DIST.AR.NET – Distributed Archival Network

Simon Margulies
Imaging & Media Lab
University of Basel
Basel, Switzerland

Ivan Subotic
Imaging & Media Lab
University of Basel
Basel, Switzerland

Lukas Rosenthaler
Imaging & Media Lab
University of Basel
Basel, Switzerland

Abstract - *The growing production of digital data challenges archiving institutions with new needs for a secure preservation of the cultural heritage of our time. The main subject of the research project 'Distarnet' is to define a protocol for a distributed system for long-term preservation of digital data. The various problems of archiving digital data are analyzed and an implemented solution is presented, taking into account the special needs of archives, museums and libraries being the holders of preservation and distribution of historical source material. Therefore closest attention is paid to the preservation of source material for future scientific researches. This paper is based on the current state of the research project and describes the protocol version 0.4.*

Keywords: archiving, migration, P2P, protocol, distributed system, metadata

1 Introduction

The growing production of digital data (digital-born or digitized) challenges archiving institutions with new problems and with new needs for a secure preservation of the cultural heritage of our time.

The preservation of digital data is different from the traditional way to preserve data, because digital data itself is meaningless to the naked human eye. Without meaning data has no information. To become understandable for humans, digital data needs to be interpreted and presented by a computer system. Therefore not only the data itself needs to be preserved to guarantee a future readability, but at least also the description for its interpretation by a computer system.

To fulfill these processes archiving institutions need to apply various approaches as outlined in [1]. In summary, a successful solution for the long-term preservation of digital data can only be achieved by a combination of data-carrier migration, data-format migration, emulation and data description. Data-carrier migration and data description are the preconditions for a successful long-term preservation of digital data, because they preserve the data itself and the description needed for its future presentation and interpretation through emulation or data-format migration. Distarnet presents a solution for automated data-carrier migration and supports data description.

Its main subject is to define a protocol of a distributed system for the long-term preservation of digital data. On one hand the various problems of archiving digital data are analyzed, on the other hand a tangible, implemented and tested open-source solution will be presented. At this the special needs of archives, museums and libraries are taken into particular consideration being the holders of preservation and distribution of historical source material.

2 Distarnet and OAIS

The Open Archival Information System (OAIS) Reference Model [2] is a widely accepted and used terminology to describe the various processes involved in an archiving institution. It does so by providing frameworks and concepts that are needed for understanding the long term preservation and access of digital data. The OAIS functional model consist of various entities interacting with each other, as displayed in Figure 1.

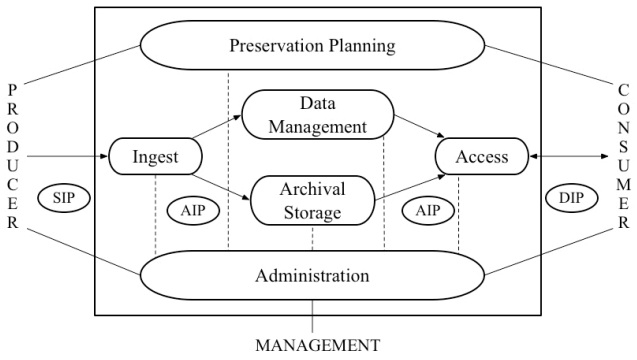

Figure 1. OAIS Functional Entities

The DISTributed ARchival NETwork, Distarnet, is a protocol for a distributed system that offers persistent storage and retrieval for digital data. It corresponds to the OAIS entity Archival Storage, that provides the services and functions for the storage, maintenance and retrieval of AIPs [Archival Information Package]. Archival Storage

functions include receiving AIPs from Ingest and adding them to permanent storage, managing the storage hierarchy, refreshing the media on which archive holdings are stored, performing routine and special error checking, providing disaster recovery capabilities, and providing AIPs to Access to fulfill orders. As depicted in Figure 2.

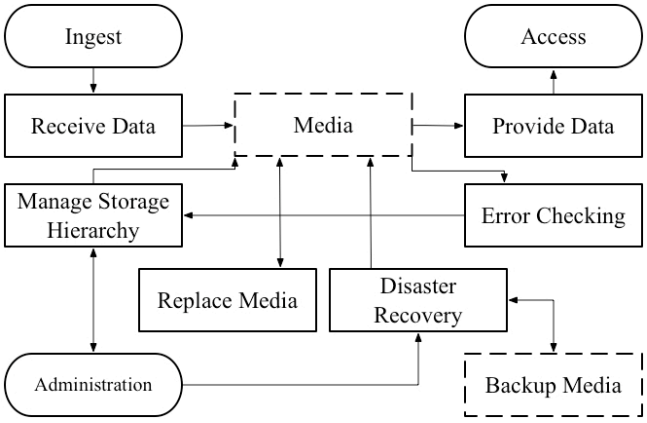

Figure 2. OAIS Functional Entity: Archival Storage

To guarantee a high security of the AIPs, the risk of data loss must be minimized and the independence of the AIPs to technological developments must be ensured. Therefore Distarnet supports various levels of data description (Metadata) to make future steps of archiving like emulation or data-format migration, possible. These processes remain external to those of Distarnet. Being the protocol of a distributed application, Distarnet puts the OAIS definitions of Archival Storage into concrete terms and sets the rules for a distributed system that successfully archives digital data.

3 Distarnet Protocol 0.4

Distarnet is defined as an XML communication protocol and a set of rules for a distributed system. Its schema will be shared as open source. The system architecture of Distarnet meets the following:

The secure tradition of the data is achieved by building a P2P architecture with strong encryption, controlled redundancy and fault tolerant recovery of network and data. Every node of a Distarnet network communicates in encrypted mode and on top of the TCP/IP protocol. The network stores every AIP in a defined and stable redundancy on different nodes at distant geographical places. If required, every node communicates with every other node. The network is fully distributed. All nodes are absolutely equal, so that there is no single point of failure. Status queries to control the availability of stored AIPs are sent periodically between nodes. If a node has lost its data, or if its data appears to be corrupt, the network restores the AIPs by copying them from redundant copies on other nodes. The defined redundancy is reestablished automatically and remains stable. This way not only the secure tradition of the data is assured but the complicated and cost intense data-carrier migration is automated. Carrier-migration becomes almost a non-issue as new hardware can be integrated by simply switching off the old hardware and attaching the new hardware to the network.

3.1 Processes of Distarnet

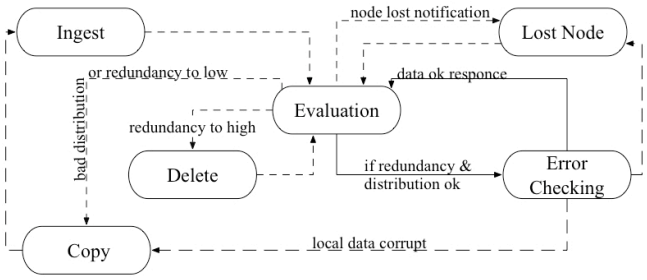

Figure 3. Processes of a Distarnet-Node

The circular flow of the Distarnet-processes starts upon data ingested (see Figure 3). In the next step the AIP is evaluated: according to security criteria, like free space, geographical distribution or node up time, the best node for an AIP is chosen. If the defined redundancy is not met or the distribution of the AIP is not ideal, evaluation starts a copy- or a delete-process. If redundancy and distribution are good the error checking-process starts controlling all local AIPs and their distant redundant copies upon integrity. If all AIPs are present and unharmed, evaluation starts again. If the error checking encounters problems, either the lost data is recopied to the node itself or the other nodes are informed about the loss of a node (Lost Node). Evaluation starts again to restore the redundancy and to prepare the copy-process. If the redundancy is determined being to high, the delete-process on the concerned node first rechecks, whether this is really the case, before finally deleting the data. These processes correspond to the OAIS model of Archival Storage as described in Table 1:

Table 1. Comparing OAIS-Archival Storage to Distarnet-Processes

OAIS	DISTARNET
Ingest/Receive Data	Ingest
Manage Storage Hierar.	Evaluation
Error Checking	Error Checking
Replace Media	Copy
Disaster Recovery	Copy

From the system behavior of Distarnet and from the fact that digital data is independent of the media it is stored on, it can easily be seen that Media and Backup Media of the OAIS- Archival Storage are a non-issue, respectively replaced by the network itself (and therefore dotted in

Figure 2). The Provide Data and Access of OAIS are discussed later in this paper in the Metadata subsection.

3.2 Security in Distarnet

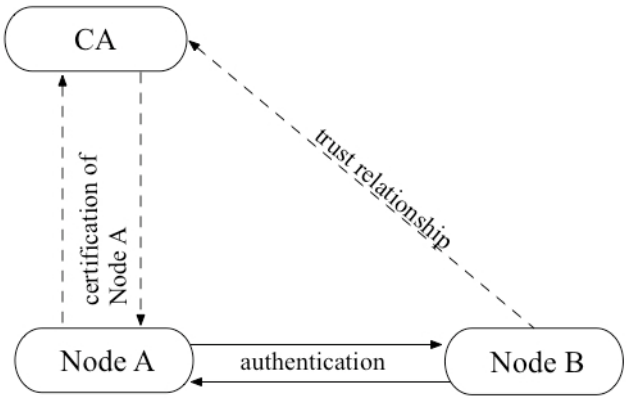

Figure 4. Process of Authentication

Distarnet requires security at its lowest level. This means it must not allow communication between nodes or access to data from nodes that are not authorized. Since Distarnet has a P2P architecture the main goal of the security will be to achieve authentication, integrity and non-repudiation in the communication of it's peers. To accomplish this, it uses Public-Key Encryption (PKE) with a Public-Key Infrastructure (PKI).

PKE is a asymmetric encryption, that allows users to encrypt or decrypt a given message without having prior access to a shared secret key. This is done by using a pair of cryptographic keys, which are related mathematically, designated as public and private key. Only the owner knows his private key, his public key is known to all other participants. A message encrypted with the public key can only be decrypted using the corresponding private key.

PKI is an arrangement, which provides third party vouching for user identities and binding of public-keys to users. It's main purpose is to manage keys and certificates, and by doing so to establish and maintain a trustworthy networking environment. The management of keys and certificates includes certificate revocation, key backup and recovery, support for non-repudiation of digital signatures, management of key histories and other features that are needed for a usable PKI. For a node to be able to authenticate itself on another node, it will need a certificate, issued and signed by a certification authority (CA), which contains the nodes public key and other specific information that can be used to verify the nodes identity. The CA has to be mutually accepted and trusted by all participants of a certain Distarnet. Such a CA issues the digital certificates for use by all other participants of a Distarnet. It is an example of a trusted third party. This permits the creation of user groups that can choose their own CA, since Distarnet can be used and run by anybody.

Figure 4 depicts the authentication process of a node. Since all nodes are the same in the respect that they are equal peers, the same process applies to all nodes. The first step is for Node A to get a certificate that allows him to participate in the network. The CA issues the certificate after a verification of Node A's identity. This certificate contains specific information about Node A and is signed by the CA. Node A can now send the certificate to Node B, who can examine the certificate to see if the contained information about Node A are correct, and if the CA, who signed the certificate is trusted. If everything results positive, then Node A is successfully authenticated by Node B. After the nodes have authenticated themselves mutually, a secure connection is established and all traffic between those nodes will be encrypted, which allows for a secure communication over unsecured channels such as the Internet.

3.3 Copy-Process in Distarnet

The copy-process in Distarnet is one of its most central processes. It must correspond to the traditional data-carrier migration of digital data. This means that every copy has to be rechecked whether it really has been successful and no data has been lost or written inconsistently during the copy process. For this Distarnet calculates checksum with a function of the Secure Hash Algorithms (currently SHA-1) [3].

Copy in a network means sending files from one node to another. Distarnet does not send the whole file at once, as files could be very big in size: If a copy is started the concerned file is virtually split into a calculated number of chunks depending on the size of the file. A chunk has a network width fixed size (currently 8388608 bytes). There are filesize/chunksize chunks of a file plus the one last chunk only containing the remaining bytes of the file, if there are any. The chunks are numbered from 1 to the calculated number. A node in Distarnet first copies all chunks of a file, then, by putting them together, reconstructs the original file. Before the copy-process the checksum of the whole file and the checksum of every chunk are calculated and send along with the chunks. The receiving node calculates the checksums of the received chunks and compares them with the ones resulted on the sending node. After the collection of all chunks the checksum of the whole file is calculated and compared to the one on the originating node(s). If a check does not result in the same checksum, the chunk or the whole file are requested again.

To speed up the copy-process and to balance the workload for the nodes involved, every node of a copy-process shares already copied chunks with other involved nodes, so that the originating node of a copy-process only

needs to copy the file once into the network. If a node receives a request for a certain chunk it sends back information about an available alternative along with the requested chunk. The proposed alternative has neither been proposed nor has it been copied from the asked node into the network before. This allows nodes to ask for actually available chunks on other nodes and therefore to optimize the copy-process.

3.4 Metadata in Distarnet

The secure preservation is the precondition of archiving data, but offers neither a guarantee for its readability nor its usability for future scientific interpretation. To fulfill these needs, different types of metadata must be preserved along with their primary data. Through administrative, technical and descriptive metadata, the retrieval, the technical interpretation and the content interpretation and consequently readability and scientific usability are made possible. The loss of only one type of metadata can bring along the loss of information about the data and consequently the loss of its readability and usability. In such a case the archiving of the data would have failed.

In a distributed system, where different data models come together, keyword queries should be semantically merged to support an overall research. E.g. if in a distributed system a first database describes John Smith as being the 'author' of a certain book, and in a second database John Smith is stored as the 'creator' of a certain book, a search like 'return all books with the author John Smith' should also return the books stored in the second database, which stores John Smith as 'creator'. Assumed that 'author' and 'creator' are semantically equal. Therefore formal mappings between different metadata standards are needed - so called crosswalks - and domain vocabularies need to be shared.

Participants of a Distarnet will form a controlled community with a common aim to preserve their data. Nevertheless the stored data can arbitrary vary and therewith the structure of its description. Although Distarnet produces its own metadata there would be little use to define an overall data model to which all participants must map their data. Being a protocol Distarnet seeks to remain independent to the content being preserved.

Embracing a controlled and rather closed community Distarnet counts, on the one side, on the will of the community to provide its data with adequate description, since without description there will not be a successful archiving. On the other side, Distarnet considers the community as being interested in sharing its data, since participants of this community collaborate in a distributed system to provide a solution for archiving digital data.

To face these needs, Distarnet stores data description in RDF (Resource Description Framework) [4] and proposes a basic set of metadata needed to adequately describe the data. Distarnet will offer a mapping of the current standards (like in [5]). It will be possible to add individual schemas and metadata of any participating archiving institution and map them to the schemas already part of Distarnet.

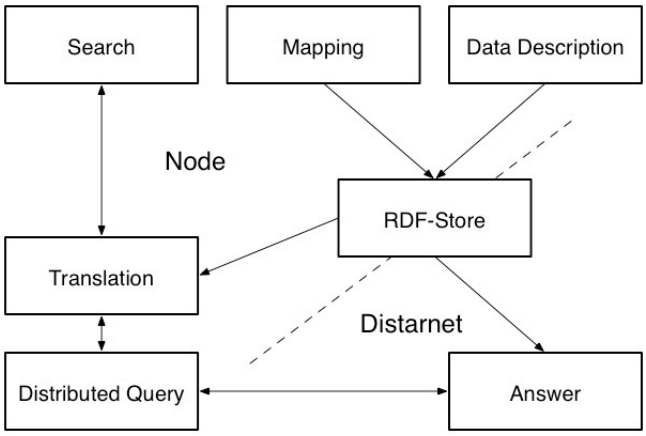

Figure 5. Distributed query translating a search by mapping between different data models

The aim is to present an easy solution to produce and use mappings between different data models for queries as depicted in Figure 5. The data description and its mapping to other data models needs to be done and stored in a RDF-Store. This happens in Distarnet on a certain participant of the network: a node. The new semantical information about the mapping is then distributed among the other nodes. From then on a search can be translated and mapped to all available data models by a querying software agent, respectively a searching person can see all data models and their available mappings and then decide, how to perform his research. In the distributed query the other nodes produce their answer by querying their own RDF-Store. The retrieved result is then shown with all found mappings to the researching person.

This way Distarnet assures the future readability and usability of the data and offers a platform for an overall schema-independent research - corresponding to Provide Data and Access of the OAIS. See [6].

3.5 Queries in Distarnet

Finding Data and researching its description are crucial to Distarnet, since there is no successful archiving process that securely stores data but cannot provide its retrieval. The collection of information in Distarnet is routed over an overlay network that stores information in a distributed hash table (DHT). Distarnet defines a distributed lookup protocol very similar to KADEMLIA [7]: Distarnet nodes hash their IP addresses to assign

themselves an unique key. The distance between two nodes is calculated by using the XOR metric on two keys. A node finds its position in the DHT by querying the network for the node with the closest hash. This way the nodes are arranged in an ascending order. Every node subdivides the DHT into periodical sections, KADEMLIA 'buckets', with the limits of 2^i und 2^{i+1} for every i, $0 <= i < j$, with j being the bit-length of the used hash function (currently SHA1: j = 160). Every node stores contact information for a limited amount of distant nodes of every distant bucket.

Finding or storing information works principally the same as finding its own position in the DHT: By calculating the hash of the searched information a node generates a key and maps it to the buckets of its DHT and sends the query to the nodes of that bucket - consistent hash functions assure that the calculated keys are distributed regularly and the load of stored information remain balanced among all nodes. A node does not need to keep track of the whole network, since a queried node, not storing the searched key, reroutes the query to closer nodes, known to him and therefore storing the information with a higher probability. A responsible node for the searched key of the DHT then handles the query and sends back the answer. Therewith lookup requires O(log N) messages, with N being the number of nodes participating in Distarnet.

4 The Distarnet Implementation

Distarnet is being implemented in Java, currently Version 1.5. The implementation is considered as proof of concept for the protocol. Therefore protocol and implementation are developed simultaneously. The funding of the project will end in December 2007, when protocol version 1.0 will go public. So far the encrypted communication between nodes, the ingest- and copy-process work and are being tested. The implementation of the KADEMLIA lookup protocol and the DHT are consistently working and tested with example data sets. Currently a basic user interface for test purposes is being implemented, which is a prerequisite for the ability to start a Distarnet in cooperation with a few selected institutions who are interested in this project. Current progress and the state of the project can be found at [8].

5 References

[1] S. Margulies, I. Subotic, L. Rosenthaler. Long-term archiving of digital data, DISTributed ARchiving NETwork - DISTARNET. In: EVA 2005 Berlin. Konferenzband, Hg. Gerd Stanke, Andreas Bienert, James Hemsley, Vito Cappellini. Berlin 2005. S. 168-174.

[2] Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1, Blue Book, January 2002.

[3] National Institute of Standards and Technology (NIST). Cryptographic Toolkit. Secure Hashing. http://csrc.nist.gov/CryptoToolkit/tkhash.html

[4] World Wide Web Consortium (W3C). Resource Description Framework (RDF). http://www.w3.org/RDF/

[5] Standards at the Library of Congress. http://www.loc.gov/standards/

[6] S. Margulies, I. Subotic, L. Rosenthaler. Archiving, Data Description and Retrieval in a Distributed System IS&T's 2006 Archiving Conference, Ottawa, Canada.

[7] P. Maymounkov, D. Mazières. Kademlia: A Peer-to-Peer Information System Based on the XOR Metric. In: Lecture Notes in Computer Science, Volume 2429/2002: Peer-to-Peer Ssystems: First International Workshop, IPTPS 2002, Cambridge, MA, USA, March 7-8, 2002. Berlin, Heidelberg 2002.

[8] http://www.distarnet.ch