Langzeitarchivierung digitaler Daten
DISTributed ARchiving NETwork
DISTARNET

Long-term archiving of digital data
DISTributed ARchiving NETwork - DISTARNET

Simon Margulies, Ivan Subotic, Lukas Rosenthaler

Imaging & Media Lab

University of Basel

Bernoullistrasse 32
CH-4056 Basel / Switzerland
Tel.: +41 61 267 04 88, Fax: +41 61 267 04 85
E-mail: simon.margulies@unibas.ch, Internet: http://www.distarnet.ch

Zusammenfassung:

Die Archivierung digitaler Daten stellt Archive und Forscher vor neue Anforderungen. Das Projekt Distarnet analysiert diese und implementiert eine Lösung zur Langzeitarchivierung digitaler Daten unter Berücksichtigung der speziellen Bedürfnisse von Archiven und Museen als Träger der Überlieferung von historischem Quellenmaterial und Kulturgut. Distarnet ist das Kommunikationsprotokoll eines verteilten Systems mit Eigenschaften eines P2P-Netzwerkes, welches Daten in hoher Redundanz und Sicherheit speichert. Die Überlieferung wird durch Fehlerverarbeitung bei Datenverlust und Einbeziehung verschiedener Ebenen von Metadaten gewährleistet.

Abstract:

Archiving digital data assigns archivists and scientists new tasks. In the Distarnet project a solution for the long-term preservation of digital data is analyzed and implemented. Special attention is given to the needs of archives and museums which preserve historical source material and cultural heritage of our time. Distarnet is a communication protocol for a distributed system with properties of a P2P-network. Its implementation stores data in high redundancy and security. Preservation is guaranteed through self-recovery from data-loss and the support of different layers of metadata.[1]

1. Einleitung: Archivierung und digitale Daten

Digitale Daten, wie sie in einem elektronischen Informationssystem genutzt werden, können von blossem Auge nicht gelesen werden. Ihre Interpretation durch den Menschen ist von einer vorangehenden Darstellung durch eine Maschine abhängig. Deshalb garantiert die "blosse" Überlieferung des Archivguts nicht mehr die Möglichkeit des künftigen Auffindens, Lesens und Interpretie-

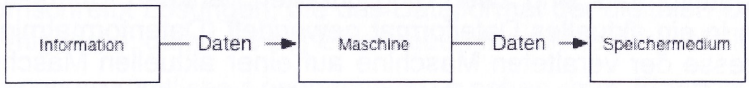
[1] Lukas Rosenthaler and Rudolf Gschwind. DISTARNET - A Distributed Archival network. In: Final program and proceedings of IS&T's 2004 Archiving Conference : April 20-23, 2004, Hyatt Regency Hotel, San Antonio, Texas. Springfield (VA) 2004. P. 242-248.

rens. Für die Prozesse der Archivierung entstehen durch die Eigenschaften digitaler Daten und der sie darstellenden Maschinen spezifische Anforderungen, die den Archivar vor neue Aufgaben stellen.

Im vorliegenden Text wird in einem ersten Teil auf die Probleme der Langzeitarchivierung digitaler Daten eingegangen und in einem zweiten Teil ein möglicher Lösungsansatz in Form des verteilten Systems Distarnet präsentiert.

1.1. Eigenschaften digitaler Daten

Digitale Daten entstehen bei der Digitalisierung beliebiger Information.[2] Dabei wird die Information in diskrete Einheiten zerlegt, die in Form von digitalen Daten meist in elektronischen Informationssystemen genutzt werden. Durch die diskreten Einheiten digitaler Daten wird ermöglicht, dass eine Kopie der Daten bei korrektem Vorgehen nicht mehr von ihrer digitalen Vorlage unterschieden werden kann, beziehungsweise, dass beim Kopiervorgang kein Datenverlust entsteht. Solche Daten bestehen aus einer vom Informationssystem abhängigen Folge von Zeichen, dem Dateiformat, das für den menschlichen Betrachter keinen Sinn ergibt.



Digitalisierung

Für ihre Wiederverwendung werden digitale Daten auf unterschiedlichen Medien gespeichert.[3] Im Gegensatz zu den herkömmlichen in Archiven gesicherten Speichermedien wie Papier oder Mikrofilm können von diesen Medien die Daten mit dem blossen menschlichen Auge nicht abgelesen werden. Deshalb wird für das Lesen solcher Medien und die Entzifferung der Zeichenfolge digitaler Daten eine Maschine benötigt, die die Archivdaten darstellt und dem Archivbenutzer zugänglich macht:

Zugang

Die Abhängigkeit der Darstellung von einer Maschine und die Möglichkeit der verlustfreien Kopie haben die Unabhängigkeit digitaler Daten von der Art ihrer Speichermedien zur Folge. Die Speichermedien können durch eine verlustfreie Kopie ihrer Daten beliebig ausgetauscht werden.

Die Unabhängigkeit vom Speichermedium, die Abhängigkeit von der darstellenden Maschine und die Möglichkeit der verlustfreien Kopie als Eigenschaften digitaler Daten stellen bezüglich der Archivierung neue Anforderungen und ermöglichen neue Lösungsansätze, wie im Folgenden dargelegt wird.

1.2. Anforderungen an ein elektronisches Archiv: Probleme der Langzeitarchivierung digitaler Daten

Zu den Hauptaufgaben eines Archivs gehört die Pflege des Archivguts. Darunter fallen auch für ein elektronisches Archiv die Minimierung des Datenverlusts und die Garantie des Auffindens, der Darstellung, des Zugangs sowie der Angabe des kausalen Zusammenhangs des Archivguts.

Um den Datenverlust digitaler Daten zu minimieren und die Überlieferung zu garantieren, sollten digitale Daten in einer inneren und äusseren Redundanz gehalten werden: Die innere Redundanz muss durch das Dateiformat gewährleistet sein, so dass bei Verlust einiger Zeichen dennoch die

---

[2] Es wird zwischen 'digitalisierten Daten' und 'digital born'-Daten unterschieden. Die digitalisierten Daten werden z. B. durch einen Scan produziert. Unter den sog. 'digital born'-Daten versteht man digitale Daten, die in der realen Welt kein materielles Original haben, welches zu ihrer Herstellung digitalisiert werden musste: Z. B. ein in einem Textverarbeitungsprogramm verfasster Brief.

[3] U.a. elektronische, magnetische, optische oder magneto-optische Speicherung.

gesamte Information der Datei wiederhergestellt und dargestellt werden kann. Die äussere Redundanz wird erreicht, indem die Eigenschaft der verlustfreien Kopie digitaler Daten genutzt wird, und Kopien der archivierten Dateien auf mehreren Speichermedien gesichert werden, die ihrerseits an unterschiedlichen Orten (geographische Redundanz) gelagert werden. Bei Verlust eines Speichermediums können die betroffenen Dateien von anderen Speichermedien zurückgewonnen werden.

Für das Auffinden der Daten ist deren Überlieferung die Voraussetzung. Um ein Auffinden zu ermöglichen, gelten für digitale Daten die gleichen Praktiken der Datenbeschreibung zur Ordnung und Verzeichnung wie für das traditionelle Archivgut. Für die Verarbeitung und den Zugang in einem elektronischen Informationssystem ist es unumgänglich, die Datenbeschreibung ebenfalls in digitaler und elektronischer Form zu halten. Man spricht dabei von Metadaten. Eine solche digitale Datenbeschreibung ist auch digitales Archivgut und muss der gleichen Behandlung zur Vermeidung von Datenverlust unterzogen werden.

Eine Darstellung kann erst erfolgen, wenn die Daten zuvor gefunden wurden. Für die Darstellung müssen die Speichermedien und das Dateiformat der digitalen Daten von der Maschine gelesen werden können. Die Lesbarkeit der Speichermedien wird mit einem periodischen Umkopieren der Daten von einem veralteten auf einen aktuellen Datenträger erreicht (Datenträgermigration). Für die Lesbarkeit des Dateiformats sind zwei[4] Möglichkeiten denkbar: Einerseits können die Daten von einem veralteten in ein aktuelles Dateiformat gewandelt (Datenformatmigration), andererseits die Darstellungsprozesse der veralteten Maschine auf einer aktuellen Maschine imitiert (Emulation) werden. Für beide Fälle wird aus der Dokumentation durch die Metadaten ausreichend Information zu ihrer korrekten Umsetzung benötigt. Bei der Datenformatmigration die Syntax und Semantik des Datenformats, bei der Emulation die Syntax und Semantik der Darstellungsprozesse der veralteten Maschine.

Durch die Darstellung der Daten wird dem Archivbenutzer der Zugang ermöglicht. Ein überliefertes und lesbares Dokument ist aber ohne die Überlieferung der kausalen Zusammenhänge seiner Vergangenheit in der Gegenwart unter Umständen nicht interpretierbar oder beinhaltet nur wenige Informationen. Die Datenbeschreibung muss deshalb eine angemessene inhaltliche Beschreibung der Daten gewährleisten. Zum Beispiel muss bei einem Bild die Beschreibung der dargestellten Gegenstände überliefert werden, damit diese überhaupt erkannt werden und einen epistemischen Wert aufweisen.

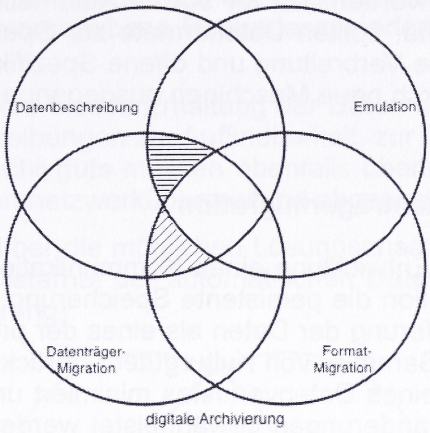
Für eine erfolgreiche Überlieferung digitaler Daten müssen sämtliche oben erwähnten Bedingungen erfüllt sein. Wird bei der Archivierung nur einer dieser Aspekte vernachlässigt, so hat das unter Umständen den Verlust der gesamten Information zur Folge. Deshalb ist eine Digitalisierung noch nicht gleichbedeutend mit der Erhaltung von Kulturgut, denn digitale Daten können nicht einfach an einem sicheren Ort weggeschlossen und nach Jahren wieder hervorgeholt werden. Sie erfordern periodische und korrekte Wartung, die nur mit dem entsprechenden technischem Wissen über die Eigenschaften elektronischer Informationssysteme und digitaler Daten durchgeführt werden kann. Die Prozesse, die für die Pflege dieses Archivguts benötigt werden, sind für ein einzelnes Archiv sehr kostenintensiv. So hängt in der heutigen Zeit die Überlieferung digitaler Daten entscheidend von ihrer Finanzierung ab. Eine möglichst grosse Reduzierung der Kosten ist deshalb von entscheidender Bedeutung für die Überlieferung digitalen Kulturguts.

1.3. Mögliche Lösungsansätze zur Langzeitarchivierung digitaler Daten

Wie oben ausgeführt ist die Umsetzung einer erfolgreichen Langzeitarchivierung digitaler Daten in einer Schnittmenge zwischen der Datenträgermigration, der Datenformatmigration, der Emulation und der Datenbeschreibung zu suchen. Die Datenträgermigration und die Datenbeschreibung sind erste Voraussetzungen für die erfolgreiche Archivierung digitaler Daten, da sie einerseits die Daten selber und andererseits Informationen über Syntax und Semantik für ihre Darstellung und Inter-

[4] Eine dritte Möglichkeit stellte die Konservierung der alten Maschinen dar. Diese Möglichkeit wird hier aber vernachlässigt, da davon ausgegangen wird, dass weder die einzelnen Bestandteile noch das technische Wissen lange genug währen, um die Funktionstüchtigkeit alter Maschinen zu garantieren.

pretation überliefern. Somit ergeben sich in der folgenden Abbildung zwei mögliche Schnittmengen, die schraffiert dargestellt sind:



Ein Lösungsansatz, der der horizontal schraffierten Fläche entspricht, überliefert digitale Daten, indem diese in einer Emulation der alten auf einer neuen Maschine dargestellt werden. Die schräg schraffierte Fläche umschreibt Lösungen, die das Dateiformat der digitalen Daten von einem alten in ein neues Format migrieren, damit sie auf einer neuen Maschine dargestellt werden können.

Wie oben erwähnt müssen sämtliche Lösungsansätze neben ihrer Eignung für eine erfolgreiche Überlieferung digitaler Daten auch auf ihre Kosteneffizienz bezüglich der involvierten Prozesse überprüft werden. Kosteneffiziente Haltung digitaler Daten wird durch die Automatisierung oder die Minimierung möglichst vieler Prozesse ermöglicht.

Für die periodische Erneuerung der Speichermedien ist, wie oben erwähnt, ihre Überlieferung durch Aufbewahrung in geographischer Redundanz Voraussetzung.[5] Die Minimierung der Prozesse der Datenträgermigration bedeutet die Wahl eines möglichst beständigen Speichermediums (eternal media).[6] Solche Speichermedien bergen häufig den Nachteil, dass technisch noch kein genügend schneller Datenzugriff gewährleistet werden kann, und somit der Zugang zum Archivgut und somit eine der Hauptaufgaben eines Archivs erheblich erschwert wird. Die Automatisierung der Prozesse der Datenträgermigration wird durch die Eigenschaften der verlustfreien Kopie und der Unabhängigkeit vom Speichermedium digitaler Daten erreicht. Die Daten werden in einem verteilten System redundant gespeichert und bei Erneuerung eines Speichermediums automatisch kopiert (medialess). Im Gegensatz zur Wahl eines möglichst beständigen Mediums hat ein verteiltes System in der Regel die Gewährleistung eines schnellen Datenzugriffs zur Folge.[7]

Für die verschiedenen Bereichen der Datenbeschreibung sind erst beschränkte Möglichkeiten der Automatisierung vorhanden. Von einer Minimierung ist abzusehen, da in der Gegenwart schwierig abzuschätzen ist, wie viel Information in der Zukunft zu einer korrekten Darstellung benötigt wird. Auf jeden Fall empfiehlt sich eine Orientierung an internationalen Standards und eine sichere Speicherung in unmittelbarer 'Nähe' zu den Primärdaten - am besten im selben Dokument. Für eine schnellere Verarbeitung und Zugriff sollte zusätzlich die Anwendung einer Datenbank in Betrachtung gezogen werden.

Die Minimierung der Prozesse für die Lesbarkeit des Dateiformats liegt in der Emulation einer alten auf einer neuen Maschine. Für Realisierung einer Emulation ist aber viel von nicht archivarischem und technischem Wissen abhängig. Die dafür benötigten externen Ressourcen sind kostenintensiv, und ihr Produkt ohne fremde Hilfe nicht anpassbar beziehungsweise für Archive selber nicht migrierbar. Vermutlich ist mit fortschreitender Zeit und resultierender Verschachtelung eine Emula-

---

[5] Daten in elektronischen Datensicherungssystemen (Backup) in unterirdischen Bunkern sind vielleicht vor unberechtigtem Zugriff sicher - jedoch auch nur so sicher wie die Verschlüsselung der eingehenden Kommunikationsleitung -, gegen ihren eigenen Alterungsprozess und eine mögliche Beschädigung durch Personen mit Zugangsberechtigung müssen trotz der dicken Wände Vorsichtsmassnahmen getroffen werden.

[6] Ein aktueller Ansatz stellt die Ausbelichtung digitaler Daten auf Mikrofilm (Haltbarkeit ungef. 300 Jahre) dar. Siehe: http://www.peviar.ch/

[7] Siehe Kapitel 2.

tion einer Emulation auf einer Emulation nicht mehr durchführbar. Die Lesbarkeit der Datenformate sollte folglich durch die möglichst weitgehende Automatisierung der Formatmigration durch Stapelverarbeitung aufrecht gehalten werden. Da für solche Automatisierungen noch keine standardisierten Lösungen vorhanden sind, sollten Dateiformate zur Speicherung digitaler Daten gewählt werden, für die durch ihre grosse Verbreitung und offene Spezifikation von einer möglichst langen Unterstützung der Lesbarkeit durch neue Maschinen ausgegangen werden kann (TIFF, PDF).

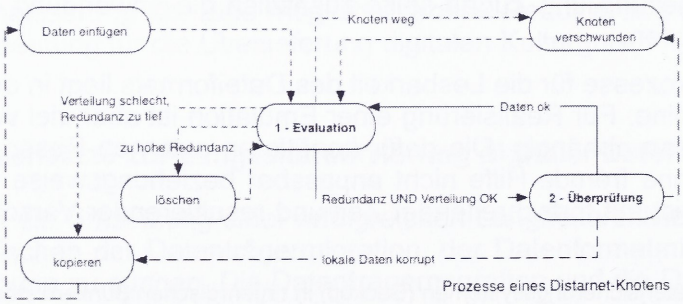
2. Distarnet: automatische Datenträgermigration

Hauptziel des Projektes ist die Entwicklung eines Kommunikationsprotokolls für ein verteiltes Archivsystem, dessen Implementation die persistente Speicherung digitaler Daten ermöglicht. Dabei wird speziell die sichere Überlieferung der Daten als eines der primären Bedürfnisse von Archiven und Museen als Bewahrer und Sammler von Kulturgütern berücksichtigt. Um eine hohe Sicherheit zu erreichen, muss das Risiko eines Datenverlustes minimiert und die Unabhängigkeit der Daten gegenüber technologischen Veränderungen gewährleistet werden. Dazu unterstützt Distarnet die verschiedenen Ebenen von Metadaten, so dass auch die weiteren externen Prozesse wie die Datenformatmigration oder die Emulation ermöglicht und somit die Lesbarkeit der Daten in der Zukunft garantiert werden. Das Projektteam besteht aus einem Wirtschaftswissenschaftler und einem Historiker, die die Bedürfnisse und Anforderungen ihrer Disziplinen bezüglich der Archivierung digitaler Daten in die Definition des Systemverhaltens von Distarnet einfliessen lassen. Kosteneffizienz und zukünftige Forschungsansprüche gegenüber digitalen Daten stehen im Vordergrund.

2.1. Distarnetprozesse: Protokoll- und Implementationseigenschaften

Distarnet wird als Kommunikationsprotokoll für XML-Botschaften in einem verteilten System definiert, dessen Schema frei zugänglich ist.[8] Es umschreibt folgende Prozesse:

Die sichere Überlieferung der Daten wird durch die Architektur eines P2P-Netzwerks erreicht, in welchem Verschlüsselung, Redundanz und fehlertolerante Wiederherstellung implementiert sind. Die einzelnen Knoten dieses Netzwerkes kommunizieren verschlüsselt und auf der Basis von Internettechnologien. Sie speichern digitale Daten auf verschiedenen Knoten verteilt in einer vorgegebenen Redundanz. Jeder Knoten dieses Netzwerkes kommuniziert mit anderen Knoten, die alle gleichberechtigt sind. Dadurch ist im Netzwerk kein funktionskritischer Knoten vorhanden, und die Skalierbarkeit gewährleistet. Statusabfragen und -meldungen zur Kontrolle der Verfügbarkeit der Daten und eine entsprechende Reaktion ermöglichen die Fehlerverarbeitung: Hat ein Knoten des Netzwerks seine Daten verloren, sind diese korrupt, setzt er neue Speichermedien ein oder existiert dieser nicht mehr, so wird automatisch wieder die vorgegebene Redundanz hergestellt. Dadurch wird nicht nur die Überlieferung der Daten gesichert, sondern auch die Datenträgermigration automatisiert:



Prozesse eines Distarnet-Knotens

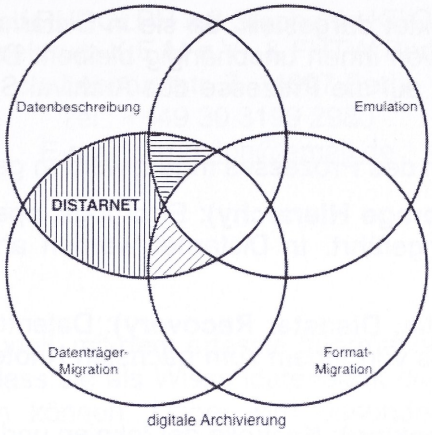
Der zirkuläre Prozessfluss kommt beim Einfügen von Daten in Gang. Es beginnt die Evaluation, die die für die Daten besten Knoten anhand von Sicherheitskriterien bestimmt und, sollten die geforderte Redundanz oder Verteilung nicht erfüllt sein, einen Kopier- oder Löschauftrag erteilt. Ansonsten wird nach der Evaluation die Überprüfung gestartet, die sämtliche lokalen Daten und ihre

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

entfernten Kopien auf deren Integrität prüft. Sind sämtliche Daten vorhanden und integer, wird erneut eine Evaluation durchgeführt. Liefert die Überprüfung ein negatives Resultat, werden entweder die Daten zurückkopiert oder andere Knoten über den Ausfall eines Knoten informiert. Die Evaluation startet erneut, um die vorgegebene Redundanz wiederherzustellen und den Kopierprozess einzuleiten.

Wie oben ausgeführt, garantiert die blosse Erhaltung der Daten aber noch nicht die Überlieferung der Information. Die Datenbeschreibungen zur Auffindbarkeit, zur Darstellung und für den kausalen Zusammenhang des digitalen Archivguts müssen ebenfalls überliefert sein. Deshalb sind sie Teil der Daten in Distarnet und können netzwerkübergreifend abgesucht werden.

Nach den obigen Ausführungen über die möglichen Lösungsansätzen einer erfolgreichen Archivierung digitaler Daten entspricht Distarnet der automatischen Datenträgermigration und der Datenbeschreibung (vertikale Schraffierung):

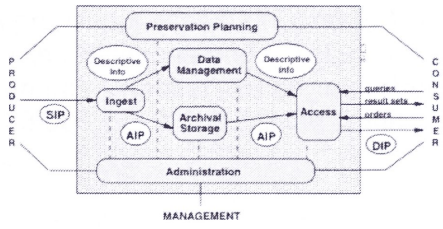


Die Emulation oder die Formatmigration sind bezüglich Distarnet externe Prozesse. Um diese Prozesse zu ermöglichen, überliefert Distarnet die benötigten Metadaten. Das Projekt orientiert sich dafür an international verbreiteten Standards wie METS und PREMIS und ergänzt diese wo nötig.[9]

Die Implementierung von Distarnet gilt als Nachweis für die Funktionstauglichkeit des Protokolls und wird plattformunabhängig als Open Source-Software in Java realisiert. Nach dem aktuellen Stand der Entwicklung steht die Kommunikation zwischen einzelnen Knoten und werden Daten hin- und hergeschickt. In den nächsten Monaten werden die einzelnen Protokollbefehle implementiert und getestet. Das Projekt endet im Dezember 2007, dann wird die erste Protokollversion definitiv und dessen Implementierung veröffentlicht.

2.2. Die Distarnetprozesse und das OAIS-Referenzmodell

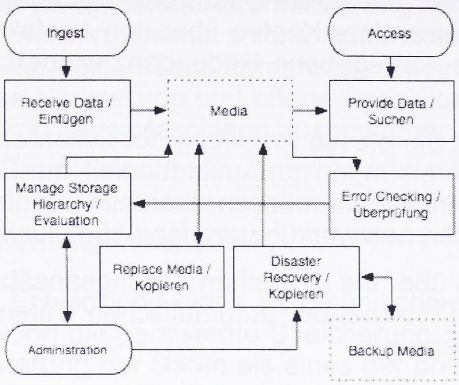
Distarnet entspricht im OAIS-Referenzmodell den Prozessen des Archival Storage und umfasst auch die übergeordneten Prozesse Ingest und Access, da Daten und Metadaten ein- und ausgelesen werden. Das OAIS-Referenzmodell[10]:



[9] Siehe http://www.loc.gov/standards/

[10] http://www.ccsds.org/documents/650x0b1.pdf

Die Prozesse des Archival Storage gemäss OAIS-Referenzmodell:



Die Speichermedien sind gepunktet dargestellt, da sie in Distarnet beliebig austauschbar und die auf ihnen gespeicherten Daten von ihnen unabhängig bleiben. Die weiter oben dargestellten Distarnetprozesse können wie folgt auf die Prozesse des Archival Storage aus dem OAIS-Referenzmodell übertragen werden:

Einfügen (Ingest): Daten des Prozesses Ingest werden gespeichert.

Evaluation (Manage Storage Hierarchy): Error logs, operationale Statisiken, Medienauswahl usw. werden durchgeführt. In Distarnet werden anstelle von Medien Speicherorte ausgewählt.

Kopieren (Replace Media, Disaster Recovery): Datenträgermigration. Entspricht in Distarnet dem Kopierprozess von einem zum nächsten Knoten. Die Datenträger sind beliebig austauschbar.

Überprüfung (Error Checking): Kontrolle der lokalen und entfernten Datenintegrität.

Suchen (Provide Data): Gefundene Daten werden an den Prozess Access überreicht.

2.3. Fazit

Durch die Eigenschaften von Distarnet wird die Datenträgermigration digitaler Daten automatisiert und als kostenintensiver Prozess für Archive eliminiert. Die zukünftige Datenformatmigration oder die Emulation einer veralteten Maschine werden durch Metadaten unterstützt, die die dafür notwendigen Informationen beinhalten. Distarnet ist konform zum OAIS-Referenzmodell.