In: Proceedings of the 16th International Conference on Database Systems for Advanced Applications (DASFAA 2011), Hong Kong, April 2011.

The DISTARNET Approach to Reliable Autonomic Long-Term Digital Preservation

Ivan Subotic[1] Heiko Schuldt[2] Lukas Rosenthaler[1]

[1]Imaging & Media Lab [2]Databases and Information Systems Group
University of Basel, Switzerland
firstname.lastname@unibas.ch

Abstract. The rapidly growing production of digital data, together with their increasing importance and essential demands for their longevity, urgently require systems that provide reliable long-term preservation of digital objects. Most importantly, these systems have to ensure guaranteed availability over a long period of time, as well as integrity and authenticity of the preserved data and their metadata. This means that all kinds of technical problems need to be reliably handled and that the evolution of data formats is supported. At the same time, systems need to scale with the volume of data to be archived. In this paper, we present DISTARNET, a fully distributed system that reliably executes pre-defined workflows for long-term preservation. Moreover, DISTARNET is designed as an open system that allows the curators of digital objects to specify new processes to cope with additional challenges.

1 Introduction

Digital data, either digitized or digital born, are increasingly gaining importance in our everyday life. As a consequence, a large spectrum of applications require that data is preserved over long periods of time — up to several years due to legal constraints in business applications or for scientific data, for the duration of the lifetime of a human in medical applications, up to potentially unlimited time spans for the preservation of cultural heritage. Approaches to digital long-term preservation are constrained by the enormous and ever growing volumes of data. In addition, long-term data archiving and preservation also needs to take into account that data has to outlive the hardware on which they are stored and the data formats in which they are represented.

Metadata is the key to providing long-term digital preservation. We will use the term *Information Object* to denote the digital data (bit-stream) and the corresponding representation information, or some other kind of metadata.

Further, the Open Archival Information System (OAIS) reference model [1] also explicitly considers the migration of digital data to new data carriers and/or data formats in periodic intervals in order to escape from the technological obsolescence of specific hardware and software products.

In short, digital long-term preservation can be defined as the task of preserving information objects, despite potential changes of the formats in which objects are stored and the underlying hardware environment. Therefore, a software

system for digital long-term preservation has to support preservation processes that guarantee i.) *integrity*: the information captured by data is not altered in any way; ii.) *authenticity*: provenance information is properly linked to each stored object by means of appropriate metadata; iii.) *chain of custody*: location and management controls are tracked within the preservation environment; iv.) *completeness*: everything that was initially stored is also available in the future and finally v.) *ease of access*: the necessary means are provided to find and identify the stored digital objects. Moreover, an essential requirement for viable long-term preservation systems is their capability to do the necessary maintenance and failure recovery in an *autonomous* way, e.g., to automatically identify when a pre-defined replication level is no longer reached and to trigger corrective actions (deploy new replicas) without human intervention.

To cope with all these challenges, we are currently developing DISTARNET (DIstributed STorage ARchival NETwork), a system for digital long-term preservation of information objects as required by archives, museums, research communities or the corporate sector. The goal is to provide a software system based on which deploying institutions can, on their own or through collaboration with others, build an autonomous, reliable, replicated, geographically distributed, Internet-based digital archiving system. Maintenance and recovery from software or hardware failures is handled in DISTARNET by means of dedicated processes that are automatically executed in a reliable way. Some of these processes encompass the necessary steps for format conversions, while other processes address failure recovery, for instance, by deploying new replicas or by migrating content from an unreliable host to a more stable one.

In the following we briefly sketch two use case scenarios that highlight the broad applicability of DISTARNET.

Scenario 1: Jim, a digital archivist at the National Museum of History & Native Art in a small European country wants to implement a new archiving solution for preserving his country's cultural heritage. This new solution should enable him to have redundant off-site replicas, although the museum itself has only one site available that can be used to deploy such a solution. However, there is a collaboration agreement between the national museums of different countries that includes access to the other institutions' computation and storage resources for deploying replicas, together with the enforcement of access restrictions on these shared data (pretty much like in a virtual organization known in the context of grid computing). Each museum deploys a DISTARNET node. When data is ingested, they will be handled according to the policies specified by their owner and will automatically be distributed across the storage resources of different museums. DISTARNET will also periodically instantiate maintenance processes and launch recovery processes when necessary.

Scenario 2: The cloud storage provider Stratocumulus Inc. plans to release DA^3S (Data Archiving as a Service), a new data management service. Essentially, DA^3S offers customers the option for long-term digital preservation of their digital assets, with dedicated quality of service guarantees on data availability (replication), integrity (regular checks) and authenticity. For this, Stratocumu-

lus Inc. installs DISTARNET on each of their data-centers. In this setting, DISTARNET will be a layer underneath the cloud, so that Stratocumulus Inc. can provide the services offered by DISTARNET fully transparent to their end users. As optional services for DA^{3}S for an extra cost, Stratocumulus Inc. offers automated data format migration. Moreover, storage location constraints can be defined through preservation policies that will allow to restrict where data will be stored since Stratocumulus Inc. runs multiple data-centers around the world.

This paper introduces the flexible and reliable DISTARNET approach to digital preservation, and in particular its metadata model and the processes that are pre-defined for maintenance and failure handling purposes. We present the challenges and risks that need to be addressed by a system for long-term digital preservation. The main contribution of the paper is the detailed analysis of DISTARNET's self-$*$ capabilities, i.e., how the system is able to automatically adjust itself to changing environments (both in terms of volumes of digital content to be archived and the available storage resources) and how it automatically recovers from different kinds of failures. The autonomic behavior also includes the compliance to quality of service guarantees for the DISTARNET users (digital archivists) such as a predefined level of data availability or specific constraints on the locality of data and replica placement.

The remainder of this paper is organized as follows. In Section 2, we summarize the challenges of digital long-term preservation. Section 3 introduces DISTARNET, in particular its metadata model and processes. In Section 4 we analyze how these processes can take care of maintenance and failure recovery in digital preservation. Section 5 discusses related work, Section 6 concludes.

2 Distributed Digital Long-Term Preservation: Challenges

The challenges that distributed digital long-term preservation systems are faced with are fault tolerance, scalability, load balancing, and security in addition to integrity of complex information objects, authenticity, data format obsolescence, long-term readability, and ease of access.

Fault Tolerance and Failure Management. The failure of one or more components in a distributed preservation system should not endanger the whole system, and should only have isolated effects. Failure or disaster situations resulting in destruction or corruption of some of the stored information objects should not lead to a complete loss of the archived data. Automated replication mechanisms should maintain a minimum number of geographically dispersed replicas (number and location defined by preservation policies) of the stored information objects. Any data loss event should trigger automated recovery processes that will reestablish the minimum number of geographically dispersed replicas. This should be done by either using the repaired failed storage nodes, or by using other available and suitable storage nodes found through resource discovery.

Management of Complex Information Objects. The long-term preservation of digital data requires the management of complex information objects, i.e. information objects that are comprised of or are part of other information objects.

The challenge lies in the automated management of such complex objects in a distributed setting. Preserving the integrity of complex objects is a twofold problem. First, the integrity of the referential information needs to be maintained, and second, the integrity of the objects themselves. Referential and object integrity checking needs to be automated. Any loss of integrity needs to trigger automated processes that will restore the integrity of the information object. If the information object cannot be repaired solely by the information it carries itself, other remote replicas need to be used. Besides, preserving integrity is an important challenge when information objects evolve (e.g., annotations or collection/subcollection information).

Scalability. The growing production of digital data that needs to be archived requires a scalable distributed preservation system that should work efficiently even with an increasing number of users and rapidly growing volumes of data that need to be stored. The addition of storage resources should enhance the performance of the system. This requires that the processes supporting the archiving operations be automated and scalable themselves.

Openness and Extensibility. A long-term preservation system should provide clearly separated and publicly available interfaces to enable easy extensions to existing components and the possibility of adding new components. The system should be able to be adapted to arising new challenges, by allowing curators of digital objects to specify new processes to cope with additional challenges.

Resource Discovery, Load Balancing and Remote Execution. In a distributed preservation system, the discovery of newly available resources, together with the monitoring and management of existing resources is very important and should be handled efficiently. The information gathered is important for the functioning of processes that provide automated replication of the information objects to suitable remote storage nodes, constrained through preservation policies. The system should be able to distribute the replicas among the available resources for improving performance based on availability, access speed, higher security, and/or reliability. Dynamically incorporating new resources or correctly handling the loss of existing resources (temporarily or permanently) should also be provided via automated processes with transactional semantics.

Security. Access to resources should be secured to ensure only known users are able to perform allowed operations. In addition, in a distributed where different institutions cooperate and share storage resources, only the data owning institution should be able to access and manage its data. However, cooperating institutions should be able to access other institution's meta-data and be granted access to the content of interest after having been authorized by the data owner.

3 DISTARNET

DISTARNET is a fully distributed system consisting of a network of collaborating nodes. The DISTARNET network consists of nodes organized together into virtual organizations (VO) [3]. The structure of the network inside the VOs is in itself organized in a P2P fashion, and thus creates what we call a *P2P-VO*. A

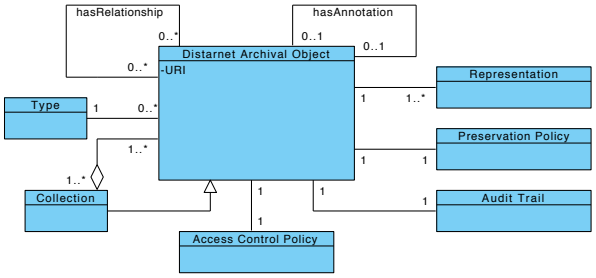

Fig. 1. Logical Data Model for DISTARNET using UML notation

DISTARNET node can be part of one or more P2P-VOs. Resources provided by nodes within a P2P-VO can only be accessed by other member nodes. Within a P2P-VO access restriction management can be used to define the allowed access characteristics to the stored content. The discovery of new resources, monitoring and management of existing resources will be done in a P2P fashion.

3.1 DISTARNET Data Model

In the context of DISTARNET the term *DISTARNET Archival Object* (DAO) will be used to denote a container holding an *Information Object* consisting of a Data Object (e.g., image, audio/video, text document, etc.) and the corresponding representation information, or some other kind of metadata. This metadata can provide additional descriptions for an information object (e.g., annotations), the descriptions of relationships between information objects (links) or information about collection/subcollection sets. Furthermore, also metadata on the object's storage and deployment can also be part of the DAO.

DISTARNET distinguishes between mutable and immutable content [12]: First, the read-only digital objects that are to be archived (e.g., images, audio/video, text documents, etc.) are considered immutable, i.e., thex cannot be modified once created. Second, the metadata of the archived digital objects is usually mutable and may exist in several versions and which can be modified (e.g., annotations pertaining to some archived digital object).

The data model used in DISTARNET allows the archiving of complex data objects. Figure 1 shows the DISTARNET logical data model in UML. In DISTARNET, every container stores one information object characterized by its type. To represent for example an annotation for an archived image, we will create a DISTARNET Archival Object of the corresponding type which will contain the annotation and make a link to the DAO containing the image – note that an annotation can be anything from text to a full-fledged DAO. The container storing the information objects corresponds to the Archival Information Package (AIP) described in the OAIS Reference Model. Although information such as annotations which are generated over time are not the original data objects that where archived, they are treated equally since they provide additional de-

scription information and need to be preserved as such alongside the originally archived data objects.

3.2 DISTARNET Processes

The goal of DISTARNET is to provide dynamic replication, automated consistency checks, and recovery of the archived digital objects utilizing autonomic behavior and predefined processes, governed by preservation policies without any centralized coordinator in a fully distributed network. Rather, the system provides two distributed repositories, namely a distributed *Replica Location Repository* that stores the degree of replication and the replica locations for each DISTARNET DAO, and a distributed *Node Information Repository* with relevant metadata on the participating nodes in a P2P-VO. Based on the information stored in these repositories, DISTARNET provides several processes that exhibit self-* properties needed for automatically dealing with the challenges of long-term digital preservation.

Self-Configuration

In DISTARNET self-configuration manifests itself in the ability of the system to automatically detect changes in the network. Events such as new nodes joining or nodes leaving are being constantly monitored and taken into account.

Node Joining Process (NJP). A node joins the network after the node credentials are configured and a number of seed nodes are added to the Neighbor Node List. The NJP then contacts the seed nodes and starts to gather information about other nodes in the network (e.g., populate the Node Information Repository). The Neighbor Node List is a dynamic list, which will contain a few *near* nodes defined as such by the P2P overlay metric.

Periodic Neighbor-Node Checking Process (PNCP). Every node will check periodically its neighbors (from the Neighbor Node List) by sending a message to which the receiver has to reply in a defined time. If the receiver does not reply, this node is marked, after some defined period of time, as lost. After that, the system will begin with the self-healing behavior.

Automated Dynamic Replication Process (ADRP). The ADRP is responsible for finding suitable storage nodes, estimating the optimal number of replicas and initiating the creation of replicas. For this, the ADRP will find – using the Node Information Repository – suitable geographically dispersed nodes for storing the replicas by taking into account possible policy-based geographical restrictions. The system will estimate the optimal number of replicas needed by taking into account the availability of nodes (based on statistics on the individual availability collected in the past) used to store a DAO and via the preservation policy imposed availability threshold of the DAO itself. This estimate will be used to raise the number of replicas if needed. To optimize the access performance the system will create if necessary additional replicas by analyzing the usage patterns of the digital objects. After evaluating a DAO regarding its overall availability in the network, ADRP will initiate if needed the *Reliable Copying Process* (RCP)

and create new replicas. The reliable copying process is a BitTorrent-like transfer mode that uses existing replicas in the network for creating new ones in a secure and efficient manner. At process level, transactional semantics according to the model of transactional processes [8] will be applied.

Self-Healing

Due to the continuous monitoring of nodes, the DISTARNET system will detect abnormal conditions or problems that may harm its proper functioning (e.g., in the case of a *Node-Lost Event* or a *Corrupted DAO Event*) and it will be able to automatically recover from the following situations.

Node-Lost Event. The system will automatically react and initiate countermeasures by reevaluating the DAOs affected by the disappeared node by the ADRP and if needed create new replicas so that the policy-defined redundancy and availability requirements are upheld again.

Corrupted DAO Event. Periodic integrity checks is done by the *Periodic Integrity Checking Process (PICP)* and if integrity is breached, will automatically trigger countermeasures like finding healthy replicas in the network and by using the reliable copying process to copy them in place of the corrupted DAOs to rectify the problem.

Obsolete Data-Format. Data formats of the archived data objects are constantly monitored and warnings are issued if a given data format is becoming obsolete. The *Data-Format Migration Process (DMPP)* can be used to automatically migrate data formats by following a predefined migration path.

Self-Learning

All the mentioned properties until now can only be provided if the system has the needed information on which it can act upon. As a consequence, the DISTARNET system must know its environment, especially the available resources, and track their changes over time. This knowledge will be continuously gathered and disseminated throughout the network and be used to autonomously manage and maintain resource allocation through ADRP (e.g., finding suitable nodes where data can be replicated to, automatic policy-based geographical distribution of data, etc.) and other processes needed for the operation of DISTARNET. ADRP and PICP are triggered periodically. The parameters that trigger these processes will be adapted dynamically by the system. They will be prolonged in the case that for a longer period of time there where no changes in the network, or shortened if there where recent changes.

4 DISTARNET Maintenance and Recovery

In what follows, we analyze how DISTARNET addresses the challenges introduced in Section 2. which can be broadly categorized in issues involving *Maintenance* and *Recovery* processes.

4.1 Maintenance

DISTARNET enforces *referential integrity* (e.g., links between DAOs, collection/subcollection information) of its complex information objects using PICP.

Authenticity, *chain of custody* and *completeness* of the archived objects are supported by the data model and supporting processes through maintaining an *audit trail*. DISTARNET processes create automatically an audit record for every operation done on an DAO, with details about who, what, when, where and why the operation was executed.

The dynamic and autonomic nature of DISTARNET encounters *hardware and software obsolescence* that at the bitstream level endanger the DAOs by means of automated "media" migration, by allowing to simply turn off the old hardware and turn on the new hardware.

The *interpretability* of the logical representation of the DAOs is guarantied by processes providing automated data format migration and is supported by the data model and the collection of extended format descriptions by using format identification and characterization tools.

Scalability, *resource discovery* and *load balancing* are supported by a fully distributed design of the network, adaptability of the triggers to lower the overall load on the nodes caused by the maintenance processes, and by the ADRP which also optimizes the usage of the storage resources provided in the network

Openness and *extensibility* are very important attributes which DISTARNET supports by allowing the curators to define new process, and by providing a flexible data model that can adapt to new needs that can arise in the future (e.g., in case of new data formats, new metadata standards).

4.2 Recovery

Due to the continuous monitoring of nodes, the DISTARNET system will detect abnormal conditions or problems that may harm its proper functioning and it will be able to automatically recover from those situations, again by means of predefined processes.

Hardware problems caused by failure (e.g., power failure, hardware failure, etc) or disaster (e.g., natural disaster, fire) can result in destruction of the whole node, or in destruction or corruption of the stored DAO. The loss of a node is discovered through the *PNCP* which triggers a *node-lost event* after some predefined period of time, because at first a *network problem* is assumed. In the case of a ode-lost event, this information is stored in the *Node Information Repository*, the *Replica Location Repository* is updated, and both are propagated throughout the network. Subsequently, the ADRP is started for the DAOs that are affected (network wide redundancy) by the lost node.

Network problems caused by lost network connection or an intermittent network connection are detected through the *PNCP*. For a lost connection to a node to be classified as a network problem, upon the return of a node it has to be verified that the node was running, only the network was down, and all DAOs are accounted for. Such discovered network problems trigger an entry into the

Node Information Repository and are taken into consideration (e.g., reliability of a node) by the ADRP.

Problems with the *content* of the archive caused through the corruption (e.g., hardware problems, malicious acts, etc.) of the DAOs is discovered through the *PICP*. Any detected corruption is logged correspondingly in the *Node Information Repository* where it speaks about the reliability of a specific node, in the *Replica Location Repository*, and in the audit trail of the DAO for future reference. Also subsequently the ADRP is triggered to resolve the problem.

DISTARNET processes are able to *self recover* from problems encountered during their execution. Every DISTARNET process is composed of single atomic tasks or other (sub)processes. The workflow engine responsible for the execution the processes, monitors and logs every step. In the event of failure of any step, the corresponding recovery process, either of the main process or of the subprocess is triggered. Further we divide the processes in *repeatable* and *non-repeatable* processes. The recovery of *repeatable* processes consist in the repeated execution of those processes either until the process succeeds or a predefined number of retries is reached. In the case of *non-repeatable* process, the recovery process consist of clean-up tasks to end the process in a consistent state.

5 Related Work

For the design of digital archives, two different approaches can be distinguished. First, a centralized approach which is followed by Kopal [5] and SHERPA DP [4]. This approach allows to easily control the archived content but implies rather high infrastructure and maintenance costs and the inflexibility to cope with rapidly increasing loads if they exceed the initial design. Second, a distributed approach followed by LOCKSS [7], Cheshire 3 [13], and SHAMAN [9] where the archiving infrastructure is distributed across the participants' sites. These systems have lower infrastructure and maintenance costs, the ability for virtually unlimited growth, and allow for a higher degree of availability and reliability. DISTARNET is built following the distributed approach.

Fedora Commons [6] and DSpace [11] are open source projects that allow the creation of Digital Repositories for managing digital objects. Those two initiatives have been merged in the context of DuraCloud [2] which offers both storage and access services, including content replication and monitoring services that can span multiple cloud-storage providers. The idea is that Digital Repositories using DuraCloud can expand their systems and provide long-term preservation capabilities, or individuals can use directly use DuraCloud as a long-term preservation system. DuraCloud and DISTARNET are similar as both provide processes for the management of digital objects necessary for digital long-term preservation. The difference lies in the flexibility of the DISTARNET processes that manage digital objects between DISTARNET nodes by using local storage or, if necessary, by using cloud based storage.

Hoppla [10] is an archiving solution that combines back-up and fully automated migration services for data collections in environments with limited

resources for digital preservation. Similar to Hoppla, DISTARNET also provides processes that support data format migration for the archived digital objects.

6 Conclusion and Future Work

In this paper, we have presented DISTARNET's self-healing and self-adaptation features. DISTARNET is an ongoing effort to build a long-term digital preservation system that will provide a fully distributed, fault tolerant archiving environment with autonomic behavior governed by preservation policies. DISTARNET autonomously provides dynamic replication, automated consistency checking and recovery of the archived digital objects. As part of DISTARNET, we have developed a highly flexible data model that takes into account user generated annotations or collections and arbitrary links between objects. Based on that, we have specified sophisticated management processes that are triggered automatically when a violation of one of the preservation policies is detected.

In our future work, we plan to run a series of test cases that will simulate the real world usage of the DISTARNET network and evaluate the autonomic behavior of the system. A further emphasis will be put on testing the scalability and performance of the system under load.

References

1. CCSDS: Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1, Consultative Committee for Space Data Systems
2. Duracloud. http://duraspace.org/duracloud.php
3. Foster, I., Kesselman, C., Tuecke, S.: The anatomy of the grid: Enabling scalable virtual organizations. Int'l Journal of Supercomputer Applications 15(3) (2001)
4. Knight, G.: SHERPA DP: Establishing an OAIS-Compliant Preservation Environment for Institutional Repositories. In: Digital Repositories. pp. 43–48 (2005)
5. Kopal. http://kopal.langzeitarchivierung.de/
6. Lagoze, C., Payette, S., Shin, E., Wilper, C.: Fedora: an Architecture for Complex Objects and their Relationships. Int'l Journal on Digital Libraries 6, 124–138 (2006)
7. Reich, V., Rosenthal, D.S.H.: LOCKSS (Lots Of Copies Keep Stuff Safe). The New Review of Academic Librarianship 6, 155–161 (2000)
8. Schuldt, H., Alonso, G., Beeri, C., Schek, H.J.: Atomicity and Isolation for Transactional Processes. ACM TODS 27(1), 63–116 (Mar 2002)
9. Shaman. http://shaman-ip.eu/shaman/
10. Strodl, S., Petrov, P., Greifeneder, M., Rauber, A.: Automating Logical Preservation for Small Institutions with Hoppla. In: Research and Advanced Technology for Digital Libraries, LNCS, vol. 6273, pp. 124–135. Springer (2010)
11. Tansley, R., Bass, M., Smith, M.: DSpace as an Open Archival Information System: Current Status and Future Directions. In: Research and AdvancedTechnology for Digital Libraries, LNCS, vol. 2769, pp. 446–460. Springer (2004)
12. Voicu, L.C., Schuldt, H., Akal, F., Breitbart, Y., Schek, H.J.: Re:GRIDiT - Coordinating Distributed Update Transactions on Replicated Data in the Grid. In: 10th IEEE/ACM International Conference on Grid Computing. pp. 120 –129 (2009)
13. Watry, P., Larson, R.: Cheshire 3 Framework White Paper. International Symposium on Mass Storage Systems and Technology pp. 60–64 (2005)