frustrating experience, the author's screen held the line:

"If you're confused, call Ellen at …"

The author did and was immediately set on the path to being able to use the system successfully.

The vision that binds the highly automated with the human being is the critical one for increasing the usability of our Earth science data centers.

3.MECHANISMS FOR DEALING WITH ISSUES

In this section, we suggest some mechanisms for dealing with the issues we have raised. First, we suggest using a maturity model for assessing the readiness of Earth science data to be placed in a digital library for curation, as well as the capability of a data center to carry out a program of scientific data stewardship and to provide future societal benefit from the data. Second, we suggest an approach for dealing with risks that threaten loss of value from data under curation. Third, we suggest developing publicly accessible and validated models for the cost of operating

continually evolving data centers that can assist in improving the cost effectiveness of data centers.

3.1Using a Maturity Model for Judging Curation Capability

A recent report by the National Research Council has strongly recommended developing metrics for judging the performance of data management efforts for climate data. Our response has been to produce an initial form for a maturity model. This model appears to have excellent potential for combining multiple metrics into a framework for assessing and prioritizing climate data records. The Maturity Model is similar to the Capability Maturity Model, Integrated, (CMMI) used to assess the progress of software development teams in moving from an uncontrolled process to a highly repeatable process.

For Climate Data Records, the Maturity Model has three basic axes:

- *Scientific Maturity*, related to uncertainty and capture of community understanding
- *Preservation Maturity*, related to managing the risk of data loss and of cost effectiveness
- *Societal Benefit/Impact*, related to the value of the data and context for long-term use and data access

The Model suggests attributes for each axis and rankings for each attribute.

For scientific maturity, the key attribute categories are

- Physical understanding of the measurement process
- Measurement of key instrument characteristics
- Public accessibility of data processing
- Rigorous validation

Within each attribute category, there is a further subdivision into individual attributes. Each attribute is ranked by representatives of the data provider (and other community representative) into a numerical scale ranging from 0 (completely immature) to 5 (definitive).

For preservation maturity, we follow the same approach, using the categories

- Systematic approach to guaranteeing preservation of data understanding
- Systematic reduction of threats to preservation
- Assurance of preservation cost effectiveness

In these categories, we expect representatives from the data center being proposed for curation to assess these attributes and ensure a satisfactory level of maturity.

For societal benefit/impact, we seek to categorize the attributes according to

- Bibliometric metrics, such as publications and citations arising from data use
- Increase in scientific community knowledge
- Economic and policy utility

Here, we seek strongest input on the maturity from representatives of the user community. Of course this last axis is among the most difficult to rank, in accord with the known difficulties of cost/benefit analysis [6].

By using the rankings, the authors hope the community will develop consensus valuations of the maturity and potential value of a proposed Climate Data Records, and thereby assist in decisions regarding archival. We do recognize that using the maturity model will be exploratory and iterative. There is no expectation that we will get exactly the right metrics or weightings into the initial version of the model. Nonetheless, we believe this approach can be useful.

It is also clear that we need to incorporate the full diversity of the community. Even within the Earth science community, we know we must accommodate a wide diversity in vocabularies, data structures, and data uses. Furthermore, we need to incorporate deep uncertainty – particularly for the societal impact/benefit axis. There is a diversity of opinion on areas of scientific controversy. Discussions in this area need a common framework and discipline. We believe that the maturity model can provide such a framework.

3.2Risk Management

The second area that we feel needs systematic development is using a systematic approach to understanding and managing the risk posed by threats to data and the context for intelligent data use. Fortunately, there is a good framework for this work using the now conventional approach represented by risk management. As outlined by Boehm [3] and many other authors [7], such a systematic approach uses the following three steps:

1. Evaluate threat mechanisms
2. Quantify (or otherwise assess) the probability of each important threat and the probable loss if the threat appears
3. Develop strategies to avoid the threat, reduce the probable loss, or insure against the loss

While some adaptation of this approach may be necessary to reduce the risk of catastrophic loss of data and context, this appears to be a good, useful strategic approach for dealing with most of the issues we have considered in this paper.

3.3Cost Modeling for Continual Evolution

Our third, and final area of recommendation is for data centers and digital librarians to develop a model for the cost of operations and for the cost of continually evolving a data center [2]. Based on our experience, we recognize that it is easy to adopt a short-term approach to managing this year's budget and developing the one after that. However, a more strategic vision is called for.

There are two key points to the cost modeling. The first is the need to identify the key workflows associated with the operations of a data center. We note that the OAIS Reference Model provides a very useful functional breakdown of the work in a conventional data center. For Earth science data centers, we need to add in additional functionality to account for data production within the center. We then recommend identifying the cost of

operations for these functions and using those as a basis for strategic decision making on the routine operations of the center.

The second point is to ensure that the description of the routine operations is placed with the context of a strategic plan for evolving the data center over a period of at least thirty years. This time period is long enough to include the fact that data will need to be migrated from medium to medium several times (six for a thirty-year plan) – with the expectation that each migration will systematically reduce the cost of storage. It is also important to ensure that the budget includes a reasonably constant category devoted to hardware and software migration of the data center's infrastructure.

4.SUMMARY AND CONCLUSIONS

In this paper, we have tried to bring to light several important issues for digital librarians who encounter Earth science data. In some cases, resolving these issues should be straightforward. In others, there are difficult issues – often sociological and political that may complicate resolution. We anticipate further discussions of these issues in a variety of forums.

At the same time, we believe that the suggestions we have made for a maturity model, for a systematic approach to risk management, and for cost modeling that incorporates continual evolution can be generalized to other areas of digital librarianship. We also hope that the issues we raise will find a place in professional curricula for digital librarians.

5.ACKNOWLEDGMENTS

Bruce R. Barkstrom is grateful for the support of NASA Langley Research Center's Atmospheric Science directorate and for the support provided by NASA Headquarters through the ESDIS Project at Goddard Space Flight Center.

6.REFERENCES

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

[2] Barkstrom, B. R., Finch, M., Ferebee, M., and Mackey, C., Digital libraries and change: adapting digital libraries to continual evolution, *Proceedings of the 2nd ACM/IEEE joint conference on Digital Libraries*, (*JCDL'02)*(Portland, OR, July 2002).

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

[4] CCSDS 650.0-B-1. *Reference Model for an Open Archival Information System (OAIS). Blue Book Issue 1*, [ISO 14721:2003], 2002.

[5] Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T., The Vocabulary Problem in Human-System Communications: an Analysis and a Solution, *Communications of the ACM*, 30 (11), 964-971, 1987.

[6] Layard, R. and Glaister, S. *Cost-Benefit Analysis*, 2nd ed., Cambridge, UK, 1994.

[7] Toigo, J. W., *Disaster Recovery Planning: Preparing for the Unthinkable*, 3rd ed., Prentice-Hall PTR, Upper Saddle River, NJ, 2003.