EDUCAUSE | CENTER FOR APPLIED RESEARCH

The Chronopolis Project: A Grid-Based Archival Digital Preservation Solution

Judith A. Pirani, ECAR
Donald Z. Spicer, ECAR

ECAR Case Study 1, 2010

4772 Walnut Street, Suite 206 • Boulder, Colorado 80301 • educause.edu/ecar

EDUCAUSE

4772 Walnut Street, Suite 206
Boulder, Colorado 80301
educause.edu/ecar

The Chronopolis Project: A Grid-Based Archival Digital Preservation Solution

EDUCAUSE
CENTER FOR APPLIED RESEARCH

EDUCAUSE is a nonprofit association whose mission is to advance higher education by promoting the intelligent use of information technology.

The mission of the EDUCAUSE Center for Applied Research is to foster better decision making by conducting and disseminating research and analysis about the role and implications of information technology in higher education. ECAR will systematically address many of the challenges brought more sharply into focus by information technologies.

The Chronopolis Project: A Grid-Based Archival Digital Preservation Solution

Preface

The EDUCAUSE Center for Applied Research (ECAR) produces research to promote effective decisions regarding the selection, development, deployment, management, socialization, and use of information technologies in higher education. ECAR research includes

- research bulletins—short summary analyses of key information technology (IT) issues;
- research studies—in-depth applied research on complex and consequential technologies and practices;
- case studies—institution-specific reports designed to exemplify important themes, trends, and experiences in the management of IT investments and activities;
- roadmaps—designed to help senior executives quickly grasp the core of important technology issues; and
- key findings—brief high-level summaries on the scope of an ECAR research study.

As part of its 2009 research agenda, ECAR recently published a study, *Institutional Data Management in Higher Education*,[1] by Ronald Yanosky. The study examines how higher education institutions are facing the challenges of institutional data management in terms of data quality, stewardship and governance, analytics, content and records management, and research data management.

Literature Review

The literature review helped identify and clarify issues, suggest hypotheses for testing, and provide supportive secondary evidence. Besides examining articles and studies from journalistic, academic, and IT practitioner sources, we consulted with practicing CIOs and data management experts to develop study objectives and survey questions.

Online Survey

We designed and administered a web-based survey that was distributed to institutional representatives (mostly senior IT leaders) at 1,733 EDUCAUSE member institutions in February 2009. We received 309 responses (a 17.8% response rate).

Interviews

We conducted follow-up telephone interviews with 23 senior IT leaders and staff from a mix of institutions to gain deeper insights into findings from the quantitative analysis and to capture additional ideas and viewpoints.

Case Study

ECAR researchers conducted this in-depth case study to complement the core study. We assume readers of this case study will also read the primary study, which

provides a general context for the individual case study findings. We undertook this case study to explore the challenges and possibilities of long-term digital preservation by chronicling the development of Chronopolis, a grid-based archival digital preservation solution. ECAR owes a debt of gratitude for their time and insights to Martha Anderson, Director of Program Management for the National Digital Information Infrastructure and Preservation Program, Library of Congress; Bryan Beecher, Director, Computing & Network Services, Inter-University Consortium for Political and Social Research; Francine Berman, Vice President for Research, Rensselaer Polytechnic Institute; Luc Declerck, Associate University Librarian–Technology Services, University of California, San Diego (UCSD) Libraries; Martin Halbert, Dean of Libraries, University of North Texas, and President, MetaArchive Cooperative; Joseph JàJà, Professor, Department of Electrical Computing and Engineering and University of Maryland Institute for Advanced Computer Studies, University of Maryland, College Park; Ardys Kozbial, Technology Outreach Librarian, UCSD Libraries; David Minor, Head of Curation Services, Data Life Cycle Services Division, San Diego Supercomputer Center; Brian E. C. Schottlaender, The Audrey Geisel University Librarian, UCSD Libraries; Katherine Skinner, Executive Director of the Educopia Institute and Program Manager for the MetaArchive Cooperative; Michael Smorul, Faculty Research Assistant, University of Maryland Institute for Advanced Computer Studies, University of Maryland, College Park; Karen Stocks, Assistant Research Scientist, Center for Advanced Computational Science Engineering, San Diego Supercomputer Center and Scripps Institution of Oceanography; Don Sutton, Chronopolis Programmer, San Diego Supercomputer Center; Dallas Thornton, Division Director, Cyberinfrastructure Services, San Diego Supercomputer Center; and James Tuttle, Digital Repository Librarian, North Carolina State University Libraries.

Introduction

As long as there have been computers, there have been digital objects that require storage and retrieval. Over time the focus has broadened to include text, audio, images, movies, software, models, simulation, and other elements in addition to numbers. In recent years these objects have been "born digital" rather than being digitized copies of analog information. The resultant growth of digital information has become almost unimaginable. A 2008 International Data Corporation white paper[2] offers a hint about the size of this "data deluge," estimating that in 2007 there were 281 exabytes (2.81×10^{21} bits) of digital data. The year 2007 was also the first in which the amount of digital data created, captured, or replicated exceeded the amount of existing storage media to store it in any currently used form (e.g., hard drives, tapes, DVDs, CDs, volatile and nonvolatile memory).

The increasingly digital nature of research contributes significantly to this data deluge. Computational tools, the cyberinfrastructure, and inexpensive data storage foster a digitally inclined research environment that generates massive amounts of data. Consequently, terabytes and petabytes are replacing familiar kilobytes, megabytes, and gigabytes in research, and will eventually grow in scale to exabytes, zettabytes, and yottabytes. As the amount of digital research data grows, the question becomes, who will sift through this data, select what is valuable, and make it accessible in the future?[3]

Preserving what we learn from the past is just as important as making research data accessible in the future. Too often research data may be downloaded, deleted, or just misplaced upon completion of a research

project, with potential great societal loss. The near disposal of Apollo 11 moon landing tapes in 2006 illustrates this point all too clearly. Nearly 100 tapes of data collected during the first moon landing from a dust detector designed by an Australian physicist ended up in a dusty basement of a physics lecture hall, despite the clear markings "NASA Manned Space Center." The tapes provide a daily record of the environmental conditions and changes taking place at the lunar site after the *Eagle* landed safely in the Sea of Tranquility. Students found the tapes while rummaging in the basement and sent a tape to NASA for evaluation. Upon review, NASA deemed the Apollo tapes the only long-term information on the lunar surface environment and as such ideal for planning future lunar missions.[4] Additionally, saving past data facilitates future research, enabling scientists to build directly upon preserved data or to integrate data collectively from a variety of sources.

Both scenarios—the data deluge and the need to preserve past research data—point to the importance of long-term digital data preservation. Selecting the data to preserve is the first step, as not every bit of digital research data needs to be preserved indefinitely; repercussions vary, too, if data is lost, depending upon its intrinsic value. Preserving NASA moon mission data has very different connotations and repercussions than saving a faculty member's scratch notes.

The actual digital preservation process is not as simple as it seems, however. "Unlike storing boxes of paper or photos in an environmentally controlled warehouse, digital data always needs to be cared for," states Ardys Kozbial, technology outreach librarian, UCSD Libraries. One can't just dump data onto a tape, hard drive, or computer and automatically expect it to survive over an extended time period. Fortunately, the NASA moon tapes emerged undamaged after languishing in a basement, and a tape drive still existed that could read the 40-year-old data.

"Long-term preservation is a different mind-set," explained Joseph JàJà, professor, Department of Electrical Computing and Engineering and University of Maryland Institute for Advanced Computer Studies, University of Maryland, College Park. "What is the guarantee that your data will be intact over the long term? What happens if the file format changes? Preservation is not for a month or two. It should be geared for 10, 15, 20 years or longer. You need special tools to focus on data integrity, multiple copies, and monitoring or you do not have preservation."

These requirements dictate a specially designed, long-term repository—known as a *deep archive*—to preserve the data for years and even decades. Data access occurs only in an emergency but with the expectation that it will exist exactly as originally ingested. This requires a technologically adaptable environment to ingest, store, and replicate accumulating stores of data and a tool set designed to tend, monitor, and access data. Archival issues come to the forefront because one must systematically select appropriate research data for long-term digital preservation, and one has to create suitable metadata for categorizing, tracking, and retrieval purposes.

Digital preservation has nontechnical requirements, too. The first of these is funding. The need for long-term data preservation begins when a research project—and its funding—ends. Preservation costs should be addressed during the proposal stage, yet there is no guarantee that the research outcomes will warrant such preservation measures. Calculating preservation costs under such circumstances is challenging at best. Ongoing projects with great societal value, such as the Human Genome Project, require digital preservation for the foreseeable future. Policy issues about data access and privacy need to be resolved, too.

The need is apparent, but long-term preservation is still emergent, and much work lies ahead. In ECAR's study *Institutional Data*

Management in Higher Education, Yanosky describes digital preservation's tentative state in higher education, stating that "although we knew that research data preservation is largely uncharted ground at the institutional level, we were still surprised at how circumscribed and uncertain a picture our survey questions uncovered."[5] ECAR research found that respondents were not confident about their institutions' ability to support the long-term preservation of research data. On a scale of 1 (strongly disagree) to 5 (strongly agree), respondents averaged only a 2.54 response to the statement that their institution had the necessary commitment to support the long-term preservation of research data, and a dismal 2.15 response to the statement that they had the necessary funding mechanisms.[6]

Auspiciously, digital preservation attracts national-level attention, with programs in place to develop viable solutions. In 2000, the U.S. Congress mandated the Library of Congress to develop a national strategy to preserve valuable digital content. The strategy developed into a program known as the National Digital Information Infrastructure and Preservation Program (NDIIPP). In 2005, the National Science Foundation's study *Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century*[7] defined issues surrounding long-term digital preservation. This work eventually led to the creation of the Sustainable Digital Data Preservation and Access Network Partners (DataNet) program in 2008, which aims to develop a set of national and global data research organizations to provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline. Over the past several years, both NDIIPP and DataNet have incubated more than a dozen digital preservation solutions, some of which now demonstrate promise for production applications.

This case study examines the issue of long-term preservation of research digital data by highlighting the development of Chronopolis, an NDIIPP-sponsored demonstration solution with production service potential. While Chronopolis is not a DataNet partner, it is characteristic of DataNet projects. It is a centrally managed, grid-based, federated service offering deep archiving preservation for participating data owners. Partners include the UCSD Libraries, the San Diego Supercomputer Center (SDSC), the National Center for Atmospheric Research (NCAR), and the University of Maryland Institute for Advanced Computer Studies (UMIACS). SDSC, NCAR, and UMIACS form a geographically dispersed grid, each being a storage node holding fully replicated collections of user data. Additionally, UMIACS is actively engaged in preservation tool development. The UCSD Libraries contribute archival and metadata expertise. Many types of data could benefit from such long-term data preservation, but Chronopolis currently hosts only research data, particularly large research data sets.

As noted, Chronopolis is still maturing as a production service. Having successfully demonstrated the technical infrastructure, the partners are now developing and fleshing out the appropriate governance, policy, and funding frameworks needed to ensure a sustainable service for the long term. By chronicling Chronopolis's development, ECAR's goal is to illustrate the complexities and possibilities surrounding long-term digital preservation and suggest to readers the form emerging data preservation tools might take. This case study presents Chronopolis from multiple perspectives: its digital preservation framework, its project sponsor, its partner organizations, its user community, and its preparations for self-sustainability.

Digital Preservation Framework and Requirements

Chronopolis provides deep archival, long-term preservation, but not all data requires such rigorous protection. Brian E. C. Schottlaender, the Audrey Geisel University Librarian, UCSD Libraries, put this issue in perspective when he observed, "Losing my digital photos would be an inconvenience, but science would stop if the Protein Data Bank is lost."

The Data Pyramid presented in Figure 1 illustrates this point explicitly. Developed by Francine Berman, vice president for research, Rensselaer Polytechnic Institute, and former SDSC director, it illustrates Chronopolis's perspective on the digital preservation environment. The pyramid is based upon the Branscomb Pyramid for Computing,[8] which illustrates the differences between the less powerful and more powerful computing platforms, and how the user community shrinks as computing platforms become more powerful. The Data Pyramid applies the same principle to data.

The Data Pyramid contains three tiers:

- The Individual Value Tier, at the bottom, consists of data that is typically stored in private repositories and may have limited long-term intrinsic value. Personal digital photos or faculty scratch notes fall in this tier.
- The Community Value Tier, in the middle, includes data of potentially longer-term value and/or of value to a larger constituency—for example, experimental data from research groups.
- The Societal Value Tier, at the top, contains data of great societal value that would be infeasible to replace. A good example is the Protein Data Bank, which evolved from a digital collection of fewer than a dozen files upon its inception in 1971 to a primary, global resource containing 61,808 structures.[9]

Data can actually move up and down the Data Pyramid. For example, digitized childhood photos of a future U.S. president start at the bottom of the pyramid but eventually progress to the pyramid's top when that person is elected president. Factors such as government regulations or business practices may influence data's place on the pyramid, too.

As data moves up the pyramid, it requires more stringent curation and preservation efforts as a result of its increased societal value, the population of people impacted by the data,

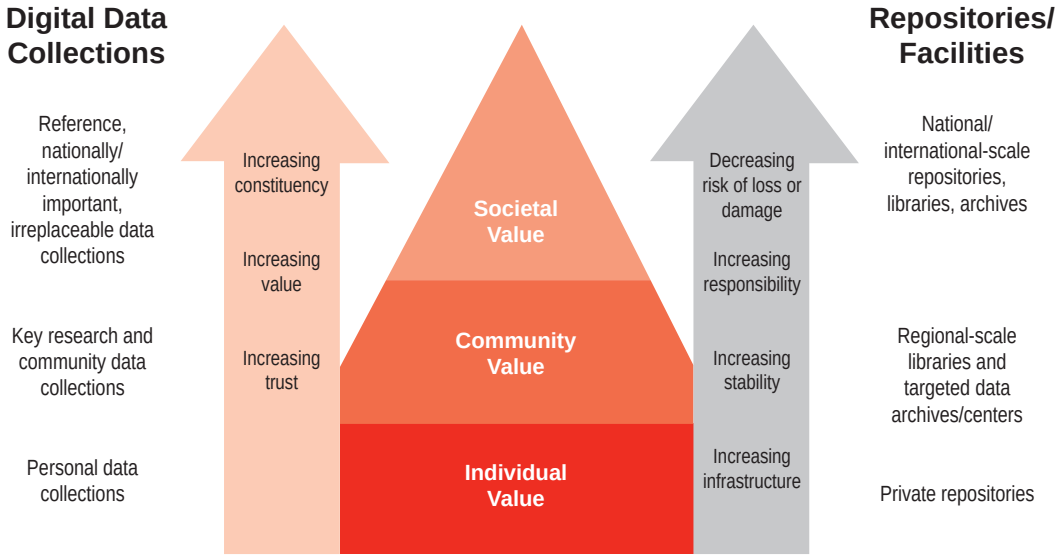

Figure 1. The Data Pyramid

Source: Reprinted with permission from the Association for Computing Machinery

and the indeterminate need for preservation. "We have to stop thinking about data as an undifferentiated thing, because it is not," stated Schottlaender. "If you use the library analogue, we invest considerable money in the form of security, environmental controls, and expert staff management in our special collections. When we get mold blooms, we get them in the regular collection, not in the special collection, because we use more resources."

Preservation of Societal Value digital data needs similar rigor. Table 1 lists appropriate digital curation and preservation services elements.

The Library of Congress's NDIIPP and the Role of Chronopolis

Recognizing the need for better preservation of valuable digital data, the U.S. Congress in 2000 mandated the Library of Congress to establish a program, the National Digital Information Infrastructure and Preservation Program. NDIIPP's mission is "to develop a national strategy to collect, preserve, and make available significant digital content, especially information that is created in digital form only, for current and future generations."[11]

The legislation called for the Library of Congress to work with other federal agencies and other stakeholders to develop a national approach to digital preservation. "The legislation's aim is to approach digital preservation broadly, indicating that the Library was to work with a wide range of public and private stewardship organizations," stated Martha Anderson, director of program management for the National Digital Information Infrastructure and Preservation Program, Library of Congress. The legislation allocated $100 million for this initiative.

Between December 2000 and mid-2003, the Library of Congress began to identify and convene stakeholders, ranging from large institutional libraries to content producers, to develop a digital preservation plan. Three areas of concern emerged:

- *Content:* Explorations around what content is at risk, who is creating it, who is taking care of it, and the preservation needs for different types of content.
- *Technical architecture or infrastructure:* Creating a digital preservation technical framework. The Library made the strategic decision to leverage preexisting resources rather than invest in the construction of a distinct infrastructure. "That is how we became interested in supercomputing," stated Anderson. "The government has invested a lot in that area and there is a lot of capacity. We were trying to understand how it might serve a digital preservation purpose for cultural heritage as well as science."

Table 1. Digital Curation and Preservation Elements[10]

Element	Activities
Data appraisal	Evaluation and selection of digital material for long-term preservation
Data ingestion	Controlled, and secure if necessary, transfer of material into the preservation archive
Data arrangement	The process of structuring the metadata into a collection hierarchy and aggregating digital entities into containers for storage management
Storage resource management	The preservation mechanisms needed to control and track the integrity of multiple archived collections
Preservation	Managing technological evolution and maintaining integrity by migrating to new media, new encoding formats, etc., as new technologies become available
Audit control	Ability to audit the contents of the various objects according to the policy set by the archive and to provide mechanisms for an independent third-party auditor to certify the integrity of any object
Data owner services	Monitoring, access, and recovery capabilities for data owners

- *Business models to support preservation:* Many organizations that feel a responsibility to preserve digital data will not have the capacity to start or to sustain their activities. Innovative business models will ensure that they can at least begin their activities.

Eventually the Library formed a network of preservation partners. "There is so much that needs to be preserved," stated Luc Declerck, associate university librarian–technology services, UCSD Libraries. "The Library of Congress understands this is a national issue and having only one solution is too risky." The approaches included an early partnership with the National Science Foundation (NSF) to undertake research into long-term management of digital information. Subsequently, the Library formed other partnerships. There are strategic partnerships with the Internet Archive and other archival organizations; organizational alliances with federations, consortia, and coalitions; and alliances with standards bodies. Finally, the Library, with the NSF and other institutions, now sponsors the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (http://brtf.sdsc.edu/), which is charged with analyzing previous models for sustainable digital preservation, developing a set of economically viable recommendations for digital information preservation, and providing a research agenda in the area of economic sustainability of digital information.

Today, NDIIPP has more than 130 partners engaged in sharing knowledge and experience regarding digital preservation. These include libraries, archives, universities, research centers, nonprofit and for-profit organizations as well as professional associations in the United States and around the world. The general foci of these partnerships include Digital Archiving and Long-Term Preservation, Technical Architecture Related to Preservation, Digital Preservation Policy, Preserving Creative America, and Preserving State Government Information.

In addition, the Library sponsored two rounds of funding, selecting eight new initiatives in each round that are aimed at preserving digital materials in the five focal areas. "We seed communities that can work together to gain some economies of scale and leverage their work," explained Anderson. Chronopolis is a Technical Architecture project funded during the second round of NDIIPP funding.

The Library works with each funded project to build its specific initiative and encourages collaboration among them. For example, the Library facilitates an annual general meeting of the projects' principal investigators so that they can learn about each other's activities and exchange ideas. "It gives you a sense of the importance of the communal effort that we are making, and these meetings breed into us a certain kind of community," stated Katherine Skinner, executive director of the Educopia Institute and program manager for the MetaArchive Cooperative, another NDIIPP-funded project.

The Library incubates these projects, with the goal of supporting their eventual transition from demonstration project to production service. As described in the Next Step section below, Chronopolis is now at that junction of its evolution, embarking on its migration to self-sustainability.

Origins of the Chronopolis Partnership

This section introduces the partners behind Chronopolis and presents a brief history of the development of the service.

Four partners currently constitute Chronopolis:

- *San Diego Supercomputer Center (SDSC):* An organized research unit of UCSD, SDSC is a national leader in creating and providing cyberinfrastructure for data-intensive research. SDSC's primary roles in Chronopolis are to ingest data into the archives, house a complete copy of

all data, provide storage and networking services, offer storage resource support, and provide management and financial oversight for the project.

- *University of California, San Diego Libraries (UCSD Libraries):* Nine libraries make up the UCSD Libraries, which support undergraduate and graduate instructional programs as well as advanced research. For Chronopolis, the UCSD Libraries provide metadata services and other advanced data services.
- *The National Center for Atmospheric Research (NCAR):* Located in Boulder, Colorado, NCAR provides high-performance computational and observational facilities and tools for atmospheric, climate, and weather-related research. For Chronopolis, NCAR provides archives, a complete copy of all data, storage and network support, network testing, and development of a user-centric data portal.
- *The University of Maryland Institute for Advanced Computer Studies (UMIACS):* UMIACS enhances interdisciplinary research and education in computing across the University of Maryland's College Park campus, conducting research programs on a broad range of areas to address both core computer science issues and fundamental problems at the interface between computer science and other disciplines. In its role as a Chronopolis partner, UMIACS provides archives, complete copies of all data, storage and network support, tool development, advanced data services, and network testing.

Several factors made Chronopolis a natural project for these partners. First is the nature of UCSD itself. UCSD was founded in 1960 during a period of expansion of the University of California System. As a new institution in a system with many mature, leading institutions, UCSD forged a new direction that complemented the older institutions. UCSD's research-intensive environment, which includes several research institutes and much research activity, generates huge amounts of data. Additionally, the institution focused its development on the high-tech industries concentrated in the San Diego and Southern California area.

Second is the UCSD Libraries' expertise in digital collection curation. Like the university as a whole, the UCSD Libraries faced decisions regarding the character of its library collections upon its inception in the mid-1960s. Since the University of California System contained several institutions with large collections of traditional materials, the decision at the UCSD Libraries was not to duplicate those collections but rather to develop a unique collection of new materials. "Rather than try to emulate or duplicate our sister libraries, we decided to put our efforts into infrastructure development," stated Schottlaender. "The print collection would be focused on delivery, and we pushed aggressively into digital library development, particularly in the last decade." Thus, the UCSD Libraries has curated in digital collections ever since such collections have become available, staffed by librarians with traditional archival training and a history of preserving collections, no matter what the medium.

The third factor contributing to Chronopolis's initial success is former SDSC Director Berman's experience in cyberinfrastructure activities, including large-scale data storage, data management, and grid computing, as well as her growing personal interest in digital preservation. A dialogue between Berman and Schottlaender commenced about digital preservation. Eventually, SDSC and the UCSD Libraries began working together on projects, including one with the Library of Congress to develop distributed preservation solutions for Societal Value data collections. The project involved multiple replications at geographically diverse locations. Subsequently, the two organizations' ideas grew grander. "Through

prolonged interaction and sensitivity to technical issues behind storage and preservation, we began thinking about the right kind of national-scale infrastructure for data preservation," recalled Berman. "You put together the need for a national data infrastructure, the importance of replication, coordination, and reliability of data—a kind of grid mind-set of technologies. Chronopolis was a mature synergy of all those ideas. Together we started creating that framework."

An additional factor is the synergy between computer and library sciences and their respective organizations. The UCSD Libraries' digital orientation, along with its proximity to SDSC, offered an unusual opportunity to combine the skills of both areas. SDSC, with its NCAR and UMIACS partners, brings insights into technology associated with preservation of digital materials, providing the infrastructure, service design, tools, and ongoing service maintenance. Yet Chronopolis's premise is built on something more—the Libraries' expertise. As a library, the UCSD Libraries is a trust agent, contributing its customer service orientation and production mentality, which are all important elements of a client-based service. It operates consistently at regular hours, helping patrons with their reference and research needs. Another important contribution is the archivist perspective, providing the long-term view of preservation, something archivists have done for hundreds of years with traditional materials. For example, the UCSD Libraries created metadata specifications to support Chronopolis services. "The library thinks in the long term, while at SDSC, when we first started working together, two weeks was a long time to save data," stated Kozbial. For example, when you really think about long-term preservation, you need metadata aspects of the data. That is when we talk with each other and use our complementary skills."

Declerck admits the synergy between these two diverse organizations did not happen overnight. "In the initial stages, there was misunderstanding on both sides," he recalled. A case in point is terminology. "Access" to a librarian means something different than to a computer scientist. "We had to clarify what the other brings to the table," Declerck continued. "At first we discussed infrastructure and technology issues, in which SDSC specializes. Now the conversations are geared toward metadata and policy issues, areas of library expertise." These metadata and policy issues are of particular significance for Chronopolis's migration to a production service.

The creation of SDSC's Data Life Cycle Services Division and the subsequent hiring of David Minor as head of Data Curation further facilitated synergy. Minor, a professional librarian with a strong IT background, is comfortable interacting in both worlds. Today, Minor and Kozbial serve as opposite liaisons between the two institutions, reinforcing the organizations' relationship.

The previous history between SDSC, NCAR, and UMIACS also contributed to the formation of Chronopolis. As a major producer of research data, NCAR has a long-term relationship with SDSC to provide mutual backup storage of each other's critical data. Additionally, SDSC and NCAR were early members of the TeraGrid, a multiyear NSF-funded effort to develop the world's first large-scale production grid infrastructure. The project brings together a national network of high-performance computation resources, high-speed networks, and large-scale storage capabilities. Based on this orientation, a grid-based approach to long-term data preservation was a natural concomitant service. Finally, UMIACS, which has a history of building software tools related to digital preservation, maintains a close relationship with SDSC. It provided a third, natural geographically differentiated replicate site to the project.

With the team in place, the four partners decided to submit an unsolicited proposal to the NSF to create a preservation grid in 2006. Even though it failed to receive support, the group decided to develop a seed pilot during 2006–2007, or "Chronopolis on a shoestring," as described by Berman. Demonstration collections ingested include NCAR observational data, a 12-terabyte data archive of social science at the Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan, and 3 terabytes of images from the National Virtual Observatory's Hyperatlas, a set of "pages" that render the sky to a common standard projection.

These projects prompted the Library of Congress to fund Chronopolis's demonstration service, with the eventual goal of maturing into a production service. As Don Sutton, Chronopolis programmer, SDSC, explained, "Chronopolis became an important resource for data grids and managing the individual objects spread across those data grids. It was a combination of having the right hardware and software in place to be helpful for the Library's preservation needs."

Technical Overview: Process and Tools

Chronopolis builds on SDSC's technology expertise regarding cyberinfrastructure to provide reliable and redundant preservation capability via a multinode grid replication process. Each partner runs a grid node with at least 50 terabytes of storage capacity. This section overviews Chronopolis's data ingestion process. See http://chronopolis.sdsc.edu/publications.html for more in-depth technical papers and presentations. Also, see the sidebar "Digital Curation and Preservation Tools."

Two Fundamental Principles

According to UMIACS's JàJà, Chronopolis is built on two principles. The first is platform independence; neither hardware nor software is tied to any proprietary solutions. The second is modularity: "You have to be able to adapt to new standards, new protocols, new technologies," explained JàJà. "We did not want to tie all the pieces together so we could incorporate new standards without having to modify the entire service. The idea is that as these standards and protocols change, it will not be easy to modify these pieces. But in most cases, you would only have to modify one piece. You can use Chronopolis's tools in any combination—one, two, all of them, together, independently."

Both principles facilitate scalability to incorporate new service providers and provide a continuing technical outlook that a long-term preservation service requires.

Data Ingestion into Chronopolis

Figure 2 provides a generalist's view of the Chronopolis service. Chronopolis users prepare their data and then push their collections to the data staging area at SDSC, where the data is verified for integrity and security. The service can ingest any digital object. The key feature of Chronopolis is that after ingestion at SDSC, data is then replicated at the NCAR and UMIACS sites via existing high-speed educational or research networks. It typically takes a day to replicate a terabyte of data to the partner sites. A set of preservation tools, described in the sidebar, tends the data while it is preserved in Chronopolis.

This service design provides a heterogeneous and highly redundant grid-oriented storage environment. The nodes' geographic dispersion enhances disaster recovery, as it is highly unlikely that an ice storm in Maryland and an earthquake in California would occur simultaneously. The model is scalable beyond three replication sites. Each site supports complete copies of all data stores and has the capability for full curatorial, audit, and access services for each Chronopolis user.

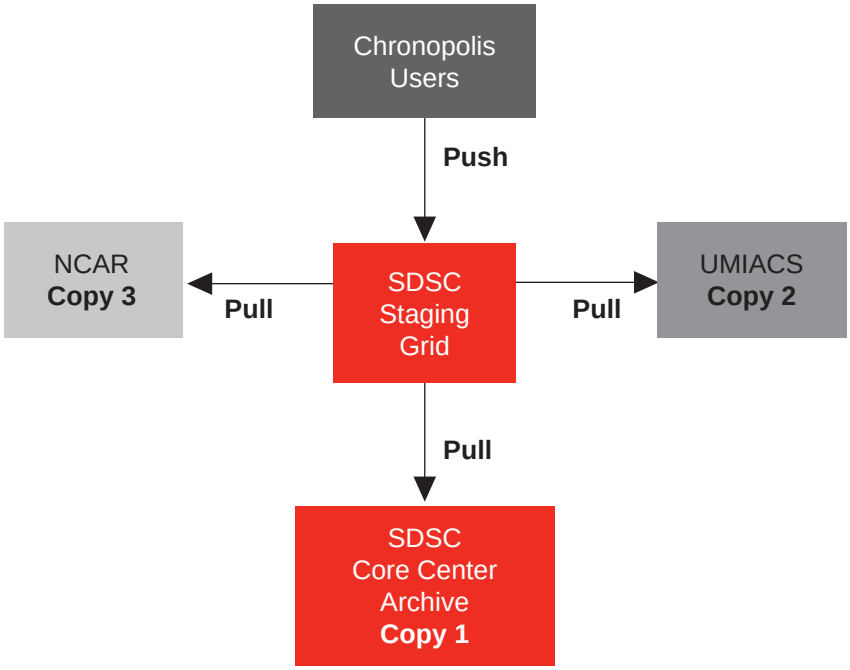

Figure 2. Chronopolis Service Schematic

Because Chronopolis is a deep archive, the service is designed to tend data through the years rather than to provide real-time access. Thus, users are not allowed to recover their data directly; for example, the service is not designed for users to place a front end on Chronopolis-ingested data and distribute it over a website. However, users can use a web-based tool to examine the directory structures of their collections down to the individual object level and to monitor their data collections' status.

The Chronopolis Data Provider Experience

A Chronopolis data provider enlists the service to preserve its digital data. Because it is a demonstration service, the relationship between the service provider and the data provider is more collegial than formal; for example, currently Chronopolis charges no fees for its services. The data provider community developed organically through the NDIIPP program and/or the original Chronopolis pilot. Four clients preserve their data in Chronopolis currently:

- *The California Digital Library (CDL):* CDL manages programs relating to digital collection management and digital management preservation. CDL uses the service as a deep archive of more than 12 terabytes of data.
- *The Inter-University Consortium for Political and Social Research (ICPSR):* ICPSR was a pilot participant and still uses Chronopolis. It maintains a data archive of millions of research files in the social sciences, some of which date from the 1960s.
- *The North Carolina Geospatial Data Archiving Project (NCGDAP):* NCGDAP is a joint project of the North Carolina State University Libraries and the North Carolina Center for Geographic Information and Analysis. The project focuses on the collection and preservation of digital geospatial data resources from state and local government agencies in North Carolina. Currently NCGDAP has collected 6 terabytes in geospatial and ancillary files.
- *Scripps Institution of Oceanography (SIO) at UC San Diego:* A part of the University of California, San Diego, SIO is a leading

Digital Curation and Preservation Tools

From its inception, Chronopolis's goal has been to support tool development and use existing tools, resulting in a suite of sophisticated tools to ensure intact data delivery and preservation. This sidebar lists the tools sequentially in the digital preservation process, from ingestion to data management and finally to data provider review. Each tool maps back to the previously described preservation elements in Table 1 in the main text. Each corresponding preservation element appears in parentheses after each tool's listing.

BagIt File Packaging Format (To Support Data Ingestion)

People typically download a kilobyte or megabyte of data from their hard drive to a flash drive in a matter of seconds. Chronopolis, on the other hand, preserves multi-terabyte-sized data collections, which require days to transmit. BagIt, developed by the California Digital Library and the Library of Congress,[1] facilitates this process. It is a hierarchical file packaging format designed to support disk-based or network-based storage and transfer of generalized digital content. Files are added to a "bag" for transmission, which contains two "housekeeping files"—version and inventory of the content collection—beyond the content itself. Bags can be very large, but to enable fast parallelizable network transfers, a large bag can be transferred with "holes" in it, known colloquially as "holey bags," that contain URLs or pointers to data, not the actual data object. The housekeeping data files maintain an inventory of the content collection and have transmission checking capabilities, so files can be missing in any component of the transfer but can be retrieved subsequently to ensure a final complete transfer. Open-source code is available that instantiates up to 16 parallel processes for a transfer.

Storage Resource Broker, and Integrated Rule-Oriented Data System (To Support Storage Resource Management)

To access ingested files across the three service nodes, Chronopolis relies currently upon the Storage Resource Broker (SRB),[2] a data handling middleware package that provides uniform access to data collections stored within a data grid, which may consist of heterogeneous storage devices distributed across multiple physical locations. It enables Chronopolis to manage collections in a single uniform manner across its digital preservation grid. SRB acts in part as a distributed logical file system that manages multi-organization file system namespaces.

Chronopolis plans to replace SRB, which has been widely used for almost a decade, with a more advanced tool, Integrated Rule-Oriented Data System (iRODS),[3] which allows the building of sharable virtual data collections and their preservation over a long time, even if the data collections are distributed across different projects, locations, hardware, and software. iRODS has been developed by the same organization as SRB, the Data Intensive Cyber Environments Center at the University of North Carolina at Chapel Hill, and is intended to be an evolution of SRB.

Storage Resource Broker Replication Monitor (To Support Storage Resource Management in a Distributed Grid)

Manually copying millions of files across the Chronopolis nodes is not a feasible proposition. There are too many files and the process is too time-consuming. Therefore, Chronopolis completes this task automatically, employing a web-based application called the SRB Replication Monitor.[4] "BagIt is the gateway for Chronopolis services users to ingest data, and then the Replication Monitor takes the data and pushes it to the other sites," explained Michael Smorul, faculty research assistant, University of Maryland Institute for Advanced Computer Studies (UMIACS), University of Maryland, College Park. Each collection within Chronopolis has one or more

replica sites with local replication policies. The Replication Monitor watches registered directories and ensures that copies exist at designated mirrors. The monitor stores enough information to know whether files have been removed from the master site and when a file was last seen.

In addition, any action that the Replication Monitor takes on the files is logged. In part, this is necessary due to the time and number of retries that are necessary to copy millions of files and terabytes of data. "When you are copying millions of files, transfer times can take weeks," continued Smorul. "You have to build automatic retries into the system to take into account different network conditions and server statuses. It is not feasible to restart data-sending every time you have a hiccup." The Replication Monitor recognizes errors and retries operations in an effort to complete a file copy.

Auditing Control Environment (To Support Audit Control)

Joseph JàJà of UMIACS described the Auditing Control Environment (ACE)[5] as the most important tool because it monitors the data's integrity while archived in Chronopolis. ACE incorporates a new methodology using rigorous cryptographic techniques. It captures information about the data during ingestion and then continuously audits the contents of the various objects according to the policy set by the archive, and provides mechanisms for an independent third-party auditor to certify the integrity of any object. "Each site maintains [its] local copy of ACE independently, so if there is an issue at one of the nodes, it does not affect the other two," stated Smorul. "The local sites monitor ongoing condition of the data after we do the initial check during original ingestion to ensure the data is intact. Each partner site can set ACE as appropriate to its respective environment. We do not try to link everything together and create a big piece of software."

ACE consists of two components, the first being an Audit Manager that checks files locally to ensure they have not been compromised. The second part, the Integrity Management Service, issues tokens that the Audit Manager can use to verify that its local store of file digests has not been tampered with. Chronopolis undertakes regular audits among the three sites to ensure the entire collection is intact. Audits can also be initiated by an archive manager or by a user upon data access.

Advanced Access Portal (To Support Data Owner Services)

The final component is Data Provider Review. Chronopolis is designed for infrequent data access, so it requires expert knowledge of the system components, something few—if any—data providers possess. In light of Chronopolis's transition to a production service, enhanced data provider tools are a higher priority, to give data providers better data management capabilities. Under development is a set of tools that will integrate the current components into a single, easy-to-understand portal for current data providers and project stakeholders. The Next Step section examines this tool in greater detail.

Endnotes

1. J. Kunze et al., "The BagIt File Packaging Format," NDIIPP Content Transfer Project (2009), http://www.cdlib.org/inside/diglib/bagit/bagitspec.html.
2. Data Intensive Cyberinfrastructure Environments (DICE), "Main Page [about Storage Resource Broker]," http://www.sdsc.edu/srb/index.php/Main_Page.
3. Integrated Rule-Oriented Data Systems (iRODS), "IRODS: Data Grids, Digital Libraries, Persistent Archives, and Real-Time Data Systems," http://www.irods.org.
4. University of Maryland Institute for Advanced Computer Studies, "ADAPT: An Approach to Digital Archiving and Digital Preservation Technology, n: Replication Monitor," University of Maryland, https://wiki.umiacs.umd.edu/adapt/index.php/Replication:Replication_Monitor_2.0.
5. University of Maryland Institute for Advanced Computer Studies, "ADAPT: An Approach to Digital Archiving and Digital Preservation Technology, ACE:Main," University of Maryland, https://wiki.umiacs.umd.edu/adapt/index.php/Ace:Main.

institution of oceanography and marine technology physics, chemistry, geology, biology, and climate. Principal investigators evaluate data storage and preservation for research data on a project basis. Stephen P. Miller, head, Geological Data Center, and his team use Chronopolis to preserve meteorological and geological information collected from their ocean research voyages.

When discussing Chronopolis's role within their organizations, users viewed it as a single component in a broader digital preservation strategy. Karen Stocks, assistant research scientist, Center for Advanced Computational Science Engineering, SDSC and SIO, and a member of Miller's research team, described her area's digital preservation strategy as a spectrum. "Chronopolis is at one end, and what is stored on your computing device is [at] the other. In between are locally maintained servers. In terms of access, we use solutions designed for 24 × 7 bits user access, and on the other end of the spectrum there is something like Chronopolis."

Data preparation for staging and ingestion at the SDSC site requires some, but not overwhelming, technical know-how. Using the BagIt ingestion tool, for example, "does require some expertise, but it is not like learning a new language like Java," explained Sutton. "The complexity lies in the data's file structure—whether they are contained in one file system or whether they need to be pulled from a number of hardware sources." If BagIt is beyond the user's technical expertise, the data files can always be moved onto tapes and mailed to SDSC directly for ingestion. Overall, clients have felt the data transfer time was acceptable, especially considering the days required to ingest terabytes of data.

File preparation is important. The data's format and type are irrelevant to Chronopolis, but clients have to determine which files to ingest and then ensure their preparedness for ingestion. For example, if one pushes data to Chronopolis in holey bags, they must contain the correct URLs to accurately point to the data for ingestion into the service, a problem one user had to address.

Upon ingestion, Chronopolis's tools automatically monitor and audit the data and can even identify previously undiscovered problems with the data collections. For example, after ingesting one data collection, the Auditing Control Environment (ACE) tool discovered a significant number of overlapping files. "The files were under a different directory," explained Michael Smorul, faculty research assistant at UMIACS. "At the collection's first levels, it looked unique, but six, seven, eight directory levels deep, similarities appeared." Chronopolis team members have not culled the data, as that is a client-level decision.

Data provider needs varied regarding hands-on data monitoring requirements. The Scripps team works with Chronopolis's Sutton on other projects, so he is a natural conduit for their Chronopolis data. "It is not the same as going to a website, clicking 'submit,' and that is all we know about it," stated Stocks. "We feel we have a sense of what is going on with the project. Don [Sutton] would let us know if there was a problem." On the other hand, Bryan Beecher, director, Computing & Network Services, ICPSR, prefers more direct data management. The Advanced Status Portal, currently under development and described later, is meant to address this concern.

Chronopolis's mission is digital preservation, which implies infrequent data access. Only one data provider, the NCGDAP, reported ever retrieving data from Chronopolis, doing so only after three separate, unsuccessful attempts from three different locally stored backups. James Tuttle, digital repository librarian, North Carolina State University Libraries, e-mailed a list of the required files and paths to the Chronopolis team, which in turn e-mailed him a "bag" containing the file links. He pulled down the files using the BagIt downloader, restoring his locally stored

data collection. "It was super simple," Tuttle reported. "I do not know how it could be any easier."

Besides direct interaction with Chronopolis team members, several communication conduits exist for data providers. They are invited to participate in the Chronopolis team's weekly conference call to keep abreast of project activities and to review occasional presentations. Broadcast e-mails inform participants about hardware upgrades and other technical issues. A SharePoint site serves as the project archive. Despite these resources, client participation varies. "It is a project about us [users] providing a [feedback] service and doing what we can," stated Tuttle. "I do not have the bandwidth to participate in the project's governance or strategic direction. I would expect many clients who use Chronopolis would be in the same boat. After all, they use the service due to lack of time, expertise, or resources to preserve digital data."

Next Step: Migration to a Self-Sustaining Production Service

The goal of the Library of Congress's NDIIPP is to incubate proof-of-concept projects, eventually fostering them into self-sustaining production services. As such, Chronopolis's development is at a turning point. Having successfully demonstrated the technical viability of its digital preservation service, the team's next step is to focus on its long-term sustainability. Driving this transition, too, is funding, as the Library's support for Chronopolis as a demonstration project ceases this year, though it may choose to provide bridge funding if the Chronopolis team develops a sustainable business plan.

Until now, Chronopolis's activity has focused primarily on the technical issues—building its preservation infrastructure and tool set. But the partners recognize that self-sustainability requires a whole new set of nontechnical skills, especially to cultivate a user community beyond its NDIIPP base and to develop underlying business models, policies, and governance. Achieving self-sustainability requires completion of the latter tasks. As Schottlaender observed, "Why would someone want to be a prospective client of our preservation service if we do not have a long-term business model? After all, Chronopolis is about the long term." In addition, partners must ceaselessly ensure the service's technical viability.

The "to do" list is lengthy, but the Library of Congress's Anderson is optimistic about Chronopolis's ability to become self-sustaining. Numerous resources lie at the team's disposal to assist with the transition. Partner UCSD Libraries contributes its direct experience with policy making, service production, and community outreach to this effort. Partners NCAR, SDSC, and UMIACS offer their technological experience.

Anderson is working directly with Chronopolis on data provider community identification, business plans, and organizational development. The Library of Congress matched the Chronopolis team with the MetaArchive Cooperative (http://www.metaarchive.org/), a digital preservation cooperative formed in 2004 consisting of 11 institutions with cultural heritage collections and a fellow NDIIPP project, to exchange organizational and technical ideas. Each service approaches digital preservation differently. MetaArchive Cooperative utilizes LOCKSS, an open-source, peer-to-peer decentralized infrastructure;[12] Chronopolis is based on the Storage Resource Broker (SRB) tool. The Library of Congress promoted the collaboration because both are at fairly similar points in their organizational development. MetaArchive Cooperative has transformed itself into a production service, becoming what Anderson described as "a poster child as to how we would like to see all the NDIIPP projects grow up to as nonprofit entities." Sharing its lessons learned helps

the Chronopolis team accordingly. "We are both creating new kinds of institutions," said Martin Halbert, dean of libraries, University of North Texas, and president, MetaArchive Services Group. "A lot of what we have discussed revolves around organizational context and infrastructures."

The relationship offers potential technical synergy, especially as the partners develop an interface between MetaArchive Cooperative's LOCKSS-based and Chronopolis's SRB-based digital preservation solutions. "We wanted to explore redundancy with another system from an exit strategy perspective—something that all of us need to have as a preservation solution," stated Skinner. "It is another piece of our bigger preservation strategy, as Chronopolis's supercomputer grid environment increases our own capacity for preservation services."

Policy: Formalizing Trust

As with many projects in higher education, Chronopolis began as a fellowship of people with similar interests working together toward a common goal. As the project moves to a production service, the current partners realize the importance of formalizing their communal trust. "We are all friends working together with a general sense about the project," said Schottlaender. "We trust each other, but we all realize that as we move into production, we have to codify our relationship."

Consequentially, work is under way to create memoranda of understanding (MOUs) and service level agreements (SLAs) that define all the technical, support, and administrative obligations that participation entails. Currently, Chronopolis partners have completed individual agreements. SDSC and the UCSD Libraries formalized their Chronopolis relationship with an MOU and an SLA. SDSC maintains separate MOUs with UMIACS and NCAR, reflecting their individual Chronopolis activities, but the partners want to extend these agreements to cover all Chronopolis collaborations. The goal is to create a more permanent set of formal trust agreements.[13]

Given Chronopolis's multi-institutional nature, codification of such agreements can become quite complex. Due to their greater experience and resources at hand, SDSC and the UCSD Libraries opted to take the lead role in this effort, creating a skeletal project term sheet with NCAR's and UMIACS's input, from which the UCSD campus counsel will draw up an MOU for endorsement by the remaining partners' institutions. The variety of contexts in which the partners reside complicates codification further. At one end of the spectrum, a research group, such as UMIACS, can commit on its own, but at the other end a federal entity, such as NCAR, must get approval through a hierarchy.

The transition from gratis to fee-based service changes the data provider relationship dynamic, requiring formal definitions of Chronopolis and client responsibilities. Of particular importance are policies to alleviate data providers' concerns about data access and data sharing. Again, complexities come into play. Not all data has the same needs in terms of privacy and access control. Policies must reflect the legal, business, intellectual property rights, and ethical issues surrounding data.

Eventually Schottlaender foresees a natural move to a more formalized governance structure, tentatively envisioned as a steering committee with an advisory board, to include clients formally in policy issues. "As you get more clients, it is only reasonable to expect that they will want to have some input," he stated.

Business Model Development

Until now NDIIPP funded Chronopolis activities, but to go forward Chronopolis must create a self-sustainable business model. "You can't have an Information Age without information, and yet the preserva-

tion and access of that information costs money," stated Berman. "A critical part of the problem is that functioning infrastructure is unmemorable and costly. You do not think about the fact that the lights stay on in your building until the room goes dark. Data cyberinfrastructure is the same. The disks, tapes, people, systems, and uninterruptible power systems all cost money."

Business plan creation requires expertise not typically found in the library or technology communities that constitute Chronopolis. The service's long-term nature further complicates the process, as partners ponder such questions as "Who pays for preservation of data for decades?" "Does one pay up front for perpetuity or pay as you go?" Current research funding mechanisms, which tend to focus on relatively short-term projects, aren't designed for this. The project life cycle mentality of research funding complicates the issue, too. A researcher receives a grant, creates a data set, and then publishes the research findings. The funding ends at the project's termination, with no long-term funding for preservation. On the other hand, it is becoming increasingly evident globally that future research will depend in part on historic data. Also, initiatives such as the Protein Data Bank, the Human Genome Project, and the Shoah Education Project are built with a huge investment and will need to be preserved into the foreseeable future.

Several information sources will factor into Chronopolis's service fee structure. Both SDSC and CDL completed separate market analyses on the cost of storing a terabyte of data, with similar conclusions. The NSF's Blue Ribbon Task Force on Sustainable Digital Preservation and Access is studying these issues on a larger scale. The fundamental premise of the task force is that once data is lost, it is lost, and thus digital preservation is a worthy investment. Other suppositions predicate their research as well: infrastructure is not free, confusion exists about preservation funding responsibilities, current institutional and granting agency funding models don't address long-term preservation, and general complacency about digital preservation stems from the common belief that current practice is good enough. The task force will release its final report on this issue in 2010 after almost two years of study.

Community Outreach

Chronopolis built its data provider community from preexisting SDSC or NDIIPP relationships. Current data providers acknowledge Chronopolis's need to transition to a fee-based service, and depending upon its costs, they're inclined to continue with the service. But long-term survival depends upon Chronopolis's ability to draw in new clients beyond SDSC and NDIIPP.

Team members admit that potential clients are rather specialized. "Chronopolis will appeal to those organizations that have preexisting infrastructures and can define their digital holdings as those that lend themselves to long-term preservation in terms of unchanging documents, used data sets, and other things not subject to change," explained Smorul. Data collections with those characteristics are most likely to fall in the Community Value and Societal Value portions of the Data Pyramid.

The Chronopolis partners' next steps are to identify and reach out to specific communities with corresponding data holdings. The Library of Congress is assisting with community identification efforts; outreach efforts include a series of presentations at relevant conferences and symposia.

Chronopolis has identified one particularly viable market for its services: the preservation of data sets that support published research in printed and electronic journals. Such preservation facilitates their use in subsequent projects or lets clients track their use for tenure and other applications. "This is a large concern because the government spends a lot of money on scientific research," stated Anderson. She cited a

Library of Congress study that discovered a data loss rate of more than 50% in social science projects over the last 15 years. "Sometimes researchers possessed the data on a disk somewhere," Anderson continued. "Other times it was deleted at the project's termination. The NSF has signaled a high interest in trying to solve this problem."

SDSC and Scripps Institution of Oceanography's Stocks concurred with this view. "For many research projects, the end product is a publication that tells the answer," she explained. "That's the data's life cycle; it fulfilled its purpose when the paper came out. Now the world is changing and researchers are investigating many long-term, large-scale questions. Data may be reused for these purposes instead of languishing in file cabinets. Data preservation is a problem that keeps getting raised among some of my colleagues."

To attract potential users, the Chronopolis team promotes its service's transparency, identifying openly the service's data center locations, equipment, and software. Such transparency is not typical of commercial alternatives, nor do these services focus on truly long-term data preservation. In addition, the team plans to apply for Trustworthy Repositories Audit & Certification (TRAC)[14] by the National Archives and Records Administration (NARA) in 2010. TRAC audit covers three categories: Organizational Infrastructure; Digital Object Management; and Technologies, Technical Infrastructure, and Security.[15] The TRAC audit is highly respected and a likely user selection criterion.

Ultimately, outreach could apply to potential partners or service providers. Additional partners will enhance the service's geographic dispersion and distributed storage of data copies. But the barriers to entry—technical expertise and adequate bandwidth—dramatically shrink the pool of potential partners.

Technical Issues

Chronopolis's move into production shifts technical priorities, too. The Chronopolis team focused initially on the implementation and refinement of the service's infrastructure, operation, and digital preservation tools. But now issues like enhanced user tools and long-term technical sustainability assume precedence.

Data remains inactive while ingested in Chronopolis, so users may find the service's interface less than ideal for hands-on data management. Beecher described the current state of affairs: "Right now Chronopolis is very much a black box. There is no real tool that would allow me to interrogate Chronopolis to locate and retrieve specific digital objects."

The imminent Advanced Status Portal will address this issue. It aggregates information in a way to allow the data providers to get feedback and to receive standardized reports on the status of the data. For example, the information that is required for monitoring the status and error conditions can currently be found in Chronopolis components if one knows where to look—something team members know, but data providers most likely do not. A web-based status display will pull status information from all Chronopolis components and integrate it so that users can quickly ascertain the state of collections of interest, find any replication or verification errors, then drill into the information to discover the cause. This interface will also provide access to collections' metrics and reports.[16]

Given its long-term nature, Chronopolis was always designed to be as independent as possible of any current technology and to incorporate the latest best practices. Both principles enhance Chronopolis's technical sustainability. Technology is never static; input, output, media, and networks are all destined to evolve, and Chronopolis must do so accordingly. A case in point is Chronopolis's current shift from SRB to iRODS for storage resource management.

Client data collections' characteristics will change, too. David Minor characterizes current data collections stored in Chronopolis as "small- [to] medium-sized collections that range from a single terabyte to a dozen terabytes." The team believes a fairly large portion of data collections falls into this range, but they foresee collection sizes scaling up fairly quickly in response to the escalating generation of digital content. Such a scenario presents new demands on infrastructure, bandwidth, and technology.

Bandwidth especially worries Minor. "The simple mechanism for transferring and moving data geographically around the country is not as robust as people assume it to be. This has a big impact on everything that we try to do. When you get above a certain data collection size, it gets easier and easier to ship data tapes by truck or plane." One measure under consideration is the local ingestion of data, enabling all three partners—not just SDSC—to function as initial staging areas for clients' data collections and subsequently to push them to the other nodes. Local ingestion presents a more convenient option for data providers, making the service potentially more attractive to prospective clients.

Other Sustainability Scenarios

While not actively under consideration, other options do exist to ensure Chronopolis's financial sustainability. One possibility may be to expand Chronopolis from a UCSD resource into a University of California resource. As a huge generator of data, the University of California System has a natural affinity with Chronopolis's digital preservation services. Such a designation may bring potential resources to the project, enhance credibility, and offer greater liability protection.

Commercial services partners, as for example Amazon and Google, are another option. Both companies offer web-based storage services currently, aimed primarily at Individual Value data. Both services provide flexible access capabilities but offer no guarantees regarding long-term preservation. Chronopolis would expand both vendors' service offerings. Additionally, Google has ties with higher education already through the Google Apps for Education offering.

Lessons Learned

As this case study shows, digital preservation is complex and multifaceted, and not easily reconciled with its challenges. Chronopolis's experiences offer several lessons for higher education institutions as they ponder their digital preservation options.

One size does NOT fit all.

More and more of the information generated daily is in digital form, falling into a spectrum ranging from the purely personal to data of long-term national importance. A similar spectrum exists for access and continuity of data needs, ranging from easy, regular access to limited access; extending from ever-changing versions to guaranteed, enduring versions. The Library of Congress recognized this need early on, forming a network of diverse digital preservation partners in which Chronopolis falls in the limited-access, long-term-preservation end of the spectrum. Similarly, institutions could organize their own digital preservation networks, utilizing a spectrum of services to safeguard their data accordingly, ranging from Chronopolis to personal data storage options. Whether managed centrally or locally (in colleges and schools) or some combination of the two, such networks will rely upon each institution's culture. But a formalized strategy will help to organize an institution's digital preservation activities.

Create synergies between IT and the library.

The Chronopolis project brings to light the library's important role in digital preservation—for example, in applying archival practices to

data collection prioritization. Yet working style, terminology, and general cultural differences could have stymied interactions between SDSC and the UCSD Libraries. The two organizations worked hard at forging their relationship, admitting that building their relationship was a long-term endeavor. "There is no magic bullet," stated Declerck. "Senior administrators can play a role and create expectations." Appointing liaisons between the two organizations created specific touch points, too. As the library curates more digital collections, its relationship with IT will grow only more vital.

Technology is just the first step.

Deep archival preservation is an emergent area, and when assessing its progress to date, all Chronopolis partners agree that preservation technology development is the easy part of the project. Rather, the difficulties lie in the nontechnical issues of policy, business models, and community outreach.

Digital preservation requires development of new funding models.

Research grant funding normally has expectations that funds will be expended by the end of the grant. Institutions are not used to taking responsibility for data collected during research projects. Yet, some data has value long after a project is concluded. There needs to be a system for priority setting of data and funding for long-term preservation of the most valuable.

Promote digital preservation education.

During case study conversations, it became apparent that lost data resulted at times from lack of forethought about the long-term consequences of digital data. Researchers, faculty members, and others never considered the value of their digital data past the life of their research project, course, or other activities. In other words, once the paper is published, the data is downloaded to disks, tapes, or flash drives that may end up in a desk drawer. Researchers often need help at the beginning of the research project in order to ensure the data coming out of the project is manageable, so that they don't face a costly and time-consuming—or even an intractable—mess at the end of the project.[17] Education about prioritizing data and the importance of digital preservation may encourage colleagues to consider the long-term consequences of their data. Libraries may be a natural venue when one considers their ongoing involvement in educating the institutional community about safe online research practices and similar topics.

Prioritize and evaluate data collections.

It is easy and inexpensive to store digital files, so people tend to save everything. But preservation requirements are far more rigorous. Thus, it is important to weed out the inconsequential data collections from those that are worthy of the investment in digital preservation. Starting this practice now will only reap more benefits in the future as the amount of digital content generated continues to escalate. A little up-front preparation will identify potential problems before data is ingested into a digital preservation system by preemptively exposing duplicate or missing files as well as recursive directories. In addition, such practices will force the data provider to consider optimal data formats to ensure future accessibility. The aforementioned education efforts could spur collective and individual actions.

Digital preservation requires a long-term outlook.

The typical technology planning window may extend three or five years, but digital preservation involves longer-term thinking in such issues as cyberinfrastructure evolution, transition of the data, and changing industry standards. Obviously, the Chronopolis team does not have a definitive view of the future, but they keep abreast of changing conditions and have designed an adaptive service that can respond to technological change.

Conclusion

Long-term digital preservation is a societal imperative with local implications to safeguard an increasing digital repository. This is especially true for colleges and universities, with their sizable inventory of Societal Value collections of research and scholarly data, to avoid an incident similar to the basement discovery of Apollo moon tapes.

But underlying technology and policy issues may complicate individual digital preservation solutions. One alternative, as digital preservation visionary Clifford Lynch describes, is services ensuring that the data is properly documented, that it is correctly placed into a well-known and well-defined format (using the standards of appropriate scholarly communities when available and applicable), and that it is preserved over suitable periods of time by the use of redundant managed storage and, when necessary, format migrations. And, most important, some organization must take responsibility for the data—technically, legally, and financially—and do what's necessary to look after it.[18] Chronopolis is a promising project that suggests a similar future of deep archival data preservation.

Endnotes

1. Ronald Yanosky, *Institutional Data Management in Higher Education* (Research Study 4, 2009) (Boulder, CO: EDUCAUSE Center for Applied Research, 2009), available from http://www.educause.edu/ecar.
2. John Gantz, "The Diverse and Exploding Digital Universe," International Data Corporation (2008), 3, http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf.
3. Michael Witt, "Institutional Repositories and Research Data Curation in a Distributed Environment," *Library Trends* 57, no. 2 (2008): 191–201, http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1126&context=lib_research.
4. Carmela Amalfi, "Space Week: Lost Moon Landing Tapes Discovered," *Cosmos*, November 1, 2006, http://www.cosmosmagazine.com/node/818.
5. Yanosky, *Institutional Data Management,* 128.
6. Ibid., 130.
7. National Science Board, *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century* (National Science Foundation, NSB-05-40, September 2005), http://www.nsf.gov/pubs/2005/nsb0540/.
8. Francine Berman, "Got Data? A Guide to Data Preservation in the Information Age," *Communications of the ACM* 51, no. 12 (2008): 50–58, http://mmlab.ceid.upatras.gr/dstructures/files/A_guide_to_data_preservation_in_the_information_age.pdf. The Branscomb Pyramid first appeared in the final report from a 1993 National Science Foundation panel, led by Lewis Branscomb, about the future of high-performance computing. The pyramid illustrates a base of least powerful computing platforms moving to a tip of most powerful computing platforms. As one moves up the pyramid, it represents the increasingly specialized nature of the computing platforms and the decreasing population of correspondingly capable computing devices.
9. Reported as of December 1, 2009, at Research Collaboratory for Structural Bioinformatics Protein Data Bank, http://www.rcsb.org/pdb/home/home.do.
10. The list of digital curation and preservation practices was compiled from the previously cited Berman article, "Got Data? A Guide to Data Preservation in the Information Age" (*CACM*, December 2008) and Chronopolis documentation found at http://chronopolis.sdsc.edu/about.html.
11. The Library of Congress, "National Digital Information Infrastructure and Preservation Program," http://www.digitalpreservation.gov/library/.
12. LOCKSS stands for Lots of Copies Keeps Stuff Safe and is an open-source solution digital preservation infrastructure based at the Stanford University Libraries. More information is available at http://www.lockss.org/lockss/Home.
13. Fran Berman et al., "The Need to Formalize Trust Relationships in Digital Repositories," *EDUCAUSE Review* 43, no. 3 (May/June 2008): 10–11, http://net.educause.edu/ir/library/pdf/ERM0835.pdf.
14. "Trustworthy Repositories Audit & Certification: Criteria and Checklist," Online Computer Library Center Inc. (OCLC) and the Center for Research Libraries (CRL) (2007), http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf.
15. Ibid., Table of Contents.
16. David Minor et al., "Chronopolis Digital Preservation Network" (paper presented at 5th International Digital Curation Conference, London, England, December 3–4, 2009); to be published in the *International Journal of Digital Curation*, June 2010).
17. Clifford Lynch, "The Institutional Challenges of Cyberinfrastructure and E-Research," *EDUCAUSE Review* 43, no. 6 (November/December 2008), http://www.educause.edu/EDUCAUSE+Review/EDUCAUSEReviewMagazineVolume43/TheInstitutionalChallengesofCy/163264.
18. Ibid.

Citation for This Work

Pirani, Judith A., and Donald Z. Spicer. "The Chronopolis Project: A Grid-Based Archival Digital Preservation Solution" (Case Study 1, 2010). Boulder, CO: EDUCAUSE Center for Applied Research, 2010, available from http://www.educause.edu/ecar.