Archiving, Data Description and Retrieval in a Distributed System

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

Abstract

Information retrieval in distributed systems is currently subject to scientific discussions and researches. The following paper describes general issues and technical possibilities to be considered for data description, search and retrieval. It focuses on the role of retrieval for archives and presents the architecture for data description, search and retrieval in the distributed archival system Distarnet.

Introduction

Archives are institutions, which administrate and preserve an amount of documents important to the historical coverage of the past of their sponsorship or a certain theme of their institution. The continuously growing information society produces more and more data, that needs to be archived and needs to remain retrievable.

Archiving institutions for digital data are confronted with various problems. Due to short-livedness of computer systems, data formats and data carriers, the retrieval and the readability of digital data in the future are at stake. To confront the problem of the unstable data carriers basically two approaches are being followed: In the first, digital data is stored on very endurable media, and in the second, digital data gets automatically migrated from old carriers to new ones. The latter can be achieved by building a distributed system [1]. In distributed systems multiple computers in remote locations are coordinated to accomplish a common objective or task. They offer a high fault tolerance and big calculation power. Despite the possibility of very heterogeneous distributed systems with a variety of different standalone systems working together, a distributed system should remain transparent and scalable. All aspects that render them suitable to confront the problem of the unstable data carriers. But the secure preservation of the archived data is only the precondition of a successful archiving. The archived data needs also to remain readable and retrievable, otherwise it will not be useable in the future. Therefore metadata must be preserved along with its primary data. Through administrative, technical and descriptive metadata, the retrieval, the technical and content-interpretation and consequently readability and scientific usability are made possible.

For the present paper a growing interconnectedness between archives providing online access to digital databases is assumed. Archives will be a part of a distributed system or will use distributed systems like Distarnet [1], to spread their data and to make its tradition more secure. Retrieval plays a crucial role in a successful archiving, as not found data is lost data. The present paper points out the shared connections between data structure and semantics, archiving and retrieval. Techniques will be presented, that provide archives with new possibilities to support retrieval and to render usability of the data for scientific research in the future more secure.

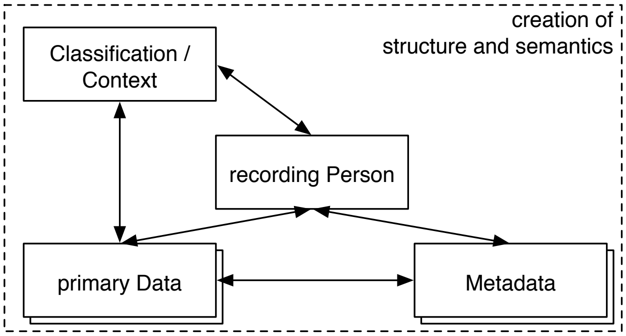

Figure 1. *Dependencies on data-ingest*

This paper is organized as follows: First we will outline the role of data description and the basics of retrieval in such a distributed system. Then we will examine what possibilities for technical solutions there are and which of them may be reasonable. Finally we will present the system architecture for data description and retrieval in the distributed system Distarnet.

Research in a Distributed System

Distributed systems exist in a variety of forms and store information in many different ways. Apart from technical aspects, like the amount of distribution (i. e. amount of technical hierarchies), retrieval in such a system depends very much upon the way information is stored. 'Garbage in, garbage out' is a common saying in information sciences, especially if it is intended as wrong formulated database queries. But successful retrieval does not only depend on correct database queries. The process of a successful retrieval starts by ingesting the data into the system. The way the data is ingested, described and stored, is crucial to its later use.

Data description

By ingesting data into a system the recording person classifies the data according to its subjective contexts. Thereby the recording person creates metadata that gets stored along the described primary data. This means, that the recording person describes the data in the way he understands and interprets it. The result is a subjectively encoded information. This is often called the 'semantics' of the data. The according process is illustrated in Figure 1.

The classification of metadata is fixed before data is ingested: The stored information is organized by structure and content of the data fields of the metadata. Data fields are named entities of information. E. g. elements like 'author', 'title', 'datum'. The

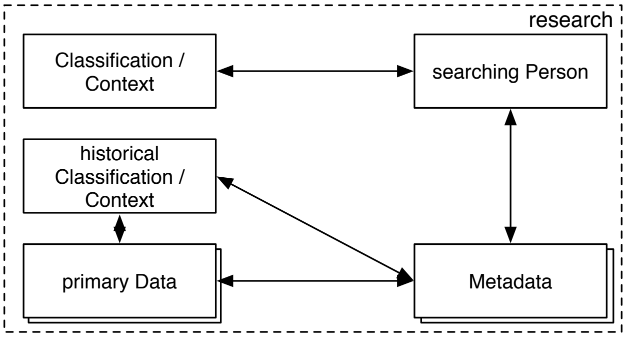

***Figure 2.** Dependencies during research*

structure describes the order and the rules, how data fields are arranged. E. g. if a book can have several titles or just one. Information sciences call this a data model. Data fields and their structure are only the framework of metadata, since ingesting metadata needs consistent use of rules for its content too: Such rules define e. g. whether a date is composed in the order YYYY-MM-DD, MM-DD-YYYY or DD-MM-YYYY or what keywords are used to classify the content. Usually norm-data is used to consistently apply keywords to name entities of information.

As depicted in Figure 1, such classifications depend on the recording person. In a broader sense, they are domain-specific. This means that it is possible, that every domain, or an institution like an archive, would have used a different classification and other keywords to describe the data - their data has different semantics.

Data retrieval

A researching person, who wants to retrieve the stored information, has its own idea and understanding of the data and its description. The researcher needs to have sufficient knowledge about the classification of the data undertaken by the recording person to be able to use the system for retrieving information. In Figure 2 these processes are outlined: A searching person gets access to the searched primary data by researching the metadata. This metadata was ingested by the recording person in Figure 1. As explained before the recording person created the metadata by subjectively classifying it. This, being happened in the past, is the reason, the classification of the primary data and of metadata is marked as 'historical' in Figure 2.

To find the searched data a researcher needs to have prior knowledge and understanding of the data model and the vocabulary used during the ingest of the data. By mapping his own classification of contexts to the historical classification a researcher tries to interpret the latter to retrieve the searched data. That are the basics of every research in a database. For a successful retrieval it is crucial to know, how the stored information is structured, and what the terms are, which were used to describe it. E. g. the ArchiviaNet on-line research tool of the Library and Archives Canada web-services [4] offers a variety of research tools, which render the vast materials obtainable by keyword-search. For every collection there is an online help, describing what actually can be found by keyword. So one knows, that in ArchiviaNet, title and location of a photograph can be searched by keyword, though the title provides a general description of any photograph. Some sections, for instance the 'Canadian Nurses', are also divided into themes with specific names: Knowing that possible keywords are e.g. 'Military Nurses' or 'Public Health Nursing' lets a researcher know, how this section has been organized and where he can find the searched information.

If various domains share their data in a distributed system, a researcher needs to semantically link between different classifications. This gets easier, the more consistent metadata was ingested and the more a researcher knows about its classifications and its vocabulary. If the researcher knows the structure and semantics of each domain the mapping of the different classifications is apparent to him. With similar subjects of the domains, their semantics are closer and their mapping can be done easier. On the contrary information retrieval in a distributed system with an uncontrolled user community and no control over the data inserted, like the Internet, is more difficult and error-prone to a distributed system used by a defined community, which shares only specific kind of data and a common interest in accurately describing the data. So successful data retrieval in a distributed system can be a difficult task. To make it easier, the various data models and their vocabulary used should be transparent and system-width accessible or shared across domains. Mappings could be formally described to make support from software agents possible.

Technical Issues

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.

To support overall queries in a distributed system basically three technical solutions are possible:

- An overall data model is used, which integrates all data. Every participant needs to map his data model to the overall model.
- Many models are present. Every participant uses his own data model and maps it to other available models.
- No mapping at all is performed and only simple full-text search supported.

The first approach offers more consistency of the structure for the data description: A researcher can count on the same data fields being always available. But such a structure usually embodies the greatest possible common denominator of a community. This means, a reduction in expressiveness and thereby a loss of information, respectively a probable loss of consistency of the content of data description, as different rules are mapped into the same data fields. A remarkable example of an attempt to establish such a model with only 15 data fields is the Dublin Core Metadata Initiative [2]. Such a structure can be useful, if all participants store similar data and if no precise data description is requested. But with a lot of different kind of data and a rich data description, it can even be impossible to map certain data fields of the own data model to Dublin Core. Major advantage of this approach is, that it can easily be implemented, and that scalability of queries in the distributed system will not depend upon the complexity of the data model.

The second approach differs from the first one by the fact, that each data model is not mapped to only one other data model, but that single data fields from one model are mapped to other data fields of other models. Such a structure remains more flexible, as there are more possibilities to find a match for a data field in another model. Respectively the information of a data field with no correspondent field in another model does not need to be reduced, as a match may be found at a later date. With the 'Semantic Web' the W3C [11] are heading in this direction, offering the Resource Description Framework (RDF) to describe data in a machine-understandable way. By defining a formal and explicit specification of a shared and common conceptualization, a so called 'ontology', a community can provide software agents with the needed information to deduct meaning and context of different source data. Such a conceptualization is layered top of the data description and provides software agents with the possibility to find matches between semantical entities, and to be able to present these matches to the user. E. g. if something matching the concept 'book', then it has an 'author' and a 'title'. 'Author' is of type 'person' and equivalent to 'creator'. A 'writer' is a sub-conceptualization of an 'author'. Having stored the 'author' 'John Smith' of a certain 'book' somewhere in a database, and researching for a 'book' by the 'writer' 'John Smith', would provide a software agent not only with the possibility to present the result, but to deduct that the researcher does not mean 'John Smith' the 'bookseller', who is another person. Of course the software agent needs to inform its user about the mappings done. In this approach, more expressiveness goes along with a possible lack of scalability and a complex implementation.

In the third approach a simple full-text search upon the all metadata is performed. As there is no mapping involved, the semantics of the data and the research remain hidden to the machine. Consequently the researcher cannot be supported. Retrieval is left to the logic of some algorithm, which needs to rate information according to the amount of appearances of keywords or according to the amount of the linking between different information sources. This is the case because of, as a machine, such a search engine cannot distinguish between apples and oranges. The retrieved result does not depend upon the searched information content but upon the popularity of the information source, as the usual search machines for the internet show. For a controlled community with scientific intents to archive data in a distributed system such ratings are not a reasonable solution, as in the scientific context ratings need to be done by the researcher and not by a machine.

Data Description and Retrieval in Distarnet

The DISTributed ARchival NETwork, Distarnet, is the protocol of a distributed system. It describes the rules for automated data carrier migration and storing data with high security in a network. As stated before the secure preservation is only the precondition of archiving data. To support retrieval, readability and scientific usability data needs to be described, and the according data description needs to be preserved as well. Through administrative, technical and descriptive metadata, retrieval, technical and content-interpretation and consequently readability and scientific usability are made possible [16]. The loss of only one type of metadata can bring along the loss of information about the data and as a consequence the loss of its readability and usability. In such a case the preservation process would have failed. This is why a distributed system, that archives digital data, like Distarnet, needs to define rules for data description and data retrieval.

Data Description

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. Thus Distarnet supports the above described second approach by defining minimum requirements about data description and facilities for mapping between data models.

At the moment archives usually provide their primary data with various layers of metadata. Experience shows, that every archiving institution develops its own, rather elaborate data model, even if it orients itself on standards like [5]. A rather common practice within archives nowadays seams to use XML for the data description. An often read argument in favor of XML is its 'human readability' holding structure and content in plain text files. Most available standards for XML-Metadata offer a hierarchical data model for information objects. The known drawbacks of hierarchical models like the lack to define many-to-many relationships or the lack of referential integrity are not considered. As a consequence of the hierarchical approach, data gets highly structured by XML. The underlying semantic entities of single parts and especially their context to other information objects remain hidden to querying software agents. Additionally the more complicated the data description gets the less 'human readable' the XML becomes and the more difficult and intransparent a mapping gets.

Maybe as a consequence, there is no established standard to express mappings in a formal and therewith machine-readable way. Defined mappings can usually only be found on web-pages expressed in HTML tables, the only exception can be found on [7], described in [6]. To support the possibility of semantical mappings Distarnet stores the data description of its data in the

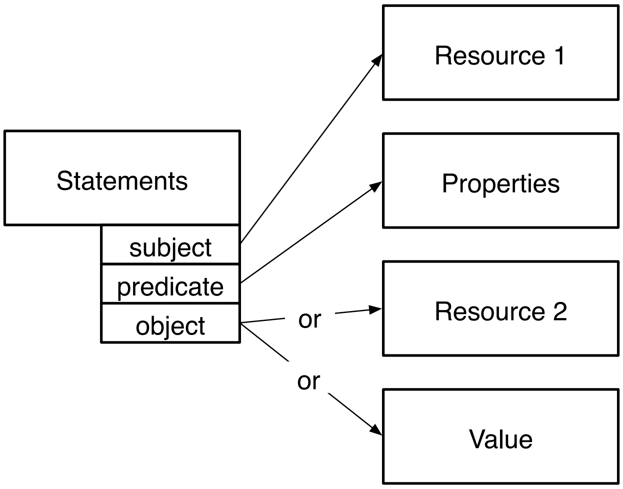

Figure 3. *RDF data model*

RDF/XML syntax [9]. By offering the very simple data model of triples formulating statements about resources. A statement consists of a subject, an object and a predicate as illustrated in Figure 3, almost like a sentence in human languages. Therefore RDF is intuitively understandable and will remain understandable as long as the concept of todays languages won't change. This renders it suitable to a protocol describing a system for long-term preservation of digital data. With the possibility of an object being either the subject of another statement or a final value almost anything could be described in RDF. Contrary to XML semantical entities do not get nested in other semantical entities but remain independent entities themselves. Such resources can be linked as RDF builds upon XML and URI technologies. Therewith it supports 'human readability' just as well as naming and addressing being fundamental to retrieval.

Search and Retrieval

To take advantage of the RDF data model the participants of Distarnet must provide facilities to query RDF statements. Additionally there needs to be the possibility to describe mappings and to execute queries according to them. The core functionality of search and retrieval is the unique and constant addressing of a resource: Every stored resource in Distarnet gets a unique identifier corresponding to the checksum, calculated with a function of the secure hash algorithms (currently SHA-1) [8].

Being under development the exact protocol specifications for queries and mappings need further examinations. Mainly there are three possibilities:

- Using a top level ontology and relate the RDF data descriptions to its concepts as proposed in [3].
- Defining mappings between properties using a combination of ontology languages like RDFS [10] and OWL [11] likely to the proposed solution in [15].
- Producing mappings with the XSLT technology [12] transforming different XML descriptions, likely to the proposed solution in [13].

Implementing a top level ontology as the one presented in [3] seams extremely difficult and its scalability questionable. Main drawback by using a top level ontology would be the limitation of Distarnet to the basic concepts of the domain, which defined the ontology. Even being a protocol for archiving institutions this approach rather resembles the first rejected possibility presented under technical issues above. These are the reasons why the last two approaches are currently discussed and evaluated for the reference implementation and the protocol specifications.

Defining XSLT transformation rules mappings hides difficulties: XSLT is designed for transforming XML documents into other XML documents. This means that data fields are transformed into other data fields. XSLT provides a solution for translating structure rather then semantics. Sometimes mappings depend on content of the XML-data fields and have rules, which apply only for parts of the content of such fields. To formally describe such mappings XSLT has only limited possibilities as outlined in [14]. The biggest drawback of using XSLT is considered the transformation viewing a translation as a one to one process: In such a translation the original semantics can be lost, e.g. semantical subtleties as the difference between a writer and an author would be lost during transformation and not just translated for the machine in the very moment, if a writer is one to one translated as being an author. In a distributed system for every existing data model there would have to be an XSLT transformation to every other data model (and for every version of every data model). Otherwise several transformation would be needed: To transform a query from a data model A to a data model C without the according XSLT transformation rules, the data model A would have to be transformed to a data model B, for which transformation rules to A and C exist, to finally being transformed to data model C. Thereby semantics would get reduced even more. Nevertheless XSLT transformations are currently being considered for Distarnet, because they are supposed to scale well for much data too.

Additionally, researches about machine supported retrieval often implicitly presume that the presented result of such a retrieval needs to be final. I. e. a searching person needs to formulate one query, which is resolved completely and presents always the exact searched answer. This can be proven to be an erroneous assumption, because for a scientific research, as it is carried out in archives, results of supported retrieval must always be judged by the researching person, which according to the result will most probable perform another search. Such a researcher needs to be able to rate the data and its description. If by an XSLT translation former semantical meaning gets reduced, a researcher could not be able to trace back the transformations. Therewith he could not be able to rate the meaning of the found data. Respectively the researcher gets a result, that is described by another data description as the one of the data when it was ingested. This would produce wrong assumptions about the data. As a consequence the result of a retrieval in a distributed system should rather show the mapping used by the software agent than hiding it. Final retrieval will always be a 'burden' of the human researcher.

Defining mappings by using ontology languages like RDFS and OWL, connections can be formally described in a more precise way than by XSLT transformations. The possibility of describing various relations like classes and subclasses, intersections and equivalency etc. provides an accuracy that meets with the real world semantics. Consequently there are considerations about the scalability of such technologies, especially if a high ex-

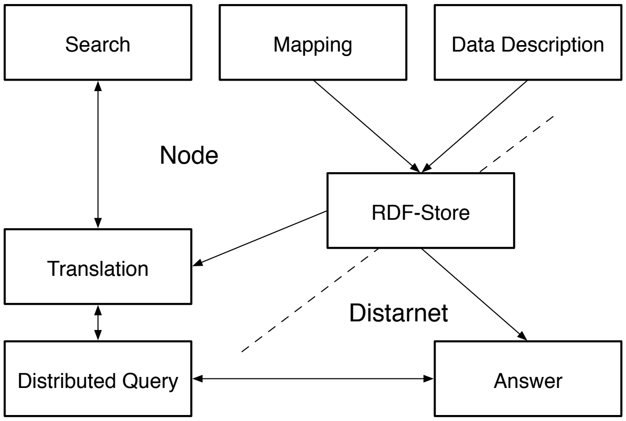

Figure 4. Distributed query translating a search by mapping between different data models

pressiveness is needed. The fact that these ontology languages build on and are expressed in the RDF/XML syntax makes them very interesting for Distarnet and assure them a longtime human readability. The main advantage is considered to be the ability to better show semantical mappings and interconnections of data than XSLT.

Technical Issues

The aim is to present an easy solution to produce and use mappings between different data models for queries as depicted in Figure 4. As described earlier, before a search can be performed, the data description and its mapping to other data models needs to be done and stored in the 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.

To provide efficient queries, the collection of information in Distarnet is routed over an overlay network that stores information in a distributed hash table (DHT) as shown in Figure 5. Distarnet defines a distributed lookup protocol similar to CHORD [17]. Nodes in Distarnet form a circle by hashing their IP addresses and arranging themselves in an ascending order. By calculating the hash of the searched information a key is generated and mapped to the DHT. The responsible node for that part of the DHT then handles the query and sends back the answer. This responsible node can be found by asking any node of the searching nodes own shortcut table, CHORDs finger table. This table stores some distant nodes, which are responsible for distant hash keys. Finally the responsible node is found by rerouting queries from a distant node to the one actually responsible for the answer. Therewith lookup requires O(log N) messages, with N being the number of nodes participating Distarnet [1].

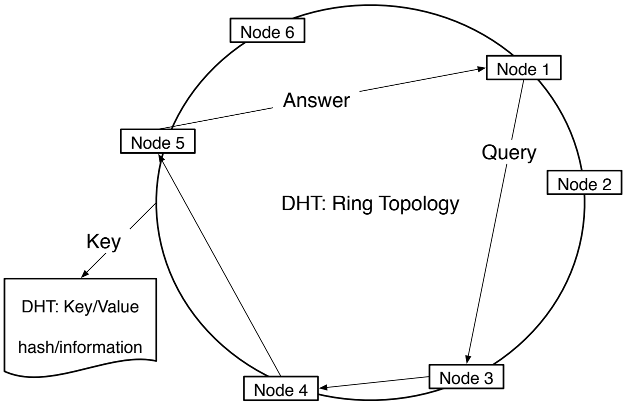

Figure 5. DHT - Distributed hash table in Distarnet

Conclusions

To support a researching person and to render retrieval successful an appropriate data description is needed. In distributed system this poses various problems as different classifications of information result in different data descriptions with different data models. Therefore mappings between these different classifications must be performed during a research. A distributed system should provide facilities to define formal mappings. The more heterogeneous data and different data description are present and used, the bigger and the more complex the amount of data to be handled gets. With defined formal mappings software agents can deduce contexts from different resources and present them to the researching person as alternatives. For Distarnet precise protocol definitions and technologies used are currently subject to further investigations.

References

[1] Currently available at: http://www.distarnet.ch/

[2] Dublin Core Metadata Initiative. Currently available at: http://dublincore.org/

[3] ICOM, International Council of Museums. The CIDOC Conceptual Reference Model. Currently available: http://cidoc.ics.forth.gr/

[4] Library and Archives Canada. ArchiviaNet. Currently available at: http://www.collectionscanada.ca/archivianet/0201_e.html

[5] Library of Congress (USA). Standards at the Library of Congress. Currently available at: http://www.loc.gov/standards/

[6] OCLC. About the metadata crosswalk repository. Currently available at: http://www.oclc.org/research/researchworks/schematrans/default.htm

[7] OCLC. OCLC SchemaTrans Crosswalk Catalog. Currently available at: http://errol.oclc.org/schemaTrans.oclc.org.search

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

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

[10] World Wide Web Consortium (W3C). RDF Vocabulary Description Language 1.0: RDF Schema. Currently available: http://www.w3.org/TR/rdf-schema/

[11] World Wide Web Consortium (W3C). Semantic Web. Currently available at: http://www.w3.org/2001/sw/

[12] World Wide Web Consortium (W3C). XSL Transformations (XSLT). Currently available at: http://www.w3.org/TR/xslt

[13] Godby, Carl Jean. Young, Jeffrey A. Childress, Eric. A Repository of Metadata Crosswalks. In: D-Lib Magazine, December 2004, Volume 10, Number 12.

[14] Godby, Carol Jean. Smith, Devon. Childress, Eric. Two Paths to Interoperable Metadata. Paper presented at the 2003 Dublin Core Conference, DC-2003. 2003.

[15] Hunter, Jane. Lagoze, Carl. Combining RDF and XML Schemas to enhance Interoperability between Metadata Applications Profiles. In: WWW10, May 1-5, 2001, Hong Kong. P. 457-466.

[16] Margulies, Simon. Subotic, Ivan. Rosenthaler, Lukas. 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.

[17] Stoica, I. Morris, R. Krager, D. Kaashoek, M. F. Balakrishnan, H. Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications. SIGCOMM 01, August 27-31, 2001, San Diego, California, USA.

Author Biography

Simon Margulies studied History and Computer Sciences at the University of Zürich and Rome. At the University of Basel he is working on his Ph.D. in the field of archiving digital data. Together with Ivan Subotic he develops Distarnet, a DISTributed ARchival Network. He analyzes the impact of digital data as source material for the historical research, develops Distarnet and advises various projects and companies on data modelling, metadata and data retrieval.

Ivan Subotic studied Bussines and Economics at the University of Basel. He is working on his Ph.D. in History in the field of archiving digital data with emphasis on legal and economic issues, i. e. the tension between legal obligations and economical possibilities of digital archiving solutions and their implication on Distarnet.

Dr. Lukas Rosenthaler, studied Physics, Mathematics and Astronomy at the University of Basel, and got a Ph.D. in Applied Physics in the field of Nanotechnology building a Scanning Tunneling Microscope. From 1988 to 1992 he worked as postdoc at the Swiss Institute of Technology in Zürich in an interdisciplinary project about the understanding and the computational simulation of the vision system of human beings. From 1992 to 2001 he developed new methods for the restoration of damaged movie films, in affiliation with the Scientific Photography Lab of the University of Basel. Since 2001 Lukas Rosenthaler is a full time staff member of the Imaging and Media Lab of the University of Basel, Switzerland. The main research topics are the restoration of movie films and the long-term preservation of digital images. He leads the project team of Distarnet.