Criteria for a Storage Concept in a P2P Archival System

Simon Margulies, Ivan Subotic and Lukas Rosenthaler; Imaging & Media Lab, University of Basel; Basel, Switzerland

Abstract

The history of tradition of historical source materials shows a wide range of paths preserved information has undergone to arrive in present time. In the first part the various possibilities of preserving information on the basis of the tradition of Homer's Iliad are demonstrated. According to the results the second part defines a storage concept for an archival storage (OAIS) being built by a distributed system. By implementing this concept a good chance of successful archiving could be achieved. In the third part the actual Distarnet protocol [1] is presented. The remarks at hand are focused on the continuing definition of the Distarnet protocol and present new results of its implementation studies. Even though the elaborated criteria could be used for any kind of an archival storage.

Tradition of Homer's Iliad

Which parts of the ancient Greek epic poem exactly are the work of one author, Homer, or have been compiled by various other characters will likely remain a secret to history. Even if Homer was the first man to write down the epic story and by this the first to start the written tradition, this first written representation of the Iliad is not preserved from ancient times. What we nowadays now as the Iliad is a version compiled from various fragments and other compilations, which again are based on other (partly) lost fragments or compilations. The original words are misty and remain subject to many present and future studies [2]: The oldest documents from the past are ancient papyri, epigraphical and archaeological remains [3]: Various quotations in epigraphical works of other ancient authors provide traces to the 'original' Iliad. The papyri mainly contain smaller text fragments, but also scholarly commentaries, abstracts and lists of words providing newer and more common translations of older terms used in the poem. From the middle ages many manuscripts have survived. Mainly the 10th centuries Codex Venetus A is a vast source of information: It combines a text version with lots of scholia, providing commentaries, glosses and lemmata from the past. Medieval scholia often give clues to many other now lost manuscripts. All these sometimes preserved, sometimes lost source materials have expanded into modern editions starting with the editio princeps in 1488 by Demetrios Chalkondyles in Florence. Figure 1 enlightens the connections of various pieces of information and material preserved from the past: It can easily be seen that the tradition of this work depends upon various circumstances: Text versions on physical representations have been copied or included in other works unnumbered times. As a consequence comparisons between different sources can be undertaken, which render the reconstruction of the original work more probable. Additionally all these documents have been kept in various distant places. The continuous and repeating use of the work through times may have supported such copying. In terms of digital preservation these documents were migrated and often accessed. Wordlists and commentaries preserved additional information. They now offer many possible backwards deductions to the original work. In terms of digital preservation wordlists and commentaries embody metadata.

Criteria for a peer-to-peer archival storage

To establish and balance a distribution of a certain redundancy for any data container a peer-to-peer storage, like Distarnet, needs to evaluate storage places according to various criteria. Every evaluated node scores points in every criterion. Additionally to points, all criteria have a weight. The total of points a node has scored is calculated as the sum of all multiplies of points by weight of each criterion. The criteria are:

Ingester: The ingester node of a data container in Distarnet is the node on which the data was ingested. If this node is live in the network, there must be a copy of the data on it at all times.

Geographical Location: The geographical location of a node is split into the Cartesian coordinates of latitude and longitude of the earths geographic coordinate system. The best node is the one with the geographical location that shows the biggest distance to the geographical location of the ingester node. If the ingester node is not available anymore, the best node is one of the two nodes of the whole network having the biggest distance between them. The next best node is determined as the one holding the biggest sum of distances from his geographical location to all previously determined nodes. By recursively calculating next best nodes an order of storage nodes results,

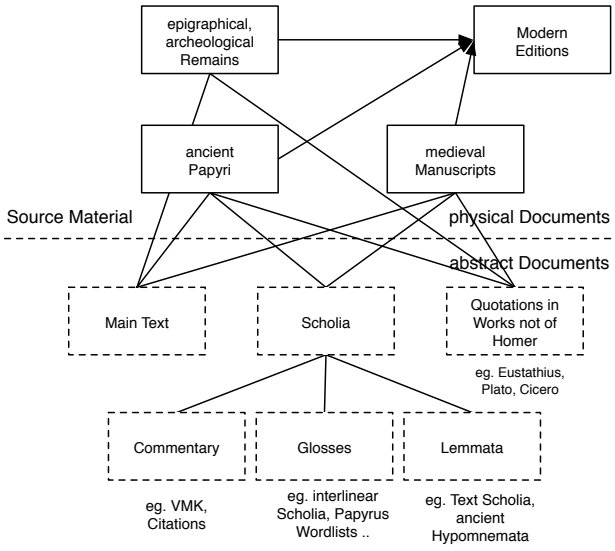

Figure 1. *Different source materials*

which ensures the widest possible geographical distribution of data containers in the network.

Space Ratio: The space ratio is the proportion of free space to total storage space of a node. To distribute data containers regularly and to not overload one node, a space ratio of 0.75 is considered as the minimal value: If a node shows a space ratio higher than 0.75 it is considered as a worse storage place than nodes with a space ratio under 0.75. As such a node with no containers stored yet and thus free total storage is considered as the most attractive storage node for containers.

Uptime: Uptime describes the milliseconds a node is running since his last restart. The evaluating node stores all uptimes he gets from distant nodes and calculates the arithmetic average for every node. The bigger the average the longer a node usually stays up and offers access to its data. The best node results as the one with the highest average uptime.

Speed: Speed describes the connection speed of a node to the network. Since access is one of the most important criteria to ensure the tradition of data, the time matters that passes until data of a node is retrieved.

In the Distarnet protocol a node transmits the values of these criteria with every message to another node. A receiving node stores this information to be able to calculate the ranking of storage places. This process is equivalent to the Evaluation process of the Distarnet protocol shown in the following section.

Distarnet protocol 0.5

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 Archival Information Package (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.

Processes of Distarnet

The circular flow of the Distarnet-processes starts upon data ingested. In the next step the AIP is evaluated: according to the criteria outlined in the first section 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:

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.

Security Distarnet

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

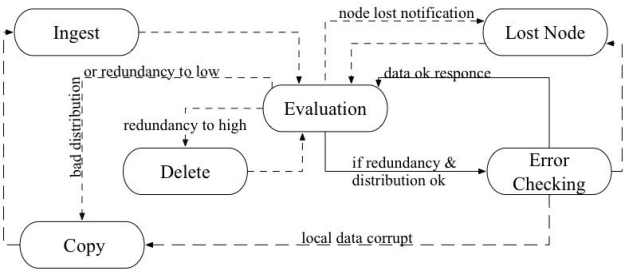

Figure 2. *Distarnet Processes*

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.

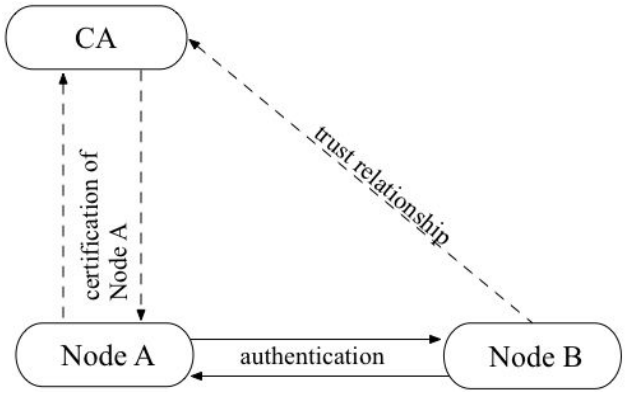

Figure 3. *Distarnet Processes*

Figure 3 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.

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 [4].

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.

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) [5] and proposes a basic set of metadata needed to adequately describe the data. Distarnet will offer a mapping of the current standards (like in [6]). 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.

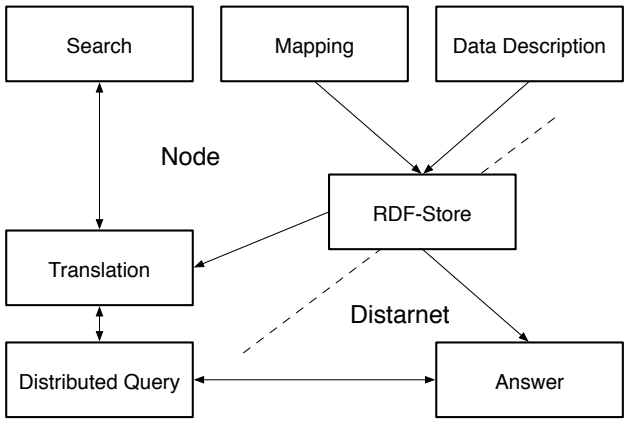

Figure 4. *Distributed query translating a search by mapping between different data models*

The aim is to present an easy solution to produce and use mappings between different data models for queries as depicted in Figure 4. 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.

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 2i und 2i+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.

Conclusions

The analysis of the tradition of Homers Iliad shows basically three different criteria that make a successful preservation more probable: geographical distribution, continuous reuse and citations in other works. These are considered by the Distarnet protocol as follows:

The geographical distribution of data is achieved by evaluating the most widely spread distribution of storage places.

The continuous and repeating use of the data cannot be implemented by an archival storage, but its access can be supported: Therefore the Distarnet protocol considers the ingester node, speed and uptime as criteria for a storage place having a strong influence on access of the data. The ingester is most likely an entity, which will be again interested in its data. Thus it is considered as the most important or the absolute criteria. The weights of all other criteria are subject to the ongoing research.

Citation and commentaries in other works can be seen as metadata about the original work. The tradition of Homers Iliad shows its importance to the preservation of the original information. So far the Distarnet protocol defines metadata, as it is common these days, as having to be stored right next to its data. Under the present results this assumption has to be reconsidered. Instead it could be argued that it is more probable to lose all information if data and metadata are stored in the same place, as it would be, if metadata was geographically reciprocal stored to its data.

References

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

[2] http://www.stoa.org/chs/

[3] West, Martin L. Geschichte des Textes. In: Homers Ilias, Gesamtkommentar. Ed. Joachim Latacz. Prolegomena. München, Leipzig 2002. p. 27-38.

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

[5] World Wide Web Consortium (W3C). Resource Description Framework (RDF). http://www.w3.org/RDF/
[6] Standards at the Library of Congress. http://www.loc.gov/standards/
[7] P. Maymounkov, D. Mazires. 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.

Author Biography

Simon Margulies studied History and Computer Sciences at the University of Zurich. He is working on his Ph.D. in History in the field of archiving digital data. Together with Ivan Subotic he develops Distarnet, a DISTributed ARchival Network under the supervision of Lukas Rosenthaler.