Building Preservation Environments with Data Grid Technology

Reagan W. Moore

Abstract

Preservation environments for digital records are successful when they can separate the digital record from any dependence on the original creating infrastructure. Data grid technology, which supports the management of records that are located on multiple storage systems, provides the software needed for infrastructure independence. This article provides a description of how data grid technology can be used to support preservation processes and of existing preservation environments that are based upon data grids.

At the conclusion of its first phase, the InterPARES project (International Research on Permanent Authentic Records in Electronic Systems) issued a number of requirements for authenticity of digital records and methods of selection and preservation.[1] The conceptual foundation of these products is exemplified in an *Intellectual Framework for Policy Development*. Among the key principles expressed in the framework are the following:

This article is the result of the work that I have carried out as a co-investigator in the InterPARES 2 research project. The knowledge contained in this article has been collectively produced within the InterPARES project. The results presented here were supported by the InterPARES project, NSF NPACI ACI-9619020 (NARA supplement), the Persistent Archives Testbed (NHPRC grant number 2004-008), the NSF NSDL/UCAR Subaward S02-36645, the DOE SciDAC/SDM DE-FC02-01ER25486 and DOE Particle Physics data grid, the NSF National Virtual Observatory, the NSF Grid Physics Network, and the NASA Information Power Grid. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation, the National Archives and Records Administration, or the U.S. government, or the final findings of InterPARES, because they are still in the testing phase. More information about the example permanent archive based on the SDSC Storage Resource Broker can be found at http://www.sdsc.edu/srb/ and http://www.sdsc.edu/NARA/.

[1] The first phase of InterPARES spanned the period 1999–2001, and the products mentioned above are included in the book that constitutes the project's final report: *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project* (San Miniato, Italy: Archilab, 2005). The text of the book is also available on the project's Web site at http://www.interpares.org/book/index.cfm.

- It is not possible to preserve a digital record as a stored physical object, but only the ability to reproduce it; and
- The preservation of authentic electronic records is a continuous process that begins with the process of records creation and whose purpose is to transmit authentic records across time and space.[2]

In light of the above, it is essential to identify clearly the digital entity that needs to be preserved when we talk about an "authentic electronic record." The InterPARES definition of an electronic record is "a digital object that possesses a fixed documentary form, a stable content, an archival bond with other records, and an identifiable context."[3] In addition, InterPARES recognizes "that the physical and intellectual components of an electronic record do not necessarily coincide, and that the concept of digital component is distinct from the concept of documentary form."[4] A digital component is "a digital object that contains all or part of the content of an electronic record, and/or data or metadata necessary to order, structure, or manifest the content, and that requires specific methods for preservation."[5] The documentary form is "the rules of representation according to which the content of a record, its administrative and documentary context, and its authority are communicated."[6] According to the InterPARES findings, an electronic record is presumed authentic if certain metadata related to its identity and integrity are inextricably linked to it. The necessary metadata that express the unique identity of a record are

- Date record is made
- Date record is transmitted
- Date record is received
- Date record is set aside (i.e., filed)
- Name of author (person or organization issuing the record)
- Name of addressee (person or organization for whom the record is intended)
- Name of writer (person or organization responsible for the articulation of the record's content)
- Name of originator (electronic address from which the record is sent)
- Name of recipient(s) (person or organization to whom the record is sent)
- Name of creator (person or organization in whose archival fonds the record exists)

[2] *The Long-term Preservation of Authentic Electronic Records*, Strategy Task Force Report, principles 6 and 8.

[3] *The Long-term Preservation of Authentic Electronic Records*, Strategy Task Force Report, principle 1.

[4] *The Long-term Preservation of Authentic Electronic Records*, Strategy Task Force Report, principle 7.

[5] *The Long-term Preservation of Authentic Electronic Records*, Strategy Task Force Report, principle 7.

[6] *The Long-term Preservation of Authentic Electronic Records*, Strategy Task Force Report, principle 7.

- Name of action or matter (the activity in the course of which the record is created)
- Name of documentary form (e.g., e-mail, report, memo)
- Identification of digital components
- Identification of attachments (e.g., digital signature)
- Archival bond (e.g., classification code)

The metadata necessary to assess the integrity of a record are

- Name(s) of the handling office/officer
- Name of the office of primary responsibility for keeping the record
- Indication of annotations or comments
- Indication of actions carried out on the record (e.g., removal of digital signature)
- Indication of technical modifications due to transformative migration

It is possible to maintain records with their metadata intact in the environment in which they are created if proper measures are taken to protect the records from malicious or accidental change, that is, if the records are managed in a trusted recordkeeping system. A *trusted recordkeeping system* has been defined as "a type of system where rules govern which documents are eligible for inclusion in the record-keeping system, who may place records in the system and retrieve records from it, what may be done to and with a record, how long records remain in the system, and how records are removed from it."[7]

Technological obsolescence, however, makes it necessary to migrate records to new hardware and software environments over time. Data grid technologies are designed to address this issue of technology evolution.[8] The expectation that data grid technology can be used to manage technology evolution is based on the observation that at the point in time when new technology is brought into a preservation environment, both the new and the old systems are present. A software system that supports simultaneous access to multiple types of storage systems can be used to manage technology evolution, ensuring that the preservation environment will be able to take advantage of new, more cost-effective technology. The challenge is designing a data management system that is able to support evolution of all of its constituent parts, from the storage system, to the database technology, to the authentication mechanisms, to the access mechanisms. Data grid technology, and in particular the Storage Resource Broker (SRB) data grid, provides these capabilities. Data grids are used for all scales of data management, from small collections that have a few thousand files and that are managed using a personal computer,

[7] Margaret Hedstrom, "Building Record-Keeping Systems: Archivists Are Not Alone on the Wild Frontier," *Archivaria* 44 (Fall 1997): 57.

[8] Arcot Rajasekar, Michael Wan, Reagan Moore, "mySRB and SRB, Components of a Data Grid," 11th High Performance Distributed Computing conference, Edinburgh, Scotland, July 2002.

to massive collections that have a hundred terabytes of data and fifty million files.

The SRB data grid has proven the concepts required for infrastructure independence, the ability to manage authenticity and integrity of records independently of the choice of storage technology and access method. The National Archives and Records Administration (NARA) research prototype persistent archive has demonstrated evolution of storage systems (added new types of storage archives), managed evolution of access methods (added graphical information systems interfaces and digital library interfaces), managed evolution of databases, and even managed evolution of the SRB data movement mechanisms to handle network devices such as firewalls. The NARA Electronic Records Archive solicitation requirements were based in part on the concepts proven in the SRB.

The SRB data grid is generic software that is used to support preservation environments, digital libraries, real-time data systems, and shared collections, ensuring that preservation systems can build upon advances developed for all types of data management environments. Data grids put authenticity and integrity under the control of the archivist, freeing the preservation environment from dependencies on the capabilities of particular storage systems. The management of security for personal and protected information, copyright, and access is made a property of the preservation environment rather than a property of the storage system. The implementation of a preservation environment is discussed in detail in this paper, with demonstrations of the capabilities that are needed to manage technology evolution while preserving authenticity and integrity. In particular, this paper characterizes what is needed to preserve digital components.

Support for Storing Digital Components[9]

A digital component is created in a software and hardware environment that provides storage attributes needed for its management and control. The storage attributes that are typically provided by the storage system include

- Storage system name (i.e., network address)
- File name (i.e., location within the storage system)
- Names of file management properties (e.g., size, file creation date, file modification date)
- Names of users (e.g., owner of the file and others having access)
- Access privileges (e.g., allowed operations on a file by a user)

[9] It is understood that what is preserved is records and not digital components. However, InterPARES 1 showed that to preserve records, one must carry out operations on digital components. This is the only reason why the expression "digital component" is preferred to "digital record" within this section of the article. The writer never loses sight of the fact that the data grid must be used to preserve records.

These attributes make it possible to identify a digital component on the basis of its location and to permit access to it. The location is a combination of the storage system name and the file name. The storage system automatically updates the storage attributes to track operations on the digital component such as changes in size.

These attributes are noticeably insufficient to satisfy the InterPARES requirements for the maintenance of the identity and integrity of the records over the long term. Therefore, any storage system chosen to host records and their digital components must support the additional metadata attributes required by InterPARES. It is possible to modify a storage system to accommodate these attributes, which would mostly fit under the existing category of file management properties.

Because the naming conventions of the storage attributes are dependent upon the storage system, however, all storage attribute names may change when a digital component is moved to another storage system. For example, the storage system name and the file name may change. Even worse, the file properties may be reset. If the digital component is moved to another site or institution, the names of the users may also change. This means that access privileges must reflect a new set of user names. Moving digital components to new storage systems, therefore, can change all storage attributes associated with a record and affect the capability to manage identity and integrity metadata needed to ensure the continuing authenticity of the records.

Data grids are software systems that automate the management of storage attributes when digital components are moved between storage systems and between sites.[10] Indeed, management of the storage attributes is required to manage software technology evolution. Because future software and hardware systems will not be the same as those used to create the digital components, we need to manage technological change within the preservation environment itself in order to preserve records. This fact forces us to pay special attention to

- the creation of an infrastructure-independent software preservation environment; and
- the management and control of each record within its archival aggregation.

A successful digital preservation environment is one in which each digital record is separated from the software and hardware technology that was used for its creation.[11] This means that the digital record can be preserved in a storage system different from the original and can be accessed and displayed

[10] Reagan Moore, et al., "Data Intensive Computing," in *The Grid: Blueprint for a New Computing Infrastructure*, ed. Ian Foster and Carl Kesselman (San Francisco: Morgan Kaufmann, 1999).

[11] Reagan Moore, et al., "Collection-Based Persistent Digital Archives, Parts 1 and 2," *D-Lib Magazine* 6 (March and April 2000) at http://www.dlib.org/.

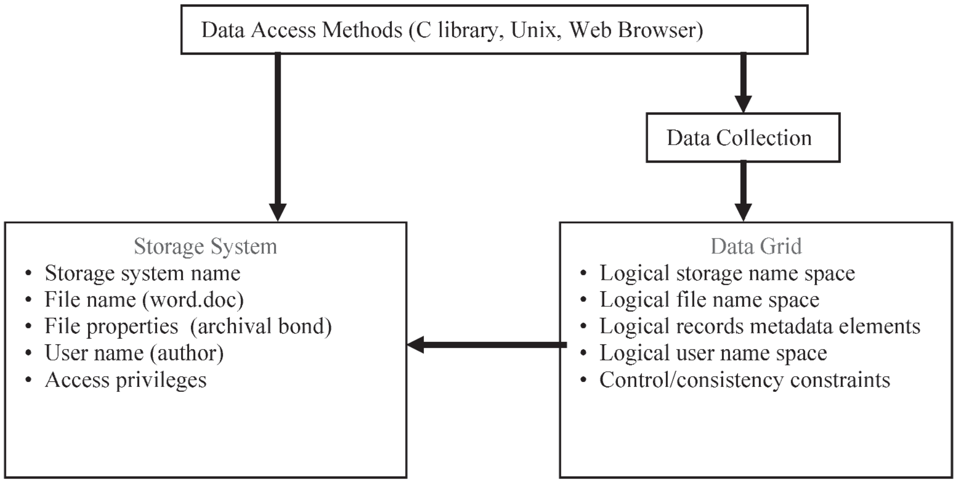
Data Access Methods (C library, Unix, Web Browser)
Data Collection
Storage System
• Storage system name
• File name (word.doc)
• File properties (archival bond)
• User name (author)
• Access privileges
Data Grid
• Logical storage name space
• Logical file name space
• Logical records metadata elements
• Logical user name space
• Control/consistency constraints

FIGURE 1. **Capabilities provided by data grids**

through applications different from those originally used, while remaining intact with its own metadata. The ability to extract digital records from the infrastructure in which they were created is needed to ensure that the records will be accessible on future hardware and software systems.

How can a record be separated from the digital environment in which it was created while maintaining its authenticity? The answer is based upon the insertion of a data management software infrastructure (a data grid) between the storage system where the digital components are hosted at any given time and the access applications used to discover and retrieve records.[12] Each of the attributes provided by the original storage environment must be managed by the data grid to ensure the ability to maintain storage metadata. Figure 1 shows how the storage system attributes are mapped onto the logical naming conventions used by the data grid, referred to as logical name spaces.[13]

The data grid creates a logical name space for each of the naming conventions used by the original storage location. The logical name spaces provide global, permanent identifiers for the attributes that were originally associated with the creator's recordkeeping system. The global identifiers make it possible for data grids to manage digital components independently of the naming conventions imposed by a particular vendor's storage product, as follows:

- **Logical storage name space.** Used to create permanent storage location identifiers, the logical storage name is mapped to the network address

[12] Reagan Moore, "The San Diego Project: Persistent Objects," *Archivi & Computer*, Automazione e Beni Culturali, l'Archivio Storico Comunale di San Miniato, Pisa, Italy, February, 2003.

[13] Note that the XML community uses the term *namespace* to reference a registered naming convention. The data grid name space is used to structure names in a collection hierarchy.

corresponding to the actual storage location by the data grid. The physical storage system can be changed, but the preservation environment continues to use the invariant logical storage name. In practice, this is accomplished by copying the records (reproducing their digital components) into a new physical storage system and then changing the mapping of the logical storage name to point to the new physical storage system.

The logical storage name can represent a list of multiple physical storage systems. When a record is "written" to the logical storage name (that is, a request is made for the creation of a copy), the reproduction process is completed only when a copy is deposited on each physical storage system in the list. This makes it possible to automate reproduction of the same record on multiple storage systems.

- **Logical file name**. Records and/or digital components are typically stored as files in a computer file system or tape-based storage system. Data grids map from the logical file name to the physical file name under which the digital component was reproduced.[14] If the digital component is moved to another storage system, the mapping from the logical file name to the physical file name and storage system is automatically updated. This means that migration[15] of digital components to new technology can be automated. A typical approach is to add the new physical storage name to the list of physical storage systems associated with the logical resource name. A synchronization command is then executed to force creation of a copy on the new physical storage system. The logical file name can be used to manage copies of the digital component. The location of each copy is stored as metadata, along with the time that the copy was produced. This makes it possible to keep a copy in a tape/disk-based storage system, accessible only by a trusted custodian[16] (an archivist, for example), and a copy in an access environment for use by the public.
- **Logical records metadata elements**. Data grids associate metadata elements with each digital component that is registered into the logical file name space, to which in turn the storage attributes are also linked.

[14] Storage systems provide a naming convention for files, called a physical file name, which is a combination of the file system directory path name and the file system name for a file.

[15] The term *migration* is used in this document to denote copying of data onto another storage system. The term *transformative migration* is used to denote the conversion of the encoding format of a document to a new standard.

[16] A trusted custodian is a physical or juridical person entrusted with independently maintaining the records of electronic data interchange (EDI) partners, "The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project," http://www.interpares.org/book/index.cfm, Strategy Task Force Report, principle 5.

In addition, preservation integrity attributes are updated whenever an operation is carried out on the records and their metadata for the purpose of preserving them. An example is the maintenance of an audit trail that records all operations performed on the logical records. The preservation attributes are also linked to the logical file name space. Therefore we can talk of an entity called *logical attributes*, which includes the records metadata, the storage system attributes, and the preservation attributes. All are linked to the logical file name, which is the key to the whole information system.

Data grids also support the digital equivalent of a cardboard box that contains paper documents. Data grids provide a mechanism to aggregate digital components (equivalent to paper documents) into a single digital file (equivalent to the cardboard box) before they are stored. The file is called a *container* and represents a completely arbitrary grouping of digital components. The storage attributes would then include the location of each digital component within the container, as well as the name of the container and the storage location of the container. Each digital component still retains its identity and can be referenced and accessed even after aggregation into a container. This means that the way the digital components are stored may be different from the way they are logically organized in the logical file name space. Presentation of the digital component to the user is based on the structure of the logical file name space, not the physical order within a container.

Data grids provide a logical name space for containers. The trusted custodian can choose to aggregate the records according to their archival arrangement. This means that the name for each container may be the name of the archival aggregation of records (e.g., file, series), or of a subaggregation (e.g., subseries), or of a group within the aggregation identified chronologically or alphabetically. A logical container represents a set of physical containers. When a physical container is filled, the data grid opens a new physical container under the same logical container name and stores the next records.

- **Logical user name**. Data grids provide a uniform identification mechanism for users that supports access across sites and administrative domains. The data grid maintains a unique name for each person who is authorized to perform preservation procedures on digital components. When the digital components are moved to another institution, the same user names can still be used to identify the actions that were performed before by a preserver. Thus a custodial and preservation history can be consistently maintained.

  Data grids also support assignment of user names to groups that can represent the individuals allowed to perform selected custodial functions. When the individuals that provide the custodial functions change, the new individuals can be assigned to the appropriate group.

- **Control/consistency constraints**. Data grids manage access privileges in the logical user name space for each logical file name and each logical record metadata element. This means that the access controls do not change when the digital component is moved to another storage location. In addition, data grids manage consistency controls on how the logical records metadata elements are updated after a preservation procedure. An example is the validation of the digital signature or checksum of a record. A checksum is generated by an algorithm that combines all of the bits in the record to provide a reduced representation. A checksum can be evaluated and compared with the value stored in the logical attributes. If the record becomes corrupted, for example by damage to a tape on which it was stored, the data grid can replace it with a copy that has been previously verified as correct.

  The consistency between the original records and the copies and between the records metadata and the records is managed by the data grid. The logical attributes can be updated immediately after an operation (for example the update of an audit trail) or can be deferred to a later time (the validation of the digital signature). In both cases the data grid provides the mechanisms to bring all storage system attributes into a globally consistent state across all copies.

In Figure 1, for instance, the original storage system might be accessed directly through a Unix shell command. In such a case the user had to know the physical file name under which the record was stored, and he or she had to have access permission to the storage system. By using data grid technology to manage the logical attributes, it becomes possible to insert data collection management technology between the user-access mechanism and the storage system. The data collection management technology is an integral part of data grids and is required to manage the logical attributes. At the same time, it makes it possible to issue discovery queries on records metadata, storage metadata, or preservation metadata; find the relevant digital record; validate the integrity of the digital components; and access the appropriate storage system and retrieve the records without having to know the physical file name or storage location. In the data grid, the access can be made through use of the logical file name or by a query on the logical attributes. The data grid handles the mapping from the logical file name to the physical storage system location.

Data Grids

As shown above, data grids are software systems that provide generic software infrastructure for managing distributed records (and records metadata).

Data grids are used to support all types of data management environments, from preservation systems, to digital libraries, to shared collections, to real-time observational data.[17] The capabilities provided by data grids are essential for automating preservation processes, mitigating risk of data loss through reproduction of digital components, assuring the permanent association of identity and integrity metadata with records, and supporting retrieval and access. At the same time, data grids are designed to manage digital entities stored in any type of storage system, while providing access through a very wide variety of access mechanisms. This ability to interact simultaneously with multiple types of storage systems and access systems forms the core of data grid support for technology evolution.

Data grids manage changes in software and hardware systems and simplify the incorporation of new technology into a preservation environment. They make it possible for a trusted custodian to take advantage of advances in technology without risk to the authenticity of the records.

Figure 2 shows the software components of a data grid. There are six principal layers. The storage systems where the digital components actually reside comprise layer 1. Data grids provide a standard mechanism (layer 3) for interacting with the storage systems, that is, a standard set of operations that can be performed upon the digital component(s) registered into the logical file name space. Typical operations include the ability to read and write a digital component in the storage system. When manipulating a thousand digital components, it is much faster to register and move them using bulk operations than it is to issue commands one at a time for each component. Bulk operations support the registration and movement of entire directory hierarchies. The bulk operations are also required for scalability, to ensure that the digital material that is preserved can grow to tens or hundreds of millions of digital components while providing good interactive response. The management of the bulk operations is done in layer 4, as part of the consistency controls.

The storage system access mechanism is implemented by writing a new storage system driver for each new type of storage system as shown in layer 2. Current data grid technology, such as the San Diego Supercomputer Center Storage Resource Broker (SRB),[18] is capable of storing digital components in Unix file systems (Linux, AIX, IRIX, Solaris, Mac OS X), Windows file systems, binary large objects in databases, disk/tape-based storage systems, object ring buffers (ORB), and the like. If a new storage technology becomes available, a new storage system access mechanism is written that understands how to

[17] Reagan Moore and Chaitan Baru, “Virtualization Services for Data Grids,” in *Grid Computing: Making the Global Infrastructure a Reality,* Fran Berman, Geoffrey C. Fox, and Anthony J. G. Hey, ed. (New York: John Wiley and Sons, 2003).

[18] C. Baru, R. Moore, A. Rajaseka, M. Wan, “The SDSC Storage Resource Broker,” Proc. CASCON ’98 Conference, 30 November – 3 December 1998, Toronto, Canada.

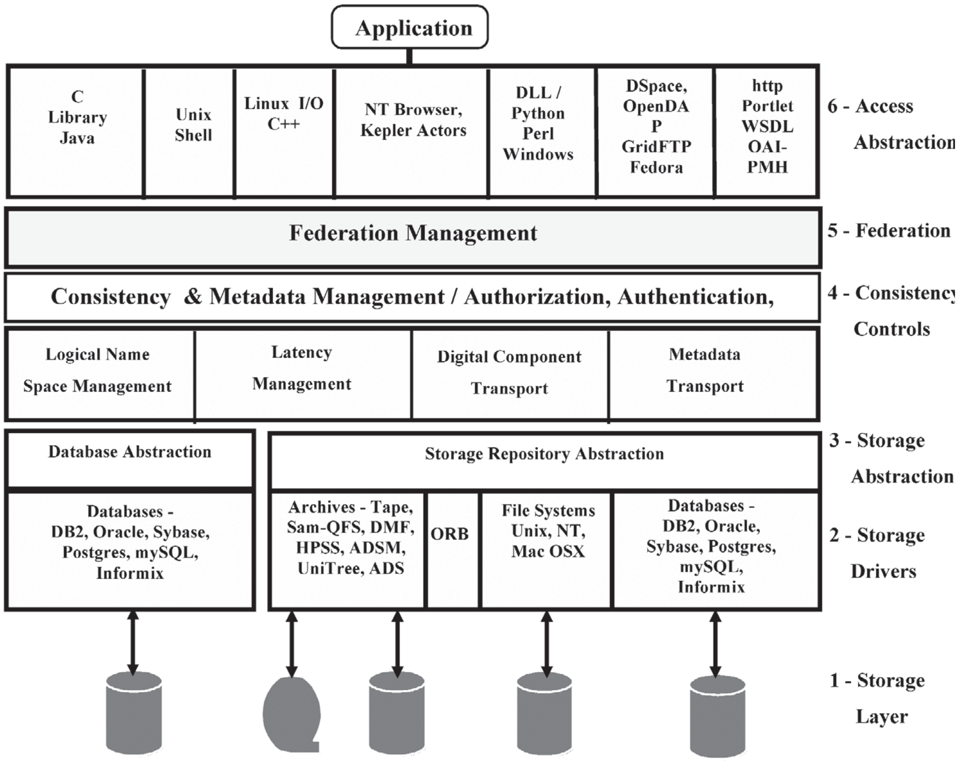

FIGURE 2. Data grid architecture

interact with the storage system and map from the standard set of operations supported by the data grid to the operations provided by the storage system.

Data grids store the logical attributes in a database. To ensure that the data grid can take advantage of new database technology, a standard database control mechanism (shown in layer 3) is used to define the set of operations required to manage sets of metadata in a database. This mechanism, called a database abstraction in Figure 2, makes it possible to use a wide variety of databases for storing the logical attributes. The data grid maps the standard operations to the access mechanism required by the particular database, whether it is Oracle, DB2, Sybase, Informix, PostgreSQL, or mySQL. The standard operations maintain each metadata set intact and preserve its link to the related record. The technology that implements the standard set of operations is called MCAT (for Metadata CATalog). It supports schema extension (mapping of new metadata onto the logical file name space), automated generation of the Standard Query Language used to specify queries to databases, import and export of metadata from XML files, bulk metadata loading, access controls on metadata, and so on.

Standard attribute names are used by the data grid to manage storage attributes across the multiple types of storage systems. This means that the records metadata name representing the owner of a digital component is the same across storage systems. It is independent of whether the digital component was stored as a binary large object in a database or as a file in a file system. In turn this means that queries on the logical attributes will return uniform results across all types of storage systems.

Layer 4 is the heart of the data grid, where consistency and constraint policies are implemented to guarantee that storage, records, and preservation metadata can be preserved and linked to each digital component. This level is also able to support an archival descriptive system that includes metadata related to the various levels of aggregation of the records. The ability of data grids to update the logical attributes after any preservation activity is ensured by having the data grid "own" the digital components managed by the data grid. In typical storage systems, ownership is asserted by storing data under an account identifier or Unix user ID. Only the person who can authenticate himself or herself to the storage system for that Unix user ID can access the digital components.

Data grids manage ownership by storing all data under an account identifier that is created for the data grid itself. Only the data grid has the ability to authenticate access to digital components in the remote storage system under this account. This means that all accesses to the digital components must be done through the data grid software. Consequently the data grid can track all operations that are performed, maintain audit trails of all accesses, and automatically update the storage location metadata when a digital component is moved to another location.

Data grids support the concept of multiple access roles. A trusted custodian can be assigned access privileges that permit the execution of preservation processes, while the public is only given the permission to read selected digital components. The access permissions can be set separately for each digital component. Access permissions can also be set for each set of logical attributes. For example, the public may be given permission to read the digital components and the records metadata, but not the audit trail information.

An implication is that the data grid must provide access controls for each digital component that is registered. Since the data grids manage the logical name spaces for users as well as digital components and logical attributes, the data grid can implement access controls that are permanent over time. Data grids effectively build a characterization/representation of the archival aggregation that is managed independently of the choice of storage technology. The archival aggregation is the logical arrangement of many records through use of the logical file name space.

The four lower software layers manage the digital components. The upper two software layers manage access to the preservation environment. New access mechanisms that become available over time can be interfaced to data grids through a standard set of operations that encompass all the activities that trusted custodians and users might wish or need to carry out. Regardless of the choice of activity, the data grid guarantees that the records and their metadata will remain as accurate and authentic as they were when acquired.

The standard set of operations for interacting with storage repositories includes opening files, reading files, writing files, closing files, and manipulating the location of the files. The file manipulation operations are usually implemented in standard access interfaces including C library calls, Unix shell commands, and Java classes as shown in layer 6. Applications use C library calls to read and write data. Interactive support for reading and writing data is done through Unix shell commands. Web-based access is typically done through Java applets that can be downloaded to a Web browser. These standard interfaces are then used to interface a preferred access mechanism to the data grid. An example is the recent emergence of the Web Services Resource Framework (WSRF) for using Web services to access archived data. The WSRF environment is implemented using Java classes. Once the WSRF environment is interfaced to the standard operations, it can then be used to access data in any of the storage repositories, including legacy systems that were acquired before the WSRF environment was developed. In fact, this is one of the main advantages of data grids: they make it possible for modern access methods to be used on legacy storage systems.

Layer 5 of the data grid environment deserves explanation. Data grids require the careful management of logical attributes in a central database, which is called the *data grid registry*. This is both a strength (consistent control can be kept on all logical attributes) and a weakness (the data grid registry represents a single source of failure and also can be a bottleneck for high level of use). Data grids use a “federation” of data grid registries to overcome these limitations. A federation makes it possible to replicate logical attributes across multiple data grid registries. In practice, each data grid and its data grid registry comprise an independent data management environment. A data grid federation is the controlled sharing of the logical name spaces among data grids, with the specification of consistency mechanisms on the update of the storage and preservation attributes. Consider the federation of three independent data grids, with the first data grid devoted to a preservation environment controlled by a trusted custodian, the second data grid devoted to public access, and the third data grid devoted to a “dark” archive that is not accessible by the public. The trusted custodian can choose which records will be replicated into the public data grid. It is possible to build multiple public repositories, in which selected records and related metadata are reproduced into additional data grids. Such an environment can be used to support public access by multiple simultaneous users, avoiding bottlenecks. The reproduction of the storage attributes assures the ability to access digital components based upon queries to any of the public data grids.

A security data grid (dark archive) can be implemented that reproduces all of the logical attributes and digital components, while restricting any access by the public. This is a data grid in which access is limited to trusted custodians, where only the trusted custodian names are recognized and no storage system is shared with other data grids. In effect, the security data grid provides a higher degree of assurance that the risk of loss of digital components can be minimized. In the security data grid, the original authentic copy of records may be stored with the records' metadata. These authentic records allow the verification of the authenticity of the archival aggregation across the federated data grids over time as technology evolves.

Process for Accessioning Records into Data Grids

When a trusted custodian acquires digital records destined for continued preservation, the following approaches can be used to import the records into a data grid:

- The trusted custodian who has data grid technology installed on his or her storage system can issue a data grid command to register the records of the creator from the original recordkeeping system into the data grid registry. The registration procedure preserves both the names and organizational structure given to the records by the creator. Data grids also support contemporary registration of records metadata. The records metadata can be extracted from the recordkeeping system if they are inextricably linked to the record. If they are not, and they exist in an XML file, they can also be registered into the data grid registry.
- The trusted custodian can receive the transfer of the records from the creator on his own storage system in a staging area. The trusted custodian may maintain the original records names and organizational structure. The custodian may appraise the records and define the subset that will be archived and the standard encoding format that will be used. The registration procedure can then follow the prior approach, and the trusted custodian may register the material into the data grid registry. The use of a staging area is most appropriate in cases where the creator does not hold the records in a trusted recordkeeping system, but in a variety of record-making applications. It is also recommended for the preservation of Web sites, because the records need to be extracted from the Web site, relinked, and organized before being registered in a data grid registry. The issue of authenticity must be addressed. If a Web page is relinked so it will point to other Web pages within the preservation

environment instead of the original site, the process used to do the relinking should be documented. Relinking of crawled Web pages is an example of the extraction of a digital record from the environment in which it was created. Note that the relinking can be applied each time the Web page is accessed or can be done once and saved.

- If the trusted custodian is dealing with more than a thousand records, then there is the challenge of uploading time. In this case, a single command can be issued to load into the data grid an entire directory either from the creator's system or from the custodian's own storage system. The data grid manages the aggregation of the files in the directory before transmission, manages the movement over the network, and then disaggregates the files for storage in the data grid. Contemporaneously, the data grid aggregates the storage attributes, moves them over the network as a single file, and then bulk loads the storage attributes into the data grid registry using parallel bulk database load commands.

Whichever approach is chosen, the records and their metadata must be verified after transfer into the data grid. Before the records are moved over the network, a checksum is computed for each record and related metadata. After the checksum and record have been transmitted, a verification occurs of the checksum against the record received in the data grid. If the correspondence is perfect, then the record and the checksum are registered in the data grid registry, and the latter is included among the storage attributes of the record. The checksum can be verified at any point in the future to detect possible corruption of the record from media or software system failure.

Most storage systems have a limited ability to manage small files, because they can usually manage at the most twenty million names before their performance starts to degrade. Thus if the trusted custodian needs to store a large number of records with sizes less than 30 megabytes, he or she can aggregate them before storing them through use of data grid containers. This minimizes the number of files seen by the storage system to the number of physical containers that are created by the data grid.

If the encoding format of the records that are being acquired by the trusted custodian is in danger of becoming obsolete, a transformative migration to a custodian-selected current format may be required before or after registration into the data grid. Since data grids can manage multiple versions of digital components, both the obsolete and current versions of the record can be stored. The purpose of doing so would be to maintain the ability to do the next transformation from the original and to maintain authenticity. However, if the encoding format of the original is completely obsolete, this could be an expensive and time-consuming endeavor.

Preservation Environments Based on Data Grids

Data grid technology is being used to support research projects on preservation for communities ranging from state archives in the United States to the National Archives and Records Administration, and actual preservation activities for communities ranging from the multicampus California Digital Library, to the National Science Foundation–National Science Digital Library. Data grid technology is also used to manage scientific data collections that have millions of files and hundreds of terabytes of data. Table 1 lists only the aggregations of digital material housed at the San Diego Supercomputer Center, sorted into three major categories: data sharing environments, data publication environments, and records preservation environments. All three categories are implemented on top of data grid technology.

The list of projects spans preservation environments that manage small archival aggregations (a few thousand records whose total size is a gigabyte) to digital libraries that have tens of millions of files whose total size is greater than 100 terabytes. The number of trusted custodians (persons with access controls or ACLs) is on the order of twenty to one hundred people for a given data management system. Across the listed projects, the SRB data grid at SDSC manages over 600 terabytes of data and over a hundred million files. The SDSC data grid uses an Oracle database for the data grid registry, a Sun F15k server to support the Oracle database and SRB servers, an IBM High Performance Storage System (HPSS) and Sun Sam-QFS file system to manage files written to tape, and grid brick technology to provide on-line disk caches for interactive access to stored records. *Grid bricks* are hardware systems built from inexpensive commodity disks used in personal computers, at a cost of $1,200 per terabyte of storage. The grid bricks are managed by the SRB data grid, which makes it possible to expand the storage capacity by adding bricks as needed. Each grid brick is treated as a separate storage system. A single logical storage name is used to reference all of the grid bricks. The data grid distributes records uniformly across the grid bricks whenever a record is written to the grid brick logical resource name.

SDSC makes extensive use of containers to aggregate small files before storage in the permanent archives. The National Science Digital Library permanent archives uses containers to aggregate digital records retrieved from Web sites. Since the average size of the retrieved records is 100 kilobytes, about three thousand records are aggregated into a 300 megabyte container before storage. The internal URL links between the Web pages are replaced with SRB data grid logical file names. Records previously stored on a Web site can be retrieved from the permanent archives through a Web browser, with the data grid automatically traversing the internal links between the Web pages.

The NARA research prototype permanent archives is a collaboration between NARA, SDSC, and the University of Maryland. Each site runs an independent data grid with a separate data grid registry. The records and logical attributes are reproduced on the sites at NARA and the University of Maryland. The site at SDSC serves as the security data grid while the one at NARA serves as the public data grid. The risk of data loss is mitigated by having at least three copies of each record and three copies of the logical attributes

Table 1 Data Grid Collections Managed at SDSC

	GBs of data	1000s	Users with
Storage Resource Broker (SRB) Data Grids at SDSC (2 Feb. 2006)	stored	of files	ACLs
Data Sharing Environments			
NSF/ITR—National Virtual Observatory[19]	93,252	11,189	100
NSF—National Partnership for Advanced Computational Infrastructure[20]	34,452	7,235	380
Hayden Planetarium—visualizations of the evolution of the solar system	8,013	161	227
NSF/NPACI—Joint Center for Structural Genomics[21]	15,703	1,666	55
NSF/NPACI—Biology and Environmental collections	104,494	131	67
NSF—TeraGrid, ENZO Cosmology simulations	195,012	4,071	3,267
NIH—Biomedical Informatics Research Network[22]	13,597	13,329	351
Data Publication Environments			
NLM—Digital Embryo image collection[23]	720	45,365	23
NSF/NPACI—Long Term Ecological Reserve	236	34	36
NSF/NPACI—Grid Portal	2,620	53	460
NIH - Alliance for Cell Signaling microarray data[24]	733	94	21
NSF—National Science Digital Library SIO Explorer collection[25]	2,452	1,068	27
NSF/NPACI —Transana education research video collection[26]	92	2,387	26
NSF/ITR—Southern California Earthquake Center[27]	153,159	3,229	73
Records Preservation Environments			
UCSD Libraries image collection	190	208	29
NARA—Research Prototype permanent archives[28]	2,703	1,906	58
NSF—National Science Digital Library permanent archives[29]	5,205	50,586	136
NHPRC Persistent Archives Testbed	101	474	28
TOTAL	655 TB	106 million	5,383

[19] National Virtual Observatory (NVO), http://www.us-vo.org/.

[20] NPACI, National Partnership for Advanced Computational Infrastructure Data Intensive Computing Environment thrust area, http://www.npaci.edu/DICE/.

[21] JCSG—Joint Center for Structural Genomics, http://www.jcsg.org/.

[22] Biomedical Informatics Research Network, http://nbirn.net/.

[23] Visible Embryo Project, "Human Embryology Digital Library and Collaboratory Support Tools," part of the Next Generation Internet Initiative and funded by the National Library of Medicine, http://netlab.gmu.edu/visembryo.htm.

[24] Alliance for Cell Signaling, http://www.signaling-gateway.org

[25] SIO Explorer Digital Library Project to provide educational material from oceanographic voyages in collaboration with NSDL, http://nsdl.sdsc.edu/.

[26] Transana is an education research tool for the transcription and qualitative analysis of audio and video data, http://www.transana.org/.

[27] Southern California Earthquake Center, http://www.scec.org/.

[28] NARA Persistent Archives project, http://www.sdsc.edu/NARA/.

[29] National Science Digital Library (NSDL), http://www.nsdl.org/.

reproduced among the three sites. The goal is to handle media corruption by accessing a copy of the records in one administrative domain, handle risk of operational error by accessing a copy within a second data grid, handle risk of system software error by accessing a copy stored on a different type of storage system, handle risk of natural disaster by accessing a copy on the other side of the continent, and handle risk of malicious users by maintaining a copy in a security data grid. Even though five types of risk are specified, the five types of copies can be three actual copies stored in the three sites. For example, the University of Maryland provides both a remote site and a different type of storage system.

The UCSD Libraries image collection is actually housed on grid bricks maintained by the UCSD Libraries. A copy of the images is reproduced under data grid control onto the HPSS archival tape storage system at SDSC to mitigate against risk of data loss.

The multiple projects using the data grid technology at SDSC are a small subset of the total number of projects worldwide that are using data grid technology. A SRB data grid is being run independently of SDSC by the Biomedical Informatics Research Network (BIRN) that is sharing data between National Institute of Health research projects. The BIRN data grid shares data among seventeen university and hospital sites, under the HIPAA[30] patient confidentiality management policy. The National Partnership for Advanced Computational Infrastructure is an even larger data grid that manages files distributed across eighty-seven storage systems. The largest data grids are international in scope. The high-energy physics BaBar[31] experiment manages experimental data that are distributed between Stanford and Lyon, France, in BaBar's own data grid. They are using data grid federation technology to link two independent data grids, one located in California, and the other in France.

The multiple projects include very small data grids that manage a few thousand records and a few gigabytes at a single site. The complete SRB data grid including a data grid registry implemented on a Postgres database can be installed on a Mac or Linux laptop in about eighteen minutes. The installation includes the Postgres database, a SRB server for interacting with the laptop file system, a SRB data grid registry that manages the logical attributes housed in the Postgres database, and the SRB clients. A typical use of such a configuration is to manage records that are being made accessible through a Web browser.

The data grid technology makes it very easy to share storage systems and databases. For the projects listed in Table 1, the three approaches listed in Table 2 are used.

[30] HIPAA, Health Insurance Portability and Accountability Act of 1996, http://www.hep-c-alert.org/links/hippa.html.

[31] BaBar is a B meson detection system at Stanford University, http://www.slac.stanford.edu/BFROOT/.

Table 2 Approaches for Data Grid Implementation

Project	Database instance	SDSC storage	Remote project storage
NARA permanent archives	Separate Oracle database	Grid Brick, HPSS	Separate data grid running Informix
NPACI users	Separate Oracle instance	Sun Sam-QFS	None
NHPRC Persistent Archives Testbed	Shared Oracle instance	Grid Brick, Sun Sam-QFS	Grid Brick

Some projects install separate sets of logical attributes that are managed within the same database in a shared hierarchy for the data grid registry. Other projects have separate sets of attributes that are managed as distinct database instances within one database, and still other projects use their own, entirely separate database. Some projects rely on SDSC for all of the storage, in effect outsourcing storage, other projects combine use of their own storage system with use of SDSC storage, and still other projects rely on only their local storage systems. An implication is that the data grid technology that is used to implement a preservation environment can be completely outsourced (both the management of the data grid registry and the data grid storage), partially outsourced (with local storage systems or a local database being used), or run entirely within the preserving organization.

Preservation Environment Operation

If a project decides to run an independent data grid, care must be taken that the data grid system is adequately managed. A data grid is an additional level of software that requires administrative support. Data grids integrate networks, storage systems, and databases into a uniform data management system. Not surprisingly, all of these components must be managed. For small archival aggregations, a complete data grid can be installed on a single laptop, and its management can be done by a single person. For large systems, managing millions of records, the following administrative support functions are needed:

- **Storage system administrator.** SDSC typically uses two full-time staff to manage disk/tape storage systems that currently house 2 petabytes of data, equivalent to about 40 million files. The administrator installs new versions of the storage software, manages the storage systems, and responds to storage failures (e.g., corrupt media).
- **Database administrator.** SDSC typically uses one full-time staff to manage twenty Oracle database instances. The administrator maintains backups of the databases, installs new versions of the software, and tracks database performance (e.g., table indices, CPU utilization).

- **Data grid administrator.** SDSC typically uses one full-time staff to manage twenty data grids. The activities include installing SRB servers, maintaining the logical name spaces for storage, users, files, and attributes, adding new resources and new users, responding to questions, and managing problems (network outages, storage system outages, server failures).
- **Network administrator.** SDSC typically uses two full-time staff to manage the network connections to the external world and the internal networks. Activities include managing the network Domain Name Service, managing network routers, and maintaining firewalls.
- **Security administrator.** SDSC typically uses one full-time staff to track security incidents, maintain reference systems, and monitor usage of the systems.

While the list of activities appears daunting, most of it comprises the skill set that is needed to manage current storage environments. Thus, an archival program of a living organization, such as a city or bank, would already have most of these functions provided for and would only need an additional person to be the data grid administrator. For small programs, the data grid administrator doubles as the database administrator, since many of the activities require similar expertise.

The Storage Resource Broker data grid software is mature technology that has been in development for more than eight years. The software represents about seventy-five person years of development and application support. The current release, SRB version 3.4.0, is available for academic use at universities and nonprofit organizations.[32] A commercial version is marketed by Nirvana Storage.

Conclusion

Data grids provide generic data management infrastructure that can be used to implement records preservation environments as well as other distributed data management systems. Data grids excel at managing reproduction of records across multiple sites, essential for mitigating against risk of record loss. Federations of data grids provide the essential capabilities needed to implement high-security environments for preservation of authentic copies. Data grids are being used today to manage aggregations that total hundreds of terabytes of data and tens of millions of files. Data grids enable the incorporation of advances in technology, ensuring that preservation environments can take advantage of more cost-effective technology as it becomes available.

[32] SRB—"The Storage Resource Broker Web Page, http://www.sdsc.edu/srb/.